telektronikk 3.04

100 YEARS ANNIVERSARY

100th Anniversary Issue

Perspectives in telecommunications

Telektronikk
Volume 100 No. 3 – 2004
ISSN 0085-7130

Editor:
Per Hjalmar Lehne
(+47) 916 94 909
per-hjalmar.lehne@telenor.com

Editorial assistant:
Gunhild Luke
(+47) 415 14 125
gunhild.luke@telenor.com

Editorial office:
Telenor ASA
Telenor R&D
NO-1331 Fornebu
Norway
(+47) 810 77 000
telektronikk@telenor.com
www.telektronikk.com

Editorial board:
Berit Svendsen, CTO Telenor
Ole P. Håkonsen, Professor
Oddvar Hesjedal, Director
Bjørn Løken, Director

Graphic design:
Design Consult AS (Odd Andersen), Oslo

Layout and illustrations:
Gunhild Luke and Åse Aardal,
Telenor R&D

Prepress and printing:
Gan Grafisk, Oslo

Circulation:
4,000

Networks on networks

Connecting entities through networks – in technological, societal and personal terms – enables telecommunication. Networks occur on different levels, form parts of larger networks, and exist in numerous varieties. The artist Odd Andersen visualises the networks on networks by drawing interconnected lines with different widths. Curved connections disturb the order and show that networks are not regular but are adapted to the communication needs.

Per H Lehne, Editor in Chief

Contents

100th Anniversary Issue: Perspectives in telecommunications

100 years

BERIT SVENDSEN

Berit Svendsen is Executive Vice President Technology / CTO of Telenor ASA and chairman of Telenor R&D

Telektronikk's forerunner, *TECHNICAL INFORMATION from the Telegraph Administration*, appeared in April 1904, with the intention "to impart knowledge, to all personnel, of the State Telegraph Administration technical installations, and to supply the more important news in the areas of telegraphic and telephonic technology."

This modest beginning took place at a time when telecommunications was dominated by the telegraph. Since the establishment of the Norwegian Telegraph Administration in 1855, the countrywide telegraph services had in 1904 grown to be a necessary instrument, a must, in all kinds of commercial activities. The telephone service was in its early youth; it existed mainly within the cities with poor coverage. Long distance telephone calls might be dispatched only on a few routes.

Radio communications had hardly come to Norway. The first trials with radiotelegraphy had taken place in 1903 in the northern part of the country.

There is a huge gap between the state of the art in 1904 and the technology of today and the wide portfolio of advanced services which are now being offered. In a number of short articles in this issue, *Telektronikk*'s editor, Per H. Lehne, gives an overview of the technology and how some events have been reflected in the journal. In all the years after the introduction of the journal idealistic engineers and technologists have presented information about innovations in telecom with emphasis on applications in Norway. In older issues you will find presentations about the bigger paradigm shifts in technology and new technical installations in the network. The journal has in that way kept up with the original idea to update the personnel in the technical field. Looking back in the old volumes a historian may find detailed information about the first long distance network made up of open-wire lines on poles, the development in telephony, the manual systems, conversion to automatic systems, automation of the long distance network, the introduction – and the end – of the telex network, the large scale deployment of micro wave systems in the trunk network, satellite communication systems, introduction of fibre technology, and not to forget the digitalisation of the telecom network.

In 1967 the Research Institute was established as part of the Telegraph Administration and opened for an influx of young and ambitious researchers, which led to a new keenness in all professional operations of the company. A natural development was that this new personnel category gradually left their mark on the journal and raised the quality of the content. Furthermore, the journal could also serve as a channel to external readers, within and outside the country, and might thus promote the company as a modern technological enterprise. As a consequence articles written in English appeared in the journal.

The objective of promoting the company was even more emphasized when the state enterprise in 1994 was converted to the private company Telenor (although fully state owned) and later introduced to the stock exchange. A new editorial policy was introduced for the journal, with more systematic selection of topics for each issue, aiming at readers both at home and abroad. As a consequence, it was decided that *Telektronikk* should have full production in English.

So, what is ahead? A leading engineer and later Director of the Telegraph Administration, Sverre Rynning Tønnesen, formulated his vision for the telephone service in the 1930s as: *"Anyone, located anywhere, should at any time easily get access to a telephone, and be able to establish a telephone call with good quality to any other person, located anywhere in the world."*

Certainly, it can be said that this vision is a reality today. But this is not the end of the development and the demands from the users. Mobile communication has revolutionized the possibility for any person to be accessible at any time and wherever the person might be, not only for telephone conversations, but also for all other kinds of communication which may be realized face-to-face. The demand for portability and higher capacity is increasing and must be expected to increase furthermore in the years to come, as a prerequisite to realizing new and enhanced services. While fixed-line technology generally offers higher capacity than mobile communication, mobile offers portability. You can easily predict that much creative research may be invested in efforts to bridge this kind of gap between the two technologies.

The modern society of today depends on efficient and reliable telecommunications services to enable business operations and public services to society. Over

ISSN 0085-7130

the last two decades information technologies and the Internet have been transforming the way companies do business, the way students learn, the way the government and communities provide services to the citizens, and communication between individuals. Digital technologies have already proved to be a powerful engine for economic growth. Some people claim that it is a change no less than the Industrial Revolution of the 18th century. By boosting economic growth information and communication technologies have great potential for creating new and better jobs, and generating greater prosperity.

On the road towards the knowledge-based society today's vision for the future would be: *"Any service, any terminal, anywhere, anytime, anyone"*.

There is a continuing challenge for *Telektronikk* to maintain the old intention, to continue to bring information about technology development, but now as the leading Norwegian telecommunication journal and with a view to promoting Telenor as a professional and outstanding company in the field of telecommunications.

We hope you have enjoyed the presentations in the journal up to now and invite you to future presentations of exciting new developments in telecommunications in the years to come.

Berit Svendsen is Executive Vice President Technology / CTO of Telenor ASA and chairman of Telenor R&D. Since joining Telenor in 1988 Berit Svendsen has held various positions; among them Divisional Director of Data-services and Project Director of FMC. Since February 2002 she has been a member of an advisory group for the EU Commission for the entire ICT research in EU's framework programme. She is also chairman of Simula Research Laboratory and Data-Respons ASA as well as a member of the Board of Directors of Ekornes AS. Berit Svendsen holds an MSEE from the Norwegian University of Science and Technology and a Master of Technology Management from the Norwegian University / Norwegian School of Business Administration / MIT.

email: berit.svendsen@telenor.com

Enlightening 100 years of telecom development

From 'Technical Information' in 1904 to 'Telektronikk' in 2004

PER H. LEHNE

Per H. Lehne is Research Scientist at Telenor R&D and Editor in Chief of Telektronikk

When the Telegraph Director in 1904 initiated the monthly sheet *TEKNISKE MEDDELELSER fra Telegrafstyrelsen (Technical Information from the Telegraph Administration)*, it is doubtful that he could imagine that it would exist 100 years later as the international journal *'Telektronikk'*, covering virtually all aspects of modern telecommunications technology. The first issue in April 1904 consisted of four pages starting with an article called 'On the protection of telephone cables from power lines', an article spanning the first two issues. 100 volumes of the journal has proven to be a fantastic source of Norwegian telecommunications history, written 'by those who saw it happen' and actually took part in the development.

1904 – The year it started

The first issue of *TEKNISKE MEDDELELSER fra Telegrafstyrelsen (TECHNICAL INFORMATION from The Telegraph Administration)* came in April 1904.

We do not know what happened, or what was said on the occasion. In another article in this Anniversary issue of *Telektronikk*, *Harald Rinde* explains how it seems to be just one initiative of a broader set of reforms to make Telegrafverket more capable of meeting all the new technological changes that occur from the 1880s onwards [1].

The technological evolution had emerged rapidly on the telegraph and telephone areas. Three years earlier, Marconi had demonstrated the wireless telegraph across the Atlantic Ocean and Telegrafverket was already planning Norway's first installation [2].

Jonas Severin Rasmussen, Telegraph Director from 1892 to 1905, came from the position as Reader (Senior Lecturer) at the Norwegian Naval College and his background was right for the time.

Rasmussen found a good man for the job as editor in *Magne Hermod Petersen*[1)]. At the time Petersen was Headmaster at the Telegraph Administration's training Institute in Christiania[2)], but he had also been doing pioneering work on wireless telegraphy. In fact, Hermod Petersen was a well-reputed expert on wireless communications. He had the educational as well as the technological skill to fill the position as editor. Later, he worked his way upward in the company to become Head of the Radio Department from 1920, and in charge of the Technical Department from 1931, and he finally himself became Director of the Telegraph Administration in 1935, a position he held until 1938.

1904

The Norwegian National Assembly (the 'Storting') discusses the establishment of a separate Norwegian consulate service. The underlying discussion was about whether Norway should have its own foreign policy under the Swedish-Norwegian union, something that would drain the union of almost any content. A unilateral resolution in the Storting in May 1905 eventually led to the dismantling of the union from June 7, 1905.

John William Strutt (Lord Rayleigh) receives the Nobel Prize in Physics "for his investigations of the densities of the most important gases and for his discovery of argon in connection with these studies". Lord Rayleigh is known to wireless engineers for the 'Rayleigh distribution', which models the signal variations of a wireless transmission subject to reflections and non-line of sight (Rayleigh fading). He also developed the theory of Rayleigh scattering, which is the phenomenon giving us the blue sky, and also applies to light diffusion in optical fibres.

The distress signal 'CQD' is established. CQD (-.-. —.- -..) was the first standard international Morse code distress signal and was originally proposed and adopted by Marconi on January 7. It only lasted until 1906, when the German standard 'SOS'-code was adopted.

John Ambrose Fleming invents the vacuum tube. The 'thermionic valve' was a diode valve able to rectify alternating current. This is the pre-amble of the radio valve. Fleming had among other things also assisted Marconi with his transatlantic experiment in 1901 [2].

---

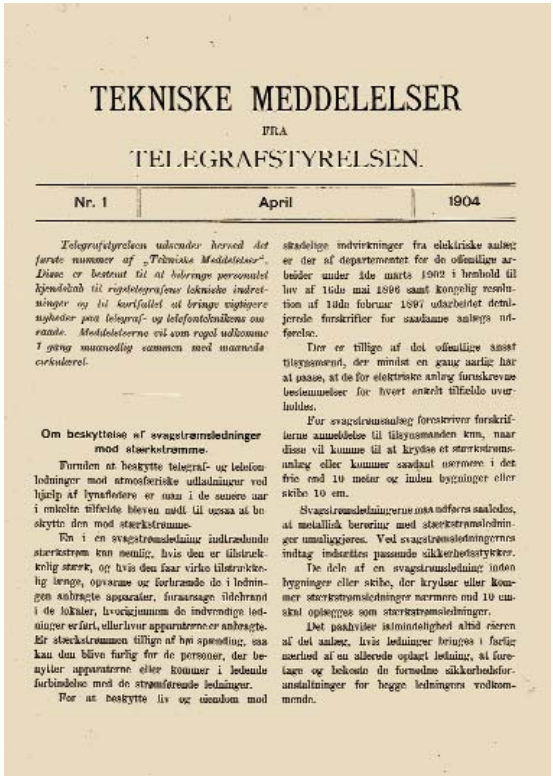
1) *In most other sources except Thorolf Rafto's history of the State Telegraph of 1955, he is called only* Hermod Petersen, *which will be used throughout this article.*

2) *The Norwegian capital Oslo was named* Christiania *until 1925.*

ISSN 0085-7130

Photo: Norwegian Telecom Museum

(Magne) Hermod Petersen was the first editor of 'Technical Information' from 1904 to1909. He had a teaching background as Headmaster of the Telegraph Administration's Training Institute in Christiania. He later worked his way up to become Director of the Telegraph Administration in 1935



TEKNISKE MEDDELELSER

FRA

TELEGRAFSTYRELSEN.

Nr. 1 | April | 1904

Telegrafstyrelsen udsender herved det første nummer af „Tekniske Meddelelser". Disse er bestemt til at bringe personalet kjendskab til regelelegrafens tekniske indretninger og tel kortfattet at bringe vigtigere nyheder paa telegraf- og telefontekniskens om raade. Meddelelserne vil som regel udkomme 1 gang maanedlig sammen med maanedscirkulæret.

Om beskyttelse af svagstrømsledninger mod stærkstrømme.

Formaalet at beskytte telegraf- og telefonledninger mod atmosfæriske udladninger ved hjælp af lynafledere er saa i de senere aar i enkelte tilfælde bleven mødt til ogsaa at beskytte den mod stærkstrømme.

En i en svagstrømsledning indtrædende stærkstrøm kan nemlig, hvis den er tilstrækkelig stærk, og hvis den faar virke tilstrækkelig længe, opvarme og forbrænde de i ledningen anbragte apparater, forraarsage ildebrand i de lokaler, hvorigjennem de indvendige ledninger er ført, eller hvor apparaterne er anbragte. Er stærkstrømmen tillige af høi spænding, saa kan den blive farlig for de personer, der betjener apparaterne eller kommer i ledende forbindelse med de strømførende ledninger.

For at beskytte liv og eiendom mod skadelige indvirkninger fra elektriske anlæg er der af departementet for de offentlige arbeider under 16de marts 1903 i henhold til lov af 16de mai 1896 samt kongelig resolution af 14de februar 1897 udarbeidet detaljerede forskrifter for saadanne anlægs udførelse.

Der er tillige af det offentliges anset tilsynsmænd, der mindst en gang aarlig har at passe, at de for elektriske anlæg foreskrevne bestemmelser for hvert enkelt tilfælde overholdes.

For svagstrømsanlæg foreskriver forskrifterne anmeldelse til tilsynsmanden km, naar disse vil komme til at krydse et stærkstrømsanlæg eller kommer saadant nærmere i det frie end 10 meter og inden bygninger eller skibe 10 em.

Svagstrømsledningerne maa udføres saaledes, at metallisk berøring med stærkstrømsledninger umuliggjøres. Ved svagstrømsledningernes indtag indsættes passende sikkerhedsstyrker.

De dele af en svagstrømsledning inden bygninger eller skibe, der krydser eller kommer stærkstrømsledninger nærmere end 10 cm. skal optages som stærkstrømsledninger.

Det paahviler selvfølgelig altid eieren af det anlæg, hvis ledninger bringes i farlig nærhed af en allerede oplagt ledning, at foretage og bekoste de fornødne sikkerhedsforanstaltninger for hegge ledningers vedkommende.

The first issue of 'Technical Information from the Telegraph Administration' *appeared in April 1904. It consisted of four pages with an article* On the protection of telephones cables from power lines, *an article which span the first two issues. The sheets appeared monthly, and the whole of 1904 comprised 74 pages, about half the size of one issue in 2004*

Editors of *Technical Information* and *Telektronikk*	
Hermod Petersen	1904 – 1909
Johannes Storstrøm	1916 – 1921
Sverre Rynning Tønnesen	1927 – 1942
Julius Ringstad	1942 – 1957
Nils Taranger	1957 – 1978
Bjørn Sandnes	1978 – 1991
Ola Espvik	1991 – 2003
Per Hjalmar Lehne	2003 –

The first issue was introduced by a short statement:

> *The Telegraph Administration herewith produces the first issue of the newssheet "Technical Information". It is intended that this will impart knowledge, to all personnel, of the State Telegraph Technical Installations, and will, in short, supply the more important news in the areas of telegraphic and telephonic technology. As a rule, the newssheet will appear once a month, at the same time as the monthly circular.*

In 1995, *Telektronikk* published an article in connection with the journal's 90th anniversary [3]. In this article, *Henrik Jørgensen* goes through the development of the journal on a backdrop of how the State Telegraph – 'Statstelegrafen', later 'Telegrafverket' – evolved into 'Televerket' – the Norwegian Telecom – and from 1995, Telenor, a modern telecommunications company. Jørgensen's article also contains comprehensive biographies of the editors until then.

This article will review briefly the journal's development, the main source for this being Jørgensen's article from 1995.

Unstable start, 1904 – 1921

In the first year, the production starts with monthly issues. As mentioned, the first issue only had four pages; however the full year gained 74 pages. Already the next year, the production rate decreases to six issues, and thereafter it ebbs away to only one issue each year in 1908 and 1909. It then stops altogether until a new editor is appointed in 1916 *(Johannes Storstrøm)*. Hermod Petersen may have been too busy with the work on establishing Svalbard Radio, which opened in 1911, to be able to concentrate on the production of the sheets.

The general situation until 1992 is that the position of editor is performed during his spare time. It is on the occasion that the journal's editorial office is transferred to Televerket's Research Institute (TF) that this changes, and it becomes part of the regular work.

The first years, the articles were anonymous, but from 1916 the writers' names appeared. It is possible from the context, subject and other sources however to draw some conclusions about the writers in the early periods, as can be seen in this issue's article on the wireless telegraph in Lofoten in 1906 [2]. The intention was probably that the sheets were to express the view and voice of the Administration, not the writers.

Again in 1921, the production stops; however the last issues in the production run were apparently printed in 1925. It seems that the reason for the journal's disappearance was that it did not perform sufficiently [3], and that the choice of material was too biased.

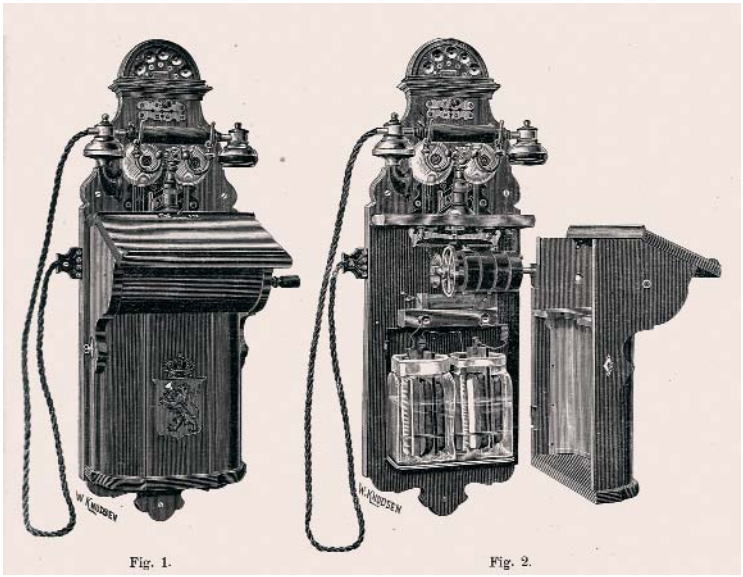


Illustration from 'Technical Information', issue 6, 1904

One of the first articles in 1904 treated the new standard telephones introduced by Telegrafverket in 1899. This was a model from LM Ericsson, which from 1902 was manufactured by Elektrisk Bureau (EB) in Christiania

Stable production starts in 1927

After the second break, it seems that a more obvious editorial control is introduced. Payment for contributions is also introduced as an incentive.

The new editor in 1927, *Sverre Rynning Tønnesen*[3)], took the task seriously and declared that [2]

> *"In those intervening years when* Technical Information *did not appear, there have been significant technical developments in both fixed and wireless telegraphy and telephony."*

> *"In order to secure a common level of knowledge within our readership, we must lay the foundation first in order to build on them in future articles. So we must provide comprehensive explanations of developments within all the different branches of telegraphic and telephonic technology, from the earliest beginnings to the later developments, as well as providing other items of interest."*

Technical Information now came very regularly, even during the Second World War, which had very little apparent effect on the journal.

For natural reasons, contributions in the first years after the war were sparse as all qualified professionals were fully occupied with rebuilding the telecommunications networks.

At the end of the 1940s and into the 1950s, the journal, edited by *Julius Ringstad*, starts receiving articles from new sources. From this time employees from Telegrafverket start going abroad on study tours.

We can early read about the transistor (1950), mobile telephony (1950), Pulse Code Modulation (1951) and new switching principles related to the telephone exchanges ITT 8A[4)] (1953). The technical and operational achievements of Telegrafverket during the VIth Olympic Winter Games in Oslo in 1952 are of course described, and in 1955, a large centenary issue for Telegrafverket is published.

From Information to Journal

Then in 1959, it was time for a name change. *Nils Taranger* had taken over as editor in 1957, and the journal had become *Telegrafverket's Technical Journal*.

The old name, *Technical Information from the Telegraph Administration* was considered too long and

In 1927, payment is introduced for contributions to the journal

"For original work, which is assumed to be based on the author's own research"	per column 15.00 kroner
"For articles based on both original work and non-Norwegian sources"	per column 10.00 kroner
"For items translated from non-Norwegian sources"	per column 7.50 kroner

3) *Sverre Rynning Tønnesen later became Director of Telegrafverket from 1942 to 1962.*

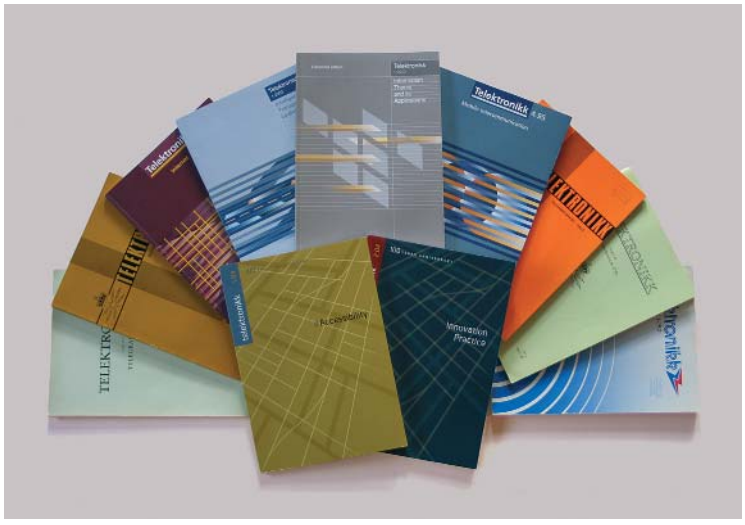
4) *The ITT 8A and 8B-switches were manufactured by Standard Telefon og Kabelfabrik A/S (STK), previously a part of Western Electric. These were introduced in the 1950s and were the first combined mechanical-electronic switches in the Norwegian telephone network.* Telektronikk *has over the years published several articles on the 8B-system.*

ISSN 0085-7130

Photo courtesy of Lucent Technologies Inc./Bell Labs

In 1950, an article from a study tour to USA the year before presents both mobile telephony and the newly invented transistor. The transistor invented by Bardeen, Brattain and Schockley in 1947 was of the point-contact type. The original is on display at Bell Labs lobby exhibit, open to the public, at Lucent Technologies headquarters in Murray Hill, NJ, USA

cumbersome, and did not cover the journal's contents which was no longer of a purely technical nature. Thus a name which was shorter and more up to the demands of the time was necessary. This was not easy to find, and a competition was launched among the employees of Telegrafverket. The winner was *'Telektronikk'*. The name was considered a truly new word in the language. It contained 'tele', a word that started to obtain a footing in the language, and 'elektronikk' (electronics), a term most common to the readers. At the same time, the journal changed appearance and was produced with cover pages.



The evolvement of cover pages from 1962 until 2004

From now on, the journal expands to contain articles commissioned from the editor as well as spontaneous contributions. Important contributions are the regular proceedings from the biennial Norwegian Telephone Engineers' Conferences (Norsk Telefoningeniørmøte – NTIM) and also lectures and reports from Telegrafverket's District Engineers' meetings.

When *Bjørn Sandnes* takes over the journal in 1978, it continues to present high quality articles when Televerket enters a new era. New telecom services appear (mobile telephone and data, paging, tele- and videoconferencing, cable-TV, etc.), and are presented in *Telektronikk* in the 1980s. Also, the new wind of liberalisation leaves its mark, and the journal has several articles on the subject from 1983 to 1986.

To some extent we see thematic issues appearing from time to time. The centenary of the telephone in Norway is the subject of a full issue in 1980; satellite communications is treated in 1981. In 1986 and 1987, when the important decisions are made on GSM (see also the special section in this issue), several articles appear in two issues describing the status as well as technical input and decisions.

When Televerket's Research Institute (TF, now Telenor R&D) was created at the end of 1967, we see a new personnel group being the source of articles to the journal. It was perhaps a logical evolvement that the journal's editorial responsibility was moved here in 1992.

New conditions

Televerket stood on the threshold of a total transformation to become Telenor AS, a limited company, from 1995. This also changed the conditions. Before this, Televerket, as part of the Norwegian state, 'had no secrets' and articles about long term planning, budgets and detailed technical solutions could be published. Under the new constraints, the journal had to change also. In 1992 it was handed over to the Research Institute, which appointed *Ola Espvik* as the new editor. His background was on traffic, reliability and operational control of telecommunications networks, however, an equally important qualification was his experience as a Lecturer and Study Advisor in telecommunications and computer science at UNIK – the University of Oslo's graduate Study Centre at Kjeller, and later as the Director for the Joint College Centre at Kjeller

Thus, a new set of objectives were formulated [2]:

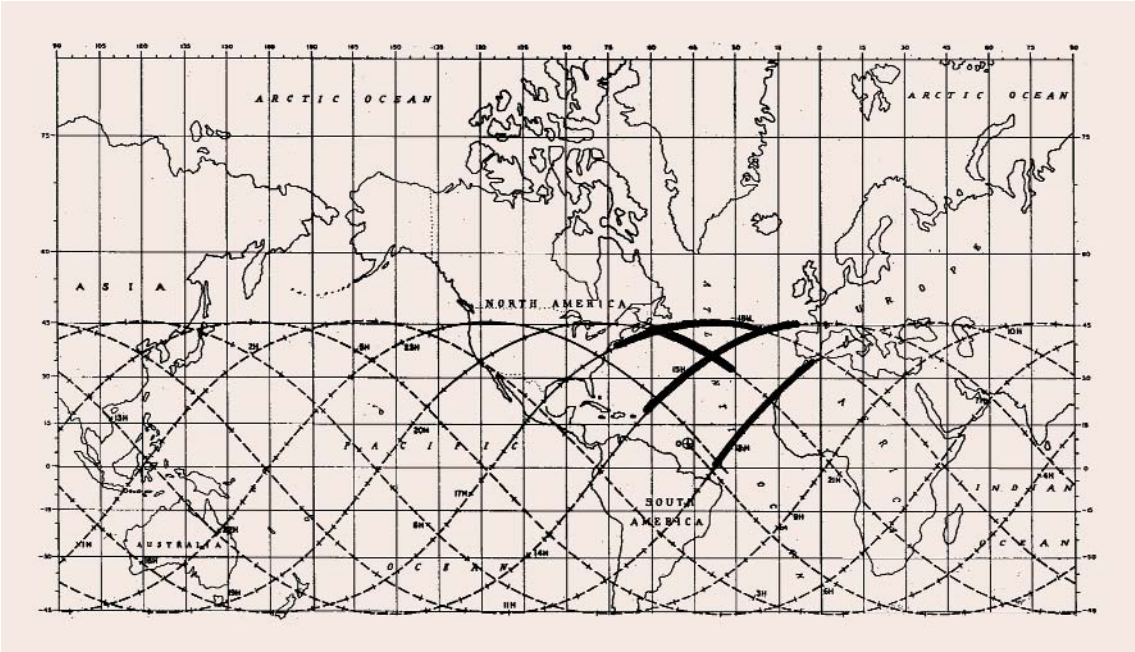


Illustration from Telektronikk, issue 1-2, 1964

Already in 1964, two years after TELSTAR I, the first commercial communications satellite, was launched; Telektronikk brings an article about the new possibilities of telecommunications via satellite. The picture shows a projection on TELSTAR I's first nine orbits, i.e. the first 24 hours. The broad lines show the parts of four orbits, or passages, simultaneously visible from Andover, Maine in the USA and Plemeur-Bodou in France. Results from TELSTAR I even indicated that reception was possible down to an elevation angle of 0.5°, already preparing the ground for the possibility of communication to high latitudes like e.g. Svalbard (approx. 3°) via satellite

Telektronikk shall be the leading Norwegian telecommunications journal.

Telektronikk shall through its choice of subjects and presentation format contribute to synchronizing professionals on the Norwegian telecommunications scene with reference to the development of telecommunications techniques.

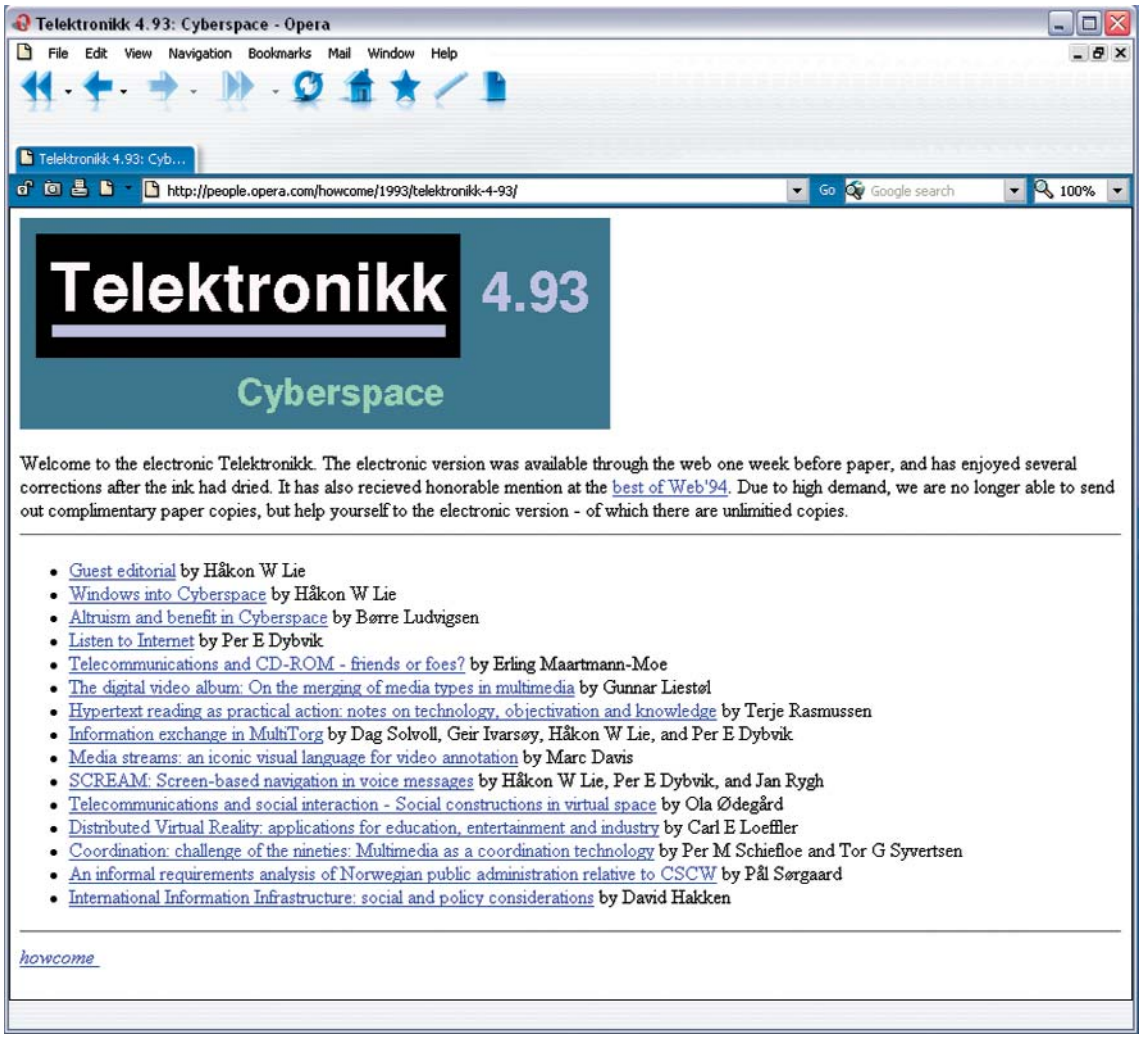
The editorial line now was systematically transformed into a thematic form, which has proven successful for presenting features like satellite and mobile communications, traffic engineering, switching, network planning, and much more, including non-technical issues like telework, user and social aspects, and telemedicine. From now on, each issue is headed by a feature editor appointed by the editor in chief.

From 1995, the journal is produced solely in English to meet the increasing internationalization. The educational profile is improved and several issues are being used as lecturing material at e.g. the Norwegian University of Science and Technology (NTH/NTNU). It is also increasingly positioning itself as a representative of the *entire* Norwegian professional environment. A pioneering event is also taking place in entering electronic publishing. Already in 1993, a special issue on 'Cyberspace' is transformed into the new web-format. According to *Håkon Wium Lie*, feature editor of the 4.1993 issue, this was a kind of green-field operation, because a lot of the procedures for converting documents into HTML and adding pictures and hyperlinks were not established at the time [4].

Illustration from Telektronikk, issue 4, 1978

In 1978, Telektronikk publishes two articles on the project of developing a new standard telephone set for the Norwegian telephone network. The 'Tastafon' was the last such standard telephone. It introduced a new design and push buttons to enable new services in the telephone network. The picture shows the prototype, or 'A'-model, which was delivered – in a number of two – from Elektrisk Bureau to Televerket in August 1976

ISSN 0085-7130



The 'Cyberspace' issue (volume 89, No. 4, 1993) was the first Telektronikk *to be put on the web in complete form, initiated by the feature editor, Håkon Wium Lie. It received an 'Honourable Mention' in the competition 'Best Document Design' at the 'Best of Web'94'*

It is under Espvik's editorship the journal takes on its current format, including a significant improvement of the layout. From 1992 until 2003, each issue has a cover page designed to reflect the subject. The cover art is done by Odd Andersen, a well-reputed designer and artist with whom Ola Espvik collaborated closely in order to bring out the 'core' of the subject in an abstract style. This process has truly produced a series of high quality art works.

2004 and looking ahead

In 2003, Ola Espvik chose to withdraw from the position as editor of *Telektronikk*, and the writer of this article was asked to take over the baton. Taking over the responsibility of this project is both intimidating and fascinating. During *Telektronikk*'s lifespan, there have been moments demanding new thinking and a new course. Stabilizing the journal under better editorial control in 1927 was one of them, and the transformation in 1992 into a journal for the Norwegian professionals with an aim to reach an international audience is another.

2004 does not demand a radical change. The objectives of 1992 are valid also today with the interpretation that the journal also must embrace the new members of Telenor's professional family around the world, thus English is the only natural choice of language.

The thematic form will continue, as it has shown be the best format for making the journal issues useful and interesting even after the immediate publication date. Themes will be chosen to demonstrate what are the important topics and developments in telecommunications, as well as giving insight to new readers in basic knowledge. Whether the current editor will succeed in keeping the professional level high and also be able to present the important features at the right times, our readers of the future will have to judge.

The web version of *Telektronikk* from 1993 has been a stand-alone happening up till now. A web page has existed for quite some years, but only giving the table of contents and the guest editorial of the issues from 1992 and onwards. From 2004, we launch our new web site where we will offer full article text availability.

In co-operation with our layout designer since 1992, Odd Andersen, we have also chosen to change the cover art into a permanent design, showing how telecommunications is networking both society and technology on many levels.

We have used paper as the primary medium for 100 years. It is my guess that it will take a long time before electronic media have the properties making them attractive and practically feasible to really replace the paper for a journal like *Telektronikk*. The most important benefits of paper are namely these: No batteries required, no decoding software needed, anywhere, anytime.

I started this article by postulating that the Telegraph Director and editor in 1904 could not have foreseen what *Telektronikk* would evolve into during the first 100 years. It is highly probable that they did not reflect on that at all. Knowing that it actually has existed for a century, it is easier to be struck by the thought of the next hundred. I can easily read the issues from 1904 and understand the contents, as well as gain insight into the technology development, problems and solutions of the time. They also tell a lot about how Telegrafverket, Televerket and Telenor organised work and prioritised technical challenges. They have truly 'enlightened the history'.

Will the readers of 2104 be enlightened? Will *Telektronikk* exist at all in 2104? What will it look like? What will be the name?

The questions are many.

Acknowledgements

The author of this article wants to thank the former editors of the journal, *Bjørn Sandnes* and *Ola Espvik* for interesting and valuable discussions and viewpoints on *Telektronikk*'s history and significance, as well as commenting on early versions of the manuscripts. Librarians *Arve Nordsveen* and *Anne Solberg* at the Norwegian Telecom Museum have also been very helpful in giving access to the archives of *Telektronikk* and providing additional information on the subjects treated in this and other articles in this issue.

References

1 Rinde, H. Why 1904? A time of rapid and radical technological change. *Telektronikk*, 100 (3), 10–12, 2004. (This issue)

2 Lehne, P H. The second wireless in the world. *Telektronikk*, 100 (3), 209–213, 2004. (This issue).

3 Jørgensen, H. 90 years of Televerket's technical journal – an overview. *Telektronikk*, 91 (4), 193–210, 1995.

4 Lie, H W. How Telektronikk changed the Web. *Telektronikk*, 100 (3), 82-85, 2004. (This issue)

Other sources

Århundrets Hvem Hva Hvor. Oslo, Schibsted, 2000.

Collet, J P, Lossius, B O H. *Visjon – Forskning – Virkelighet. Televerkets Forskningsinstitutt 25 år*. Kjeller, Norwegian Telecom Research (Telenor R&D), 1993.

Norwegian Telecom Museum (Norsk Telemuseum). September 22, 2004 [online] – URL: http://www.norsktele.museum.no/index_ieoff.html

Rafto, T. *The history of the Norwegian State Telegraph 1855–1955 (Telegrafverkets historie 1855–1955)*. Oslo, Telegrafverket, 1955.

Per Hjalmar Lehne (46) is Research Scientist at Telenor R&D and Editor in Chief of Telektronikk. He obtained his M.Sc. from the Norwegian Institute of Science and Technology in 1988. He has since been with Telenor R&D working with different aspects of terrestrial mobile communications. His work since 1993 has been in the area of radio propagation and access technolgy, especially on smart antennas for GSM and UMTS. He has participated in several RACE, ACTS and IST projects as well as COST actions in the field. His current interests are in the use of MIMO technology in terrestrial mobile and wireless networks and on access network convergence, where he participates in the IST project FLOWS.

email: per-hjalmar.lehne@telenor.com

Why 1904?
A time of rapid and radical technological change

HARALD RINDE *)

Harald Rinde is Researcher at the Centre for Business History at the Norwegian School of Management, BI

Telektronikk's predecessor originates in a period of rapid and radical technological change. The Norwegian Telegraph Administration, the dominant Norwegian operator, had to adopt a new and more systematic approach to the development of its knowledge base. Introducing *Technical Information from the Telegraph Administration* was only one initiative among a broader set of reforms. Right from the beginning, it became an important vehicle in the development of a knowledge based telecom industry.

Introduction

Historically, the journal known today as *Telektronikk* started out in April, 1904, as a monthly publication entitled *Technical Information from the Telegraph Administration*. The new publication was distributed to the employees of the Norwegian state-owned telegraph and telephone operator together with the monthly administrative circular from the management. To explain the origin of *Telektronikk*, then, we must understand why the management of the Norwegian Telegraph Administration (NTA) decided to initiate such a publication at just this time. In the following, I will argue that the initiative should be seen as part of a broader shift around the turn of the century, when the old telegraph administration took important steps towards becoming a modern, knowledge-based telecom operator.

Sadly, no evidence has been uncovered in the NTA archives on the specific background for the initiative that led to the new publication. Neither does the *Technical Information* series itself shed much light on its origins. In the very first issue, it was stated briefly that the purpose of the new series was to supply employees with knowledge of the technical equipment used by the NTA. Furthermore, the series would be used to pass on important news on the development within the field of telegraph and telephone technology. In short, the purpose of the *Technical Information* series was to disseminate technical information. Big surprise. But why did the task of keeping the staff up-to-date on technological issues become important enough to warrant such a series at just this time?

The situation around 1900

In 1904, the Norwegian Telegraph Administration had been around for nearly 50 years without requiring such a series. Furthermore, the NTA was in a particularly tight financial situation in the first five or six years after the turn of the century. In the aftermath of the crash on the Christiania (Oslo) stock exchange in 1899 the Norwegian economy experienced a sharp cyclical downswing, slowing down the otherwise rapid growth in revenues from telegram and telephone traffic and forcing the National Assembly to cut back on its appropriation of grants to the NTA. Given these circumstances, it seems unlikely that the *Technical Information* series may have been merely the realisation of a stray idea developed within the NTA management. To justify the allocation of scarce resources in a time of cut-backs and budgetary restraints, there must have been a strong belief in the need for such a publication within the organisation.

To understand how such a need could arise, it is necessary to take a closer look at the changes that the Telegraph Administration were going through around the turn of the century. With the partial exception of the first few years of the new century, the period from the 1890s until 1920 was a time of unprecedented expansion for the NTA, both in terms of staff and tasks. The number of white collar employees rose from a modest 466 persons in 1895 to 1277 persons in 1905, and 4024 persons in 1920. If we add blue collar workers, extras, and the privately employed staff at the state telephone stations in rural areas, the figures more than double.

This rapid expansion reflected the multiplying of tasks as the new telephone and wireless telegraphy services were added to the NTA's traditional task of operating the national telegraph system. As late as in 1890 the NTA owned a total of only 26 telephones, and all telephone subscribers in the country belonged to private networks. From the mid-1890s, however, a new policy was initiated. The construction of a national long-distance telephone system began, and important local telephone networks in cities and

*) *Co-author of a new three-volume History of Norwegian Telecommunications, to be published by Gyldendal Fakta in January, 2005.*

ISSN 0085-7130 © Telenor ASA 2004

towns such as Trondheim (1898) and Oslo (1901) were taken over by the NTA. In addition, from 1901 engineers at the central staff of the NTA were involved in the navy's experiments with radio telegraphy. Two years later, a civilian system of wireless telegraph stations was tested in Lofoten in the Northern part of Norway, and in 1906 the first two such stations were opened for regular service. By 1920, wireless telegraph stations had taken over the transmission of telegrams between Norway and America as well as the important new task of communicating with ships at sea.

The change from a small and focussed telegraph administration to an expansive multi-task telecommunications operator enhanced the need for a more systematic approach to the development of technical skills and competence within the organisation. For most of the nineteenth century, it would hardly be adequate to characterise the Telegraph Administration as a knowledge-intensive institution. The job as a telegraph operator was considered a practically oriented occupation, and until 1898 the formal education consisted of a three month course programme, focussing mainly on the practical mastering of sending and receiving telegrams. Most middle managers in the NTA were recruited among the telegraph operators, and until the 1880s the Telegraph Director was the only member of the organisation who was formally trained as an engineer.

Jonas Severin Rasmussen was General Director of the Telegraph Administration from 1892 until 1905. His background before that had been as a senior lecturer from the Norwegian Naval College. His appointment helped speed up the process of adopting a new policy on recruitment, education and competence building (Photo: Norwegian Telecom Museum)

Telecommunication becomes scientifically based

Thus, the organisation was hardly in a position to handle rapid and radical technological change. From the late 1870s and increasingly through the 1880s and 1890s, the telegraph technology changed profoundly with the replacement of the traditional manual Morse system by automatic telegraphs and the introduction of multiplex telegraphy, in which two, four, and even six telegrams could be sent simultaneously. In the rapidly evolving fields of telephony and radio telegraphy, technological change was even faster. To make things worse, the new innovations tended to be more directly linked to scientific theory than previously. For instance, radio telegraphy was based on James Clerk Maxwells theory of the electromagnetic field, and the early stages of the development of the new technology took place in the laboratories of physicists such as Heinrich Hertz, Édouard Branly and Oliver Lodge. Similarly, in the case of long-distance telephony the seminal introduction of pupinisation (1900) and the Krarup cable (1901–02), were based on the abstract mathematical theories of Oliver Heaviside and Aimé Vaschy, which contrary to the established view of practically oriented telegraph engineers implied that the transmission quality of telephone cables or wires could be improved by increasing their self-inductance. Of course, scientific knowledge was only one element in the innovation process. The eventual commercial success of the wireless, for instance, depended crucially on the efforts of British-Italian entrepreneur Guglielmo Marconi, who brought the new device out of the laboratories and developed it into a practical communications device. Similarly, the development of the Krarup cable was linked to the practical implementation of the new principles in the construction of new submarine cables connecting Denmark with Sweden and Germany.

To handle multiple technologies, each of which were rapidly changing, the NTA had to adopt a new policy on recruitment, education, and competence building. The appointment of the dynamic Jonas Severin Rasmussen as new Telegraph Director in 1892 helped speed up the process. In the next few years, a number of electrotechnical engineers educated at leading German polytechnical universities were recruited to the Telegraph Director's central staff. Furthermore, the acquisition of the private telephone network in Oslo

ISSN 0085-7130

in 1901 meant that the country's leading milieu of telephone engineers was brought into the organisation. At the same time, an educational reform was initiated. While the practical education of female telephone operators was decentralised to the local exchanges, the educational plan of 1898 established a new Telegraph School in the capital for the male telegraphers and technicians. Instead of a three-month practical course in telegraphy, the new education lasted for $1^1/_2$ year and included both theoretical and practical subjects. The exclusion of females soon came under attack, and from 1910 both sexes were admitted on an equal basis. At the same time, the 18-month training programme was split between a 'lower' and a 'higher' course.

Still, the question of how to keep the regular staff continually up-to-date remained unsolved. In 1888 the first regular journal for telegraphers was published. Seven years later the journal changed its name to *Elekroteknisk Tidsskrift*, signalling a gradual development towards a general electrotechnical journal. There can be no doubt that this journal represented an important infrastructure for the distribution and discussion of information on recent development within the field. However, in the heated struggle in the 1890s about the future organisation of the Norwegian telephone system, *Elektroteknisk Tidsskrift* strongly favoured private telephone companies over state ownership, frequently offering stark criticism of the Telegraph Director and the NTA. It is highly probable that this made the journal less attractive to the NTA management as a suitable source of information for its employees. Furthermore, from the turn of the century the journal gradually devoted more and more attention to the emerging electrical power business at the expense of the coverage of telegraphy and telephony. The introduction of the *Technical Information* series in 1904, then, may be seen as an attempt to fill this gap in the existing information infrastructure.

Summary

We have traced the origin of *Telektronikk* to a period of rapid and radical technological change, when the dominant Norwegian telecom operator was forced to adopt a new and more systematic approach to the development of its knowledge base. As one initiative among a broader set of reforms, *Technical Information* was introduced to distribute up-to-date information on recent technological developments within the telecommunications field. Right from the beginning, then, *Telektronikk*'s predecessor became an important vehicle in the development of a knowledge-based telecom industry.

Selected references

Brittain, J A. The Introduction of the Loading Coil: George A. Campbell and Michael I. Pupin. *Technology & Culture*, 11, 1970.

Kragh, H. The Krarup Cable: Invention and Early Development. *Technology and Culture*, 35, 1994.

Rafto, T. *Telegrafverkets historie 1855–1955*. Oslo, Telegrafverket, 1955.

Rinde, H. *Et telesystem vokser fram. Norsk telekommunikasjonshistorie, bind 1, 1855–1920*. Oslo, Gyldendal Fakta, January 2005. Forthcoming.

Strømberg, E. *Utdanningen i Televerket 1855–1950*. Oslo, Teledirektoratet, 1984.

Harald Rinde (39) is a Researcher at the Centre for Business History at the Norwegian School of Management BI. He is a Dr.Art. (2004) from the University of Oslo with a dissertation on the organisation of national telephone systems in Scandinavia in the 1880–1900 period, and is author of the first volume in a new three-volume history of Norwegian telecommunications which will be published in January, 2005. Rinde has published numerous books and articles within the field of business history and the history of technology. He is currently working on a research project about the history of the lawyers' profession in Norway.

email: harald.rinde@bi.no

Internet Protocol – Perspectives on consequences and benefits by introducing IP

TERJE JENSEN

Terje Jensen is Senior Research Scientist at Telenor Research and Development

The Internet Protocol (IP) has been subject to an overwhelming interest during the last decade or so. One may raise the question of why this has been the case. Broadly, the answer can be that IP represents a shift in the philosophy of telecom business besides being a fairly easy and widespread protocol. This article makes some reflections on issues related to IP development – mostly based on technical aspects, but also linking these to a broader business perspective.

1 Introduction

The Internet Protocol (IP) has been around for several decades, formally managed by IETF (Internet Engineering Task Force). Just 10–15 years ago the commercial interest for this protocol emerged from the telecom provider's side. Several factors took place to support such a development: i) the volume of IP implementations in different terminal equipment (PCs and host computers); ii) the development of web browser applications providing an intuitive user interface for accessing information; iii) email applications taking the step into the residential sphere; iv) gradual deployment of the IP suite in other terminals than computers; v) a growing political and commercial interest for broadband service offerings. Naturally, other factors also contributed in different regions. Overall, it seems too simple to claim that a single event or factor resulted in the huge interest for IP-based services. Therefore, one may wonder whether it was the *open arena* for discussing IP-related work progress provided by IETF, the range of *deficiencies* seen for the IP suite, or the growing commercial interest – and potential *return on effort*, that inspired this huge effort. Possibly the actual explanation is a combination of all these as well as a few others.

Making a distinction between IP as a protocol and the "Internet movement" is essential. Principally, IP is just a protocol defining a set of header fields and capabilities. However, no Internet offering is seen without the presence of IP. Hence, IP and Internet are these days considered to go hand-in-hand. On the other hand, IP-based networks are also implemented that do not allow for public traffic. Examples are enterprise-internal networks, e.g. within and between enterprise sites. Networks for management activities are also regularly realised by IP.

For a future-looking strategy one has to be open for the introduction of other protocols – complementing and/or replacing the IP suite. Although during the last years several other candidates have been promoted for different areas, the IP suite now seems to be the undisputable winner at the network layer. In fact, as its sheer volume and momentum grows the challenge grows for alternative solutions to take over that position.

Objectives of this article are to place the IP deployment into a commercial context and to outline trends and potential benefits. Some technical aspects will be treated as well, as the commercial opportunities follow from the technical possibilities.

As several papers have described the IP history the following chapter will rather briefly outline some of the events. The intention is to provide a basis for further descriptions in subsequent sections. Chapter 3 gives an outline of the IP suite, giving a general overview of protocols and mechanisms related to IP and found in IP-based networks. Then, current trends for implementing IP-based networks are described in Chapter 4, together with drivers and consequences as seen today. No perspective article would be complete without reflections on what is next to come. This is the topic of Chapter 5; both discussing next steps related to the IP suite and issues that may ask for radical changes or even replacement of IP. An efficient operation of networks and systems requires that past, current and future(s) are linked in a smooth manner. This is treated in Chapter 6, also giving concluding remarks.

2 Take a look over your shoulder – reflecting on the past

2.1 IP key characteristics

Returning to the conception of IP, one of the main goals was to provide a protocol that was able to transport the information to its destination. This was an essential requirement even during failure situations. Hence, emphasis was placed on the eventual arrival at the destination, less on strict real-time delivery. Without entering the discussion of whether this was originally intended for having a network surviving a nuclear attack, a nuclear power plant accident, reaching vessels under various conditions, or some other

 ISSN 0085-7130

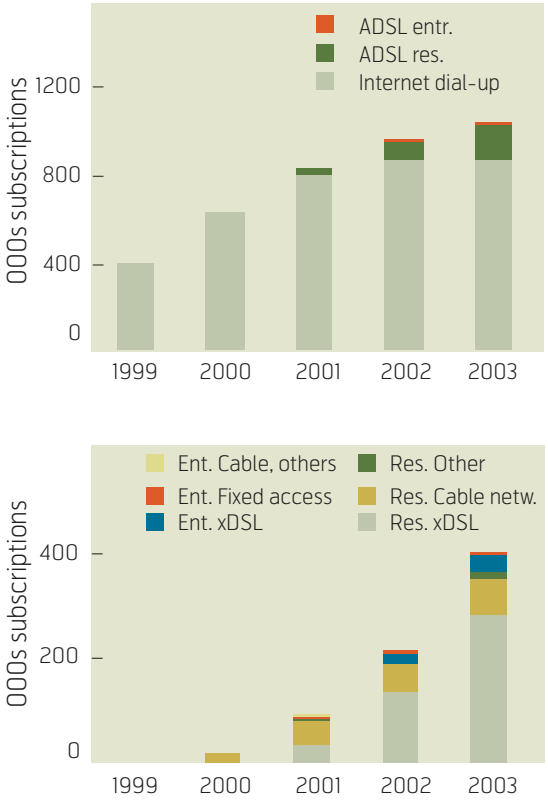


Figure 1 Top – Telenor Internet subscriptions Norway – dial-up and ADSL access (from the Telenor Annual Report 2003); Bottom – Broadband access growth in Norway (from Norwegian Post and Telecommunications Authority)

cases, it seems clear that the robust transport capability was given high priority amongst the requirements.

This is also reflected by the two *key characteristics* that identify the IP format:

- Each packet carries the addresses of its destination and source;
- Each packet can be forwarded individually, independent of routes of preceding and following packets.

2.2 Volumes and industry events

Preparing the foundations in the late 1960s, the IP principles being elaborated in the early 1970s have survived for several decades. The more rapid growth came after the establishment of WWW (World Wide Web) in the beginning of 1990s. The intuitive user interface "click and get" dramatically lowered the threshold for information collection. Emails accompanied the usage, although such message transfer capabilities had already been proposed in the 1960s. However, other phenomena allowed such a growth, like the relative price reduction of computers and the information made available by different sources.

In the late 1990s, however, it seemed as if the telecom industry was becoming too optimistic in their estimates of traffic volumes and revenue potentials in this area. The so-called bubble-burst left several companies out of business and others going through drastic downscaling. This also influenced the progress of IP-related developments. As pointed out in [RFC3869], more recent events in different parts of the world have led to increased insecurity and renewed interest in having dependable telecom infrastructure. The players pulling through the bubble burst also seem to have more sensible expectations to the industry. It is still necessary to join forces in order to make progress in the IP-related areas. This is also a motivation for the organisations and initiatives, such as ITU, IETF, ETSI, 3GPP, DSL forum, Infranet initiative, and so on. In later years a steady increase in the traffic has accompanied the users' interest in the Internet. Figure 1 illustrates the situation regarding Internet subscriptions and broadband accesses in Norway.

2.3 IP on-time to meet a need in service portfolio

Considering the outset and initial motivation, one could raise the question of how this protocol has managed to maintain its position. That position is manifested not only during the transition from a single-class (best-effort) to a number of differentiated classes, but also during the shift from academic and internally controlled environments to commercial operations with a range of players involved.

One explanation for the success of IP may be that it belongs to a simple solution that was presented to the market at the right time. In the early 1990s there were few services around providing information collection and messaging to private persons. Then, the combination of browser and email neatly covered that demand. This was accompanied by more affordable PCs and higher modem bit rates.

The commercial interest in IP came around a time when broadband systems were defined. As such, there was a recognized need for having a system and protocol suite supporting these services. Several of the telecom providers had therefore joined forces to specify B-ISDN (Broadband Integrated Digital Network). In a fairly traditional (telecom) manner one started out by defining services and control regimes. As an outsider to this the IP suite had a simple implementation, requiring neither parallel signalling nor traffic declaration schemes. As we all know, it turned out that the relevant applications were actually well-

matched with the simple best-effort regime offered by the IP suite.

Some may claim that several of the proposals related to IP (see following chapter) are to "repair" the *initial deficiencies*. Hence, one could ask whether it would not be better simply to find a replacement. Still, it is a fact that the IP formats have broadly maintained their interpretation for several decades. In fact, this may well be one of the causes for its popularity; it would be more discouraging to have to adapt to different versions and protocols that were developed and steadily changed.

On the other hand, one cannot neglect all the suggestions that have been made related to improving certain aspects of having an IP network. In retrospect, it seems nice that IP as the core protocol has been left unchanged, and most of the suggestions might be seen as modules that could be added on to the core protocol. Such a *modular-based philosophy* supports the flexibility of proposing new modules replacing others to allow for more services, more efficient network operation, and so forth.

3 The IP suite in a broader perspective

3.1 IP features

The most fundamental IP service is to offer an *unreliable, best-effort, connectionless packet transport* between a source and a destination. The service is called unreliable as delivery of the packet is not guaranteed (no verification of correctness taken care of by IP). Connectionless comes from the fact that each packet is treated independently from other packets in the same flow. Best-effort comes from the fact that no packet is assumed to be discarded or distorted on purpose.

A protocol stack is used when transferring information between two points. IP is commonly referred to the network layer as packets can be inspected in intermediate routers between the sender and the receiver. Version 4 of IP is the one most applied today. This adds a 20 byte packet header to the information, see Figure 2. The packet header contains sufficient description for the packet to arrive at its destination and for the information flow to be reconstructed by the receiver. Some pieces of the header also allow packets to be favourably treated (priority field) and avoid that packets stay in the network following too long routes (time-to-live field).

IP version 6 is also promoted by several sources. A basic argument is that more addresses are supported (128 bits address fields compared to 32 bits for IPv4). However, an effect of this is that the header of IPv6 packets has grown to 40 bytes. On the other hand, the header supports more flexibility and enhanced options.

Error and control messages are incorporated at the IP layer. These allow routers and terminals/hosts to exchange operational and maintenance information. One motivation is to inform that a destination is unreachable, which could be detected by the time-to-live header field being decreased to zero.

3.2 Transport protocols

Above the IP layer, two transport protocols have found their place; one for fast un-acknowledged transfer – UDP (User Datagram Protocol), the other for acknowledged transfer – TCP (Transmission Con-

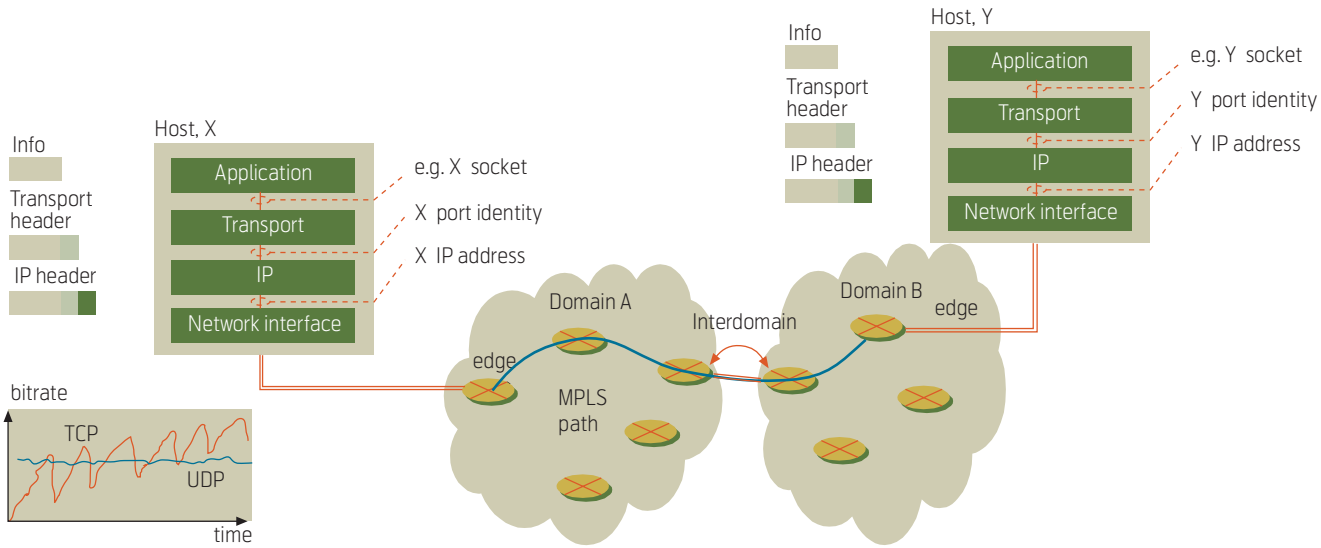


Figure 2 Illustrating IP suite

ISSN 0085-7130 © Telenor ASA 2004

trol Protocol). The former is commonly applied for transferring data having strict real-time requirements, where retransmission for correcting an erred or dropped packet is not relevant as that is assumed to arrive too late. Examples are transfer of real time voice and video. TCP is widely deployed for transfers of data where correct information is requested. Hence, it has been used for file transfers, email transfers, web browsing, and so forth. In order to support this a connection between the end-points has to be established and acknowledgements have to be exchanged.

Due to an initial idea of IP that end-systems are in charge of the information transfer, it is a challenge to find the appropriate transfer rate (kbit/s) for a host/application. TCP has an incorporated mechanism for estimating the proper rate to apply (slow-start and rate adjusting by increasing and decreasing factors). In principle, the inherent TCP mechanism would pursue the proper rate by increasing its transmission rate until a packet is lost (or arrives too late). Measurements have shown that this may result in an oscillating transfer rate. However, a maximum window size can assign an upper limit of the information unacknowledged and thereby a maximum transfer rate. This maximum window size is utilised in some systems, e.g. for UMTS (Universal Mobile Telecommunication System) to avoid the oscillating behaviour of TCP.

3.3 Underlying protocols

The IP layer may utilise several underlying protocols, including PPP (Point-to-Point Protocol), MPLS (Multi-Protocol Label Switching), and Ethernet. In fact, one of the advantages of IP is its ability to make use of different link protocols. In the core network, MPLS appears as a current main candidate. This protocol assists to separate traffic flows by offering a tunnelling mechanism. Hence, MPLS tunnels can be managed by intermediate nodes relieving the IP routing and processing capacity. Commonly MPLS is used to implement VPN (Virtual Private Network) services.

Several people have investigated relations between IP and high-capacity systems, e.g. optical networks. This deals with questions like:

- Architecture, e.g. should the IP layer and the optical layers interact as peers exchanging information as between equal parties, or should IP simply be an overlaying client?

- Which framing protocol should IP use – if any – as container when carried by an optical network?

- Should redundancy and protection schemes on IP and optical layers co-operate or carry out their actions independently?

3.4 Routing and naming

Different routing protocols are often applied – with a division between the domain-internal protocol and the protocol used between different domains. Here, a domain may represent a set of routers (a single operator may define a number of domains). A main difference between the intra- and inter-domain routing protocols is the detailed information exchanged. Between domains this information is limited not to reveal too much on the domain internal structure. In addition, limiting the information also results in less data to be exchanged and stored in the routers. Between domains, BGP (Border Gateway Protocol) provides mechanisms for aggregating routes and ranges of IP address prefixes. This supports efficient connection between networks.

There is also a distinction between routing and forwarding; the former used to describe the exchange of routing information (by routing protocols), and the latter used for sending IP packets according to the routes. When MPLS is used as tunnel, the IP forwarding is by-passed, allowing those packets to travel along different paths to what the routing would tell. Today, this mechanism could be used for load-balancing a network.

There are different sets of addresses and identifiers in a network. The IP address gives a source or a destination referring to the IP layer, see Figure 2. Then, transport protocols, e.g. TCP or UDP give a port identity. Between a transport layer and an application, a socket identifier is commonly applied. This will then point to the application, such as a file transfer application. A hierarchy of applications can also be used; for example the web browser can invoke the file transfer application. An example of identification at the application layer is URI (Universal Resource Identifier). A DNS (Domain Name System) service is applied to translate between URI and IP address, such as from www.telenor.com to 129.134.11.24. Then, clicking on a link (described by URI) can be converted to an IP address (and the server) where the information is found.

3.5 Traffic handling

Different traffic handling mechanisms are also defined for IP. DiffServ (Differentiated Services) is a way of defining a number of traffic classes where IP packets in the different classes receive different priorities. The result is that high priority IP packets may be forwarded before low priority packets. An argument is to give flows with strict real-time require-

ments, e.g. voice, higher priority to limit the delay, and delay variation. The mechanisms also mean that packets must be classified, the corresponding flows monitored and queueing disciplines have to be implemented.

To fully support ensured services, however, additional mechanisms must be installed. Examples are admission control and resource reservation. Commonly statistical guarantee can be promoted where some of these mechanisms are more loosely applied.

3.6 Related protocols

The traffic handling is motivated by combining high resource utilisation at the same time as requirements from applications are fulfilled. Typically, these mechanisms are fully deployed on the edge of the IP network. An example is the service edge routers where IP packets are inspected after being generated by a DSL (Digital Subscriber Line) user. PPP may for example be utilised as security means for each user – offering a tunnel between the residential and the service edge router. To provide an IP address and authentication of the user equipment, RADIUS (Remote Authentication Dial-In User Service) could be used between the service edge router and an authentication server. Other protocols are also proposed for service control and management, such as SIP (Session Initiation Protocol) and COPS (Common Open Policy Service). SIP, including some enhancements, has currently much support for controlling services for both fixed and mobile/wireless access.

A quick visit to the IETF web pages gives an impression of topics that occupy those who are engaged in that forum. However, there are also other organisations working on essential aspects of running IP-based operations. Examples are requirements described by ITU and implementation and interoperability issues described by a number of fora such as the DSL forum.

4 Current implementation trends and consequences

4.1 IP central in all strategies

Every major operator is currently deploying broadband – and narrowband – services utilising IP. Most operators see IP as an essential element in their network and system development. This should therefore be carefully planned and implemented as part of the strategic plans, not as an isolated activity.

Looking more closely at the operators' strategies, we see that a wide range of services utilising IP are to be provided. Examples of services are TV broadcasting, video-on-demand, voice/telephony, Internet access, VPNs. There is also a clear trend that IP is introduced in all major network areas. Besides the wired/fixed network, bodies working on mobile systems and broadcasting systems introduce IP to take a growing position.

The steady growth of IP-based networks is motivated by a *number of drivers*:

- New applications and user groups; these place additional requirements on network capabilities. One example is that low-revenue users also want access to networks and information, requesting sponsored or low-cost solutions.

- New technologies; more technical solutions are developed related to IP-based networks, supporting efficient configuration and operation as well as supporting more user demands and application types.

- Increased load and network expansion; increasing the broadband offering in a region commonly asks for more distribution of equipment. For example, when the access rate on copper lines increases, shorter lines are often used implying that equipment is located closer to the users.

- Higher network dependency and more providers involved; steadily more actors are depending their businesses on available IP network services. The up-time requirements are therefore becoming stronger. At the same time, the number of players involved in some of the services is growing, requesting arrangements between the players to sort out duties and responsibilities.

This could be taken as proof that IP has matured from simply a best-effort solution to being able to support business-critical applications and differentiation of services.

4.2 Consequences accompanying IP suite

Broadly, one could recognize two *"schools of thought"* driving the IP development: i) those wanting to introduce ensured and differentiated services, and, ii) those wanting to keep a single IP transport service class. On a generalized level, this could also be seen as the battle between keeping most of the service control (referring to IP transport service) within the network as for school i), and keeping the control in the end-system as for school ii). In case school ii) should be the right one to follow, this may be argued on the grounds that basic IP transport services are the only ones asked for. That is, these services must be

incorporated into the portfolio – a sub-argument is that faced with possible performance bottlenecks, more capacity should be added. Some may also claim that this is less expensive than implementing advanced traffic handling mechanisms. However, most providers lean towards school i). They mainly apply the following arguments:

- Differentiation: a range of different users and applications is to be supported. These have different requirements (such as delay, loss and availability), intuitively asking for different services.

- Cost reduction: being able to support differentiated services in a common physical network allows for cost savings. That is, it should be less expensive to invest in and operate a common network supporting a range of services, compared with having different physical networks. A principle scale argument as known from traffic studies supports this argument. Applying traffic theory also shows that better resource utilisation can be achieved when service differentiation is implemented.

- Support on new service packages: having a common IP network should alleviate developing and offering of so-called converged services. That is, service packages crossing different traditional network accesses are more swiftly available with a common IP network as basis. The cost and speed of service offering is an important argument. In several of today's incumbent operators' system portfolio, the level of complexity due to interaction between different systems seems to dramatically slow down the service development speed.

Summarised, two main groups of argument are driving the IP deployment from an operator's perspective; the *cost saving* and *the option to offer new service packages*. The former is further backed by the observation that "IP type of equipment" has been falling faster in price compared to others. This is expected to continue; driving deployment of IP into new areas of telecommunications. The second, new service argument is also a trend expected to continue. In some respect services currently provided by other platforms are implemented on the IP platform, e.g. voice/telephony. There are also more services added; on-demand, gaming, multi-party messaging. Even if some of these are introduced by other access forms (e.g. mobile), the network-internal implementation is commonly based on IP.

The availability of an "IP interface" also lowers the threshold for other parties to offer services. This is therefore a driver for other providers to get access to an IP-based interface, being able to provide the services to users in an easy manner.

From the outset, users might not care about technical implementations. However, the Internet and IP philosophy has also accompanied the wide deployment of personal devices such as mobile handsets and computers, as well as other devices. Therefore end-users also benefit from the wide spread of IP-enabled devices. Naturally, security mechanisms must also be installed avoiding non-authorized intrusion and revealed personal information.

4.3 Some technical deployment trends

In most IP core networks, MPLS is utilised as a tunnelling mechanism. There is also a push towards introducing DiffServ for service priority and efficient resource utilisation. Underlying Ethernet interfaces are argued for because of cost-savings. The overall principle is then a common physical network allowing for virtual separation of traffic classes. The shift from IP version 4 to version 6 has not found a definite answer in Europe and USA. One cause is that the address space is not that restricting in these regions. There are growing interests in finding beneficial solutions which can explore capabilities of IP version 6 – other enhancements supported in addition to greater address space.

For service and session control, several international bodies have promoted SIP. Similar solutions have been promoted for future releases of mobile systems.

On the access side, alternatives to PPP are investigated. One argument is to arrive at functionally similar solutions for all access forms, whether based on cable or air.

Incumbent operators evaluate how their network portfolio should evolve. 4–8 years ago there seemed to be a hype that every service would be based on IP fairly soon. These days, however, a more healthy view is observed; that only services efficiently realised on IP-based networks should be produced in such a way. This does not however pose too many obstacles for the rapid growth of IP traffic. Still, to satisfactorily meet the accompanying requirements, more solutions must be implemented – some described in the following.

5 So, what's next?

5.1 Overall issues

As part of the strategy work, it is always reasonable to ask questions as to which trends that dominate the evolution, as well as to which phenomena are challenging the current way of running the operation.

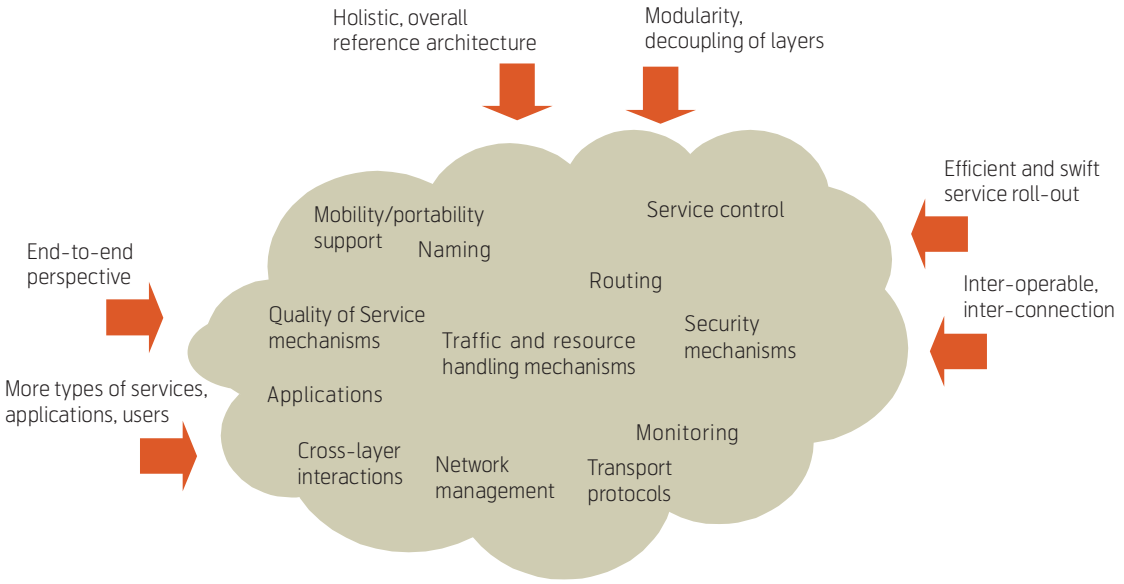


Figure 3 Samples of some topics and mechanisms to be addressed during further work on IP-related development

Therefore, it is sensible to look at both the *evolution within the IP suite* and which factors may lead to *replacements of parts of the IP suite*.

Addressing the former question, work is on-going in several organisations. Examples are the next generation networks activity in ITU and the "all-IP" target in 3GPP, ref. Figure 3. IETF is also pondering on issues that should be researched to ensure a sane evolution of the IP suite.

Considering that so-called next generation networks are based on IP, a number of related trends and corresponding challenges are:

- Holistic perspective: Providers want to offer more "advanced" services on IP. This implies that a reference architecture must be in place. Finding which components to apply for the different modules in the architecture has not been completed. Protocols and mechanisms elaborated in IETF must be put together in an efficient manner. As part of this area, interworking with other networks has to be solved. However, it is essential that the IP network does not import the complexity such interworking could lead to.

- Separation, or decoupling, of control and bearer/transport functions: This allows for defining service platforms to some extent independent of the underlying IP network. A clear interface definition has to be in place, though. The decoupling of bearer and service control supports a modular-inspired architecture. One of the service platform candidates is a solution developed by 3GPP; IMS (IP Multi-media Subsystem). This solution has been promoted by several organisations and operators to potentially offer services on most access types, both fixed and wireless. This requires that requested functionality is supported by the IP-based network.

- Offer a wider range of services and applications: The types of applications assumed to be supported by an IP-based network are steadily growing wider. Real-time and business-critical applications are placed on the network in addition to web browsing, email and file transfers. This intensifies the dependability, performance and security aspects. Features such as multicast, mobility, unified access for users, community networking, and so forth, must be supported at IP layer for efficient realisation.

- The end-to-end perspective: Considering the applications, an IP network is utilised for more demanding purposes. For commercial operations, the users require ensured service levels to justify the payments involved. For example, breaking a real-time video presentation is not acceptable when it is paid for. Hence, mechanisms for ensured service delivery is required. This implies invocation of traffic and QoS management.

5.2 Some technical issues

On the more technical level, several issues have been identified, e.g. ref. [RFC3869], [ITU_IP]. All of these have relations with the four areas i) overall architecture/functionality, ii) traffic/performance, iii) security, and, iv) network evolution. The efficiency and

scalability must also be taken into account. Some of the issues to address are:

- Naming: allocation of IP addresses, domain names, and others (e.g. ENUM, Electronic Numbering), has to be co-ordinated on a global level. Discussions are on-going including ITU, IETF and others. A transition from version 4 to version 6 of IP is also considered.

- Routing: Protocols for routing may need to be enhanced or replaced considering the new needs emerging from commercial operations. Examples are more routing metrics including business matters, faster routing convergence to recover from unexpected events, interconnections between domains, multi-homing of devices, and mobile/ad-hoc environments.

- Network management: The growing traffic and users require that a (logical) network perspective must be supported for efficient management. Otherwise, the complexity cost would likely outgrow the scale gain. This also includes monitoring and measurement arrangements. How and what to measure are essential questions – balancing the accuracy and scalability aspects. Management and control of customer equipment would also be requested allowing end-users to be relieved of intricate configurations. Management of any intermediate system (sometimes referred to as middleboxes) should also be supported. Autonomous procedures would likely be efficient in this manner, naturally also asking for performance and security aspects to be solved.

- Quality of service: How the different mechanisms actually influence the real end-user experience is not understood. The user sees the effect of the total system, asking for understanding of the overall architecture, likely involving several domains and actors and different combinations of equipment configurations (such as queueing disciplines).

- Efficient resource utilisation: An example is congestion control mechanisms. In-built features of TCP have been utilised to avoid congestion. New applications and new environments (e.g. wireless) demand other control mechanisms. A new transport protocol is under development, DCCP (Datagram Congestion Control Protocol). Besides such a protocol, information exchanges across layers could also be utilised. For example, intermediate IP routers could signal load levels to the end-systems.

- Application capabilities: Higher requirements are placed on intuitive user interfaces. Depending on the services provided by the IP network, easily comprehended feedback to users can also be required. An example is to inform a user that a load situation somewhere (within the user network, access network, server side, etc.) results in a degraded service level.

5.3 Challenging IP's position?

Then, one may ask whether there are any candidates to replace IP. It is a broad opinion these days that no strong candidate exists. On the other hand, a number of issues may motivate for not using IP in all circumstances:

- Lack of holistic overall architecture. If such an architecture is not defined, solutions will likely lead to inefficient configurations.

- The overhead added by the IP protocol suite. In case the protocol features are not utilised, low capacity links do not favour this overhead.

- No coherent mechanisms for ensuring services. The strictest requirements place rather tough demands on the network solution, where other networks currently may meet these demands in a more efficient way.

- Security matters may inspire for increased use of "hidden" identifiers. That is to avoid that attacks and information not asked for can be distributed.

... and, in principle, any other solution that allows for more cost-efficient operation offering timely services. The combination of inexpensive solutions and easy installation has an urge of deciding the winning technologies. The sheer volume of IP-based devices and applications, however, make if difficult to believe that any alternative would gain significant position in the medium term. Naturally, increased traffic volume, simple transport network functions and treatment of aggregated traffic mean that the IP level may not be inspected by all intermediate nodes.

6 Present questions – bridging past and future

6.1 Immediate challenges

Based on the current status for commercial operation of IP-based networks, there are several challenges to face in the short term in order to prepare for the future. These challenges are mostly formulated from a network operator's perspective. In broad terms they are grouped into:

- *Business* concerns: Although perhaps provocative, it is fair to say that few of the major operators have really incorporated processes and systems for swift offering of IP-based services. Several of the newcomers seem to be much more streamlined, though. These processes refer to every phase from defining customer-related systems, deciding on pricing structures, allowing bundled services, supporting self-service variants, and so forth.

- *Service* concerns: Communicating what service is actually provided seems to be challenging. SLA (Service Level Agreement) could be elaborated correspondingly. This also asks for monitoring apparatus documenting that SLA conditions are met. Offering more value-added services is also expected to be on the operator's wish list. Then, it is necessary to integrate service control and management. A particular aspect is that several systems and players may contribute to complete the service delivery. Additional mechanisms are also required to allow this to take place in a controlled manner.

- *IP transport* concerns: Interplay between different systems, within and partly outside an operator's domain requires (standardised) protocols. In some areas these are in place, such as routing between domains. For interacting with customer equipment, several options are currently proposed. Configuring IP resources is also a challenge on the same level, in particular in view of multiple services assumed to be supported. Monitoring is a challenge on the IP network layer. It is not much of a problem defining a monitoring scheme collecting huge amounts of data. But to define a scheme, which captures the traffic and performance in an efficient way, is still an unresolved issue. Defining and tuning traffic handling mechanisms is also a necessity to deliver ensured services.

6.2 … and, in summary

The steady growing presence of IP in almost all telecom operations is evident. No matter which access networks and services that are looked at, IP-based implementations are seen in the current or future roadmaps. This *IPfication* is an expression of the packet-oriented implementations of telecom services. No operator/provider should neglect this trend and proper steps should therefore be prepared in a timely manner. A basic argument for an incumbent operator is that these steps should also assist when reducing its internal complexity, which likely has grown for several years. That is, the *IPfication* allows a drastic cleansing of an operator's system portfolio. This implies reduced cost base and more swiftly service offerings. The whole telecom industry is currently very occupied with finding means to lower costs and allow for enhanced service portfolio. Society as a whole would also benefit from the lower cost base and improved service offerings. In fact, the technical solutions fuel the so-called global information and knowledge-based society as defined by United Nation's bodies.

Major challenges rest within areas of defining an overall reference architecture, traffic and resource handling, security and efficient evolution schemes. Target reference architecture allows for describing modules that should be in place to support service offerings. The modular architecture corresponds with trends of introducing open interfaces and decoupling different layers. This is also a competitive edge considering the growing heterogeneity and dynamics foreseen for service environments.

In some respect the *IPfication* could be considered as Internet philosophy brought into a commercial context. Here, a principle of "good enough" seems to be dominating. That is, not too much, neither too little. What is then "good enough" in a more specific interpretation? Well, that is a multi-billion question continually examined by most players involved.

References

[ITU-IP] International Telecommunication Union. *ITU and its Activities Related to Internet-Protocol (IP) Networks, ver. 1.1*, April 2004. Work in progress. [online] – URL: www.itu.int

[RFC3869] Atkinson, R, Floyd, S. *IAB Concerns and Recommendations Regarding Internet Research and Evolution*. IETF Request for Comments 3869. August 2004. [online] – URL: www.ietf.org

Dr. Terje Jensen (42) is Senior Research Scientist at Telenor Research and Development. In recent years he has mostly been engaged in network strategy studies addressing the overall network portfolio of an operator. Besides these activities he has been involved in internal and international projects in various areas, including network planning, performance modelling/analyses and dimensioning.

terje.jensen1@telenor.com

Personal communication fabrication in the Lyngen Alps

NEIL GERSHENFELD AND MANU PRAKASH

Neil Gershenfeld is the Director of MIT's Center for Bits and Atoms

The 100th anniversary of Telektronikk is an auspicious time to look back at Telenor's pioneering role in providing access to telecommunications, and look ahead to the possibility of expanding the availability of not just communication technologies but also the means for their development.

An early goal of telecommunications was universal service. Access was seen as an essential enabling tool for participation in a modern economy, and codified in legislation such as the 1934 Communications Act in the US. This was followed by activities that have sought to provide access not just within but across economies, a vision that has been pioneered by Telenor's support for Grameen Phone in Bangladesh. This provides communications via village-scale micro-businesses selling time on a mobile phone, showing that a self-sustaining business model can bring advanced technologies to where they are needed in some of the least-developed places in the world. Now, users are beginning to participate in deploying their own telecommunications infrastructure via mesh wireless networks. Commodity 802.11 access points and radios can be modified to serve as low-cost routers, and high-gain antennas can extend their range to tens of kilometers. Such a system still requires a conventional carrier for the long-haul links, but it does expand not just access but also deployment of the local loop when the economics might not otherwise be favorable (for either the user or carrier).

Manu Prakash is a Master's Candidate at MIT and a research assistant at MIT's Center for Bits and Atoms

An example of this is Telenor's "Electronic Shepherd" project [1] in the Lyngen Alps (Figure 1). Here, far above the Arctic Circle, the goal is to assist the traditional practice of Sami nomadic herding with nomadic data. Driven by changes in habitat and land use, aims include tracking the animals and their predators, monitoring their health, bringing them down when needed, and providing continuity of information throughout the animal products supply chain. Components of the system include short-range wireless tags for sheep or reindeer, GPS "bell" tags for the lead animals, and tag readers at salt licks interfaced to a multi-hop 802.11 network.

While this activity is on the current frontier of community-scale communications, it is also looking beyond access, and deployment, to an even greater possibility: the local development and production of advanced telecommunications technologies. It is now possible in the lab to print three-dimensional structures, displays, sensors, actuators, and logic [2]. Putting all of these capabilities into a single printer promises to create one machine thay could make any machine, including itself. The relationship between such a personal fabricator and today's industrial machine tools is analogous to that between a personal computer and a mainframe, bringing the programmability of the digital world to the rest of the world.

Figure 1 The Electronic Shepherd project

ISSN 0085-7130 © Telenor ASA 2004

While this is a long-term research goal requiring many years of work to come on the enabling materials and mechanisms, it is possible today to approximate the functionality of a personal fabricator with a small set of parts and tools. Two things make this possible. First is the intersection of consumer electronics with machine tools. The metrology that allows a CD player to position a read head to a micron can do the same for a cutting tool. Inexpensive table-top NC milling machines can now provide tenth-mil (0.0001") positioning resolution and use few-mil tooling, making it possible to fabricate features down to micron scales. Applications include subtractively machining surface-mount circuit boards, along with making associated electromagnetics, packaging, and interfaces. And second, a $1 microcontroller can have a clock cycle shorter than 1 us, which is fast enough for microcode implementation of communication modulation, instrumentation signal chains, and video drivers. Together, the accessibility of microns and microseconds means that with a few thousand dollars in capital equipment and a few dollars in components it is possible to design and build complete communicating systems.

These capabilities are being deployed by MIT's Center for Bits and Atoms in field "fab labs" in locations ranging from inner-city Boston to rural India, to coastal Ghana, to the Lyngen Alps. The goal of the fab lab project is to bring into the field today the tools that will eventually be integrated in tomorrow's personal fabricators, in order to understand why and how they will be used. The initial embodiment uses a collection of commercially-available materials and machines, connected by software developed to implement engineering workflows. Instead of a formal curriculum, the educational model is "just-in-time", based on project-based peer-to-peer training. Instructional materials and project documentation are generated by the users as they learn, and shared through a collaboratively-edited Web resource [3]. This can be thought of as open-source approach to hardware as well as software.

As an example, high-gain 2.4 GHz 802.11 antennas are essential for the Electronic Shepherd project to bring data down from the mountains. Because of the rough terrain many antennas are needed to route around obstructions, representing a significant cost at commercial prices. Furthermore, the links vary in their need for gain vs coverage, requiring a range of antenna types. Hence one of the first fab lab projects there is developing antennas that can be locally produced and characterized.

The design is based on micropatch arrays (Figure 2) [4–6]. Their development requires sophisticated elec-

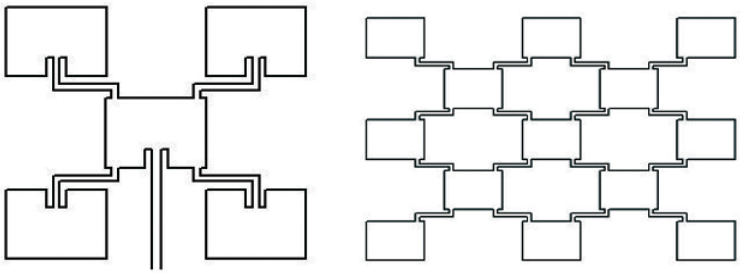


Figure 2 Edge-fed 5-patch and series-fed 13-patch arrays

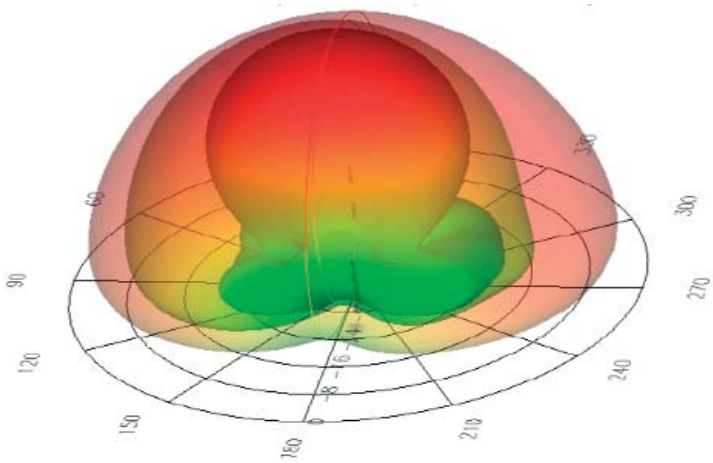


Figure 3 Calculated normalized radiation patterns for 1, 5, and 13 patch series-fed arrays

	Angle (deg)	Directivity (dB)	Gain (dB)
1 patch	175.68	6.12	6.12
5 patch	69.45	10.15	10.14
13 patch	25.17	14.56	13.62

Table 1 Calculated dependence of the radiation divergence angle, directivity, and gain on the number of patches

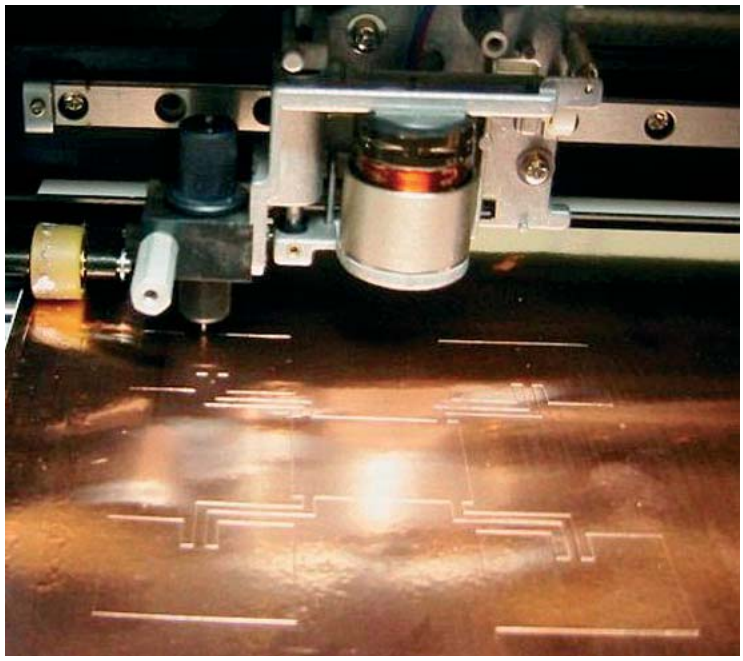


Figure 4 Plotting a micropatch array

ISSN 0085-7130 © Telenor ASA 2004

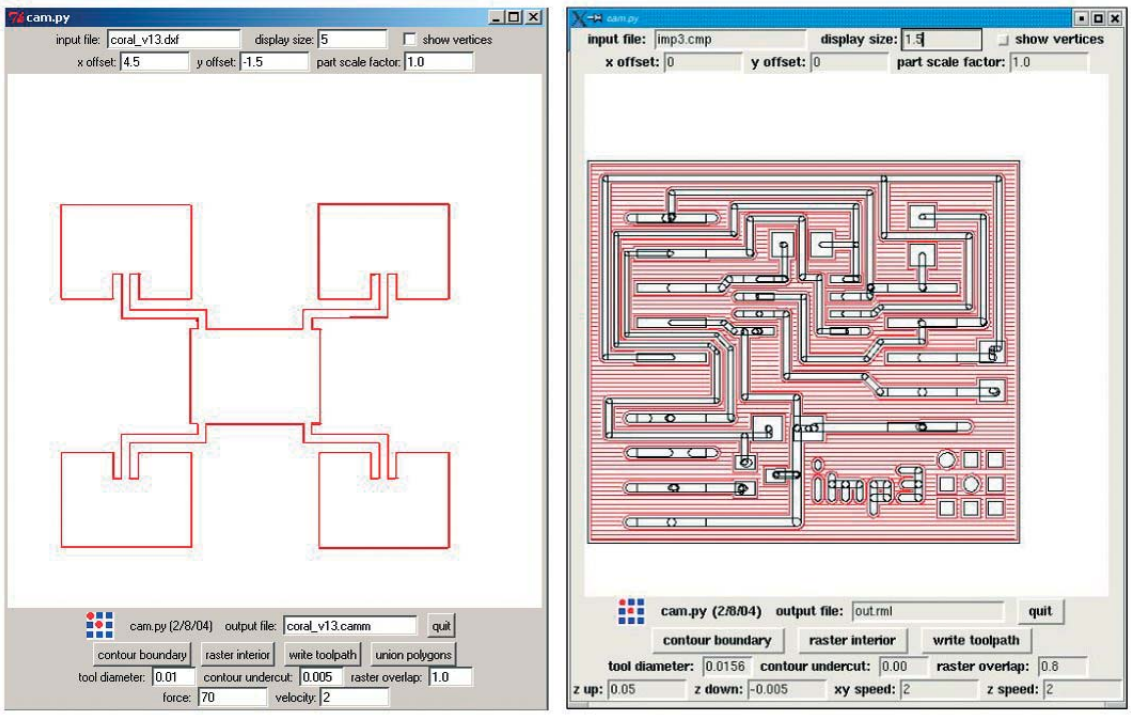


Figure 5 Toolpath generation for micropatch cutting and PCB milling

tromagnetic modeling, and they must be fabricated with tolerances that are a small fraction of the wavelength (sub-mm for 12.5 cm 2.4 GHz radiation), but they can then be made out of inexpensive copper roll stock, and the antenna gain is easily chosen by varying the number of patches (Figure 3). For larger arrays a series-fed design was used to minimize resistive losses from fanout in the impedance-matching feed network, and the antenna performance was modeled with the method of moments (Table 1).

The patch arrays and their feed lines are cut out with an inexpensive computer-controlled sign cutter (Figure 4) [7], using an application that was developed for toolpath generation across fab lab processes (Figure 5).

Copper-clad low-loss RF substrates are expensive, but with this process the materials choice is greatly simplified because the antenna can be applied to a substrate with a transfer adhesive (Figure 6). We chose window glass as an inexpensive easily-available material with acceptable performance.

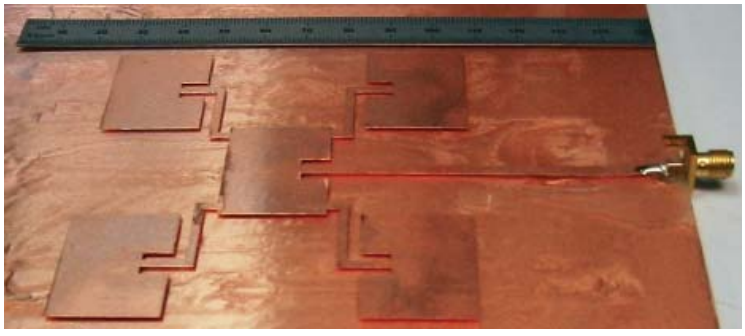


Figure 6 Assembled micropatch array and feed line

For a candidate substrate material it is necessary to measure the dielectric constant to size the patches, and the parallel resistance to evaluate the loss. These parameters can be determined from the complex frequency response, or in the time domain from the step response to an applied voltage. The step responses are measured by an inexpensive microcontroller with single-cycle RISC instructions [8]. This has a relatively slow A/D (~10 kHz) but a fast sample-and-hold amplifier (<1 us), which can be used to undersample the step response by iteratively applying a charging voltage, waiting a variable number of clock cycles, sampling, and converting (Figure 7). The capacitance and the series and parallel resistances can be found from the initial value, initial slope, and asymptote of these curves, and the measurement parasitics determined and removed by varying the electrode size.

The circuit board is made in the fab lab by subtractive machining rather than wet etching, to avoid chemical waste and minimize consumables. A table-top milling machine with sub-mil metrology [9] is used with a 1/64” four-flute short-shank center-cutting end mill [10], to make features within a 0.050” surface-mount pitch (Figure 8). Toolpaths are generated with the same application used for the micropatches (Figure 5).

The reflected power S11 for a 5-patch array is plotted in Figure 9, showing the primary antenna resonance. The materials cost for this antenna is ~$1; a corresponding commercial antenna costs ~$100–200. The substrate characterization instrumentation uses ~$2 in parts, and the equipment to produce everything costs ~$6k. The incremental fabrication cost per antenna could approach the materials cost with roll-to-roll stamping, but the real value of this production process is its flexibility and locality rather than its throughput. Each antenna made can be optimized for the link where it will be used, and a similar workflow can be used to develop and produce other parts of the system, including the animal radios and packaging, and embedded network nodes.

Over time, the goal of the fab lab project is to replace the commercial components with tools that can be made in the lab, until eventually the labs themselves are self-reproducing. The current capabilities of the Lyngen fab lab can be thought of as being comparable to the minicomputer era in the evolution of computation. Their cost and complexity are on a scale appropriate for use by a workgroup, but like minicomputers that already makes them accessible for demonstrating and developing applications driven by the needs of individuals instead of industries. Rather than being on the remote receiving end of technologies developed elsewhere, Sami herders in the Lyngen Alps and their counterparts around the world can create local solutions to their local needs, leading a global revolution in personal fabrication.

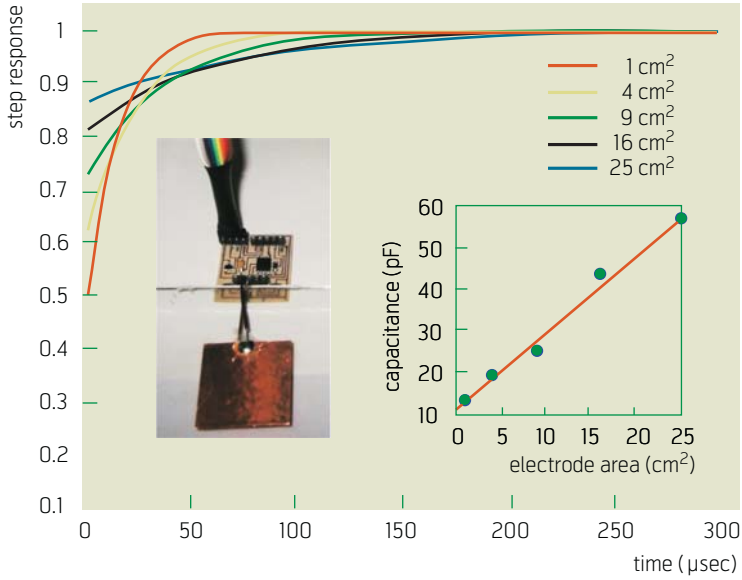


Figure 7 Step response measurement of the substrate dielectric constant

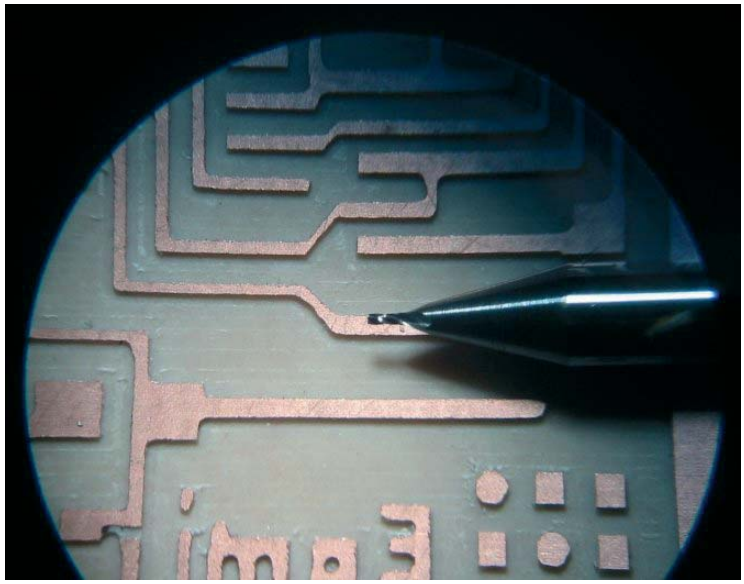


Figure 8 Subtractive milling of the impulse response instrumentation

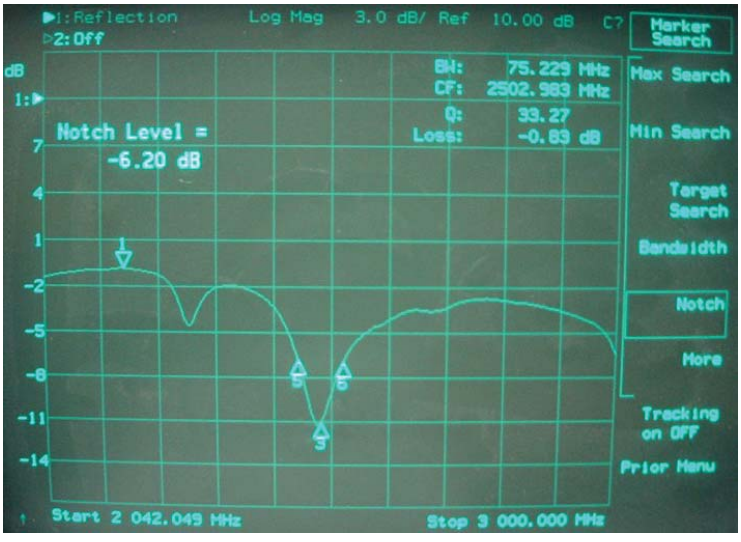


Figure 9 Reflected power for a 5 patch array

Acknowledgements

This project grew out of collaboration with Haakon Karlsen Jr. from Solvik Gård, and Bjørn Thorstensen and Tore Syversen from Telenor. Preliminary antenna designs were developed by Matt Reynolds and Rehmi Post. This work was supported by the Center for Bits and Atoms (NSF CCR-0122419).

References

1 Thorstensen, B et al. Electronic Shepherd – a Low-Cost, Low-Bandwidth, Wireless Network System. In: *Proceedings of the 2nd International Conference on Mobile Systems, Applications and Services – MobiSys*, Boston, MA, 2004, 245–255.

2 Ridley, B, Nivi, B, Jacobson, J. All-Inorganic Field Effect Transistors Fabricated by Printing. *Science*, 286 (5440), 746–749, 1999.

3 *The Center for Bits and Atoms*. August 30, 2004 [online] – URL: http://fab.cba.mit.edu

ISSN 0085-7130

4 Legay, H, Shafai, L. Planar resonant series-fed arrays. *IEE Proceedings Microwaves, Antennas & Propagation*, 144 (2), 67–72, 1997.

5 Balanis, C A. *Antenna Theory: Analysis and Design*, 2nd ed. New York, Wiley, 1997.

6 Pozar, D M, Schaubert, D H. *Microstrip Antennas: The Analysis and Design of Microstrip Antennas and Arrays*. New York, Wiley-IEEE Computer Society Press, 1995.

7 *Roland CX24, CAMM 1 desktop vinyl cutter*. October 19, 2004 [online] – URL: http://www.rolanddg.com/products/cx2412.html

8 *Atmel AVR ATtiny15L*. October 19, 2004 [online] – URL: http://www.atmel.com/dyn/general/tech_doc.asp?doc_id=7222

9 *Modela MDX20*. October 19, 2004 [online] – URL: http://www.rolanddg.com/products/mdx20a.html

10 *SGS Tools product number #30101*. October 19, 2004 [online] – URL: http://www.sgstool.com

Prof. Neil Gershenfeld is the Director of MIT's Center for Bits and Atoms. His unique laboratory investigates the relationship between the content of information and its physical representation, from molecular quantum computers to virtuosic musical instruments. Technology from his lab has been seen and used in settings including New York's Museum of Modern Art and rural Indian villages, the White House/Smithsonian Millennium celebration and automobile safety systems, Las Vegas shows and Sami reindeer herds. He is the author of numerous technical publications, patents, and books including "When Things Start To Think", "The Nature of Mathematical Modeling", and "The Physics of Information Technology", and has been featured in media such as The New York Times, The Economist, CNN, and the McNeil/Lehrer News Hour. Dr. Gershenfeld has a BA in Physics with High Honors from Swarthmore College, a Ph.D. from Cornell University, was a Junior Fellow of the Harvard University Society of Fellows, and a member of the research staff at Bell Labs.

email: gersh@cba.mit.edu

Manu Prakash (24) is currently a Master's Candidate at Massachusetts Institute of Technology and a research assistant at Center for Bits and Atoms at MIT. Manu is passionate about technological advances for developing countries. He believes the hardest technological challenges lie at the most remote places in the world. Manu is an active member of several organizations including Association for India's Development, Thinkcycle and Design that Matters. Before joining MIT, Manu started an educational program in Indian schools; to instill the passion of tinkering with technology; titled "Build Robots Create Science". He was recently awarded the Ideas Prize at MIT on his work on Battery Vending Machines.

Manu's research interests span from physics of computing, microfabrication techniques and RF design. He holds a Bachelor's Degree in Computer Science and Engineering from Indian Institute of Technology, Kanpur.

email: manup@mit.edu

Perspectives on the dependability of networks and services

BJARNE E. HELVIK

Bjarne E. Helvik is Professor at the Norwegian University of Science and Technology

We are likely to meet a future where we rely on an increasingly wider range of information and communication services in our private, social and professional life. In addition to what we are familiar with today, some services will be "invisible" and provided by an ambient intelligent network. A part of the services will not be critical, while a continuous functioning of others will be mandatory for our productivity and well-being. All of these services are intended to be provided by one integrated communication infrastructure. This requires that attention is paid to availability, continuity of service, safety, i.e. dependability issues, in its design and operation. The objective of this paper is to identify some challenges and indicate tentative developments to cope with these dependability issues. One major class of challenges is posed by the steadily increasing complexity and heterogeneity with respect to services as well as service requirements, the technical installations, and the broad specter of parties providing these services. A foreseen development to deal with this is towards autonomy in (re)configuration, interaction between network entities and fault management. Another major class of challenges is posed by the necessity to have a stable core transport network and service provision. Within these areas, a change of technological generation has started combined with a merge between "Internet technologies" and "traditional telecom technologies". These developments are outlined and discussed.

Many foresee a future where we will have a "digital" existence in cyberspace that will be an integrated part of our private, social and professional life. At the same time, we will be surrounded by intelligent networked devices that will ease and support our living. Even if this future scenario does not become true in its full extent, it is beyond doubt that our personal, social and economic welfare is and will become increasingly more dependent on information and communication technology. However, can we trust our welfare to this technology? What means may be taken to ensure the trustworthiness of services delivered and at what cost?

1 Introduction

It is beyond the ambition of this paper to give comprehensive answers to the questions outlined above or to discuss the issue to its full extent as illustrated in Figure 1. I will focus on the system issues and the characteristics of the service provided and not go into detail, neither on the dependability of the hardware, software or "humanware" components that go into the system, nor into the economic, social and "human" impact of failures. What will be done is to outline what are the threats to the trustworthiness, what will be the challenges in future systems and to discuss the shift in technical solutions. The paper consists of two parts, which may be read separately. The first, general part in Chapter 2 will discuss the ongoing evolution of networks and telecommunication services with respect to technology, provision and use. History has shown that dependability must be built into the systems and cannot with success be added afterwards. The current "ruling" technologies, the Internet and SPC telephony had this in mind. Do we today make the correct choices for the next generation networks and service provision? A number of challenges are posed by the increasing technological and operational complexity, the approach toward ambient intelligence, as well as the integration of services with widely differing requirements in the same system. The second part of the paper, i.e. Chapter 3, focuses on technological issues in the core transport network and 'centralized' service provision. The dependability of these parts of the network is of great

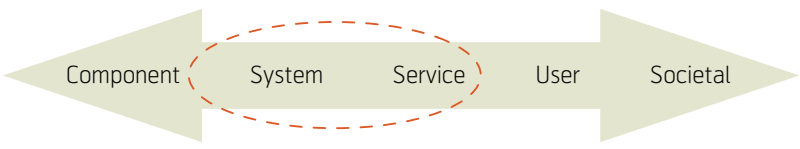


Figure 1 The span of issues related to dependability is wide. The paper concentrates on the system and service levels, having the failures of HW, SW and "human" components in mind as well as the requirements of and implications on the users and society

ISSN 0085-7130

Figure 2 Dependability ontology based on concepts and terms from IFIP WG10.4

importance since the consequences of outages are large. A high dependability (or synonymously robustness, resilience, survivability or fault-tolerance) is in the current systems typically ensured by a high usage of dedicated spare resources and special designs. We have a trend toward more dynamic and flexible use of spare capacity, less special design, and at least for the transport network, a less complicated redundancy structure and fault management. Can we trust these solutions in the same way as the proven technologies? Before we come to these issues, it is necessary to introduce some concepts and give some background.

1.1 What is dependability?

In the above section the term trustworthiness is used to describe the property we are seeking. Dependability is a term which is adopted in standardization and elsewhere in order to define all aspects of a system related to this property, cf. for instance [15, 10]. Dependability may be defined as the trustworthiness of a system, in the sense that we (justifiably) may rely on the services it delivers to its users. A user may be another technical or human system, as well as an end user which interacts with the former, i.e. definition wise, the concept of system and service is used wider than indicated in the levels of Figure 1. For a more

comprehensive definition and introduction to dependability concepts, it is referred to [3].

A top level ontology of the dependability area is shown in Figure 2. From this figure we see one of the reasons for introducing a new term, the common term *reliability* has also been given specific technical meanings, the ability to provide uninterrupted service and the probability of doing so for a specific period. Hence, the new term to avoid ambiguity. Dependability consists of:

- ***Threats*** or impairments to dependability: faults, errors and failures, as well as their causes, consequences and characteristics. These may arise from physical imperfections in the system, physical influence and damage from the outside, human-logical faults made during specification, design, development, installation and operation, etc. Threats may also be intentional and hostile. The latter kind is commonly regarded as security issues.

As illustrated in Figure 3, there is an overlap between the random and intentional threats, and it may be advantageous to have a unified view of the trustworthiness. However, in this paper just the random threats will regarded, cf. Section 1.2 below.

- ***Means*** to achieve dependability are: 1) avoiding that faults occur or are introduced into the system, 2) finding and removing faults from the system, and 3) tolerating faults in the system by preventing them from resulting in failures. Means like testing and evaluation by modelling and mathematical analysis or simulation, are also needed for the cost-efficient design and dimensioning of systems with respect to their required attributes and to validate and reach confidence in the result. See also Figures 2 and 3.

- ***Attributes*** of a system, which characterise its properties with respect to dependability. The attributes may be regarded as the following "abilities". *Avail-*

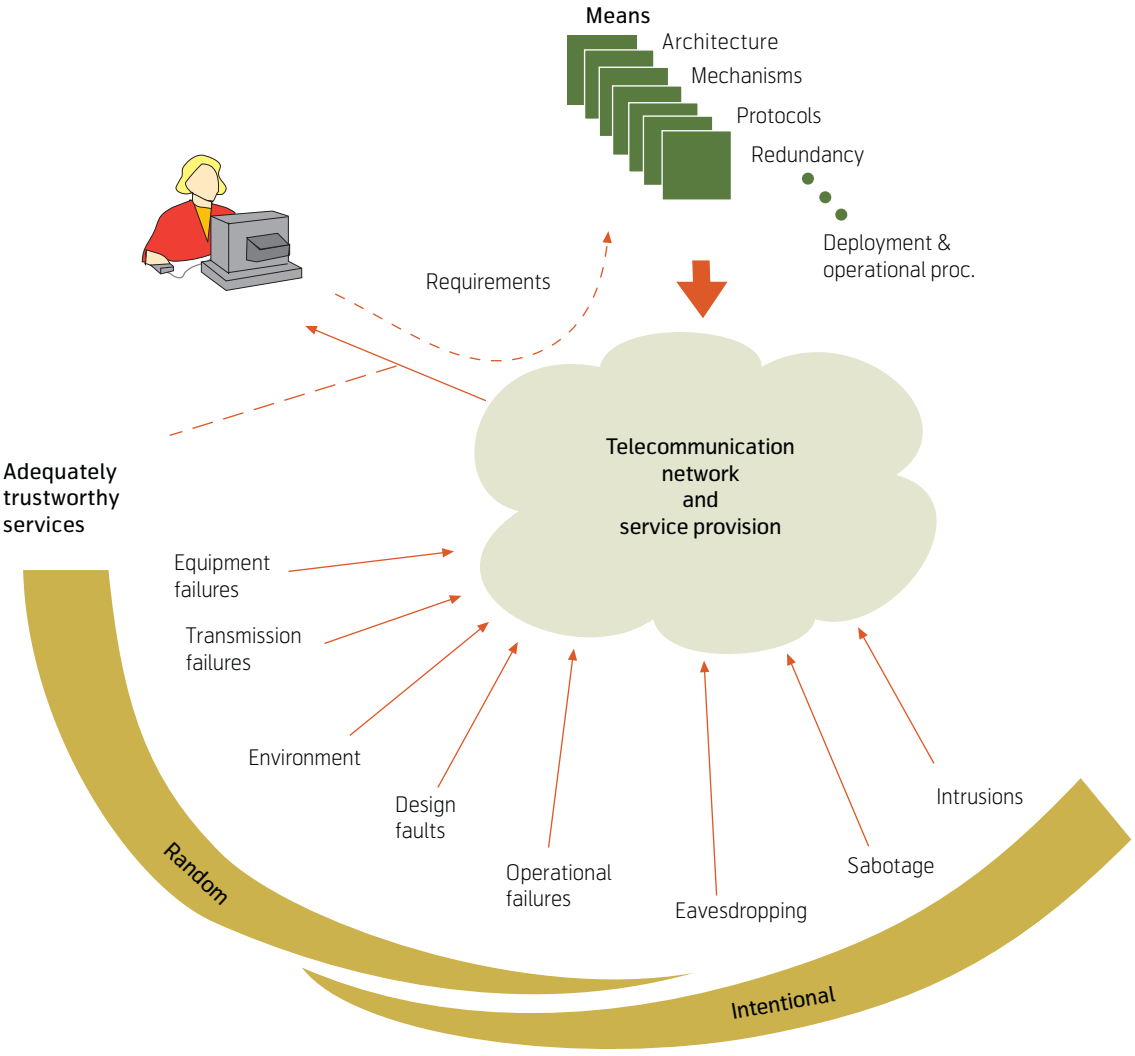


Figure 3 Trustworthiness subject to random and intentional threats

ISSN 0085-7130

ability: to provide a set of services at a given instant of time or at any instant within a given time interval. *Reliability*: to provide uninterrupted service. *Safety*: to provide service without the occurrence of catastrophic failures. *Maintainability*: the ability of the system to support fault removal and service restoration, as well as undergo change. *Confidentiality*: to prevent unauthorized access to and/or handling of information. *Integrity*: the absence of improper alterations of information and system state. The attributes are characterized by measures, some included in Figure 2.

It is seen that the quantification of all dependability attributes, i.e. the measures, except for maintainability, are significant QoS parameters for describing the services provided by the system, see for instance [12]. Some of the characteristics are also related to security, which deals with intentional threats to systems trustworthiness, cf. Figure 3. Dependability is also strongly related to the performance and traffic handling abilities of a system. A system cannot be said to be available or work reliably unless the services are provided with a sufficient capacity, and have response time and delay characteristics as required.

1.2 The real threats to dependability

The dependability (trustworthiness) of a system should be envisaged similar to the common concept of the trustworthiness of a person. It is whether the person delivers services as agreed/expected and otherwise behave according to the common norms that matters, not the reason why they do or do not, and ideally, not related to how the person is built or what they look like. The conception of dependability (reliability) in a technological context has, however, been coloured by its use when it emerged as a technical discipline in the 1950s. At that time the focus was on mechanical and relatively small, simple hardware dominated information and communications systems. At that time, permanent physical faults were the dominant cause of failure.

This is no longer the case. Generally speaking, the major cause of lack of dependability in today's information and communication systems are faults due to human error. Logical failures are made during specification, design, and implementation, some of which prevail when systems are put into operation. These are often perceived as 'software faults'. However, with the increasing hardware and overall system complexity there is an increasing number of logical faults in hard- and firmware as well. In addition to the logical faults, there may be glitches in the adaptation of a (sub)system to its co-operating (sub)systems and the environment. Mistakes are made during configuration, operation and maintenance. However, the preconception that "only" hardware failures reduce system dependability led for a long time to less awareness of these threats to system dependability. Human-introduced faults in systems has not got the proper awareness as system complexity has grown and has in some cases been "accepted".

Published fault and failure data are scarce[1)], so it is difficult to be precise concerning the relative impact of the various kinds of causes. The relative impact will of course also vary depending on the kind of system and its design, e.g. fault-tolerant or not. The reader is referred to [14, 21, 37, 31] for a rough overview. To summarize into a few simple statements:

- Equal attention should be paid to faults stemming from
  - physical causes,
  - inadequacies in specification, design and implementation, and
  - operation and maintenance.

- The technology to deal with physical hardware faults are better developed. There are fewer such element failures and the systems designed to tolerate them are in most cases able to do so. However, rectifying failures due to these faults requires manual intervention, which results in longer down times, and the contribution to system unavailability may become significant.

- Logical faults in the software (and hardware) is a significant contributor to the failure intensity of system elements. Fortunately the failures typically have a limited effect. Rather few of them lead to major system outages, but they have the potential to do so, for instance by being a common cause throughout a system and by error propagation.

- Larger outages are typically caused by destruction of the physical infrastructure by nature and humans, or by operation and maintenance mistakes.

- It seems as operation and maintenance related faults have a large and increasing importance.

---

1) *Operational failure data are costly to collect. Furthermore, organizations tend to keep them for internal use and restrict their public availability. (The Norwegian university network operator, UNINETT, forms an exception as they make part of the operational statistics public, see http://drift.uninett.no/downs/.) However, some PTTs require that QoS related failure statistics and/or significant incidents are reported, see for instance [21].*

In conclusion, the real threat to the dependability of today's and future systems is their unmanageable logical complexity yielding design and O&M faults. An important reminder is also what is pointed out in [21]; overload may be a major contributor to lack of dependability.

1.3 How to deal with the threats

There is a lot of confusion concerning the concepts of fault, error and failure, since in daily language the terms are used interchangeably and mean that something does not work, is incorrect, etc. However, in dependability engineering they denote separate phases from cause to consequence as illustrated in Figure 4. Starting with the consequence:

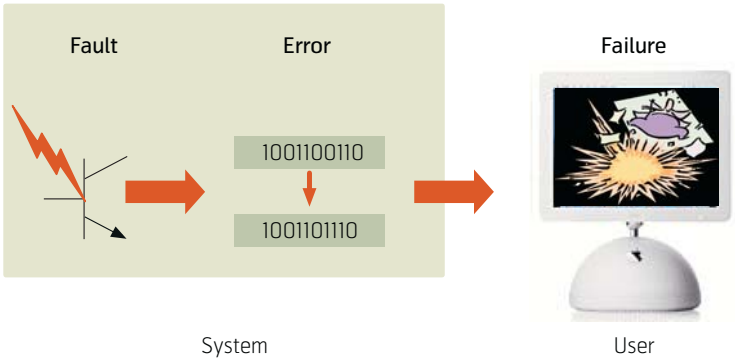


Figure 4 Relationship between faults, errors and failures

- ***Failure***[2)]: delivery of incorrect service or the transition from correct service delivery to incorrect. A failure is a manifestation of an error that we observe on the outside of a system.

- ***Error***[3)]: a system state that is liable to lead to failure, i.e. the manifestation of a fault within a system.

- ***Fault***[4)]: the cause of an error. Note that faults may have a multitude of physical and human causes and be of various kinds as illustrated to the right in Figure 5.

In composite systems, where one part of the system uses services from other parts, the failure of one part may constitute a fault in another. This is discussed together with other conceptual issues related to the failure process in [3, 10].

The basic means to obtain dependable systems are fault prevention and fault tolerance. These may be seen as barriers in the cause consequence chain as illustrated in Figure 5. The fault prevention techniques aim at avoiding the causes of failures, the faults, to enter or occur in the system, e.g. by extensive quality assurance of the software, benign operating conditions for the hardware, etc. In the fault tolerant approach, the existence of faults is accepted. However, when they result in errors, the system is given a structure and mechanisms to prevent the errors from manifesting themselves as failures. A number of mature and well known techniques exist, e.g. FEC (forward error correcting codes) or CRC (cyclic redundancy check) combined with retransmission of erred packets to tolerate transient transmission faults. In the transport network there are techniques for rerouting of traffic when network links and nodes fail. We will return to techniques in Section 3.1. Fault-tolerance in the computing functionality is achieved by replication. Various options will be discussed in Section 3.2. Note, however, that we only have generally applicable and mature fault tolerance techniques to deal with physical faults and that for faults caused by humans, we have primarily to rely on prevention.

For the completeness of the discussion a third barrier is introduced in Figure 5: contingency planning and disaster recovery; i.e. plans made and actions taken in order to limit or reduce the consequences of failures. Such plans are commonly made by companies having a business critical information and communication infrastructure. The plan deals with major failures of this infrastructure, also those caused by fires and other disasters, as well as other business critical functions. If we regard the publicly provided communication services, the traditional services and their provision are/have been independent, and the services are to some degree substitutable. For instance, if "the Internet does not work" we send a fax or make a phone call. This has contributed to the reduction of the consequences of failures. However, with the ongoing convergence of "Internet-services", fixed and mobile telephone, broadcast etc., this option may be reduced or removed due to common infrastructure and highly interwoven services. Companies that are highly dependent on communication services are commonly advised to obtain duplicate services from independent providers. However, this does not guarantee independence, since in the current telecommunication marketplace, these may buy services from common third parties.

---

2) *In Norwegian: Feilytring*

3) *In Norwegian: Feiltilstand*

4) *In Norwegian: Feilårsak*

ISSN 0085-7130 © Telenor ASA 2004

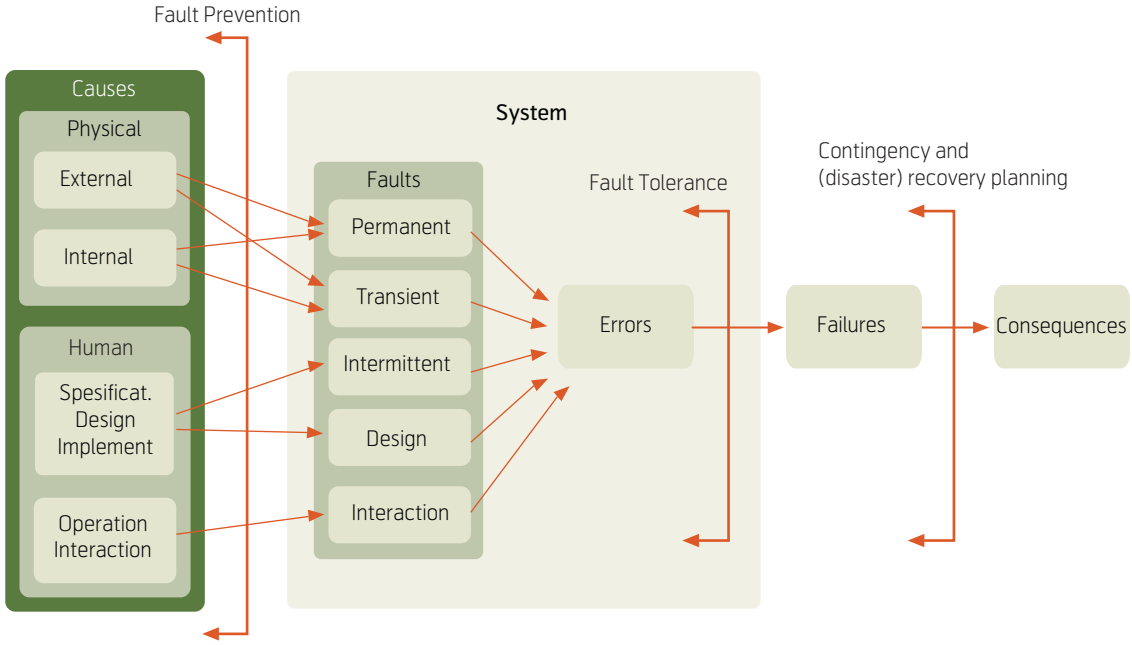


Figure 5 Approaches to dependable systems

2 The technological evolution and some of its implications

A low dependability is often accepted from new technological inventions providing a substantially new functionality. After this first "toy phase", as the use and our dependency on the functionality increase, the requirements become stricter. Regard cars and planes as examples. Furthermore, history has shown that when failures may have severe consequences, we tend to become conservative with respect to introducing new solutions when a proven one exists. In this situation the technology deployment trend tends to be more toward marginal improvements than to introduce radically new solutions. Within information and communication services, we are simultaneously at both ends, as well as in the middle, of this maturity scale. Users "toy" with new Internet services having dependability attributes orders of magnitude less than for instance the 'always working' fixed telephony service. From a dependability point of view this mix of proven and immature technology, between no and strict requirements poses major challenges. The text below will deal with some of them.

An old word of wisdom is that dependability cannot be added to a system after it is designed, it has to be an integrated aspect of the design from the beginning. This may easily be forgotten in the current eagerness to bring new services and networking concepts swiftly to the market, which may give backlashes when the requirements move from the toy stage to the necessity stage.

Several of the ongoing techno-econo-commercial trends in the evolution of information and communication systems have a significant indirect impact on the dependability of these systems, but first, a brief glimpse at the history.

2.1 History

Telephony started out as a toy, but soon reached a technical maturity and a widespread use that has lead the users to expect it to always work. The leading unavailability requirement for telecommunication systems of "less than two hours of accumulated system down time per forty years", corresponding to three minutes per year, was originally a design goal set for the ESS 1 in the early 1960s [8]. This goal was set to match the empirical unavailability of the then current technology, the electromechanical crossbar switching system. To keep the perspective; in the mid 1980s the major manufacturers of SPC switching systems reported that this goal was reached. This design goal is also presumed to be the initiator of the often claimed five nines in other systems[5)]. For further historical summary of dependability of (voice) communications, see [26].

The other "ancestor" of the communication network, the Internet, had its origin in the fear of nuclear

5) *Unavailability: 2 hours / 40 years = 3 minutes / 1 year = 5.7 $10^{-6} \approx 10^{-5}$; Availability: 1 – 10^{-5} = 0.99999*

attacks during the cold war. To cite Paul Baran in one of the groundbreaking works toward the Internet, also in the early 1960s [4]:

> "We will soon be living in an era in which we cannot guarantee survivability of any single point. However, we can still design systems in which system destruction requires the enemy to pay the price of destroying *n* of *n* stations. If *n* is made sufficiently large, it can be shown that highly survivable system structures can be built, even in the thermonuclear area."

In addition to a highly interconnected structure where any existing path between two users might be used, dependability was ensured by protection of packeted data by error detection codes and protocols handling retransmission etc. of corrupted or lost packets. In today's effort toward making the Internet the common transport platform for all services, it should be kept in mind that it was designed for high survivability and not for real time services or as a "minimum cost" infrastructure. For a history of the Internet see [24].

No quantitative dependability objectives have been formulated for the Internet. However, the ideas from its origin of inherent robustness and survivability have prevailed. Hence, it provides a fairly good dependability, in spite of its heterogeneity, rapid growth and rather uncoordinated operation and management. The unavailability of its backbone is reported to be in the order of 10^{-4}, which is still an order of magnitude poorer than the corresponding figure for the telephone network, 10^{-5} [22].

Mobile communications poses additional challenges to the dependability. With user and environment mobility we get added complexity and a more composite operator/provider setting. Adding terminal mobility, we also have an unreliable radio access channel between the user equipment and the network. Limited coverage and handover failures may also disrupt the continuity of service. In the design of current and future wireless systems, measures have been taken to deal with these specific issues, but to cite H.A. Malec: "There was no indication that a major concern for the design of the 3G was reliability performance" [26].

2.2 Integration and differentiation

We are likely moving toward a truly service integrated network, i.e. one physical and logical integrated infrastructure providing all services. However, the dependability requirements of the various services will differer widely, e.g. leisure browsing vs. emergency services and business critical services. Some of these aspects are tentatively illustrated in Figure 6. In particular, there is a difference in the various user (groups) willingness to pay for a higher quality of the same service. A high dependability comes at a high cost with respect to equipment volume, development effort, management and maintenance. The additional cost (as a rule of thumb, at least 50 % of the total cost) needed to reach today's telephone network standard, is likely to be too high for 'unimportant' services and cost aware users. At the other end of the requirement scale, since telecommunications is an important part of the public infrastructure, regulators may set requirements on the dependability of some services. Hence, an (increased) differentiation of dependability with respect to kind of service, user group or user importance/willingness to pay is expected in the future. A first step in this direction is to obtain an idea of the value of various services to various user groups as illustrated in Figure 7. How dependability differentiated services technologically should be provided, is an open issue. Further on, there are even more challenging commercial and operational issues that must be solved. To make a transition to a differentiated dependability regime,

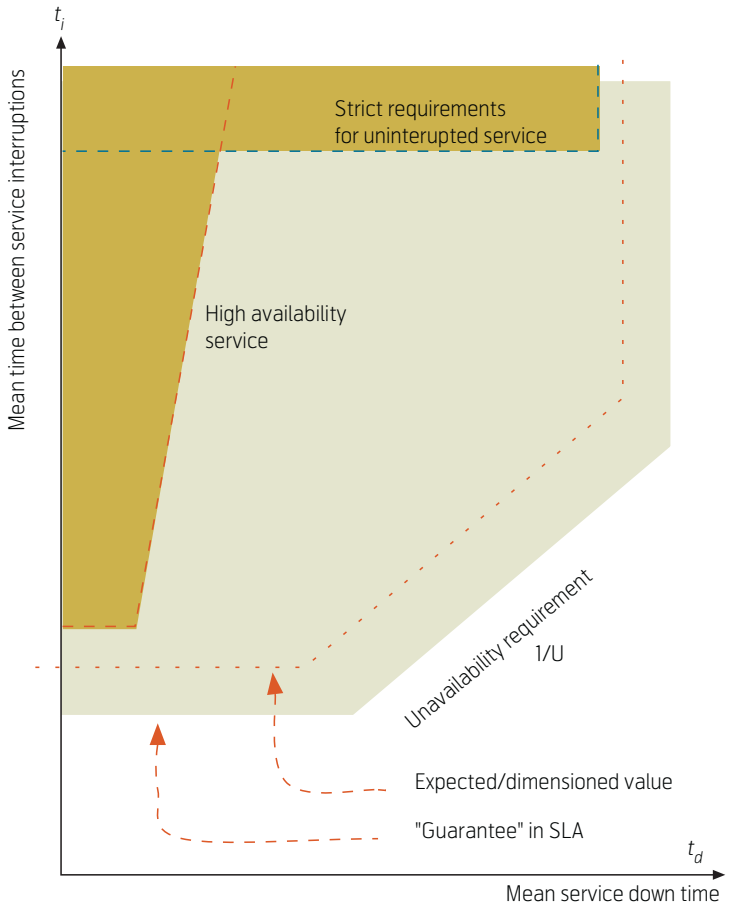


Figure 6 Illustration of differentiated dependability requirements. The shaded areas represent acceptable areas for various kinds of services

solutions adding a minimal complexity and allowing a gradual introduction are need. As corollary, solutions should be sought that inhibit inherent unreliable components and services from affecting services with stricter requirements.

Differentiation will add to the technological and operational complexity discussed in Section 2.3. In the integrated network, we will have a co-existence of

- services with highly differing requirements utilizing the same set of system components;
- system components with highly differing dependability attributes; and
- network operators and service providers with highly differing ability to deliver dependable services.

These items should also be kept in mind when ambient intelligent services are considered in Section 2.6.

2.3 Technological and organizational complexity

The increasing complexity of the systems and the consequence that human-induced faults constitute the major threat are already discussed above. Coping with the complexity and keeping the number of faults low have been a major effort in software engineering for the last decades. Although one is not entirely on top of the problems, and there are exceptions, we are reasonably able to cope with the dependability of software (sub)systems by fault prevention and removal, cf. Figure 2. Preventing failures due to design faults will be a major challenge, also in the future. However, operational faults have become the major cause of severe system failures. Operational faults and corresponding means to avoid failures have got the least research attention hereto, and should be placed in focus. The number of services, the potential service interaction problems, the heterogeneity and size in number of HW and SW network components will increase this liability for design and operational failures in the future.

A related issue to technical complexity of the systems, is the complexity of the commercial marketplace after the demonopolization of traditional telecommunication services, the growth of the commercial Internet, and the convergence with information services and entertainment. The "network" will be provided by a large number of (simultaneously) co-operating and competing actors, some having a vertically integrated network providing a set of services within a geographical area, others providing transport or services within one or a few segments or layers, some providing end-user services worldwide, etc. In general, there will be many actors involved in providing an end-user service. Mobility increases this multiple actor involvement. In this setting it will be extremely difficult to guarantee a specific dependability to the end user. A management of the dependability attributes delivered, as well as other QoS attributes, will require extensive co-operation among the actors. An anticipated element in this co-operation is extensive service level agreements (SLAs). These define the QoS that is going to be delivered among them, how to measure the delivery and reaction schemes if a party does not comply, e.g. some sort of economic compensation [18]. In this context, dependability QoS attributes pose a special challenge, since their typical or average values are impossible to measure on a short time scale, for instance in less than a year.

In addition to the professional/commercial actors, there is a trend toward private and non-commercial entities being involved in the provision of services, e.g. ad hoc and peer-to-peer networks, as discussed below. An interweaving of elements from these and "professional" services is foreseen.

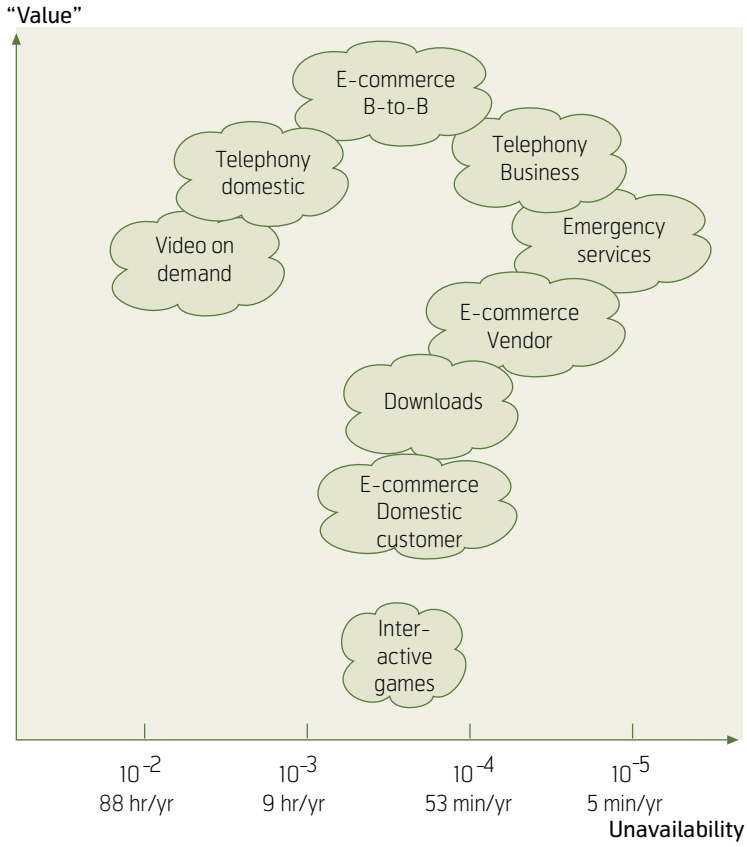


Figure 7 What are the values of the various services to their users? (NB! Do not interpret the mapping directly)

2.4 The self organizing networks

The area of information and communication systems and services is still getting more comprehensive. During recent years the fields of peer-to-peer networking, (non-military) ad-hoc networks, ubiquitous computing and grid computing have emerged and started to grow. Most of these are still at the "toy stage" and dependability is not a major concern for the pilot users. However, there is ongoing research on this kind of systems/networks where dependability is an important issue. See for instance [27], where the MAD properties Mobility, Adaptivity and Dependability are introduced as crucial for these systems. Common to self organizing networks is the lack of a supervisory, coordinating and managing entity and a stable structure. A substantially higher degree of self management or autonomy, than found in current systems, is required. These networks have no authority that ensures their dependability or other QoS attributes. Hence, an application/system function that requires a certain dependability must interact with its network environment, assess the capabilities of the environment, and use the current computing and communication surroundings in a way that ensures that requirements are met.

This, together with the manpower issues discussed below is a major driving force toward expecting future systems and system components to become more autonomous.

2.5 Manpower

Correct configuration and management are key issues for having dependable systems. As pointed out above, failures due to faults induced through configuration and management are currently perhaps the most salient contributor to system failures. Today, dealing with these tasks usually involves personnel. The required competence of this personnel increases as the systems get more complex, and competent personnel, both in terms of number and skill, become the limiting factor.

The manpower shortage anticipated for the ICT business will be driven by configuration and management effort rather than by development. The development effort grows roughly in proportion with the number of network components developed (HW and SW). However, each type of component is installed in a huge number of copies, and the configuration and management effort is at least growing with the product of the number of components and their number of copies. Hence, without a shift in configuration and management paradigm, the demand for highly skilled personnel will limit growth and the dependability of the services provided. This is one main motivation for research on autonomous systems, e.g. [19, 1]. In addition to adaptivity and flexibility, intrinsic planning and dimensioning as well as lower cost to the user may be achieved. For the future services based on ambient intelligence, it seems to be a prerequisite to have autonomous configuration and management.

2.6 The challenge of the ambient intelligence

Up to now, we have discussed information and communication systems roughly in the context we know these systems today. What is foreseen in the future has been characterized with many cognomens, e.g. ubiquitous computing and communication, pervasive computing and ambient intelligence. An integration of sensors and processors into everyday objects like clothing, white goods, toys, furniture, etc. as well as for special purposes as "health sensors" in our bodies is predicted. These objects will be enabled to communicate with each other and the user, for instance by means of ad-hoc and wireless networking. "Natural interfaces" will allow the users to control and interact with the environment by voice, gestures, etc. in a personalised way depending on preferences and context. User interaction will apply new visualisation devices ranging from projections directly into the eye to large panorama displays. Information and communication technology is expected to be everywhere, for everyone, at all times.

Some of the tasks (services) expected to be handled by the ambient intelligence are mission or safety critical; for instance, supervision and support of persons with life threatening diseases, extensive economic transactions, support of emergency operations and vehicle control. Failures may open for theft of identity and values, as well as catastrophic second order failures. For other tasks, failures may lead to various degrees of inconvenience, ranging from no food and heating to that of the favorite music not playing. Although, even for the systems without explicit dependability requirement, the dependability is expected to be important just to be profitable in a competitive low margin market – failures cost.

Regarding the dependability challenges in providing these services, it should also be kept in mind that:

- The same kind of "infrastructure" shall support all these kinds of tasks. In an (unregulated) competitive setting, it is unlikely that costly redundancy will be provided for fault tolerance.
- The hardware and software elements of these systems are expected to be inexpensive COTS (commercial off-the-shelf) components.
- The systems are characterized by a high complexity, heterogeneity, and dynamism in application mix and structure.

- There will be a wide range of network operators and service providers; from professional corporations, via inexperienced startups to unskilled private individuals.

- There will be no system administrator (in many cases not even a man in the loop to detect failures). Autonomy in (re)configuration, fault detection and most of the fault handling is a prerequisite.

It is seen that the dependability challenge of the ambient intelligence also includes aspects of what is discussed in Sections 2.3, 2.2, 2.4 and 2.5 above. If we in addition include the security challenges imposed by the ambient intelligence concept, it is seen that new architectures are required to be trustworthy. Work is going on to be more specific about what the threats are and where the research effort should be put [28].

An implementation of ambient intelligence that we may trust, rely on the ability to dependably access the core network and remote services. To achieve this, access networks must deliberately be built to enable independent diverse access channels from (critical) end user devices to the core network. Radio access with overlapping cell structures or similar, in addition to wired access, may provide this diversity. However, high dependability requires that this is taken into account in the architecture and design, as well as in the topology of the network covering an area. Handover to maintain continuity of services and in some cases diverse access is another unsolved issue. To relate this to what was discussed in Section 2.3, these coordinated overlapping access options may be given by independent market actors.

From the above, it should be evident that we have a wide range of challenges with respect to providing trustworthy end-user services in the future ambient, integrated, multi-technology and multi-provider environment. In the next part of the paper, we regard the change of generation of technology which has already started in the core communication infrastructure.

3 The core communication infrastructure

The dependability of the core communication infrastructure is very important, since outages will have immense failure effects/consequences. As pointed out in Section 1.3, there exist successful techniques for tolerating physical faults, both in the transport network, as well as in the computing platforms used to control and manage the network, and to provide services. Below, the ongoing trends for the transport network will be reviewed, and in Section 3.2 the computing platforms will be dealt with.

3.1 The transport network

A brief review of the strategies to obtain resilience, i.e. fault tolerance in transport networks is given below, before we introduce the multilayer design of core transport networks and discuss the foreseen future developments.

Resilience mechanisms

The term resilience characterizes the ability to continuously provide service in spite of complete or partial failures of network elements.[6)] The following design parameters should be considered in the choice of mechanism:

- The additional network elements and the capacity needed;

- The time from an element failure until the transport service is restored. An interrupt shorter than 50 ms is considered to be without significant impact. (Originally, this requirement stems from telephony, other applications/services may of course set other requirements);

- The organization of fault handling and resource management;

- The ability to provide differentiated dependability;

- The dependability of the mechanism itself.

This section introduces *protection*, *reconfiguration* and a few *self-healing strategies* for making networks resilient, and discusses their merits.

- ***Protection***. A dedicated spare path is established between the end nodes of the protected path or subpath. There may be intermediate nodes as indicated in Figure 8, but these do not perform any switching or rerouting of the protected traffic. Protection comes in two variants, depending on whether the spare is active or not:

  - *1+1 protection*: The information (bits, packets, cells, …) is sent synchronously on both paths towards the destination. The destination selects the stream it considers as the one with the higher integrity, e.g. fewer bit errors. Immediate (a few ms) handling of path failures, which allows them to go unnoticed by the users, is feasible.

6) *Common synonyms of resilience are network robustness, fault-tolerance and survivability.*

- *1:1 and N:M protection*: These are stand-by protection strategies, where one (N) dedicated spare (or back-up) path(s) is/are available for one (M) active primary path. In case of failure, the receiving node must inform the sending node, which redirects the data stream. This incurs a temporary loss of service (10 – 100 ms) usually tolerated by the users.

Protection is simple, fast and proven, and has less costly control. It is the "workhorse" resilience strategy of today's transmission networks and is standardized for ATM and SDH networks [16, 17]. However, it requires more transfer capacity and is less flexible with respect to dynamic changes and differentiation than the alternatives outlined below.

- ***Reconfiguration*** is a strategy based upon a centralized management of the network. A network management system, see Figure 9, supervises and controls all network resources. It keeps a map of them, their utilization and status, etc. Based on this information it sets up transmission paths between nodes in the network. When a network failure occurs the network management system reconfigures the routing through the network. See for instance [6] for a case with preplanned reconfigurations.

  Reconfiguration has the potential of being flexible and using the transmission capacity very efficiently due to its ability to perform a global optimization. It may, however, be slow, in the order of minutes, and is vulnerable, since its centralization yields a single point of failure and it depends on information transfer in a network with failure(s) not yet handled.

- ***Self-healing*** is used as a common denominator for several strategies which have distributed control and require no dedicated pre-reserved transmission capacity. See for instance [7] for additional material.

  - *Primary with backup paths*. When a path (the primary) is established through the network one or more disjoint backup (stand-by) paths are established simultaneously, see Figure 10. These do not exclusively reserve any link capacity and do not carry any traffic unless a node or link along the primary path fails. We have a research effort on finding such paths distributively by emergent behaviour, see for instance [36]. Technologies like ATM and MPLS are suited for establishing such paths, which also enable load distribution and differentiation.

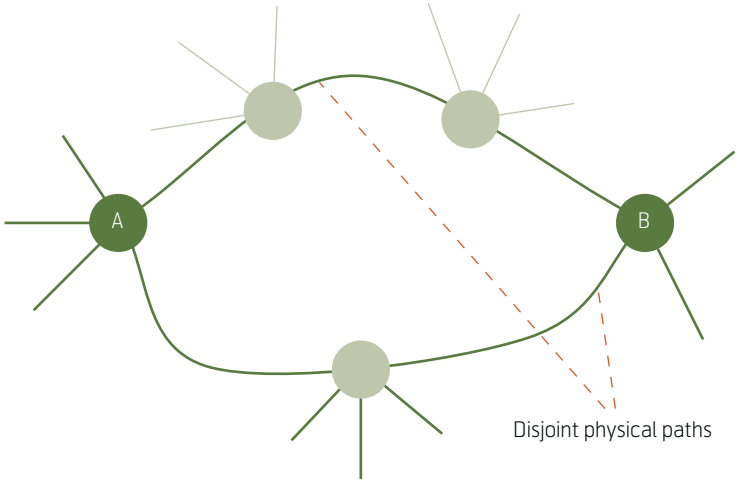


Figure 8 Protection of communication between nodes A and B. 1+1 protection when the information flow simultaneously along both paths, 1:1 protection when one acts as a back-up for the other

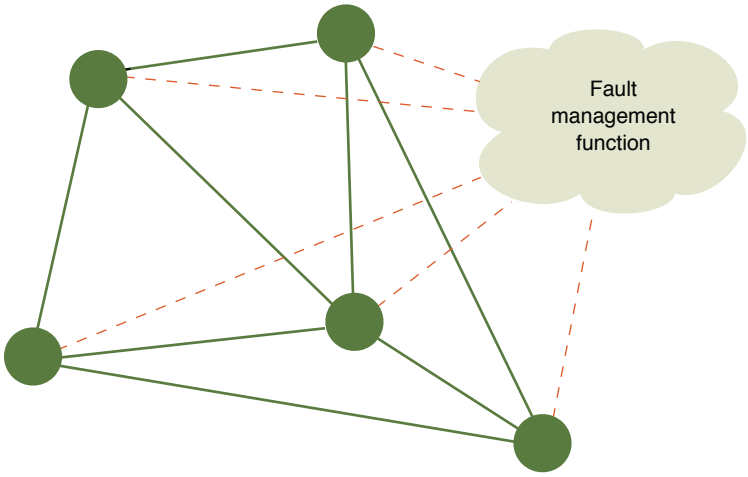
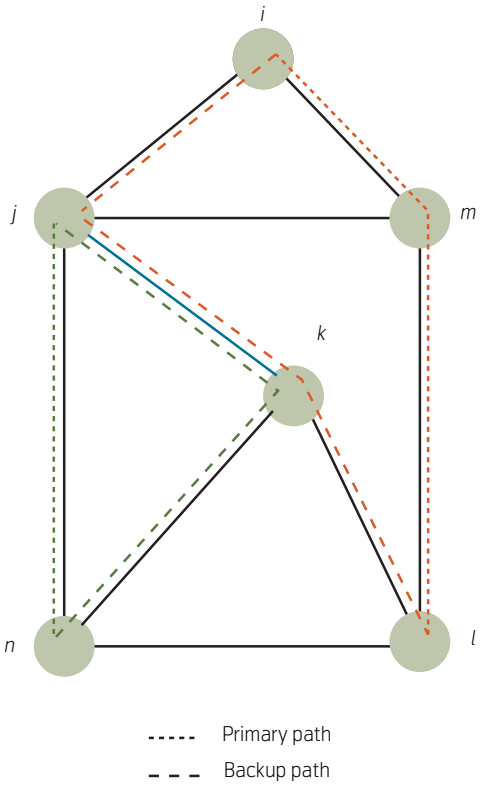


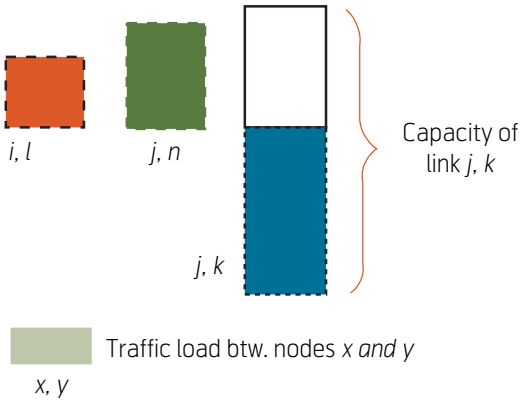
Figure 9 A centralized fault management gather information from all network elements, and, in case of an element failure, reconfigures the network

  - *Flooding* is a fully distributed network management technique to restore communication following an element failure [33, 11]. The technique floods the network with requests in a search for available capacity to replace that of a failed link. Flooding, especially end-to-end restoration poses a number of algorithmic problems. Restoration time is slow and no dependability/ QoS guarantees may be given.

  - *Rerouting* is a self-healing strategy relying on the traffic handling mechanisms of the network and the user equipment to re-establish individual packet flows/connections after they have been



a) Primary and backup paths between *i* and *l*, and *j* and *n*



b) Capacity allocation on link *j*, *k*

Figure 10 Illustration of shared capacity for backup-paths. The primary paths between i *and* l*, and* j *and* n *have no network element in common, and are assumed to fail independently. Hence, they share the unused capacity on link between* j *and* k *for backup*

disrupted by a network failure. Internet (re)routing is a well known example. This strategy differs from the self-healing strategies presented above, as well as protection and reconfiguration where dedicated fault management entities restore the paths within the network for a bulk of flows/connections. Fault handling by rerouting cannot be 'invisible' to the user and no dependability/QoS guarantees can be made. The delay before a flow or connection is restored may be significant.

Rerouting, flooding and reconfiguration without prepared plans are denoted restoration techniques as they seek to find new paths and restore the broken traffic flows after a failure, while the others have advance plans, i.e. are proactive. Some of the above strategies may be suited for both span reestablishment, where the affected traffic are diverted locally around the failed network element (usually a link), and for end-to-end reestablishment, where a path is (re)established between the source and destination node in the network. End-to-end reestablishment yields a better diversion of the traffic and has less requirements for spare capacity.

The properties of the fault handling and redundancy strategies presented above are summarized in Table 1. The entries should be considered as indicative. A wide range of values may be seen, especially for reconfiguration and flooding. All the strategies may be used to introduce dependability differentiation, however, only the primary/backup approach enables this without introducing additional control functionality and/or interactions between currently independent control and management functionality.

The "hard wired" protection schemes are fast, simple and proven, but may be costly in terms of spare resources required. They are also inflexible with respect to rapid changes in topology and traffic pattern as well as differentiation. In an Internet based platform, the advantages of protection will be bal-

	Resources			Control			Speed	Dep. diff.	Spare reqmnt.
	Dedicated	Preplanned	On demand	Centralized	Dist. Control	Dist. man.			
Protection	X					X	Fast	No	Large
Reconfiguration		X	X	X			Medium –	No	Small
Primary-backup		X			X	X	Fast	Yes	Small
Flooding			X			X	Medium +	No	Medium
Rerouting			X		X		Slow	No	Small

Table 1 Organization and properties of various network fault handling strategies

anced against the gain of rerouting and MPLS based primary backup schemes. The reduction in number of management layers discussed below will invigorate the latter alternatives.

Layering

Transport networks have a layered design, for instance as illustrated in Figure 11. The physical bottom layer provides a link between nodes in the layer above, which again may provide logical links to the layer above, etc. Note that the topology of the networks seen at the various layers may differ. Each of the layers typically has its own technology adapted to its purpose. Figure 11 indicates four layers with corresponding technologies familiar to most readers, where optical switching, OxS, is on its way into the network.

When ensuring dependability in this multilayer architecture, we are met with a number of questions: a) In which of these layers should we introduce the ability to tolerate physical faults of the transmission medium? b) Which resilience strategy and mechanisms should be used? c) How should the fault management in the various layers be co-ordinated? The answers to these questions are complicated by the fact that, in contradiction to what is illustrated in Figure 11, networks are not homogeneously and strictly layered; the structure may vary in adjacent autonomous systems and the lower layers may carry several independent transport services[7)] as illustrated in Figure 12.a. Furthermore, not all transport services (or end-users) have the same dependability requirements and hence the same need for fault-tolerance in the network.

As indicated in Figure 12.a each "layer" has its individual resource management. This hampers the coordination in provisioning dependability differentiated services, it complicates the fault management since it is necessary to avoid several layers competing to handle the same failure[8)]. An 'independent' design and operation of each layer may also introduce excessive (or missing spare) capacity. For discussions on design of multilayer networks and on which layer that it is most profitable to introduce redundancy, see for instance [20, 23, 13, 35]. Some rules of thumb:

1. The higher up in the network, the greater the flexibility, and the spare resources needed are likely to be less;

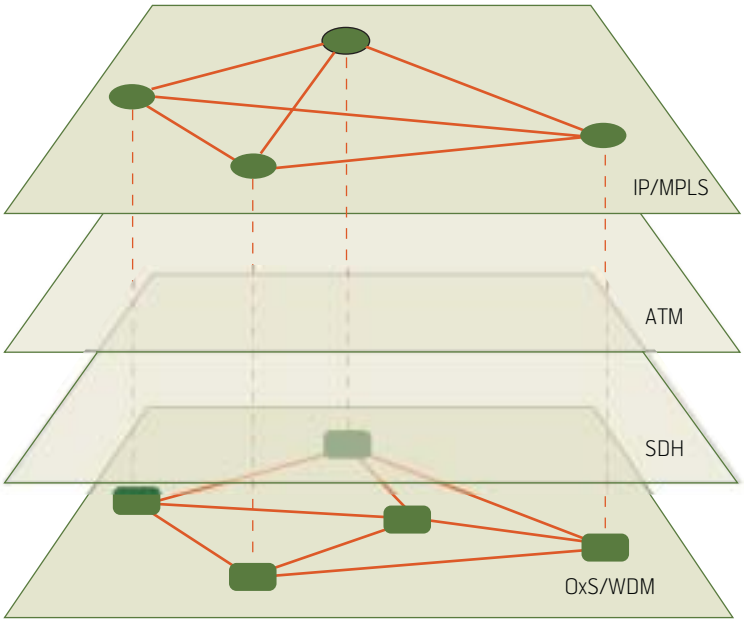


Figure 11 Layers in a backbone transport network

2. The higher up in the network, the longer the delay from the failure event until service is restored;

3. Low layer redundancy handling tends to introduce least complexity since the mechanisms may be simpler and the redundancy handling may be common to a number of higher layer network services;

4. Fault tolerance on a layer can only handle faults at its level and on the levels below. Hence, some fault tolerance is always required at the highest level.

There are strong arguments for a reduction of the number of layers and for having a co-ordination of the resource management in the layers. At the same time it is advocated that an IP based network with (G)MPLS and QoS extensions should provide all transport services, i.e. the backbone networks tend toward getting an organization as illustrated in Figure 12(b). This should in principle enable a better resource utilization and less cost. An aspect of this common management architecture is that it becomes feasible to tailor the dependability provided to that required, cf. Figure 6, by using a mixture of resilience mechanisms like the primary-backup approach and rerouting as well as preemption of low requirement traffic. However, with the observations of the current Internet in mind, e.g. [22], remembering that the current Internet to a large extent is supported by the pro-

7) *By transport service in this context is meant the transport of information through the network, not the service provided by a transport protocol.*

8) *Nested hold-off timers is the commonly recommended approach, see for instance [23].*

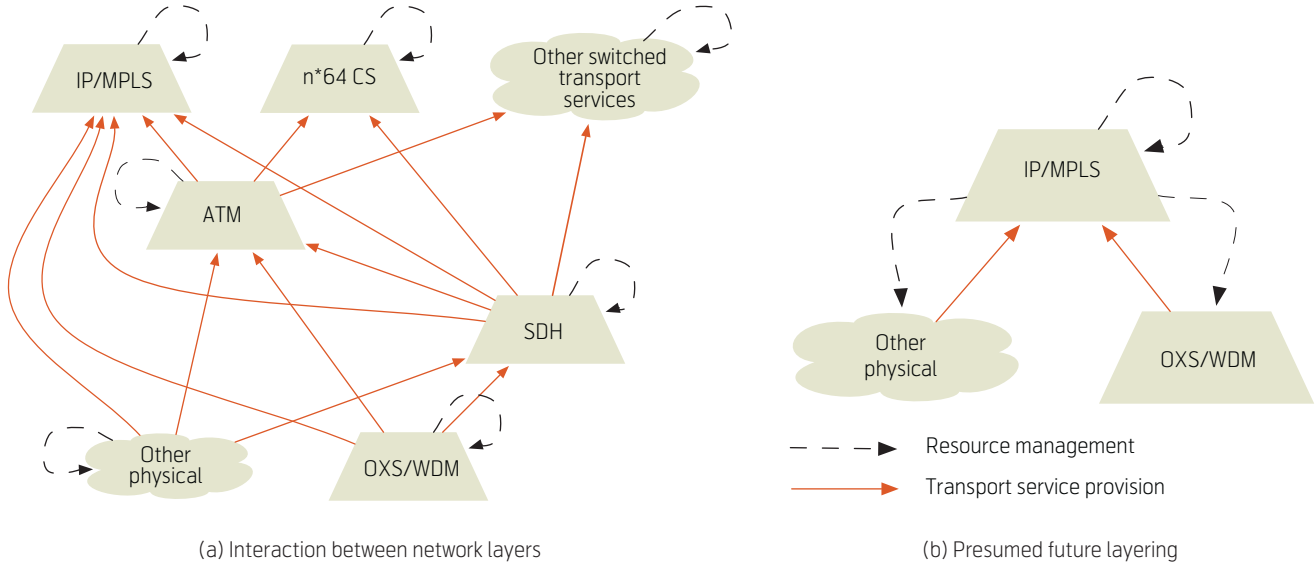


Figure 12 Current and intended future relation between transport network layers providing transport services

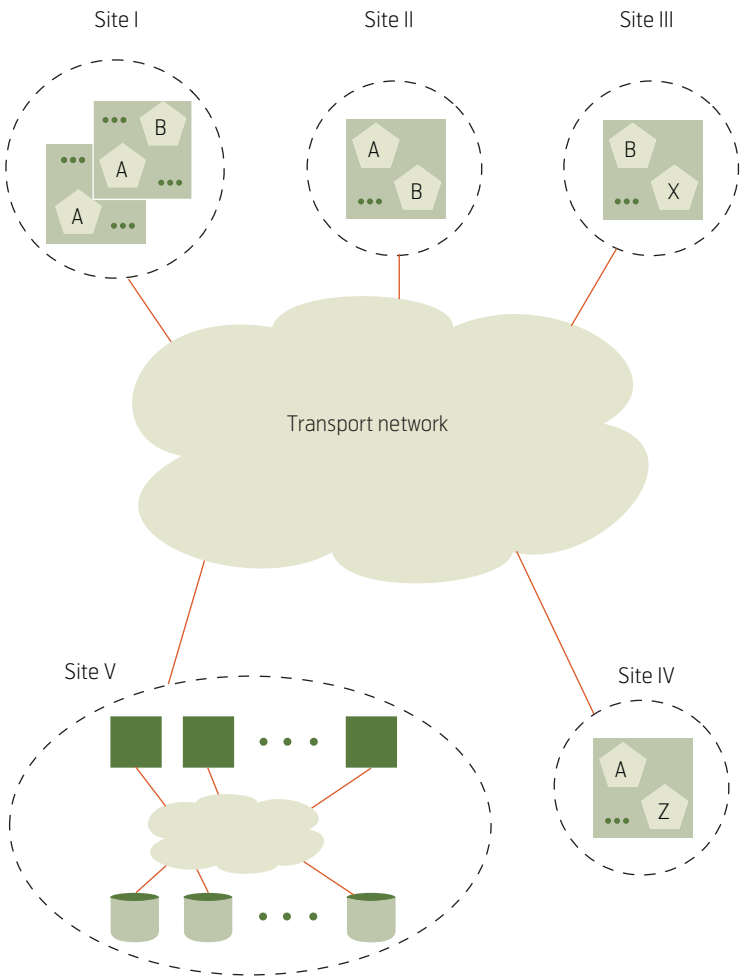


Figure 13 Various options for providing fault tolerant computing platforms. At Site I, the node itself is made fault tolerant by duplication, i.e. replicas of processes A and B run synchronously on two computers. In the next option, the distributed computing paradigm is applied, and the same processes are distributed over Sites II, III and IV. The last option illustrated is a computing cluster at Site V, where a mix of strategies may be applied

tection mechanisms in the SDH network, and that the Internet resilience technology required currently is at the R&D stage, care must be taken in this technology transition. Otherwise, we may move from a situation where physical faults in the transport network is a moderate contributor to lack of dependable services, cf. Section 1.2, to a situation where they may become a major problem.

3.2 Control, management and service provision

There are two basic strategies for achieving fault tolerance in the computing platform providing control, management and provision of services. See Figure 13 for illustrations.

1. To use *fault-tolerant network nodes*, i.e. to make the computing and database platforms supporting the functionality of the network nodes fault-tolerant, having redundant computing and/or storage capacity in each node. The replicas of the processes are executed or stored within the same node. For instance, to make an intelligent network SCP fault tolerant, a duplicated synchronous computer may be used to execute the functions.

2. To introduce *fault tolerance at the network level*, i.e. to have co-operating replicas of software objects/processes providing network functions in several nodes and thereby enable tolerance of node failures. Network level fault-tolerance is based on dependable distributed computing technologies.

The legacy system

The legacy switching nodes and network centric server nodes used in PSTN uses strategy 1. The architecture of the nodes may vary. Fault-tolerance may

for instance be achieved by micro-synchronous duplication, as in Site I of Figure 13, e.g. the design used in AXE [32]. Alternatively a distributed and mainly load-shared architecture, as in System 12 [5] has been applied. This architecture has commonalities with the cluster architecture illustrated at Site V in Figure 13, discussed below. See for instance [2] for a presentation of the design and analysis of such systems. These "legacy" architectures have proven their ability to provide highly available services. In addition, this approach has several other advantages. In most systems, it hides the fault-tolerance mechanisms for the application designer. The synchronization between replica incurs little or no delay, and the strict real time QoS requirements of communication systems are met. However, these systems are dedicated systems and are more expensive than off-the-shelf computer and database systems. For some architectures scalability may be an issue.

The distributed platform

The alternate strategy 2 is illustrated in Figure 13, Site II – IV. The computing and storing hardware of the nodes in the network are not fault tolerant. The services delivered by the network, however, may be fault tolerant by having replicas of the various service providing software objects/modules in several network nodes. For instance, in Figure 13 there are two synchronous replicas of object/module A. These replicas have consistent states and a failure of one of the replicas, or the node hosting it, may be tolerated. The replication may be handled by appropriate middleware, see for instance [25]. As in strategy 1, this requires redundant computing and storage capacity in the network. However, off-the-shelf equipment may be used, the redundancy installed for the various objects/modules may be tailored to the dependability requirements of the services they provide. Thereby, an increased flexibility and a reduced cost may be achieved.

It is a common objective that replicas shall be location and fault transparent, i.e. clients/users of the services shall not need to know where the replicas are located or whether some of the replicas have failed. Achieving a flexible replication scheme and consistent replicas across an unreliable network is no simple task. For more than two decades, considerable research effort has been put into these problems. Tool-kits and prototype systems have been developed, cf. for instance our effort, Jgroup/ARM [29]. A standard for fault tolerance in CORBA exists [30]. This approach was also an element of the Telecommunication Information Networking Architecture (TINA) [34].

So will fault tolerance at the network level and the distributed computing paradigm rule in the future? If this shall be the case, the fault-tolerance and replication strategies applied, must be tailored to the requirements and adapted to the particularities of telecommunications. The critical issue is to keep replicas which may substitute for each other consistent, i.e. synchronized, without incurring intolerable delays and increase in workload. Some sample issues:

- QoS requirements put rather strict real time requirements on many tasks, while synchronization between distributed replicas may be time consuming, especially on non-broadcast media like WANs.

- Many tasks have short processing times compared to synchronization times between replicas, hence an approach must be chosen so an immense overhead can be avoided.

- Reduce the state-information that needs to be persistent in case of failure and require synchronization. (Some of the Internet's success as a transport network may be attributed to its statelessness.)

Systems based on the distributed computing paradigm have not yet demonstrated an operational dependability which matches that of "traditional" systems. This delays the adoption of "distributed computing" in telecommunications. It should also be remembered that the middleware platform itself may become a dependability bottleneck.

The dedicated server (cluster)

In the edge centric Internet approach, dependability requirements for the end-user services have typically not been explicitly stated. The dependability objectives are normally set by the provider as a trade-off between the loss in case of failure and the cost of increased dependability. The servers have typically been built with inexpensive off-the-shelf hardware and with an architecture adapted to the functions/services that should be provided. This is also the case for fault tolerance aspects, where the design has been ad-hoc, using a mix of proprietary solutions, as well as elements from distributed computing and commodity fault tolerant subsystem like databases. As some services have become big business, the server systems have become huge, and due to size they have frequent element failures. Hence, high availability provided by fault tolerance has become an important issue, see [31] for an interesting discussion. A commonality among these systems is a stateless front end which easily may share the processing load and state-full back-ends with fault tolerant stable storage avoiding many of the synchronization problems discussed above. Located at a single site these systems are vul-

nerable to environment failures, like fire, and hence, they may have copies at other geographical sites, eventually loosely synchronized and operating in load sharing.

At the moment it seems as if dedicated server, control and management systems with architectures targeted at their purpose and dependability requirements, built on commodity hardware and subsystems, is a likely continuation of the evolution. Hence, a migration to systems where all functions are placed transparently on a general distributed platform, as for instance foreseen in the work towards TINA (the Telecommunication Information Networking Architecture) [9], is less likely in the near future.

4 Summary and concluding remarks

It is predicted that in the future, information and communication services will be far more important and interwoven into our lives and societies than they are today. That we must be able to trust many of these services is obvious, i.e. they must be dependable. The trend is towards a single future integrated network providing these services. This leads to an increased vulnerability to failures of "the network" as substitute "services" vanish. High dependability comes at a cost. Hence, it is likely that we have to be more explicit about dependability requirements, and that there will be a differentiation with respect to dependability between services in and users of the integrated network.

The overall complexity of the information and communication infrastructure is increasing, both with respect to technical solutions and with respect to the multitude of market actors in the future, converged, integrated service provisioning scenarios. To cope with this complexity is a major challenge. Hereto, complexity has primarily been considered as a system design issue. However, there are indications that managing the complexity will become even more important in operation and maintenance. Increasing autonomy in system management is the likely means to deal with this.

The widespread use of ambient intelligence for other than toy services requires a huge effort in ensuring the dependability. At present, neither the requirements nor the means are fully understood. The core information and communication infrastructure will also be of major importance. In this domain, the requirements and solutions are better understood, at least for dealing with physical failures. With respect to computing platforms, it seems that the current trend is a pragmatic approach. Commodity equipment and subsystems, elements from distributed computing, dedicated designs, etc. are used to achieve a dependability tailored to the application specific requirements. In the transport network the trend is toward fewer layers and an integrated management. Generally speaking, there is a shift in technology which yields a higher flexibility, better resource utilisation and less cost. However this may add unmanageable levels of complexity, and care must be taken not to abandon proven and stable technologies too rapidly. The need for a stable communication infrastructure favors incremental changes rather than radical shifts in technology.

Acknowledgement

The author would like to thank the Editor, Poul E. Heegaard, Svein J. Knapskog, Victor Nicola and Otto Wittner for comments on an earlier version of the manuscript.

References

1 Aagesen, F A, Helvik, B E, Anutariya, C, Shiaa, M M. On adaptable networking. In: *The First International Conference on Information and Communication Technologies, ICT'2003*. April 2003.

2 Ali, S R. *Digital Switching Systems; System Reliability and Analysis*. McGraw-Hill, 1997.

3 Avizienis, A, Laprie, J-C, Randell, B. Fundamental concepts of dependability. In: *Position Papers for the Third Information Survivability Workshop, ISW-2000*. The Institute of Electrical and Electronics Engineers, Inc, October 24–26, 2000.

4 Baran, P. On distributed communications networks. *IEEE Transactions on Communications*, 12 (1), 1–9, 1964.

5 Beyltjens, van Houldt. System 12, switching system maintenance. *Electrical Communication*, 59 (1/2), 80–88, 1985.

6 Coan, B A, Leland, W E, Vecchi, M P, Weinrib, A, Wu, L T. Using distributed topology update and preplanned configurations to achieve trunk network survivability. *IEEE Transactions on Reliability*, 40 (4), 404–416, 1991.

7 Decina, M, Plevyak, T (eds). Special Issue: Self-Healing Networks for SDH and ATM. *IEEE Communications Magazine*, 33 (9), 1995.

8 Downing, R W, Novak, J S, Tuomenoksa, L S. No. 1 ESS maintenance plan. *The Bell System Technical Journal*, 43 (5), 1961–2019, 1964.

9 Dupuy, F, Nilsson, G, Inoue, Y. The TINA consortium: Towards networking telecommunication information services. *IEEE Communications Magazine*, 33 (11), 78–83, 1995.

10 Laprie, J-C (ed). Dependability: Basic Concepts and Associated Terminology. *Dependable Computing and Fault Tolerant Systems*, vol 5. Springer, 1992.

11 Egawa, T, Komine, T, Miyao, Y, Kubota, F. QoS restoration for dependable networks. In: *Network Operations and Management Symposium, 1998 – NOMS 98*, IEEE, vol 2, 503–512, 1998.

12 Emstad, P J, Helvik, B E, Knapskog, S J, Kure, Ø, Perkis, A, Svensson, P. *Annual Report 2003; Centre for Quantifiable Quality of Service in Communication Systems*, chapter A Brief Introduction to Quantitative QoS, 18–29. NTNU, 2004.

13 Nederlof, L et al. End-to-end survivable broadband networks. *IEEE Communication Magazine*, 33 (9), 63–70, 1995.

14 Gray, J. A census of Tandem system availability between 1985 and 1990. *IEEE Trans. on Reliability*, 39 (4), 409–418, 1990.

15 ITU-T. *Terms and definitions related to quality of service and network performance including dependability*. Geneva, 1994. ITU-T E.800 (08/94).

16 ITU-T. *Types and characteristics of SDH network protection architectures*. Geneva, 1998. ITU-T G.841 (10/98).

17 ITU-T. *ATM protection switching*. Geneva, 1999. ITU-T I.630 (02/99).

18 Jensen, T, Grgic, I, Espvik, O. Managing quality of service in multiprovider environment. In: *Proceedings from Telecom99*. ITU, 1999.

19 Kephart, J O, Chess, D M. The vision of autonomic computing. *Computer*, 36 (1), 41–50, 2003.

20 Krishnan, K R, Doverspike, R D, Pack, C D. Improved survivability with multilayer dynamic routing. *IEEE Communication Magazine*, 33 (7), 62–68, 1995.

21 Kuhn, D R. Sources of failure in the public switched telephone network. *Computer*, 30 (4), 31–36, 1997.

22 Labovitz, C, Wattenhofer, R, Venkatachary, S, Ahuja, A. Resilience characeristics of the Internet backbone routing infrastructure. In: *Proceedings of the Third Information Survivability Workshop*, Boston, MA, USA, October 2000.

23 Lai, W, McDysan, D (eds). *Network hierarchy and multilayer survivability*. IETF RFC 3386, November 2002.

24 Leiner, B M, Cerf, V G, Clark, D D, Kahn, R E, Kleinrock, L, Lynch, D C, Postel, J, Roberts, L G, Wolff, S. *A brief history of the Internet*. 10 Dec 2003.

25 Maffeis, S, Smidt, D C. Construction of reliable distributed communication systems with CORBA. *IEEE Communication Magazine*, 35 (2), 56–60, 1997.

26 Malec, H A. Communications reliability: a historical perspective. *IEEE Transactions on Reliability*, 47 (3), 333–345, 1998.

27 Malek, M. The NOMADS republic. In: *Proceedings of International Conference on Advances in Infrastructure for Electronic Business, Education, Science and Medicine on the Internet (SSGRR 2003)*, L'Aquila, Italy, 2003. Telecom Italia.

28 Masera, M, Bloomfield, R (eds.). *A dependability roadmap for the information society in Europe; Part 1 – An insight into the future*. Technical Report Deliverable D1.1., Project IST-2001-37553; Accompanying Measure System Dependability (AMSD), August 2003.

29 Meling, H, Montresor, A, Babaoglu, O, Helvik, B E. *Jgroup/ARM: a distributed object group platform with autonomous replication management for dependable computing*. University of Bologna and NTNU, 2002. Technical Report UBLCS 2002-12.

30 OMG (Object Management Group). *CORBA 2.5 – Chapter 25 – Fault tolerant CORBA*. September 2001.

31 Oppenheimer, D, Patterson, D A. Architecture and dependability of large-scale internet services. *IEEE Internet Computing*, (2002), September-October 2002.

32 Ossfelt, B, Jonsson, I. Recovery and diagnostics in the central control of the AXE switching system. *IEEE Trans. on Computers*, 482–491, June 1980.

33 Struyve, K, Van Caenegem, B, Van Doorselare, K, Gryseels, M, Demeester, P. Design and evaluation of multi-layer survivability for SDH-based ATM networks. In: *Global Telecommunications Conference, 1997, GLOBECOM'97*. IEEE, vol 3, 1466–1470, 1997.

34 TINA-C. *Quality of service framework*. TINA-C (Telecommunication Informatiton Networking Architecture consortium), 1994. Technical Report TR-MRK.001-1.0-94.

35 Veitch, P, Johnson, D. ATM network resilience. *IEEE Network*, 11 (5), 26–33, 1997.

36 Wittner, O, Helvik, B E. Distributed soft policy enforcement by swarm intelligence; application to loadsharing and protection. *Annals of Telecommunications*, 59 (1-2), 2004.

37 Xu, J, Kalbarczyk, Z, Iyer, R. Networked Windows NT system field failure data analysis. In: *Proc. 1999 Pacific Rim Int'l Symp. Dependable Computing*. Los Alamitos, CA, IEEE CS Press, 1999.

Bjarne E. Helvik (51) is Professor at the Norwegian University of Science and Technology (NTNU), Department of Telematics. He is also with the Norwegian Centre of Excellence (CoE) in Quantitative Quality of Service in Communication Systems. He is Siv.ing. (MSc) from the Norwegian Institute of Technology (NTH) in 1975 and was awarded the degree Dr. Techn. in 1982. He previously held various positions with ELAB, SINTEF Telecom and Informatics and as Adjunct Professor at NTH.

His field of interests includes QoS, dependability modelling, measurements, analysis and simulation, fault-tolerant computing systems and survivable networks. His current research focus is on distributed, autonomous and adaptive fault-management in telecommunication systems, networks and services. He also deals with teletraffic issues and the architecture of networks and network based systems.

email: Bjarne.E.Helvik@item.ntnu.no

Vulnerability exposed: Telecommunications as a hub of society

JAN A. AUDESTAD

Jan A. Audestad is Senior Adviser in Telenor

During the last ten years, the information and communication technology has changed the way in which we manage society in an irreversible manner. The technical and managerial infrastructures of society have become very robust against random attacks by Nature, terrorists and others. On the other hand, the infrastructures have become extraordinarily vulnerable to attacks directed at certain functions or equipment – knowing how and what to attack may disrupt the basic services of society: banking, energy delivery, logistics, production and telecommunications. It has recently been discovered that this contradictory behaviour is a result of the structure of the physical and abstract networks interconnecting society. The structural principle is called scale-freeness. In scale-free networks most of the nodes are interconnected with few other nodes, but there is a small but significant number of nodes called hubs that are connected to a large number of other nodes. The Internet and the Web are scale-free networks. In a graph interconnecting all societal activities, ICT represents the single most important hub – if the ICT infrastructure is destructed, all of society collapses.

1 Entering the era of connectivity

We have been through different epochs in the evolution of the information and communication technology (ICT). During the 1980s we digitalised the telecommunications network and developed the modern computer; during the 1990s we explored the realm of information – how to create it, how to store it, how to disseminate it, and how to retrieve it. At the beginning of the new millennium, we have just started to explore the vast area of how to interconnect the trillions of autonomous CPUs we are placing everywhere in homes, on and in our bodies, in offices, in factories, in automobiles, in actuators and sensors of the infrastructures, in earth orbit, on Mars, and so on.

The infrastructure of society is changing rapidly, and so is the way in which we organise society. The international society becomes smaller and is knit even more tightly together than before. The separation between us is no longer only "six degrees"[1)] in acquaintanceship, but in all dimensions of society.

Evolution is taking us towards a society immensely more complex than it is today. This also introduces new threats against society. The situation is soon – or perhaps already – such that the warmonger no longer needs to deploy soldiers on foreign soil but just launch an attack taking down the infrastructure of the enemy from behind his desk. This can even be done by a single individual at war by no one else than his or her own mind.

This paper describes the nature of the new threat and what it may lead to if we are not careful. The idea is based on mathematical theories discovered during the last few years, and the analysis of the new threats and their consequences has just begun.

The paper is an odyssey through some of the developments that have taken place during the last ten years where the objective is to uncover some of the new threats against the society evolution has created. The nature of these threats must be recognised and understood before we can design the countermeasures against them.

2 The anatomy of a paradigm shift

Paradigm shift is a prestigious term that must be used with the utmost care. It is too easy to claim that a paradigm shift has taken place even if the change is small, local, reversible and hardly recognised by anyone else than a small group of enthusiasts. Contrary to what many people believe, general relativity and quantum mechanics do not represent paradigm shifts in physics. They are just refinements of classical theories: Newtonian mechanics is still valid provided that the space where the phenomenon takes place is neither tiny nor immense.

A change in the direction of science or society is a paradigm shift if something huge and important has taken place that has a lasting impact. In order to be a paradigm shift the transition taking place in society

1) *Milgram discovered in 1967 that we can pass a letter to a completely unknown person anywhere in the world via about six mutual acquaintances. This is called the small world syndrome. The degree of separation in the Web is 19 degrees; that is, the billions of web pages you can possibly reach are just about 19 mouse clicks away.*

 ISSN 0085-7130

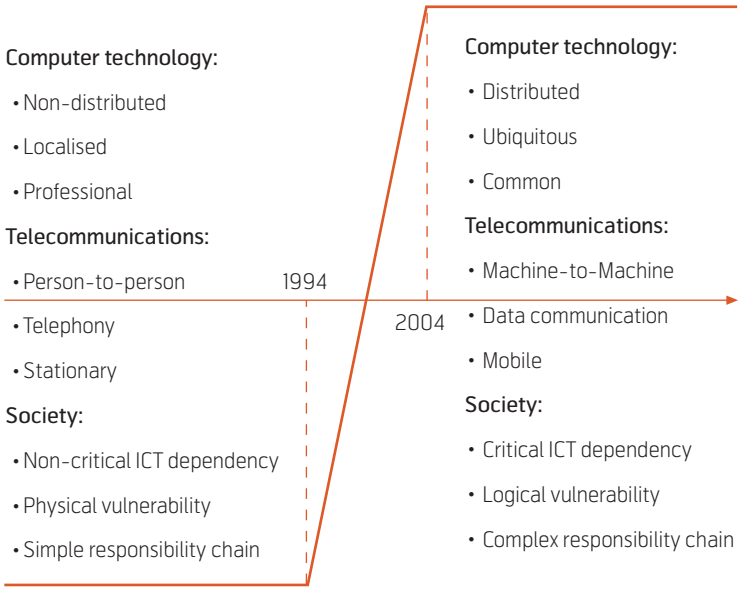


Figure 1 The extent of the paradigm shift

must be irreversible; that is, it is not possible – at least without spending much energy – to return to the state that existed before the transition started.

Such a transition has just taken place. We may call it the computerisation of society. The transition is illustrated in Figure 1.

The transition has taken place during the last ten years. Ten years may look like a long period of time but it is not. This is the time it takes to develop a new technology; it is the time it takes the market for a new technology to mature; it is the time it takes the matured market to be replaced by a radically new technology. The time constant of changes in society is thus in the order of 30 years; i.e. the time constant is of the order of one generation. Often the changes take place over much longer periods of time than this.

The change that took place during the last ten years is also much more profound than most of the earlier changes society has gone through. The gradient of the change is therefore much steeper than we are used to. The change is also irreversible: we cannot easily go back to the methods by which we previously managed society since we neither have the proper technical infrastructure nor the skills to run the earlier systems.

The extent of the change is illustrated in Figure 1. The event triggering the change was the World Wide Web. What made the change possible was that the performance/cost ratio of computers had passed a critical threshold: the personal computer had become large enough for professional applications and cheap enough for everyone.

The change would have taken place despite the Web but the start of it may have been delayed by several years and the dynamics of the change may have been less ferocious.

The major elements of this change can be described as follows.

2.1 Computer technology

Before 1994 almost all computer applications were localised to a single computer or to a small local cluster of computers. There were few applications distributed over several sites. Now almost every computer application is distributed over several computers at remote locations. The Internet makes it possible for all these computers to interact with one another.

If we define a computer to be a device containing a CPU, there were rather few devices that could be called a computer in 1994 compared with the current situation. The number of CPUs in the world has at leased increased by a factor of 1000 since 1994. It is projected that this number will increase by another factor of 10 to 100 before the end of this decennium. There will then be between 10 and 100 trillion devices containing a CPU – between 1000 and 10,000 for each of us. And this number is likely to increase steadily in the future – if not a major operation failure of the same devices brings us back to the Stone Age.

Before 1994 computers were the toys of the professionals. Now almost everyone in the industrialised world has access to a computer. Most of us cannot do a day's worth of work without using a computer directly or indirectly. At home we activate the device to play games, pay bills, buy goods or search aimlessly for information.

2.2 Telecommunications

Telecommunications has until recently been concerned with interactions between people. This is still an important application of telecommunications. However, since 1994, an increasing proportion of the traffic is taking place between people and computers. This is the application of the Internet. With the tremendous growth of the number of autonomous devices containing CPUs, the majority of traffic in the future will be between machines. All the trillions of devices will be connected directly or indirectly to the Internet. This revolutionises the whole concept of telecommunications.

Before 1994, most of the telecommunications traffic was made up of telephone calls. Now most of the traffic is data communications. This trend will continue, requiring that the telecommunications operators alter their business models in order to get revenues from the new traffic patterns. Before 1994, the idea was to merge data communication with telephony; now, telephony is merged with data communication.

Telecommunications used to take place between stationary points. Now much of the traffic takes place between mobile terminals. This trend is only increasing and it is expected that most of the traffic will soon be between mobile entities, most of these being autonomous devices containing a CPU.

2.3 Society

Society has become critically dependent on the information and communication technology (ICT). This was not the case ten years ago. Much of the remainder of this paper is about this problem.

This has caused a change in the way in which we assess vulnerability and risk. Ten years ago our focus was on natural disasters appearing randomly and intended damages caused by bombs and explosives and spreading of pathogens, poisons and nuclear waste. Now I am pleased to observe that the awareness that we may be hit by logical bombs, viruses, worms and other computer horrors causing even more damage than their physical cousins is slowly diffusing into the minds of politicians and managers.

Except for events such as pandemics, meteorite hits and worldwide war, most events causing us trouble are local and can be handled by a single organisation. Information attacks are different. They may hit all of us independent of nationality, wealth, skin colour or religious belief. The source may be a single computer run by a single person just anywhere in the world. To counter such an attack requires complex, international cooperation and coordination not just between friends but also across traditional conflict zones. To my knowledge there is no real effort to establish worldwide cooperation against information warfare. What may make such efforts fruitless is that information warfare may be part of the arsenal of weaponry a country has at its disposal – a fact you will not disclose to your enemies. This is even so on the small scale. Companies badly hit by data crime do not report such events because it may reduce their reputation in the marketplace. This type of irrationality seems to be seated in the deep abysses of our minds: lice are spread because people do not report such incidents in fear of being labelled unhygienic.

3 Emergence of a production factor: ICT and society

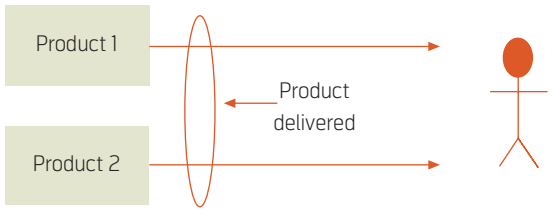
The four classical production factors are land or natural resources, labour including skills, capital in terms of assets, and entrepreneurship or the ability to exploit opportunities. These are the resources that are used in the process of production.

During the last ten years, ICT has changed the way in which goods are produced, marketed, distributed and sold. There are three aspects of this evolution: ICT has become raw material, ICT is included as a component in the product, and ICT is used as production tool.

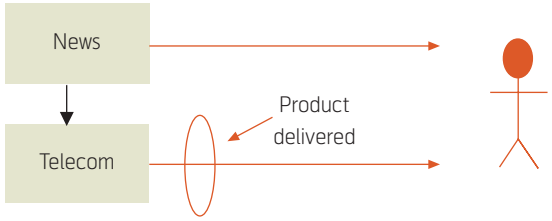
This evolution has made society critically dependent on ICT.

3.1 ICT as raw material

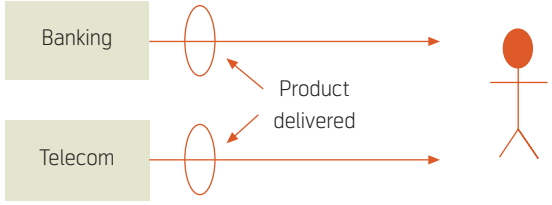
We have had products where ICT, in particular telecommunications, has been used as raw material for a long time. These are products where ICT is one of the components of the product, for example television broadcast services. However, the number of such products has increased violently during the last few years. During the dot.com boom, it became evi-



a) Actual delivery of product by cooperating networks



b) Perceived delivery of newspaper



c) Perceived delivery of electronic banking service

Figure 2 Delivery of products by independent networks

ISSN 0085-7130

dent that the pricing of these products is rather tricky and that they behave differently in the marketplace than other products. One reason is that it has been customary to regard ICT, in particular the telecommunications component, as a separate product independently of the contexts in which it is used. Sometimes the customer is not willing to pay for the product as a whole but only for the telecommunications component. This is so because the customer views the product as a telecommunications product only. The electronic newspaper is a good example.

In other cases, ICT and other components of the product are paid for separately such as in electronic banking services: the customer does not realise that ICT is an integral part the electronic banking service.

In both cases, the final product consists of components delivered in parallel by several providers. The final product is in fact put together by the customer. The customer does not always realise that this is the way it is done. In the example of the newspaper, the customer views the product as a single entity delivered by the telecommunications provider and is thus willing to pay for this single service. In the banking example, the customer perceives the services received from the bank and the telecommunications provider as independent services and is therefore willing to pay for them separately.

These principles are illustrated in Figure 2.

3.2 ICT as component of a product

The modern automobile has become a formidable computer platform containing numerous CPUs, memory chips and communication devices. The computer platform is just a part of the vehicle in the same way as the tyres, steering wheel and crankshaft. The same example holds for aircraft, television sets, buildings, household appliances and so on. ICT is just a component of the final product.

3.3 ICT as production tool

Finally, ICT is used in almost all production of goods and services. Examples are financial services, logistics, production of electric power, extraction and refining of oil and production of food. Society grinds to a halt if the ICT infrastructure stops functioning.

ICT is then taking on roles similar to the classical factors of production. In the first case it is a resource taking on the same role as land. In the third case it is an asset taking on the role of capital. In general, ICT changes the need for capital and labour.

Globalisation of the company and decentralisation of management and production are examples of the capabilities ICT offers.

4 New mathematical insight: Random networks and scale-freeness

A graph is shown in Figure 3. The graph is a very simple object but the mathematics of it is formidably complex. The mathematics of graphs is not the purpose of this paper. We will just be concerned with one particular application of graphs.

A network of any kind is a graph. Electricity grids, telecommunications networks, roads, river basins, sewage systems and freshwater supply systems are examples of networks. Acquaintanceship among people, genealogical trees, scientific co-authorship, food webs, cell metabolism, webs of political influence and industrial ownership, email, and the World Wide Web can be represented and analysed as graphs. This shifts the focus from analysing a number of different phenomena separately to study them jointly in terms of a single abstract mathematical object.

The graph consists of *vertices* that may or may not be interconnected by *edges*. Vertices that are connected by an edge are said to be *adjacent*. There may be more than one edge between adjacent vertices (not shown in Figure 3). The graph to the left in the figure is undirected. The graph to the right is directed: there is an edge from one vertex to another as indicated by the arrow but not in the opposite direction. A two-way interconnection requires two edges with arrows in opposite directions as shown.

A *path* between two vertices is a contiguous string of edges between them. The *distance* between two vertices is the shortest path between the vertices, and the *diameter* of the graph is the longest distance between any two vertices. A graph with isolated components or vertices such as the graph in the figure has infinite diameter. A graph is *path-connected* (or just *connected*) if a path exists between any two vertices.

The diameter of the acquaintanceship graph of every living human is about 6 (Milgram's six degrees). The

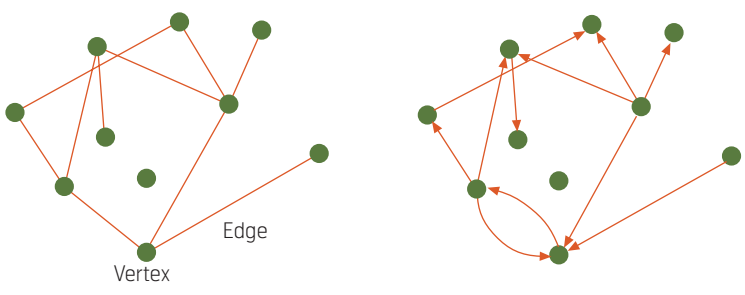


Figure 3 A graph or network

diameter of the World Wide Web is about 19. The graphs with small diameters as compared to the number of vertices are the *small world* graphs. Both the graph of acquaintanceships and the Web are small world graphs.

The number of edges at a vertex is called the *degree* of that vertex. For a directed graph we define *in-degree* and *out-degree* as the number of arrows pointing into or out of the vertex, respectively. Vertex, edge, diameter and degree are about the only terminology we need from graph theory in order to understand what follows.

We will be concerned with a particular type of graph called *random* graph. Random graphs may be generated using various algorithms. The following two algorithms are particularly easy to understand, and one of these algorithms provides us with a type of graph that is particularly important in assessing the vulnerability of the society.

The theory of random graphs is rather new. The first comprehensive work in the area is that of Erdös and Rényi (E-R) from 1959–60. Then it took almost 40 years before the next radically new step was taken: the discovery of the structure of scale-free graphs. The mathematical foundation of the theory for scale-free graphs was developed mainly by the physicists Barabási and Albert and the mathematicians Bollobás and Riordan during the last two or three years.

The E-R graph can be generated in the following way. In the starting position, the inventory of vertices of the graph is given. If the graph contains 101 vertices, these vertices can be interconnected in pairs by at most 5050 edges. Select a pair of vertices and toss an unfair coin that has probabilities 0.02 and 0.98 for showing heads and tails, respectively. If heads shows up, then draw an edge between the vertices. Continue this process for all 5050 possible pairs of vertices. A graph generated in this way then contains on average 101 edges. Since each edge contributes to the edge degree of two vertices, the average edge degree of this graph is 2 for all vertices in the graph.

For different probabilities on the unfair coin, E-R graphs with different properties can be drawn: graphs which are fragmented into isolated components, graphs which contain one giant component, graphs which contain a certain pattern, graphs that are connected and so on.

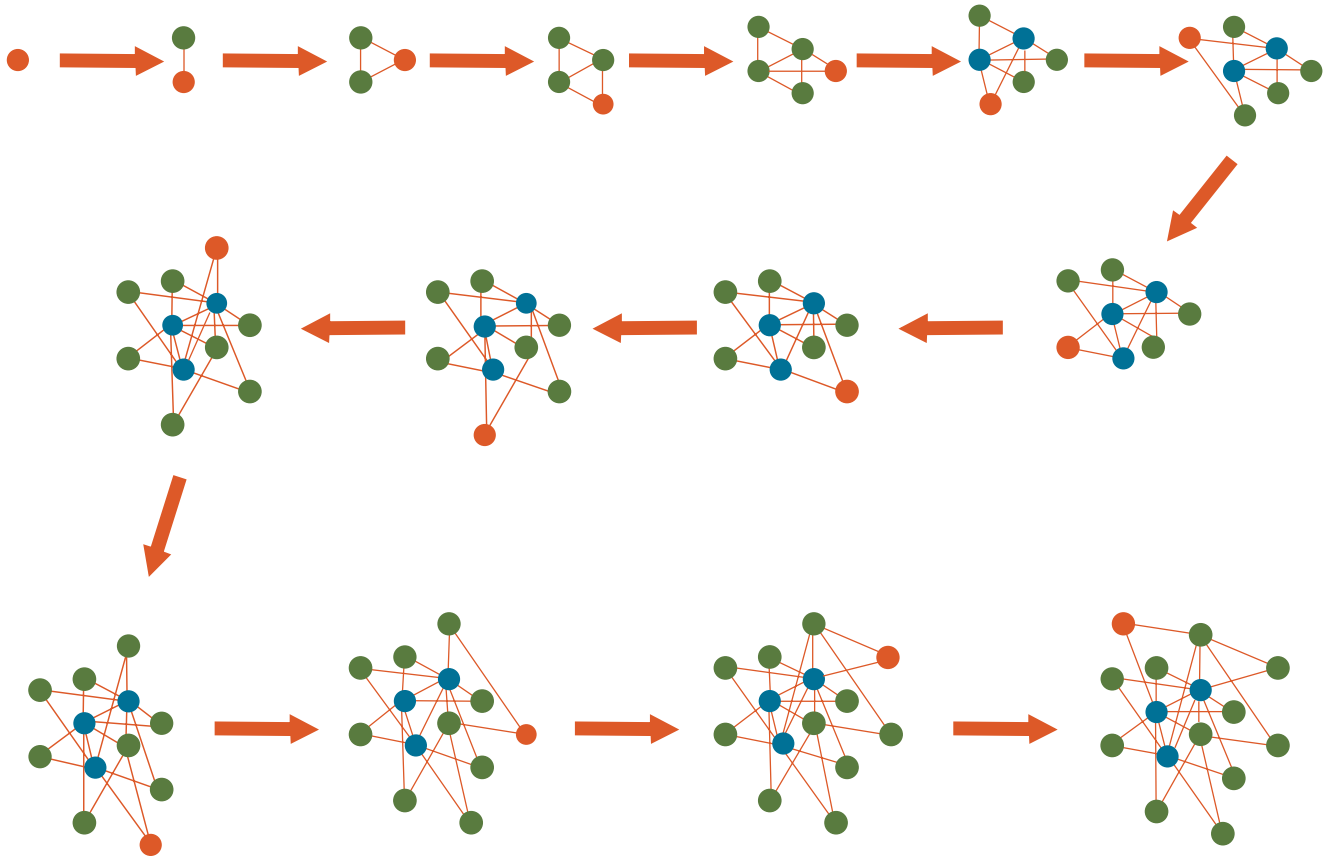


Figure 4 Growing a graph

ISSN 0085-7130

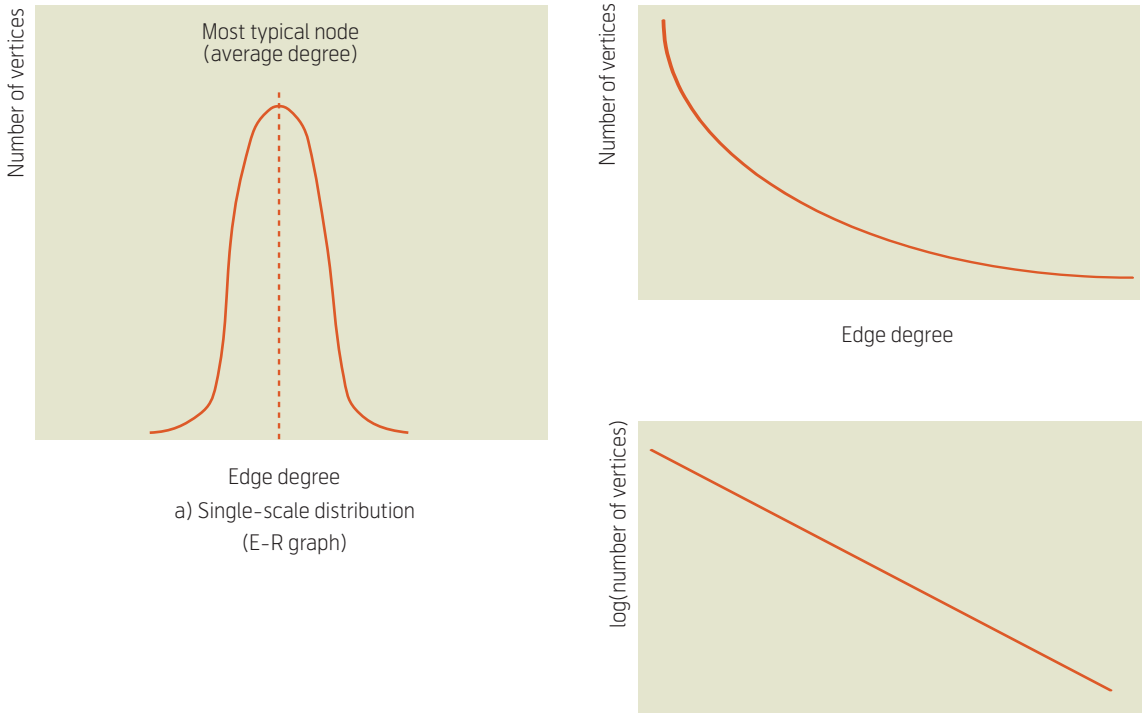


Figure 5 Distribution of edge degrees of E-R graphs and scale-free graphs

A large E-R graph is featureless in the sense that it looks the same in every section of the graph. Select any two subsets of vertices from anywhere in the graph. These subsets will look the same from a statistical viewpoint.

The other type of graph can be generated by a growth process first suggested by Barabási and later refined by Bollobás and others. The process is illustrated in Figure 4 where a graph is grown for 15 generations.

The algorithm is as follows. At each generation add a new vertex (red in the figure) and connect it to two existing vertices. The probability that there will be an edge between the new vertex and an existing one is (say) proportional to the degree of that vertex. In the fourth generation of the graph process (see the figure), there are two vertices with degree 3 and two vertices with degree 2. The probability is then 1.5 times larger that the new vertex added in the fifth generation will be connected to a vertex with degree 3 than to a vertex with degree 2.

This process results in an uneven growth pattern where the degree of some vertices increases faster than that of other vertices. A few vertices will then have a much larger degree than all other vertices. These vertices are called *hubs* (blue in the figure).

The edge degrees of a large graph grown in this (or similar) way will be distributed according to the power law $P(k) \sim k^{-\gamma}$ where $P(k)$ is the probability that the vertex has degree k and γ is a constant usually between 1 and 3. The graph is said to be *scale-free*. Observe that if $\gamma > 1$, then the average degree of a large scale-free graph approaches zero. In contrast, the degree of all vertices of an E-R graph is distributed in accordance with a Poisson distribution around the average degree of the graph. Almost all vertices in the E-R graph will then have a degree very close to the average, and the number of vertices with degrees different from the average decreases exponentially with the distance from the average. The difference between the statistical distribution of the edge degrees of E-R graphs and scale-free graphs is shown in Figure 5.

The difference between the two types of statistical distributions is striking. The phenomena resulting in E-R graphs and scale-free graphs are just as different.

Most natural graphs are not strictly scale-free; for example, the log-log graph may be slightly curved or certain edge degrees may not exist at all. It is then more fair to say that the graph is multi-scaled for which the power law distribution is a reasonable approximation enabling us to draw important conclusions and gain new insight. In analogy, it is rather uninteresting to calculate the orbit a particular leaf follows when it falls to the ground: it is more enlightening to understand which forces make it fall to the

ground in the first place and which forces prevent it from following in a straight line.

The growth process in the example above is just one of several ways in which a random scale-free graph may evolve. Phenomena that may cause scale-freeness are still being investigated.

Scale-free graphs also share several properties with fractals and critical phenomena in chaotic systems such as percolation and phase transitions in thermodynamics. This offers access to a wealth of mathematical methods and results from other fields of mathematics and other mathematical sciences. Furthermore, it has recently become apparent that scale-free graphs are essential for understanding several aspects of biological and social systems, psychology, politics, business management and linguistics. This opens for a type of cross-disciplinary insight we have never mastered before.

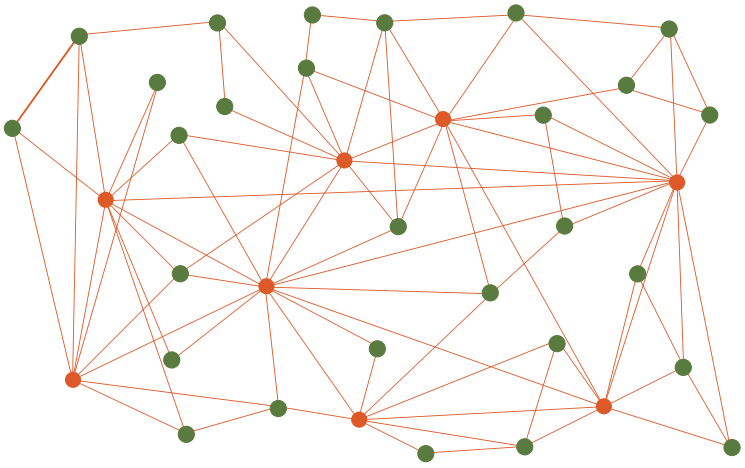


Figure 6 Example of scale-free graph

5 You cannot have your cake and eat it: Robustness and vulnerability

Figure 6 is an example of a small scale-free graph containing eight hubs (in red). This graph is used to illustrate the robustness and vulnerability of scale-free graphs. The graph is small but suffices to illustrate the most important conclusions.

Albert et al. define robustness of a graph in terms of how much the average distance between vertices in the graph increase if a certain number of vertices are removed randomly from it. If vertices are removed from an E-R graph, the average distance increases evenly. On the other hand, the average distance increases only slowly if vertices are removed from a scale-free graph.

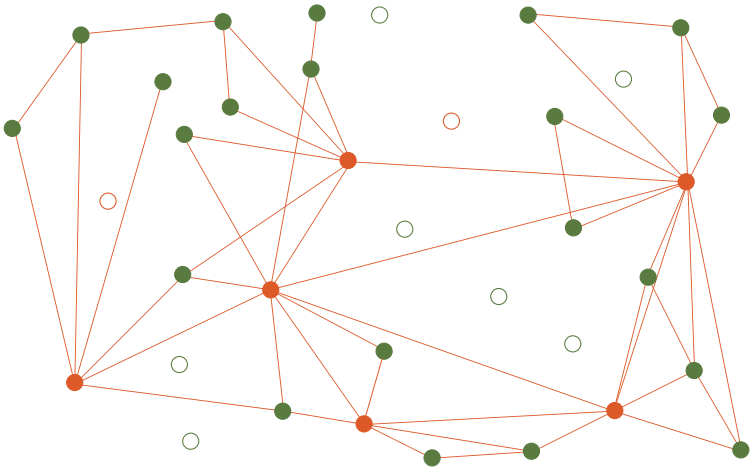


Figure 7 Removal of random nodes

Figure 7 shows the effect on the graph in Figure 6 if nine random vertices, corresponding to 25 % of the vertices, are removed. In the process even two hubs were removed. The average distance has increased from approximately 2 to approximately 3. If the two hubs had not been removed, the increase in the average distance would hardly be measurable. The general observation is that scale-free graphs are very robust against random attacks: remove a few billion pages from the Web and hardly anyone will recognise it; remove a few thousand routers from the Internet and the traffic will continue to flow unhindered.

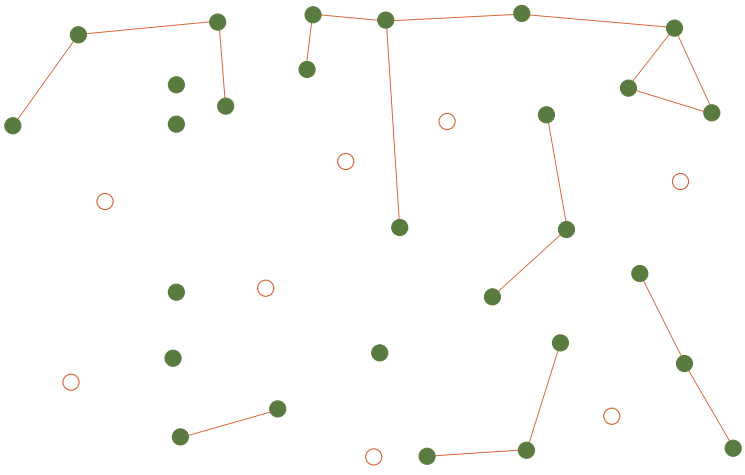


Figure 8 Removal of hubs

If the attack is directed against the hubs, the graph disintegrates into isolated components as shown in Figure 8. If terrorists want to take down the Web, then the best they can do is to take out the largest hubs, namely the search engines. The Web then dis-

integrates into myriads of disconnected islands of web pages that are impossible to find. Similarly, find the location of particularly large routers and some of the centralised functions in the Internet such as interconnect management servers and name servers and take them out. The traffic handling capability of the Internet then plummets rendering the network useless. Taking out a number of random routers using explosives has no effect at all on the network at large. This is only a nuisance and an expense for the owner of the building in which the router is placed.

If emails containing viruses are filtered out by organisations having many large address lists, the spreading of the virus is hampered. The dissemination of the SoBig virus was slowed down in this way.

These simple examples illustrate what can be proved by rigorous mathematics, namely that scale-free graphs are very robust against random attacks but extremely vulnerable to directed attacks. The latter example also shows that it may be possible to protect the network against directed attacks if the structure of the network is properly understood.

The robustness is also a clue to the understanding of other phenomena.

Nature acts blindly. An error in the chemistry of a cell caused by, say, radiation is thus a random attack on the metabolic network of the cell. In the metabolic network, the vertices are chemical substances and an edge between two substances indicates that they take part in the same chemical reaction. This network is scale-free and is thus robust against random changes. This robustness makes organisms survive even in the most hostile environments where mutations take place frequently. Neural networks are scale-free. This ensures that the brain is working perfectly well even after millions of neurons have died. Food webs are also scale-free: removing species at random from an ecological system usually does not cause total collapse of the system but removing just a single species may sometimes destroy the ecology. In the latter case, the removed species was a hub in the web.

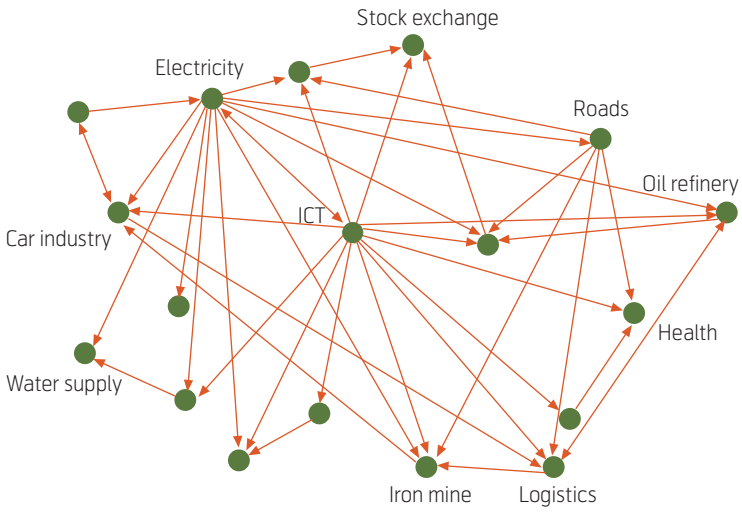


Figure 9 Society as a graph

It is not surprising that natural selection favours scale-freeness since such graphs are not likely to be destroyed by random incidents.

The problem with the Internet, the email service and the Web is that they are not attacked by someone like Nature acting blindly but by someone who can identify and take out just those few functions that cause the whole system to disintegrate. This and the extended role of telecommunications described above are the two factors that cause a dramatic change in the vulnerability of society.

6 Telecommunications as a hub of society

As shown in Figure 9, all activities taking place in society can be visualised as a directed graph. The vertices represent the activities that keep up the hustle and bustle of society; a directed edge indicates that the activity the arrow points at cannot be completed without the aid of the activity at the root of the arrow. The edges of the graph may even be weighted in various ways where the weights associated with an edge may indicate, for example, how strong the dependency is. Similarly, a separate graph may be drawn for every single activity, including service provision and industry, just in order to study how this activity depends on other activities.

Altogether it is possible to describe society in numerous ways using graphs, focusing on particular sets of activities, variables and dependencies. The discovery of the properties of scale-free graphs may stimulate such research since it is possible to draw important conclusions by just observing the structure of the graph.

The complete graph structure of society is immeasurably complex. Even the graph of a single industry is immensely complicated. However, many of the graphs of society are likely to resemble scale-free graphs. This means that we may use new insight to assess how robust and vulnerable society is to different types of incidents. One example in epidemiology is that random vaccination of people against AIDS does not slow down the spreading of the disease. This is so because AIDS spreads in a scale-free graph. The

only efficient method to bring an end to the pandemic is to identify who may be the hubs in this graph and vaccinate them. This is not just efficient and cheap but it creates ethical dilemmas since one core issue of identifying the hubs is to pick out promiscuous homosexual mails. On the other hand, random vaccination is democratic and free from ethical dilemmas, and it may give a haunted population the impression that something is done in order to help them.

Figure 10 shows one particular graph of society. The graph is three-partite and consists of a single vertex (telecommunications) at the upper layer, a number of independent clusters of electric power suppliers at the middle layer, and the rest of society at the bottom layer. The directed arrows indicate that every activity of society depends on electric power and telecommunications.

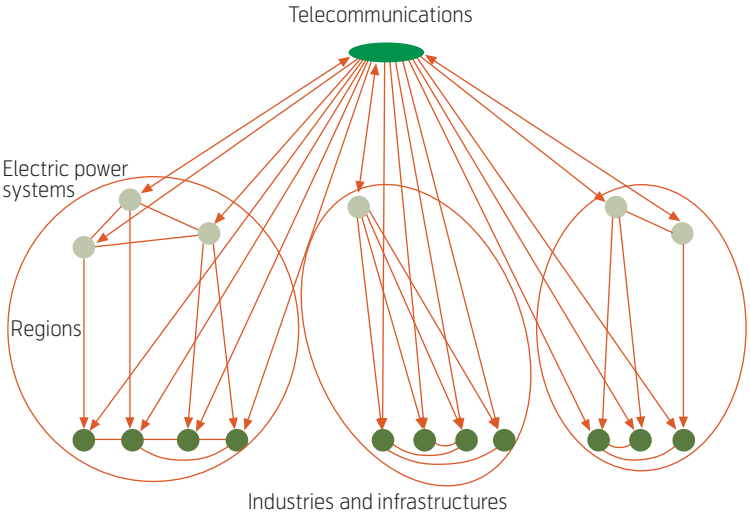


Figure 10 The hubs of society

There is only one single telecommunications system reaching into terminals and computers everywhere on the globe. Every other organisation or societal activity, except primitive agriculture and fishery, cannot function without telecommunications. Telecommunications is thus a single hub that every other activity depends upon: taking down telecommunications is the same as taking down society. In principle, it suffices to launch a single attack against telecommunications in order to take it down globally, though it may be extremely hard to figure out how this can be done. However, what is important is not to analyse how the telecommunications system itself can be destroyed but how this system can be used to destroy other systems.

The electric power system is also a hub, or rather several hubs of society. Electric power systems are regional. This means that if there is a problem in one region, the other regions are not influenced by this problem. In order to take down the power system it is necessary to launch separate attacks against each region. However, several incidents lately have shown that the regional systems are vulnerable not only to directed attacks but to any type of attack because the power grid is sensitive to cascading failures: if it fails at one point, overload may cause secondary failures in other parts of the system. The vulnerability of the power grid is thus different from that of telecommunications.

It is well known in epidemiology that a disease usually starts to spread only if a sufficiently large population is infected. The threshold depends on how virulent the disease is, how it spreads from person to person, and the density and mobility of the population. Usually, the initial phase of the epidemic develops very slowly. This is a threshold effect found in all E-R graphs. One exception is AIDS: the Euro-American epidemic spread essentially from one person.

One of the early findings concerning scale-free graphs is that these graphs possess no threshold. This explains the onset of the AIDS epidemic. Furthermore, we know from experience that a data virus is usually spread from a single source. That it takes the virus only a few hours to spread over the whole network is owing to the scale-free nature of Internet.

Therefore, the importance of telecommunications as a hub lies thus in the fact that it can be used to disseminate efficiently data viruses that destroy the computers connected to the network. These are the computers we use in order to run and manage our society. If the data virus is the bullet, the telecommunications network is the gun. So instead of taking down society by destroying the telecommunications network it is far more effective, and probably much simpler, to damage society by utilising the ubiquity and the scale-freeness of the telecommunications network to launch an attack against the computers connected to this network.

7 Stand up and fight!

During the last ten years we have made society critically dependent on computers. Everything we do now is done much more efficiently and cheaper than before; we can do things that were impossible ten years ago; and we have transformed our way of living. The price tag is that we have made society vulnerable to attack by virtually anyone.

We have seen that the source of the increased vulnerability is that the ICT infrastructure has the shape of scale-free graphs. Furthermore, we may expect that

many of the dependency graphs of society also are scale-free. These trivial observations indicate that we cannot assess the vulnerability of society unless we understand the properties of random scale-free graphs. It is only recently that we have developed the mathematical tools that enable us to study these aspects of society.

Intuitively, it can also be inferred that protecting scale-free networks require different countermeasures compared to those required for protecting networks and systems not having this property. The way in which we historically have organised our defences is not based on such knowledge and is likely to be ineffective against the new threats. These threats did not exist ten years ago, and their mere existence were recognised just a couple of years ago by groups remote from the power hierarchies of society. There has simply not been enough time for the authorities to recognise the full impact of the new threats.

Scale-freeness is not just a threat. The phenomenon exists both in nature and society because it offers many benefits, in particular, protection against the caprices of random processes. The menagerie of benefits is still being explored. These studies may even provide us with tools that can be used as efficient countermeasures against deliberate destructions.

Bibliography

Albert, R, Barabási, A-L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74, 2002.

Albert, R, Jeong, H, Barabási, A-L. Error and attack tolerance of complex networks. *Nature*, 406, 2000.

Audestad, J A. Challenges in telecommunications. *Telektronikk*, 98 (2/3), 159–182, 2002.

Barabási, A-L. *Linked: The New Science of Networks*. Perseus Publishing, 2002. This is a popular introduction to random graph theory requiring no knowledge of mathematics.

Bollobás, B. *Random Graphs*, 2nd edition. Cambridge University Press, 2001. (The theory of E-R graphs)

Bollobás, B, Riordan, O. Robustness and vulnerability of scale-free random graphs. *Internet Mathematics*, 1 (1), 2003.

Complexity, 8 (1), Special issue: Networks and Complexity. September/October 2002.

Godoe, H. Innovation regimes, R&D and radical innovations in telecommunications. *Research Policy*, 29, 2000.

Mandelbrot, B. *The Fractal Geometry of Nature*. Freeman, 1983.

Milgram, S. The small world problem. *Physiology Today*, 2, 1967.

Spencer, J. *The Strange Logic of Random Graphs*. Springer, 2001.

Stabell, C B, Fjeldstad, Ø D. Configuring value for competitive advantage: On chains, shops, and networks. *Strategic Management Journal*, 19, 1998.

Jan A Audestad (62) is Senior Adviser in Telenor. He is also Adjunct Professor of telematics at the Norwegian University of Science and Technology (NTNU), and he holds a professorship in informatiton security at Gjøvik University College, where his main task has been to build up a new master degree in information security, sponsored, amongst others, by Telenor. He has a Master degree in theoretical physics from NTNU in 1965. He joined Telenor in 1971 after four years in the electronics industry. 1971 – 1995 he did research primarily in satellite systems, mobile systems, intelligent networks and information security. Since 1995 he has worked in the area of business strategy. He has chaired a number of international working groups and research projects standardising and developing maritime satellite systems, GSM and intelligent networks.

email: jan-arild.audestad@telenor.com

Radio interface and access technologies for wireless and mobile systems beyond 3G

GEIR E. ØIEN

Geir E. Øien is Professor of Information Theory at NTNU, Trondheim

In the present paper we describe some of the radio interface and access technologies and principles that we believe will become most important for future wireless and mobile communication systems beyond 3G (B3G). We cover multiple-input multiple-output (MIMO) technology, link adaptation techniques, multi-carrier modulation (OFDM), iterative ("turbo") receiver processing, and cross-layer design methodologies. We seek to highlight the main advantages (and, if any, disadvantages and limitations) of each component technology described, with minimal use of mathematics. The intention is to give an overview without going too deeply into technical details; however, an extensive and up-to-date reference list is provided so the interested reader can know where to go for more information on a particular topic.

1 Introduction

Predicting the research focus and market development of a field that changes as quickly as communication technology, in particular in the wireless field, is not a task that can ever be executed with 100 % accuracy. Looking 15 years into the past, many of today's most important and penetrating communication technologies (e.g. GSM, WWW) were not yet in public use. Taking into account the accelerating pace of improvements and cost reductions with regard to enabling technologies, as well as the ever-increasing impact that communication technology has on the human lifestyle, it is likely that the future may hold even greater surprises than did the past. However, some basic trends are emerging.

Many different groups and organisations have made predictions and foresights of future communication technological developments; see, for instance, the *CELTIC* initiative [CEL1]. With regard specifically to wireless communications, the book [KAR03] discusses possible scenarios for the "mobile world" in 2015. Four different possible scenarios are suggested, differing widely with respect to economic, political, and technological developments. In this paper we will not assume any particular scenario in the economic and political sense, since this necessarily will depend on several factors outside our field of competence. Rather we will focus on the more "objective" truths of technology development, and discuss certain basic technological aspects of how a technologically sound and efficient future wireless/mobile communication system should be designed. For the most part, with one exception, we will restrict ourselves to what is usually called "lower-layer" technological challenges (referring to the OSI protocol layer stack).

More specifically, our goal is to provide an overview over some of the *radio ("air") interface and access* technologies we believe will be of greatest importance and have the greatest promise as basic enabling technologies in future wireless and mobile systems. We will not go deeply into technical details; rather, we will try to highlight some of the major technology trends and buzzwords and what the advantages of these technologies are. Since the scope of necessity is quite broad and space is limited, we can only scratch the surface of the various topics – as well as providing a diverse list of state-of-the-art, up-to-date references for further reading.

2 Air interfaces and radio access – an overview

As defined in the present paper, the "air interface" and "radio access" part of a communication system can be thought of as residing between the mobile or wireless terminals and the core network, and mainly encompassing technology components for:

- Channel (i.e. error control) coding/decoding
- Modulation/demodulation/detection
- Synchronization
- Channel estimation
- Equalization
- Interference management
- Multiuser access and scheduling

Note that some of these operations may, depending on the technology used, be performed in a joint and/or iterative manner, e.g. coded modulation, joint estimation/detection, iterative multiuser decoding/detection, etc.

It is not our intention to provide an exhaustive overview of *all* possible candidate technologies for implementing the air interfaces/access technologies in future mobile and wireless systems; rather, we will concentrate on a select few technology components and principles that we believe have the greatest

 ISSN 0085-7130

potential in the realization of spectrally efficient, reliable, flexible broadband multimedia wireless and mobile communications. Before going into the technologies specifically, we will however briefly review the currently more or less broadly accepted picture of how future wireless and mobile communications systems will most probably develop.

3 Wireless and mobile communications beyond 3G: An overview of prevailing trends

In the future, wireless communications will become almost ubiquitous, and the ability to communicate seamlessly, nomadically, and with increased mobility will be ever more important. Relevant perspectives on the development of so-called 4th generation (4G) – or "Beyond 3rd Generation" (B3G) – wireless technology are described in [WWRF, 4GMF].

Also, the *International Telecommunications Union* (ITU) has developed a Recommendation – ITU-R M.1645 – for B3G [ITU1], and a corresponding list of key research questions [ITU2] providing guidance for researchers wishing to contribute to B3G development. The recommendation reads as follows (IMT-2000 being yet another way of saying "3G"): *"Systems beyond IMT-2000 will be realized by functional fusion of existing, enhanced and newly developed elements of IMT-2000, nomadic wireless access systems and other wireless systems, with high commonality and seamless interworking."*

The research goals as established by the ITU for the novel B3G radio interfaces indicate ubiquitous coverage with target (downlink) data rates in the range of 1 Gb/s at low mobilities and 100 Mb/s at high mobilities. A non-exhaustive list of the key research questions considered by the ITU include (focusing on the points most relevant for air interface and radio access development) [ITU2]:

- Would there be any significant constraints in achieving the target data rates, consistent with the timelines as discussed in the Recommendation ITU-R M.1645, related to terminal speed, or to the power consumption and the consequences for the battery in the terminal?
- What would be the RF channel bandwidth(s)?
- What spectrum efficiency can be achieved?
- If the target bit rate for the uplink of the radio interface was lower than for the downlink, how would this affect its characteristics?
- To what extent could a common or adaptable radio interface be capable of meeting the goals for both high mobility such as mobile access and low mobility such as nomadic/local wireless access?
- To what extent might the evolution of IMT-2000 and other existing radio interfaces meet these goals?
- What technology developments in the radio access network might reduce the need for additional spectrum for systems beyond IMT-2000?
- What are the implications of different technologies (antenna concepts, physical layer concepts, higher protocol layers, etc.) and system architectures (such as cellular type deployments, multihop system based deployments, etc.) on the feasibility of the data rate targets, the range, the system capacity etc.?
- What is the impact of different frequency ranges and mobility targets on the feasibility of the data rate targets?
- What is the spectrum demand for initial coverage in the deployment area for different system architectures and technology advances?
- What would be the maximum size of a cell (or the coverage per fixed node)?

It is clear from this list that there are still a lot of open and nontrivial questions related to the development of radio interfaces/access techniques for novel wireless communication systems. However, it seems clear that B3G will ultimately represent a convergence between fixed wireless access, wireless mobile, wireless LANs, and packet-division-multiplexed networks [4GMF]. One integrated terminal with one global personal number will, in time, most probably be able to access freely any wireless air interface.

The radio transmission modules will become fully software-definable, re-configurable, and programmable (often referred to as *software radio* or *cognitive radio*). The network will be heterogeneous with regard to available and interacting technologies, content, and services; re-configurable and adaptive; with different sub-networks interconnected and with different capacity and bandwidth requirements, channel characteristics, transmission technologies, processing capabilities, desired and available services, and subnetwork architectures at the different levels.

The technologies, subsystems and bandwidths involved will depend on communication distance and will include Personal Area Networks (PANs) over

ultra-short distances (< 10 m), broadband WLAN technology within small cells (< 100 m), a backbone cellular system for communication over medium ranges (< 1 km), all the way up to satellite links for providing ubiquitous access to more basic services in remote and sparsely populated areas.

One of the most important issues governing future development of wireless technologies, services, and business models is the regulation of the available radio spectrum. In the US, there is currently a move towards regulatory reforms that may eventually free up enough radio frequency (RF) bandwidth to significantly influence the development of mobile telephony and wireless Internet services [SPEC]. As an example, 85 MHz bandwidth previously allocated for analogue UHF broadcasting is being released for use in future mobile communications services. Digital terrestrial broadcasting (DVB-T) exploits the bandwidth far more efficiently, facilitating both more TV programs and extra bandwidth for other services. Frequency reallocation, spectrum leases, and spectrum sharing will allow the use of several new frequency bands for future mobile communications. These frequencies will find different uses depending on range and capacity needs, with the highest frequencies being reserved for high-capacity short-range communications.

To sum up, the currently prevailing paradigm for the implementation of B3G is that the core network will evolve toward an IP-based network, serving a wireless Internet radio access based on packet switching for all services, including voice [NEW]. The major trend will be a move towards higher frequencies (above 5 GHz), leading to a nano- or pico-cell structure which will make it difficult, if not impossible, to design the network and provide continent-wide coverage on the basis of the standard cellular concept known from e.g. GSM. Rather, the network will evolve towards a structure of multi-layered ad hoc networks, not one rigid network structure. In such a structure base stations may be installed where they are needed, and connected to each other in a self-configuring way, to transfer IP traffic similarly to present Internet wired architecture [NEW].

In such a heterogeneous network architecture distributed high-speed WLANs will serve local hot spots, seamlessly inter-connected by the overlayed backbone cellular network and by a wired infrastructure. A multitude of wireless sensors will also be embedded and integrated in the network for dynamic and intelligent interaction and communication between users and devices, as well as directly between devices.

It is clear that there will be a strong demand on B3G technology to deliver much higher data rates and more diverse services than 2G and 3G systems. The design of suitable radio interfaces has to take this into account; the dominant traffic load will be high-speed burst-type traffic. This is a great challenge for all existing radio interface technologies, and novel developments must be made. In the next section we will focus on describing the air interface technology components that we believe will become central in B3G radio interfaces. Ultimately these technologies may enable a potential increase in bandwidth efficiency by a factor of 10 to 100 compared to today's wireless systems like GSM and UMTS.

4 Some important air interface and radio access technologies for B3G

4.1 MIMO technology

Multiple-Input Multiple-Output (MIMO) technology might be the single most important factor for enabling dramatically increased capacity and link reliability in future wireless and mobile communications systems, as well as for improving multi-user transmission, detection and interference cancellation. A block diagram of a generic MIMO system is shown in Figure 1.

As seen from the figure, the key feature of a MIMO system is the use of *multiple transmit and receive antenna elements* (antenna arrays). MIMO can thus be viewed as an extension of the original concept of "smart antennas" (where multiple antenna elements are used on *either* the transmitter or receiver side). For readers interested in exploring technical details beyond what is given in the present paper, an excellent tutorial on MIMO systems can be found in a previous issue of *Telektronikk* [G&A02].

It was first demonstrated by Foschini and Gans in 1998 [F&G98] that dramatic capacity increases might be enabled on wireless links by equipping *both* the transmitter and the receiver with multiple antennas, as shown in Figure 1. By so-called *spatial multiplex-*

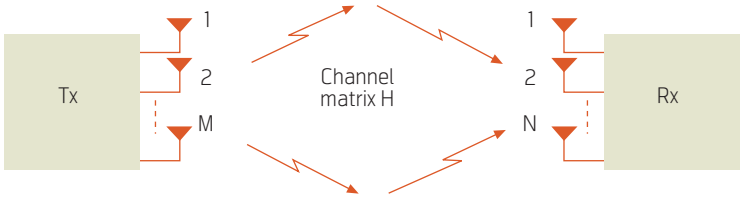


*Figure 1 Generic MIMO system with transmitter (Tx), MxN channel matrix (**H**) containing fading coefficients of all spatial subchannels, and receiver (Rx)*

ISSN 0085-7130

ing, such MIMO systems can enable the same bandwidth to be reused many times within the same cell. In effect, intelligent space-time signal processing is then used to open up *several parallel "virtual data pipes"* between a transmitter and receiver within the same bandwidth. Each of these data pipes will have the same channel characteristics as an ordinary single-input single-output (SISO) wireless fading channel. Independent data streams can be transmitted through each of these virtual SISO pipes, and the resulting capacity is then the sum of the capacities of the individual data pipes. This is called the *spatial multiplexing gain*.

The remarkable thing is that this effect comes because of, not despite of, the multipath fading properties of a typical wireless channel. Fading, which traditionally has been considered as an impairment – not an asset – of the channel, is actually exactly what enables MIMO systems with spatial multiplexing to achieve their remarkable spectral efficiencies. This may seem counter-intuitive at first, but is a simple consequence of the fact that the MIMO channel can be decomposed, by appropriate space-time processing techniques, into a set of parallel data pipes. The richer the scattering environment (i.e. the more "severe" the multipath fading), the more such independent pipes can be created.

Another possibility is to use the available degrees of freedom provided by the multiple antennas to design processing schemes to *stabilize* the quality of the wireless link (mitigate fading) and thus to achieve *more reliable* communications (lower error rate) for a given average throughput. This is called the *spatial diversity gain*, and systems exploiting all their degrees of freedom for maximum spatial diversity gain are often referred to as *MIMO diversity systems*. An illustration of how spatial diversity mitigates fading and stabilizes the channel quality is given in Figure 2. Here, the black and green lines are fading levels of two independently fading channels whereas the red line is the fading level after optimal diversity combining of the two independent channels. Thus the diversity is of order 2 (i.e. 2 receive antennas are used). It is clearly seen that the combined channel exhibits less severe fading (fewer deep fades) than each of the two individual channels.

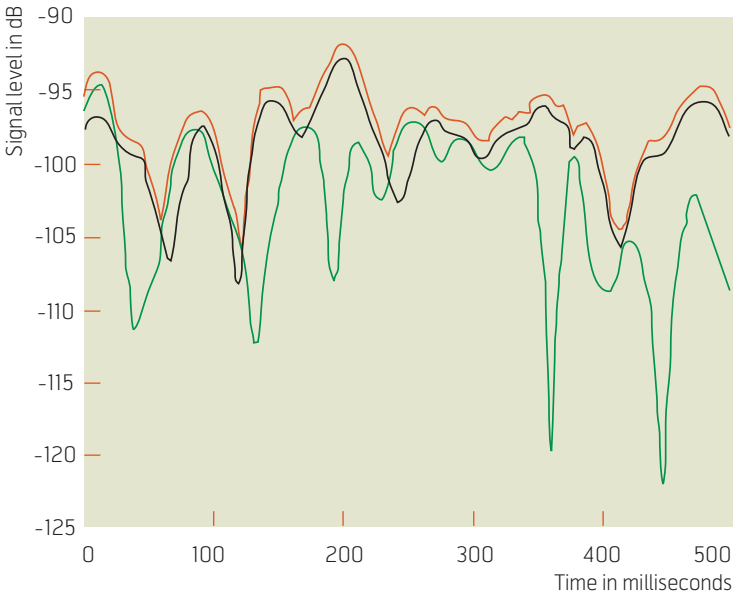


Figure 2 Illustration of spatial diversity gain (courtesy of prof. David Gesbert)

Trade-offs between the two extreme possibilities of spatial multiplexing and diversity combining, which seek to exploit partially the multiplexing gain and partially the diversity gain, are also possible.

Assuming a MIMO system with M transmit antennas and N receive antennas available, a total of MN different fading radio channels ("subchannels") are set up between the transmitter and receiver. In the original MIMO theory developed by Foschini and Gans and subsequently used and refined by a host of researchers worldwide, these subchannels were usually assumed to be mutually uncorrelated. This is usually a good assumption if the antenna array elements are placed close to half a carrier wavelength apart. The assumption provides an upper bound on the achievable MIMO spatial multiplexing or diversity gain, since all subchannels in this case fade independently of each other. However, the assumption of uncorrelated subchannels is far from true in all practical cases, e.g. when the antenna elements are placed too close together (as might be the result on a small-sized terminal with MIMO capability). If there is strong spatial correlation the individual subchannels are more prone to fade "synchronously," and subsequently the potential gain from combining them is reduced.

To explain how the gain of spatial multiplexing appears, in a straightforward way, let us now assume that all the MN subchannels in an MxN MIMO system are indeed uncorrelated. Furthermore, we assume that the multipath fading on these channels has stationary statistical properties which are also subchannel-independent. At first glance, and assuming that it is possible to resolve all the MN paths by means of some proper space-time processing scheme, one might at first think that the capacity can potentially be increased by a factor of MN compared to a traditional single-input single-output (SISO) system. However, this is not the case since in a power-limited system, as is usually assumed, the total transmit power P [W] of course has to be shared between the M transmit

antennas. Therefore, the average available power *per transmit antenna element is only P/M* [W].

Foschini and Gans [F&G98] showed, assuming Rayleigh fading statistics and Additive White Gaussian Noise (AWGN) on each subchannel, that the *ergodic channel capacity* or *Shannon capacity* (which for our purposes can be thought of as capacity averaged over time in a mobile system) of such a MIMO system grows *linearly* with the smallest number of M and N, when both M and N are large. Mathematically, we may then write the approximate ergodic capacity as

$$C = K\log_2(1 + \rho) \text{ [bits/s/Hz]},$$

where $K = \min(M,N)$ and ρ is the received signal-to-noise ratio (SNR) at one of the receiving antennas. The potential capacity increase of a MIMO over a SISO system is illustrated in Figure 3 for $M = N = 4$.

The above ergodic Shannon capacity is an asymptotic and theoretical limit on the average amount of information that can be transmitted reliably (with an arbitrarily chosen maximal error rate) through the MIMO channel over time. In practice we must of course use some transmission scheme that is implementable with finite complexity and delay, which will limit the practical capacity. For this purpose, so-called *space-time codes* have been developed. In particular, space-time *block* coding (STBC) has become a popular technique. Space-time codes can be designed either to exploit the spatial multiplexing gain [Fos96] or the spatial diversity gain [Ala98,Tar99].

The idealized MIMO theory presented above has also now been extended to include spatially correlated channels. It has been shown that even when quite high spatial correlation is present between the subchannels, a significant spatial multiplexing gain (which translates to potential rate increase) and/or spatial diversity gain (which translates to more stable channel quality) can still often be achieved. This means that MIMO technology holds a great deal of promise for a wide range of practically important scenarios, and not only in the idealized case.

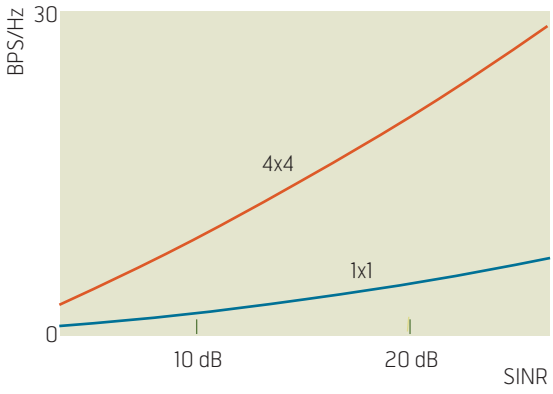


Figure 3 The ergodic capacity (in bits/s/Hz), as a function of the channel signal-to-interference+noise ratio in dB, for a 4x4 MIMO system (red) and a SISO system (blue) (Courtesy of prof. David Gesbert)

4.2 Link adaptation techniques

Link adaptation is a prerequisite for optimal capacity exploitation in temporally and spatially varying radio channels, both in the single-link and multi-user (network) case. The principle is to adapt throughput (transmit spectral efficiency) and power on a link dynamically to predicted channel conditions, to maximize the ergodic channel capacity while adhering to some given design quality constraints. In this way one avoids the "worst-case" choice of transmission schemes that must be made in non-adaptive systems, which therefore typically only can exploit a fraction of the theoretically available capacity.

In an adaptive transmission scheme, the receiver estimates the channel state, and the obtained channel state information (CSI) is fed back to the transmitter via a feedback (also called return) channel. Typically the CSI fed back will be an index indicating in which sub-range (out of a finite number of possible sub-ranges) the instantaneous channel-signal-to-noise ratio (CSNR) resides.

The transmitter subsequently uses the CSI to update the transmission parameters such as modulation constellation, error-control code, and/or transmit power level. This principle is illustrated in Figure 4. Here, n is the index of the transmitter mode (possible com-

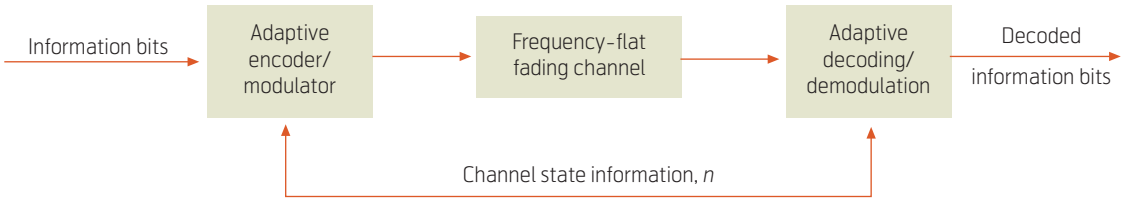


Figure 4 Block diagram of a link adaptation scheme where the channel codes and modulation constellations are adaptive

ISSN 0085-7130

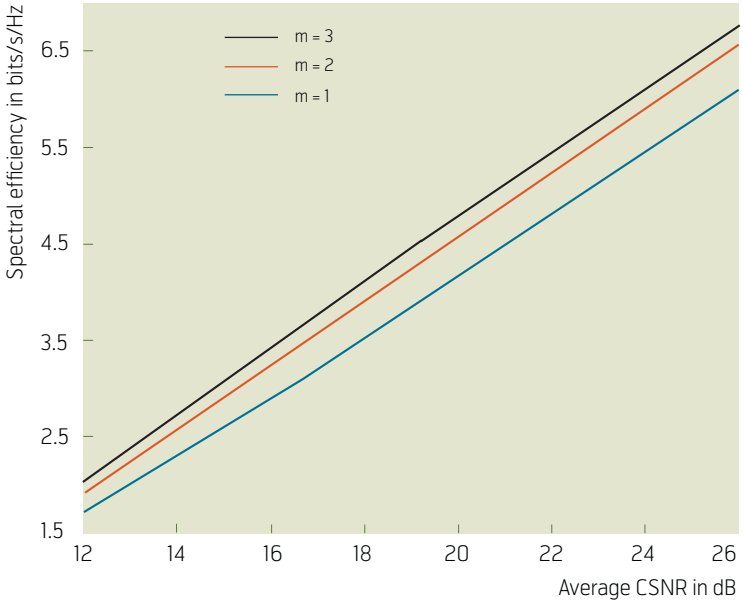


Figure 5 An example of the achievable capacity of a realizable adaptive coded modulation scheme, as a function of the signal-to-noise ratio on Nakagami-m channels (from [HOL02])

bination of channel code and modulation constellation) to be used, corresponding to the channel being in a state which corresponds to the n^{th} sub-range of signal-to-noise ratio levels.

The choice of transmitter settings can for example be done subject to fulfilment of a chosen target bit-error rate (BER) and an average transmit power constraint, so that the transmitter attempts to transmit information at the highest instantaneous spectral efficiency that is possible, while simultaneously fulfilling the BER constraint and the power constraint. Thus, for a high CSNR the transmitter will choose its settings corresponding to a high spectral efficiency, whereas it will decrease the spectral efficiency dynamically as the channel quality deteriorates. Other design criteria are also possible, depending on the application at hand.

The idea of link adaptation dates back to the early 70s, when Murakami and Nakagami suggested using such an idea to combat fading on a deep-space satellite-to-earth channel, by using the earth-to-satellite channel as a feedback channel [M&N71]. In the early 1990s Steele, Webb, and Hanzo were suggesting varying the modulation constellation size according to the quality of the channel [S&W91, WHS91].

In 1997, link adaptation was put on a firmer information-theoretic grounding when Goldsmith and Varaiya showed that the ergodic capacity of a fading channel could be attained by a certain adaptive scheme in which transmit rate and power were continuously and instantaneously adapted to the channel conditions if these conditions were assumed perfectly known at the transmitter side [G&V97]. Practical link adaptation schemes, typically in the form of *adaptive coded modulation* [G&C98, L&M98, GOE99, HHØ00, V&G00], constitute discrete-rate, discrete-power approximations to this idealized and practically nonrealizable capacity-achieving scheme. An illustration of the capacities that typically can be achieved in practice is given in Figure 5. Here, the parameter m refers to the use of a *Nakagami-m* channel model. $m = 1$ corresponds to the well-known Rayleigh fading channel whereas increasing values of m means less severe fading and the introduction of a line-of-sight (LOS) component.

Since the groundbreaking theoretical work of Goldsmith and Varaiya many researchers have made important contributions to the understanding, design and performance analysis of link adaptation schemes. An exhaustive overview over most research contributions up to 2002 can be found in [HAN02]. It is fair to say that the advantages, design choices, available degrees of freedom, and practical limitations of link adaptation techniques have become well-known, and such techniques are gradually being included in several upcoming wireless communications standards. Some of the most important recent contributions in the field deal with:

- Analysis of the effect of imperfect CSI at the transmitter and receiver side on the performance of adaptive (coded and uncoded) modulation schemes [ØHH04, C&G04].
- Optimization of parameters in adaptive (coded) modulation schemes to counteract the effects of imperfect CSI, and to maximize the average spectral efficiency for a given channel under various constraints [C&G04, HØA03, JØH04, D&Ø04].
- Combination with OFDM, antenna diversity, and MIMO concepts.

4.3 Multi-carrier modulation and access

Multi-carrier modulation and access in the shape of *OFDM (Orthogonal Frequency Division Multiplexing)* is rapidly becoming the most viable option for spectrally efficient broadband modulation and user multiplexing with easy handling of inter-symbol interference (ISI). It is currently being included in new standards for both digital subscriber lines, digital broadcasting systems (DAB, DVB), and WLANs (e.g. the IEEE 802.11 and 802.16 standards). It is currently considered a major candidate for modulation and access in B3G proposals around the world, in particular for the downlink transmission [SVE03, K&S02].

In OFDM, a high-rate data stream is broken down into many parallel, orthogonal data streams, each of lower rate. These lower-rate data streams (subchannels, subcarriers) are superimposed in time by means of an inverse Discrete Fourier Transform (IDFT, typically and efficiently implemented by means of a Fast Fourier Transform (FFT)) before transmission. They can be resolved, due to the orthogonality in frequency imposed by using the IDFT, by performing the corresponding Discrete Fourier Transform (DFT) at the receiver side. A block diagram of the transmitter in an OFDM-based communication system is shown in Figure 6, while the individual subchannel spectra is shown in Figure 7 (for the case of $N = 5$ subcarriers for clarity – in practical OFDM applications the number of carriers may typically be in the order of 50 – 1000). The receiver will be the inverse of the transmitter in the sense that it first performs the inverse transform of the transmitter, and then parallel-to-serial converts in order to recover the original time sequence.

The advantages of OFDM are direct results of the splitting of the signal into many distinct, low-rate, orthogonal subcarriers. When we have N subcarriers of equal bandwidth, the bandwidth and thus the transmission rate on each subcarrier is $1/N^{\text{th}}$ of that of a single-carrier scheme with the same bandwidth as that of the total OFDM scheme. This in turn means that the OFDM symbol length (in time) is N times longer than for a single-carrier scheme with the same bandwidth, which again implies that the radio channel's impulse response can be N times longer before the symbols start *interfering* with each other in time (inter-symbol inteference or ISI). Thus, for a given channel, an OFDM system will be much more robust towards ISI than will a single-carrier system. As the number of subchannels increase for a given overall bandwidth, eventually each subchannel can be considered as flat-fading, i.e. without any ISI.

OFDM is also a spectrally efficient technology, which has great flexibility in the sense that the individual subchannels can be individually bit and power loaded, dynamically with respect to the instantaneous quality of each individual subchannel. That is, if channel state information is made available to the transmitter, link adaptation can be performed for each individual subchannel. Also, subcarriers can be dynamically allocated to different sources and users in a multi-user, multimedia communications system, to maximally employ the available degrees of freedom. Furthermore OFDM can be combined with MIMO technology, different kinds of space-time processing, and different kinds of multiple access schemes, e.g. multicarrier CDMA. Frequency diversity, stemming from having N (more or less) independently fading subchannels. can be exploited by means of coding across the subchannels. Finally, OFDM is robust to narrowband interference since the total signal power is spread over a large bandwidth.

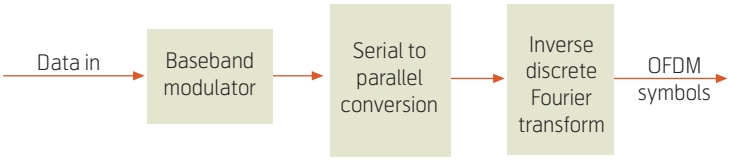


Figure 6 Block diagram of an OFDM transmitter

Of course a price has to be paid for all the above advantages. OFDM systems typically exhibit a high *Peak-to-Average Power Ratio (PAPR)*. The PAPR is defined, as implied by its name, as the ratio between the maximal and the average symbol power. Remember that OFDM symbols are created by performing an IDFT, i.e. by summation of many weighted complex exponentials. In the rare, but still possible event that all or most of the terms sum constructively (in phase) the result may be a very high amplitude. This may result in nonlinear distortion problems such as clipping noise. This is due both to finite word-length effects and to the fact that high power amplifiers (HPAs) used in wireless systems are typically nonlinear devices. The nonlinear distortion will degrade the BER performance and cause energy leakage between subchannels. There is currently a lot of research effort worldwide on mitigating the effects of high PAPR, e.g. by HPA linearization, and also

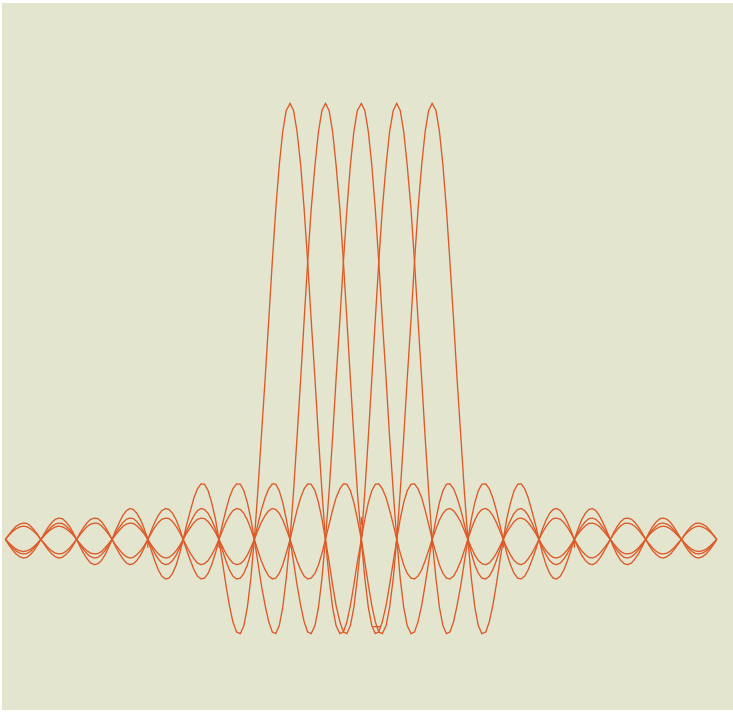


Figure 7 OFDM subcarrier spectra for N=5 subcarriers

ISSN 0085-7130

on signal processing and coding methods for PAPR reduction [RLP04, YFY03, N&L02, RJK02].

In traditionally designed (i.e. using IFFT and FFT) OFDM systems there is also some bandwidth loss due to the insertion of overhead information in the form of a *guard interval* between symbols. The guard interval is typically constructed by *cyclic extension*. It is there to avoid the "tail" of one OFDM symbol leaking into the next and thus causing ISI. The length of the guard interval must therefore be at least equal to the length of the channel impulse response.

However, as the number *N* of subchannels grow larger the overhead represented by the guard interval will become smaller and eventually insignificant, since the symbol length increases linearly in *N* while the necessary guard interval length remains constant. It is also possible to remove the need for a guard interval completely by introducing *pulse shaping* into the system, where the pulse shape is optimized for ISI removal [V&H96]. The price paid for this modification is a computationally somewhat more complex transmitter and receiver, since straight-forward FFT algorithms can no longer be used.

Lastly, OFDM systems rely on very accurate frequency and phase information to preserve the orthogonality between subcarriers (and thus the ability to resolve the subchannels in the receiver). Therefore they are more sensitive to frequency and phase offsets than single-carrier systems. However, various compensation methods exist for mitigating this problem; see e.g. [HSK01].

4.4 Iterative receiver processing

With the invention of Turbo codes by Berrou et al. in 1993 [BER93], the field of error-control (channel) coding was suddenly revolutionized. After 50 years of constructing mathematically complex codes trying to approach the information-theoretic performance bounds of Claude Shannon [ØI02], the coding community was confronted by a scheme in which a simple *parallel concatenation* of two basic *recursive systematic convolutional (RSC) codes*, separated by an *interleaver* (cf. Fig. 8), outperformed all elaborate code constructions of previous times.

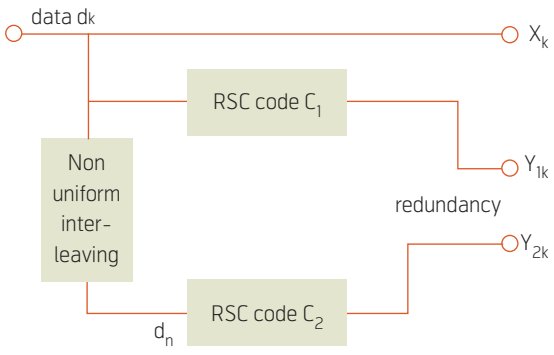


Figure 8 The parallel concatenated turbo coding scheme of Berrou et al. (encoder)

The improvements indicated by Berrou's results were in fact so large compared to classical codes that some actually did not believe the results at first. However, the results were quickly confirmed by other research groups, and a period of intense research into turbo codes followed.

Many alternative turbo code constructions, including *serial* concatenation instead of parallel concatenation of component codes, has since been proposed, and it is fair to say that turbo codes are now quite well understood. Tools from graph theory have been particularly useful in furthering the understanding of these codes [LOE04, KFL01] as well as that of related codes such as the *Low Density Parity Check* (LDPC), or Gallager, codes. An excellent tutorial overview of much of the fundamental turbo coding research from the period 1993 – 2002 can be found in a previous *Telektronikk* article by Ytrehus [YTR02].

The most significant difference between turbo codes and previous coding schemes is the decoding algorithm, which in the turbo case is *iterative* and exploits so-called *"soft" (non-quantized) information* about the correctness of decisions made during decoding. Two component decoders alternate in passing decisions to each other in an iterative fashion until the process converges.

The other distinguishing feature is the construction of *long, quasi-random codewords by interleaving of simple component codes*. This feature makes the code construction strongly resemble that of the "random coding argument" originally used by Claude Shannon in the proof of his channel coding theorem, when demonstrating the existence of capacity-achieving codes [SHA48]. Part of the genius of turbo codes is that such codes can now be constructed while still retaining a near-optimal decoding algorithm of reasonable complexity, instead of having to use the exhaustive nearest-neighbor search implicit in Shannon's construction.

The turbo-decoding algorithm is actually built on an algorithm for optimal decoding of linear codes, originally published in the mid-70s and nicknamed the "BCJR" algorithm after its authors [BAH74]. To improve reliability by means of successive iterations, one exploits the fact that when a (preliminary) deci-

sion on a symbol is made during one of the iterations, the receiver is also able to compute the *probability of this decision being correct or not*. This is a consequence of the fact that the soft information is retained throughout the decoding and passed between the component decoders. No hard decisions on the bit level are made before the algorithm converges. If the reliability of an intermediate decision is high (probability of correctness close to 1 or 0), then the decision can be strongly relied upon in subsequent processing (i.e. further iterations), whereas if it is low (probability of correctness close to 0.5) then it is relied less upon in further iterations.

Thus, simply put, the benefit of the iterations in the turbo-decoding algorithm is that they successively improve the decision reliabilities with each iteration, until one is "almost certain" that a decision made on a codeword is correct. This decision is then finally output from the receiver to the user. Figure 9 (from [O&R02]) shows a comparison between turbo codes and some classical error control codes, as well as uncoded communications, with regard to bit-error rate performance versus energy spent per information bit. It is seen that the turbo codes come significantly closer than all the other codes to the theoretical limit indicated by the capacity curve.

In the last eleven years, 'turbo'-based iterative algorithms and the use of 'soft' reliability information have in fact been shown to dramatically improve the performance of not only pure channel coding, but also joint multi-user detection and channel estimation [PAR04], multi-user decoding [CMT04], joint pilot-symbol-assisted coded modulation and channel estimation [LGW01], synchronisation [NOE03], equalization [KST04], interference suppression [WAN99], and so on. It can be used in the context of MIMO systems, or with OFDM, CDMA, space-time codes, etc. The "Turbo principle", as the rationale behind iterative receiver algorithms has been come to be called, is in fact a general principle that can be used in any setting where two components of a communication receiver have the possibility of exchanging information about the reliability of their respective decisions. The information from one component is used as extra input to the processing in the other, to improve the reliability of its decisions. This reliability information can in its turn be used in a similar fashion by the first component, and so on, until the process (in most practical cases) converges.

The price paid for the associated performance improvements is ("as usual") more complex receiver processing than is the case with non-iterative schemes. However, with the emergence of ever more powerful computing capabilities, this is already, in many instances, a small price to pay compared to the performance improvements experienced. This trend will only be strengthened in the future.

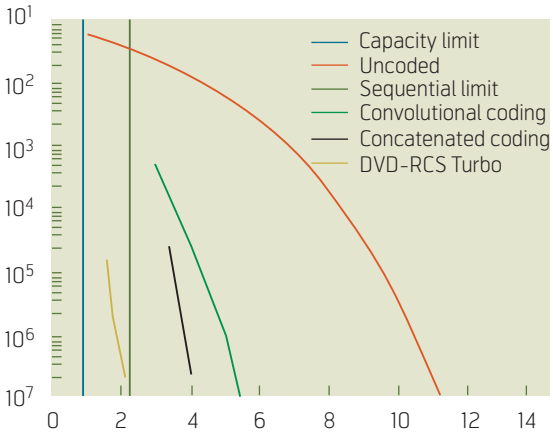


Figure 9 Performance comparison between turbo coding, uncoded communications, and some classical coding schemes. Performance is measured as bit-error rate against energy spent per information bit (Courtesy of Nera Research)

Turbo coding is already being included in existing and upcoming standards for wireless communications, such as the UMTS standard, and it is expected that the Turbo principle will be even more important and find wider use in B3G systems.

4.5 Cross-layer optimization techniques

One of the foundations of modern digital communications has been the *layered protocol architecture*. Such an architecture consists of a layered stack of protocol modules, each of which solves a specific problem by using services provided by lower-layered protocols, and providing services to higher layers. For maximum modularity and ease of design, communication in this architecture typically happens mainly between adjacent layers, and is of limited scope.

A classical example [CRM04] is the Ethernet protocol stack where the transport layer protocol TCP (Transport Control Protocol) overlays the network protocol IP (Internet Protocol), which in its turn overlays the Ethernet link layer. At the bottom of the protocol stack is the physical layer, by which everything is physically implemented. The link layer provides connectivity within one given network segment (sending and receiving frames to/from hosts), while the network layer delivers data packets across multiple networks. The transport layer adds connection-oriented reordering of packets, error recovery, flow control, and congestion control. Each layer only uses information from the layer directly beneath it in the protocol stack when operating.

As discussed repeatedly in previous sections, an important aspect of wireless channels and networks is dynamic behaviour, such as temporal and spatial changes in channel quality and user distribution. A conventional protocol structure as described in the previous paragraph is inflexible and unable to adapt to such dynamically changing network behaviours, since the various protocol layers can only communicate with each other in a strict and primitive manner. In such a case, the layers are most often designed to operate under the worst conditions, rather than adapting to changing conditions. This leads to inefficient use of spectrum and energy.

An example is that the TCP protocol in its current version is designed to interpret all packet losses as a sign of network congestion, whereas in a wireless system the losses might just as well be a consequence of the channel going into a fade. TCP reacts to such losses by decreasing the source's packet transmission rate and thus lowering the network throughput unnecessary in many cases. However, upcoming versions of TCP are expected to feature *Explicit Congestion Notification (ECN)*, in packets that are explicitly marked (one particular header bit is set to 1) by routers in the network each time a true congestion occurs. Thus TCP may only reduce the source packet transmission rate when detecting such marked packets [SRK03].

The concept of ECN, where TCP receives explicit information about network conditions on lower layers, is a simplistic example of the novel design paradigm of *cross-layer optimisation*, or *cross-layer design*, and how such design may enhance the ability of network protocols and applications to interact in a more intelligent manner and to observe and respond to wireless channel variations. Cross-layer design in general simply means that transmission schemes and protocols at the physical layer, link layer, network layer, transport layer, application layer etc. are enabled to communicate more information between themselves across layers, and that their operation – based on such information – in some sense may be adapted and optimized jointly.

In particular, knowledge of the wireless medium gained by the physical and MAC layer (e.g. through channel estimation/prediction) is envisioned shared with higher layers. If exploited properly this knowledge may lead to a much more optimal allocation and use of network resources [SRK03, CRM04, CPM04, T&G03, ZHE03].

As a more elaborate example, consider the design of medium access (MAC) protocols which are used to schedule multiple users' access to the wireless uplink channel, or the base station's downlink transmission of information to the users, in a cellular network. In such design one may improve the overall system throughput by taking into account the quality of the wireless channel and the available coding and modulation schemes. Assuming time-division multiplexing between users, and the use of link adaptation to exploit uplink capacity for each user, the overall network throughput can in fact be maximized by always scheduling the user whose uplink channel has the highest channel-to-noise-and-interference ratio (CNIR), and thus the highest channel capacity, of all the users. The throughput gain of such a method over straightforward and traditional "round-robin" time-division multiplexing of users (which is independent of the channel state of each user) is called the *multi-user diversity gain*. Effectively, the system attempts to "ride the peaks" by transmitting only at high CNIRs, by switching dynamically and adaptively between users depending on their channel quality. Examples of such mechanisms are already found in the enhanced general packet radio service (EGPRS) of the GSM system and in the high data rate (HDR) versions of the CDMA2000 3G system [SRK03].

To this can also be added fairness constraints to ascertain that all users are granted access to the channel, e.g. after some prescribed maximum waiting time, or some given percentage of the time, depending on their individual needs and on whether they actually do have anything to transmit at a given point in time. One can also take into account different fading statistics for different users, and different traffic classes with different capacity requirements and heterogeneous quality-of-service (QoS) constraints, when designing the MAC protocols.

The above are only selected examples of what cross-layer design can mean in practice. Other examples include joint optimization of source and channel coding schemes, and energy-optimization of coding and modulation schemes taking circuit energy consumption into account as well as the transmission energy [CGB03]. The latter technique amounts to cross-layer design of the physical and link layer and may be of particular importance in short-range applications where the circuit energy and the transmission energy consumption may be comparable in size, and where overall energy efficiency is of the essence (such as in wireless sensor networks).

A more complete list of hot cross-layer design research topics on a more general level currently include:

- Throughput-optimized cross-layer control of MIMO systems,

- Space/time/frequency processing techniques for improved multiple access schemes,
- Techniques for increased flexibility in spectrum use,
- Feedback channel and signal design for efficient cross-layer interaction,
- Hybrid-ARQ schemes,
- Dynamic resource allocation and QoS issues,
- Channel-state-aware protocols,
- Stability and robustness issues in cross-layer interactions.

There are also many other aspects that might – and probably will – be taken into account in future cross-layer design than those covered by the above examples [CRM04]. Such aspects include security issues, which are outside the scope of this article.

4.6 Channel characterization and implementation issues

In addition to the techniques presented so far, and as a prerequisite for many of them, the ability to accurately measure, model, characterise, estimate and predict the properties of many different types of radio propagation environments will continue to increase in importance with regard to successful design and optimisation of novel and more advanced radio interfaces. Channel knowledge is essential, e.g. for link adaptation and opportunistic multiuser scheduling algorithms.

As an example, it is illustrated in Figure 10 how the bit-error rate (BER) of an adaptive coded modulation system degrades at low signal-to-noise ratios and high feedback channel delays due to the fact that the channel state information used to adapt the transmitter mode then becomes too inaccurate [ØHH04].

Measurement campaigns, development of quantitative criteria for channel model optimization, statistical methods for fitting of observed channel data to simulation models, and procedures for optimization and characterization of model parameters are all important tools which can help us in predicting the "true" performance of a given transmission scheme, or when optimizing parameters in such schemes for the best possible performance, or when establishing theoretical bounds for system performance. The better the theoretical model or simulation model, the more accurate system performance predictions and system design choices can be made before going to the expensive and time-consuming task of building prototypes.

Finally, it should also be mentioned that the development and commercialisation of new equipment for radio communication will call for the development of hardware with improved performance and reduced component costs. In order to avoid specification of communication standards that are difficult or expensive to implement in hardware, it is important to maintain a useful dialogue between engineers working at different levels of the standardisation process.

5 Summary and conclusions

In the present paper we have described some of the major radio interface and access technologies that we believe will become of the most importance for future wireless and mobile communication systems beyond 3G (B3G). We have only given an overview without going deeply into technical details; however, an extensive and up-to-date reference list has been provided so the interested reader can know where to go for more information on a particular topic.

We have tried to highlight the main advantages (and, if any, disadvantages and/or limitations) of each component technology described. We have mainly covered multiple-input multiple-output (MIMO) technology, link adaptation techniques, multicarrier modulation/OFDM, iterative ("turbo") receiver processing, and cross-layer design methodologies, while also briefly touching upon channel characterization and implementation issues.

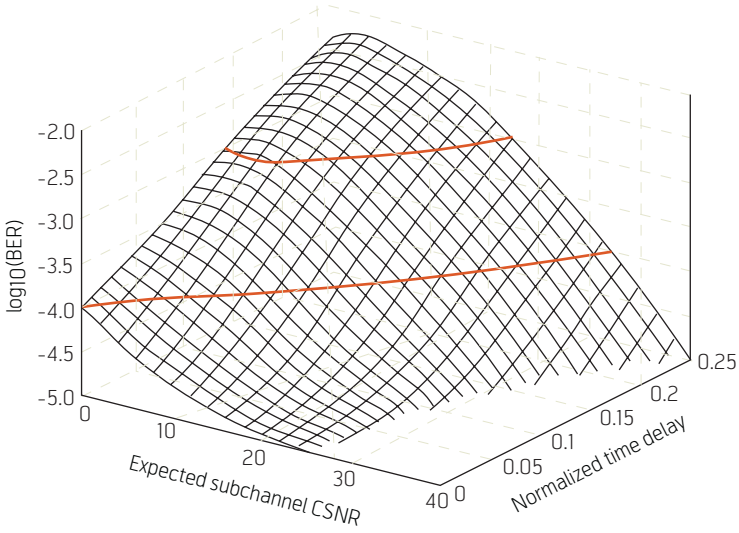


Figure 10 BER degradation in an adaptive coded modulation system due to faulty channel state information at low signal-to-noise ratios and large feedback delays. The target bit rate when designing the system was 10^{-4}

ISSN 0085-7130

There are of course other interesting research topics in wireless communications which could have been covered here as well. In particular *ultra-wideband* radio communications is currently a hot topic in the international research community, due to its potential for providing high-capacity, low-power transmission over very short ranges without interference limitations. However, the scope must be limited in some way and we believe the selection made reflects the major importance, maturity of understanding, and large potential of the chosen topics relative to others.

The impact of the 3G technology, which is currently under deployment, is still uncertain. As we have seen, B3G technology is still at the research stage, and will not be available until at least after 2010. However, through the use of techniques described in this paper, B3G has an enormous potential for increase of capacity, and is expected to have a very large impact on future mobile communications.

References

[4GMF] *4G Mobile Forum*. September 28, 2004. [online] – URL: http://www.delson.org/4gmobile/intro.htm

[Ala98] Alamouti, S M. A simple transmit diversity technique for wireless communications. In: *IEEE Journal on Selected Areas in Communications*, 16, 1451–1458, 1998.

[BAH74] Bahl, L R et al. Optimal decoding of linear codes for minimizing symbol error rate. In: *IEEE Trans. Inform. Theory*, 20 (2), 284–287, 1974.

[BER93] Berrou, C, Glavieux, A, Thitimajshima, P. Near Shannon limit error correcting coding and decoding. In: *Proc. ICC-1993*, Geneva, Switzerland, May 1993, 1064–1070.

[C&G04] Cai, X, Giannakis, G B. Adaptive PSAM Accounting for Channel Estimation and Prediction Errors. To appear in *IEEE Trans. Wireless Comm.*, 2004.

[CEL1] The CELTIC initiative – Cooperation for a European sustained Leadership in Telecommunications. *CELTIC Purple Book, Part two: Technical Scope of the CELTIC Initiative*. September 28, 2004. [online] – URL: http://www.celtic-initiative.org/Publications/purple-book.asp

[CGB03] Cui, S, Goldsmith, A J, Bahai, A. Energy-constrained modulation optimization for coded systems. In: *Proc. GLOBECOM 2003*, 372–376, Dec. 2003.

[CMT04] Caire, G, Müller, R R, Tanaka, T. Iterative Multiuser Joint Decoding: Optimal Power Allocation and Low-Complexity Implementation. In: *IEEE Trans. Inform. Theory*, 50 (9), 2004.

[CPM04] Chan, Y S, Pei, Y, Modestino, J W. On cross-layer adaptivity and optimization for multimedia CDMA mobile wireless networks. In: *Proc. 1st International Symposium on Control, Communications, and Signal Processing*, March 2004, 579–582.

[CRM04] Carneiro, G, Ruela, J, Ricardo, M. Cross-Layer Design in 4G Wireless Terminals. In: *IEEE Wireless Communications*, April 2004, 7–13.

[D&Ø04] Duong, D V, Øien, G E. Adaptive trellis-coded modulation with imperfect channel state information at the receiver and transmitter. In: *Proc. Nordic Radio Symposium 2004*, Oulu, Finland, Aug. 2004.

[F&G98] Foschini, G J, Gans, M J. On Limits of Wireless Communications in a Fading Environment when Using Multiple Antennas. In: *Wireless Personal Communications*, 6, 311–335, 1998.

[Fos96] Foschini, G J. Layered Space-Time Architecture for Wireless Communications. In: *Bell Labs Technical Journal*, 1, 41–59, 1996.

[G&A02] Gesbert, D, Akhtar, J. Breaking the Barriers of Shannon's Capacity: An Overview of MIMO Wireless Systems. In: *Telektronikk*, 98 (1), 53–64, 2002.

[G&V97] Goldsmith, A J, Varaiya, P P. Capacity of Fading Channels with Channel Side Information. In: *IEEE Trans. Inform. Theory*, 43 (6), 1986–1992, 1997.

[HAN02] Hanzo, L, Yee, M S, Wong, C-H. *Adaptive Wireless Transceivers*. New York, Wiley, 2002.

[HHØ00] Hole, K J, Holm, H, Øien, G E. Adaptive Multidimensional Coded Modulation over Flat Fading Channels. In: *IEEE J. Select. Areas Commun.*, 18 (7), 1153–1158, 2000.

[HOL02] Holm, H. *Adaptive Coded Modulation Performance and Channel Estimation Tools for Flat Fading Channels*. Trondheim, Norway, NTNU, April 2002. PhD thesis 2002:18.

[HSK01] Han, D-S, Seo, J-H, Kim, J-J. Fast Carrier Frequency Offset Compensation in OFDM Systems. In: *IEEE Trans. Consumer Electronics*, 47 (3), 364–369, 2001.

[HØA03] Holm, H et al. Optimal design of adaptive coded modulation schemes for maximum average spectral efficiency. In: *Proc. IEEE SPAWC 2003*, Rome, Italy, June 2003.

[ITU1] ITU. *Framework and overall objectives of the future development of IMT-2000 and systems beyond IMT-2000*. Geneva, ITU, 2003. ITU Recommendation ITU-R M.1645. URL: http://www.itu.int/rec/recommendation.asp?type=folders&lang=e&parent=R-REC-M.1645

[ITU2] ITU. *General focus areas for research and further study for the future development of IMT-2000 and systems beyond IMT-2000*. September 28, 2004. [online] – URL: http://www.itu.int/ITU-R/study-groups/rsg8/rwp8f/docs/focus-areas.doc

[JØH04] Jetlund, O et al. Spectral efficiency bounds for adaptive coded modulation with outage probability constraints and imperfect channel prediction. In: *Proc. Nordic Radio Symposium 2004*, Oulu, Finland, Aug. 2004.

[KAR03] Karlson, B et al. *Wireless Foresight: Scenarios of the Mobile World in 2015*. Chichester, UK, Wiley, 2003.

[KFL01] Kschischang, F R, Frey, B J, Loeliger, H-A. Factor Graphs and the Sum-Product Algorithm. In: *IEEE Trans. Inform. Theory*, 47 (2), 498–519, 2001.

[KST04] Koetter, R, Singer, A C, Tuchler, M. Turbo Equalization. In: *IEEE Signal Processing Magazine*, 21 (1), 67–80, 2004.

[K&S02] Kaiser, S, A. Svensson, A (eds.). Broadband Multi-Carrier Based Air Interface. WWRF Working Group 4 White Paper, version 1.4. In: *Proc. 7th Wireless World Research Forum Workshop*, Eindhoven, the Netherlands, Dec. 2002.

[LGW01] Li, Q, Georghiades, C N, Wang, X. An Iterative Receiver for Turbo-Coded Pilot-Assisted Modulation in Fading Channels. In: *IEEE Comm. Letters*, 5 (4), 145–147, 2001.

[LOE04] Loeliger, H-A. An Introduction to Factor Graphs. In: *IEEE Signal Processing Magazine*, 21 (1), 28–41, 2004.

[M&N71] Murakami, S, Nakagami, M. Adaptive feedback communication system in a fading environment. In: *Memoirs of the Faculty of Engineering, Kobe University*, 17, 127–147, 1971.

[NEW] *Description of the pan-European Network of Excellence on Wireless COMmunications (NEWCOM)*. September 28, 2004. [online] – URL: http://commgroup.polito.it/newcom/

[NOE03] Noels, N et al. Turbo synchronization: an EM algorithm interpretation. In: *Proc. ICC 2003*, 4, 2933–2937, May 2003.

[N&L02] Nikookar, H, Lidsheim, K S. Random Phase Updating Algorithm for OFDM Transmission with Low PAPR. In: *IEEE Trans. Broadcasting*, 48 (2), 123–128, 2002.

[O&R02] Orten, P, Risløw, B. Theory and Practice of Error Control Coding for Satellite and Fixed Radio Systems. In: *Telektronikk*, 98 (1), 78–91, 2002.

[PAR04] Park, S Y, Kim, Y G, Kang, C G. Iterative receiver for joint detection and channel estimation in OFDM systems under mobile radio channels. In: I*EEE Trans. Veh. Tech.*, 53 (2), 450–460, 2004.

[RJK02] Ryu, H-G, Jin, B-I, aKim, I-B. PAPR Reduction Using Soft Clipping and ACI Rejection in OFDM System. In: I*EEE Trans. Consumer Electronics*, 48 (1), 17–22, 2002.

[RLP04] Ryu, H-G, Lee, J-E, Park, J-S. Dummy Sequence Insertion (DSI) for PAPR Reduction in the OFDM Communication System. In: *IEEE Trans. Consumer Electronics*, 50 (1), 89–94, 2004.

[SHA48] Shannon, C E. A Mathematical Theory of Communication. In: *Bell Syst. Techn. J.*, 27, 379–423 and 623–656, 1948.

[SPEC] Staple, G, Werbach, K. The End of Spectrum Scarcity. *IEEE Spectrum*, March 2004.

[SRK03] Shakkottai, S, Rappaport, T S, Karlsson, P C. Cross-Layer Design for Wireless Networks. In: *IEEE Communications Magazine*, 74–80, Oct. 2003.

[SVE03] Svensson, A et al. An OFDM Based System Proposal for 4G Downlinks. In: *Proc.Multi-Carrier Spread Spectrum Workshop*, Oberpfaffenhofen, Germany, Sept. 2003.

[S&W91] Steele, R, Webb, W T. Variable rate QAM for data transmission over mobile radio channels. In: P*roc. WIRELESS '91*, Calgary, Alberta, Canada, 1–14.

[Tar99] Tarokh, V, Jafarkhani, H, Calderbank, R. Space-Time Block Codes from Orthogonal Design. In: *IEEE Trans. Inform. Theory*, 45, 1456–1467, 1999.

[T&G03] Toumpis, S, Goldsmith, A J. Performance, optimization, and cross-layer design of media access protocols for wireless ad hoc networks. In: *Proc. ICC-03*, 3, 2234–2240, May 2003.

[V&H96] Vahlin, A, Holte, N. Optimal Finite Duration Pulses for OFDM. In: *IEEE Trans. Commun.*, 44 (1), 10–14, 1996.

[WAN99] Wang, X, Poor, H V. Iterative (Turbo) Soft Interference Cancellation and Decoding for Coded CDMA. In: *IEEE Trans. Commun.*, 47 (7), 1046–1061, 1999.

[WHS91] Webb, W T, Hanzo, L, Steele, R. Bandwidth efficient QAM schemes for Rayleigh fading channels. In: *IEE Proceedings I: Communications, Speech, and Vision*, 138 (3), 169–175, 1991.

[WWRF] *Wireless World Research Forum (WWRF).* September 28, 2004. [online] – URL: http://www.wireless-world-research.org/

[YFY03] Yusof, S K, Fisal, N, Yin, T S. Reducing PAPR of OFDM signals using partial transmit sequences. In: *Proc. 9th APCC 2003*, 1, 411–414, Sept. 2003.

[YTR02] Ytrehus, Ø. An Introduction to Turbo Codes and Iterative Decoding. In: *Telektronikk*, 98 (1), 65–77, 2002.

[ZHE03] Zheng, H. Optimizing wireless multimedia transmission through cross layer design. In: *Proc. ICME '03*, 1, I-185–8, July 2003.

[ØHH04] Øien, G E, Holm, H, Hole, K J. Impact of Channel Prediction on Adaptive Coded Modulation Performance in Rayleigh Fading. In: *IEEE Trans. Veh. Tech.*, 758–769, May 2004.

[ØIE02] Øien, G E. Information Theory: The Foundation of Modern Communications. In: *Telektronikk*, 98 (1), 3–19, 2002.

Geir E. Øien (38) received his MSc and PhD degrees from the Norwegian Institute of Technology (NTH) in 1989 and 1993, respectively. From 1994 until 1996 he worked at Stavanger University College as associate professor. Since 1996 he has been with the Norwegian University of Science and Technology, since 2001 as full professor of information theory. Prof. Øien is a member of IEEE and the Norwegian Signal Processing Society. He is author/co-author of more than 50 research papers in refereed international fora. His current research interests are in information theory, communication theory, and signal processing, with emphasis on analysis and design of wireless communication systems.

email: oien@iet.ntnu.no

The adoption, use and social consequences of mobile communication

RICH LING

Rich Ling is a sociologist at Telenor R&D

In 2003, 100 % of Norwegian teens between the age of 16 and 19 had a mobile telephone. This surprising statistic points to the rapid adoption of a technology that was only marginally commercialised when these teens were born in the mid to late 1980s. This paper is based on analysis of survey data that has been collected by Statistics Norway in cooperation with Telenor and surveys conducted by Telenor. Over the period described here as many as 10,000 individuals have participated in the various surveys. I will look at some of the watershed transitions associated with mobile telephony, including its adoption into the broader society, teens' enthusiastic embrace of the technology and the rise of SMS – led by teens. The paper also examines some of the dynamics that have not changed including general time use on the phone, and the gendering of the technology. Finally the article considers some of the broader social consequences of mobile telephony.

Introduction

According to statistics collected by Statistics Norway in 2003, 100 % of the teens aged 16 – 20 they interviewed in their media use survey had a mobile telephone. That is to say, Norway's best organization in the area of survey research was not able to turn up teens without mobile telephones. This finding is as astonishing both in the speed with which the transition has come and the omnipresence of the technology. Only five years previous to this, a minority of the teens had a mobile telephone. However, in the intervening period the device had established itself into the teen identity as securely as any other artifact. The same survey showed that 86 % of all Norwegians over the age of 8 had a mobile telephone.

By way of a somewhat poor comparison, the telephone took 60 years to reach 80 % household penetration in the US. Electric lights took 30 years and the automobile took 60 years to reach 80 % penetration in the US. The radio reached an 80 % adoption rate in about 16 years starting in 1920. The TV took 12 years to reach 80 % and about 35 years to reach nearly universal adoption (Fischer 1992).

These are poor comparisons because many of these innovations required the development of extensive infrastructure where GSM telephony could in some respects hitch a ride on the pre-existing landline telephony system. Another difference is that the statistics cited by Fischer are for household adoption, not personal adoption. In this respect the figure of 86 % represents a much higher rate of penetration than the household items described above.

In this article I am interested in looking into the adoption and use of mobile telephony over the past decade in Norway. From the commercialization of GSM – which happened in 1993 – until the present, there has been a dramatic shift in the adoption and use of mobile telephones. Norway and indeed Scandinavia are among those areas where this revolution has been the most intense. The mobile telephone has become commonplace across Europe, in Japan, Korea, Israel and in some other portions of the world. This revolution has been visible in North America and it has been relatively unknown in large portions of Asia (outside south eastern Asia) and particularly in Africa. Thus, the experience of Scandinavia is likely to provide insight into how mobile communication will change society in other areas of the world.

Interestingly, the adoption of mobile telephony has not been uniform even within Norway. Various socio-demographic groups have, at various times, gone through the adoption process. In the early stages of adoption it was often business persons – as seen in the stereotype of the yuppie – and perhaps delivery persons who were the adopters. As the system developed, and as alternative subscription systems were developed, teens and young adults started to own and use mobile telephones. As will be shown below, there are relatively few people in Norway (about 7 per 100) who have no access to a mobile telephone. In almost all other cases there is at least the ability to borrow a mobile telephone on an irregular basis.

With this dramatic change, one can also see elements of stability as well as various forms of social consequences. On the one hand there are gendered use patterns that, in some ways, mirror the broader gender patterns in society. At the same time, there are various social consequences of mobile telephony that include the enhanced ability to coordinate activities, the provision of security, the development of identity and the strengthening of social networks.

 ISSN 0085-7130

Watersheds in the use of mobile telephony

The general adoption of GSM

The dominating trend over the last decade is the widespread adoption of GSM telephony (see Figure 1). Access to this form of communication has penetrated into the Norwegian everyday life. It was really that system that popularized the mobile telephone in Europe and in many parts of the world. On a world wide basis about 7 out of 10 mobile telephones use GSM. GSM telephones were often the telephones that allowed people to make calls whenever and wherever they wished. However, there have been at least two transitions previous to the commercialization of GSM telephony, namely manual mobile telephony and the NMT system.

Manual mobile telephony was available in Norway from the late 1960s. Using this system the caller had to call a telephone operator and request that a call be set up with the desired party. Starting in the early 1980s, however, there was the commercialization of the NMT system. It represented one of the first cellular systems wherein one could "roam" from one country to another, albeit within Scandinavia. During the mid 1980s the growth rate for NMT telephones was well over 150 % per year.

NMT (and other systems of its ilk) were those that were adopted by yuppies and other highflying users. The system was relatively expensive and the equipment was bulky and awkward to use. It often meant that one used a car-mounted device, or a large suitcase sized portable unit that was filled mostly with batteries. Eventually, the size of the devices was reduced to where it was more similar to a smallish loaf of bread. The actor Michael Douglas playing Gordon Gekko, the rogue financer in the film Wall Street, outlined the use of these devices in the popular imagination. Gekko is pictured calling to the young, uninitiated stock trader from the beach outside his home in the Hamptons. Gekko, of course uses one of the early hand held mobile telephones for the call. The imagery is that of the well-heeled stock magnate who can afford to use the latest technology. Interestingly, within a decade, the mobile telephone would be available to literally all persons, at least all Scandinavians, at absurdly low prices. Thus, the early image of the mobile telephone as a plaything of the rich gave way to the consumption of certain types of devices as opposed to simple consumption.

Indeed it was not until the mid 1990's and the introduction of GSM that the subscription rates for the NMT system began to fall. The NMT system still lives a marginal life at the time of this writing. Those who wish to have the maximum level of coverage such as hunters and those who use the wilderness favor it. However, NMT will eventually be decommissioned and the band width reallocated to other uses.

From 1993 it has been the GSM system that has dominated the market for mobile telephony. During the period of introduction the growth rates were truly phenomenal. In Denmark, TeleDanmark hoped for about 15,000 users in 1993 but was able to sell more

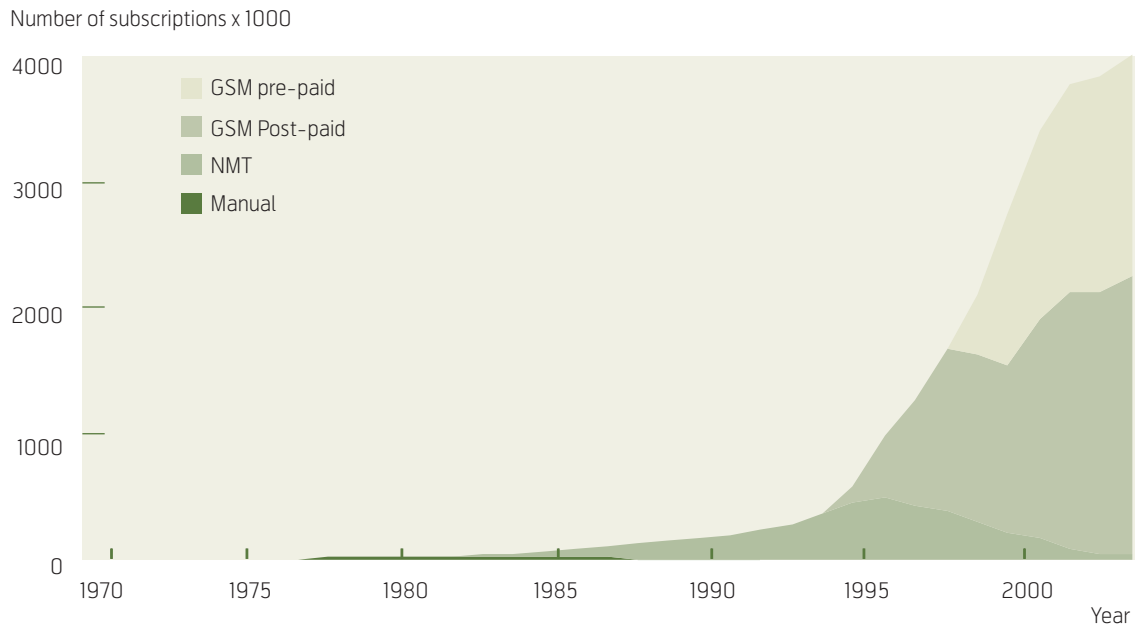


Figure 1 The rise of mobile telephony in Norway 1969 – 2003

ISSN 0085-7130 © Telenor ASA 2004

than four times as many. The same was reported by Sonofon (Haddon 1997). Indeed the growth rates for Norway were on pace with those of the other Scandinavian countries. As noted above, this transition may be one of the swiftest adoption curves for a major technology. Where the car, the traditional telephone, the TV and even the PC/Internet have taken years and decades to be adopted, the mobile telephone seems to have been swallowed whole. GSM dominates all other systems and has become the de facto world standard for mobile telephony.

The final transition shown in Figure 1 is the growth of post-paid GSM subscriptions. Starting in the late 1990s this subscription form became one of the most popular ways to start using a GSM telephone. The subscriber could purchase a complete package including the mobile terminal, a SIM card and a certain amount of pre-paid use in a simple package. These packages have been sold from newsstands, grocery stores and even from venders on the street. The traffic costs per minute for prepaid subscriptions are relatively high, but at the same time the subscriber does not need to pay any fixed subscription fee every month. Thus, these subscriptions are ideal for those who have only occasional use for a mobile telephone since they can use the mobile phone as needed, but they do not need to pay the fixed subscription fee. The other group that has welcomed this subscription form is younger teens. It is perhaps the most common form for this group. Somewhere around 70 % of all 13 – 20 year olds use this type of subscription. As the teens move out of their parents' home and establish a life of their own, they also migrate over to a traditional post-paid subscription. However, approximately 40 % of the women and 20 % of the men in the 20 – 60 year age group, and about half of retirees have a pre-paid subscription.

Pre-paid subscriptions have an immediate appeal for parents who wish to control the use of their child's mobile telephone. In addition, parents often use the pre-paid card as an early object lesson in letting the teen earn the money they need to support their mobile telephone habit. Another element in the teen's adoption of mobile telephony was access to inexpensive – and heavily subsidized – terminals. The establishment of prepaid subscriptions and the general access to inexpensive terminals resulted in the growth of the teen market as we will see below.

The final feature of the chart is the flattening of the curve after 2000. The material shows that we are approaching the absolute number of persons in Norway. There are approximately 4.58 million people in the country and there are slightly more than 4 million telephone subscriptions. There are often subscriptions associated with functions (ambulances, payment teminals, etc) and there are cases of people having multiple subscriptions. In addition there are some groups that are not part of the mobile revolution, notably the elderly and the young pre-teens. Thus, every person will not necessarily have a subscription. However, the total number of subscriptions could easily exceed the population, particularly if devices such as vending machines, heating systems in cottages and the like receive their own subscriptions. Indeed in Luxembourg and in Taiwan there are more mobile subscriptions than there are people.

Teens and their adoption of the mobile telephone 1997 – 2001

The year 1999 was a watershed year when thinking of teens' access to the mobile telephone. Our data shows that in 1997 there were relatively few teens who reported owning a mobile telephone (see Figure 2). This was in the period before the development of pre-paid subscriptions and handsets were relatively expensive. Thus, ownership was generally reserved for those teens who had jobs and who were specially interested in mobile communication. This often meant that it was the males who had mobile telephones, presumably not only because they had jobs but because they were the ones interested in mobile telephony. During this period there were significantly more males who reported ownership.

Two years later the adoption rates were much higher and gender differences had been eliminated. The material here shows that the actual profile of the curve had changed between 1997 and 1999. Where fewer than 20 % of the middle aged teens had a mobile telephone in 1997, this had risen to over 70 % in the interim. The difference was no longer between the mid and older teens, but between the younger and the middle aged teens. Interestingly, the gender gap had also been erased.

Finally, in 2001 the great preponderance of teens between 13 and 20 had a mobile telephone. In this last period the gender gap had again opened up, however this time it was the females who were the most enthusiastic consumers. In 2001 there were significantly more female than male teens with a mobile phone. By way of interpretation there are several issues at play. Early in the adoption cycle it seems that there was a fascination with the technology that motivated adoption. This technical relationship to the device seems to be more often developed among males than females. As the adoption process proceeds the mobile telephone goes from being seen as a technical device to being a tool to support social interaction. Where males seemingly excel with technologies, many have commented that women are often facile

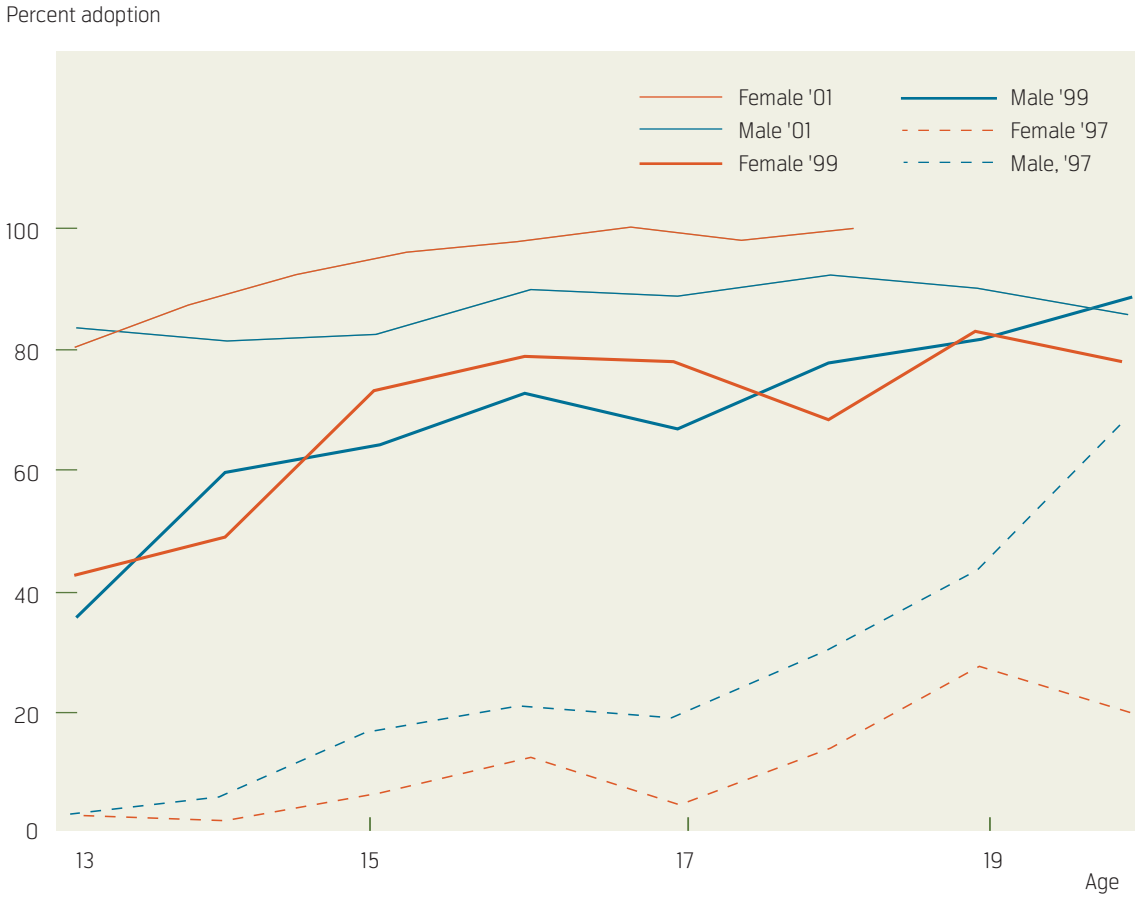


Figure 2 The adoption of mobile telephones by teens, 1997 – 2001

when considering social interaction. Indeed women are often better at the mechanics of conversation and social interaction. In addition, women often have the responsibility to organize social dealings.

Thus, to trace the gendered adoption of the mobile telephone is to trace its transition from being a technical fascination that is perhaps a status symbol to being a tool to support social interaction.

Teen adoption of SMS in the late 1990s and adult adoption in 2003

Teen users were among the first to adopt the use of SMS on a regular basis. Already in 1998 teens with mobile telephones were sending two to three SMS messages per day (see Figure 3). Thus, along with their adoption of the mobile telephone, there was the almost immediate discovery of SMS as a way to communicate. Some of this may be a legacy from the extended use of pagers by teens. Immediately previous to their adoption of the mobile telephone and SMS, the use of pagers represented the leading edge of teen technology. Teens had developed various forms of sending coded messages to one another using the landline telephone system and pagers. Thus, a "3" could mean that the gang is meeting now, a "6" might mean that I cannot come this evening etc. The discovery and adoption of mobile telephones along with the discovery of the SMS system provided a two-way version of this form of signaling. An added advantage was that originally the system was free of charge. This meant that it was a natural channel of communication. The teen needed only to learn how to access SMS and learn the somewhat clumsy form of text entry.

Teens took SMS into use quite quickly. The material in Figure 3 shows that between 1998 and 2003 the number of SMS per day had more than doubled. In addition, it shows that use had been largely concentrated among teens. This again underscores the situation seen in Figure 3 where it was teens and young adults who were the most likely to use SMS on a daily basis. Here, looking at the actual number of SMS messages being sent on a normal day, teens were the most active. Until 2003 the curve describing the number of SMS being sent had been characterized by having a rather steep peak around the late teen and young adult stage. As the reader follows the curve describing the use by adults and the elderly it falls dramatically. Thus, in many respects the use of SMS was highly focused around teen culture. It was not a part of adults' repertoire of communication channels. Teens and young adults were more likely to use SMS

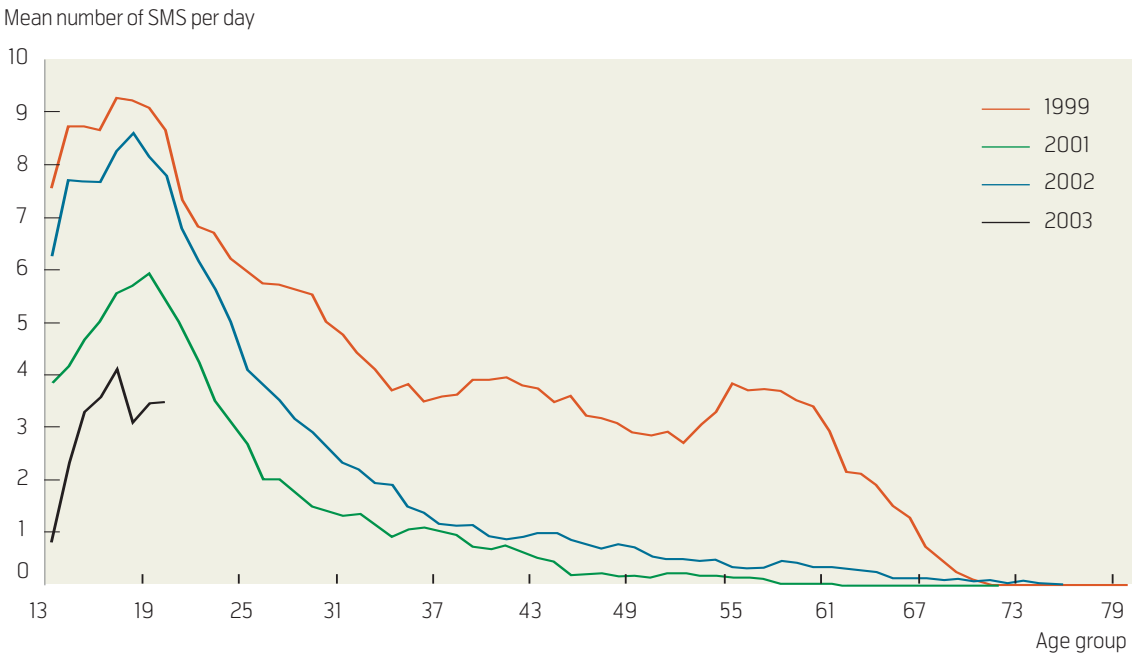


Figure 3 The mean number of SMS messages per day 1998 – 2003 by age and gender

on a daily basis and they were likely to use it many times during the day. The same was not true of other groups.

This situation however is evolving. The data from 2003 shows that the dramatic fall in daily use for adult groups has been eliminated. Rather than a direct fall, there is more of a gradual downward curve. The same is true when considering the number of SMS messages sent. In 2003 adults have gone from sending one or perhaps two SMS messages per day to sending 3 – 4 messages.

The situation in 2004

What is the situation with regard to the use of mobile communication in Norway anno 2004?

Access to mobile telephony in Norway – Increasing personalization

By almost any measure it is clear that Norway is a mature market when it comes to mobile telephony. As of the start of 2004, 87 % of all persons over the age of nine reported personally owning a mobile telephone. In 1999 slightly less than six of ten reported owning a mobile (see Figure 4). Looking at a somewhat broader measure, in 1999 about 80 % of the population had some form of access to a mobile telephone.[1)] They either reported owning one or having access to a mobile telephone that was perhaps shared with outers. Thus, in 1999 58 % owned a mobile telephone alone, about 22 % had shared access and the remaining 20 % reported having no access at all.

These numbers had changed by 2003. In that year 87 % had exclusive ownership of a mobile telephone. In addition about 7 % had some form of shared access and the remaining 6 % reported having no access. There are two general trends here. The first is the growth in ownership and the second is the increasingly individualized ownership of mobile telephones. In countries where the adoption rates are lower, it is far more common to have a shared "family" mobile telephone, particularly among the youngest and the oldest portions of the population. However, as the adoption rate increases and as use becomes more common, there is the growing sense that a particular device belongs to a single individual. Many of the features of the handsets and the subscriptions support this latter interpretation. Having a personalized calling list, a log of one's own SMS messages, various personal information on the telephone in effect personalizes the device. In addition, as features such as MMS, electronic payment and various PDA functions become more common, the mobile device will also become increasingly like a woman's purse or a man's pocket book; that is, it will become a personal effect and a repository for our personal affairs.

Beyond the functional dimensions of the mobile device, there is also a style related aspect. The owner-

1) *These statistics are in reality for all the population that is over 9 years of age.*

ISSN 0085-7130 © Telenor ASA 2004

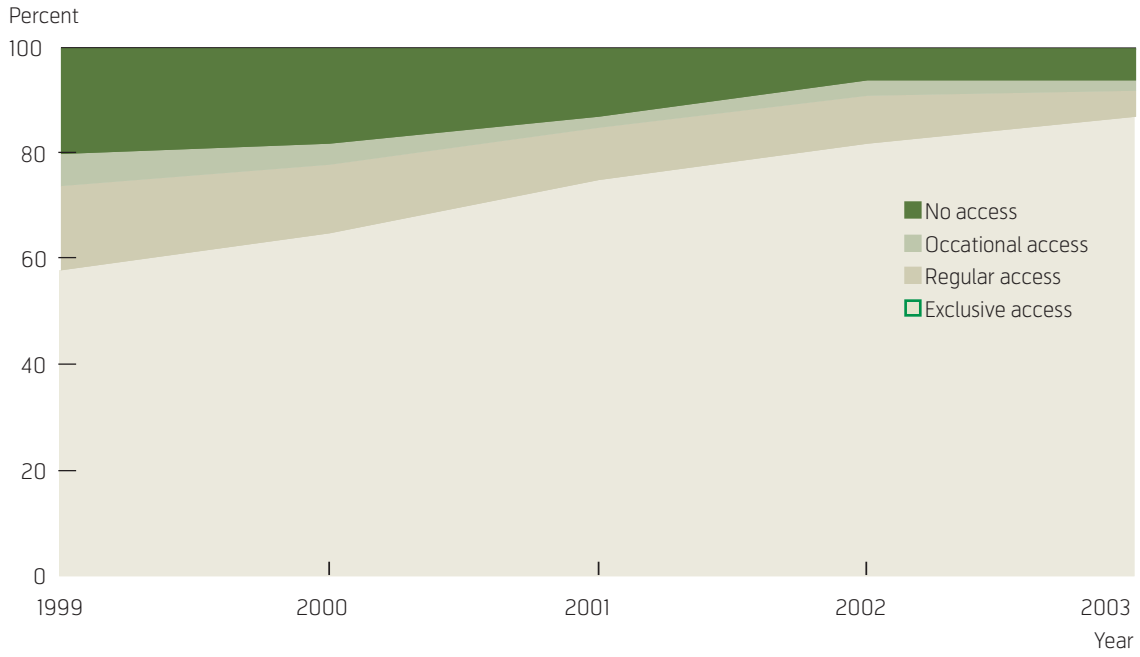


Figure 4 The individualization of the mobile telephone in Norway, 1999 – 2003

ship and display of a mobile telephone is, and will continue to be used as a type of fashion accessory. It will be selected, displayed and interpreted as an extension of the self, just as jewelry, clothing and other accessories are. Again, this plays on the growing personalization of the device.

Number of calls per day

The material from SSB shows that men use both the landline telephone and the mobile telephone more often than women (see Figure 5). In addition the material shows that women talk longer on the landline telephone, but not on the mobile. Starting with the number of calls per day men generally call more often than women (4.3 vs 3.5 times per day, respectively). In addition men are more inclined to use the mobile telephone when they call. In general men place about half of their private calls via a mobile telephone while women use the mobile telephone in about a third of all cases.

These things vary somewhat with age (see Figure 5). For both men and women, and considering both landline and mobile telephony, the golden age of use is

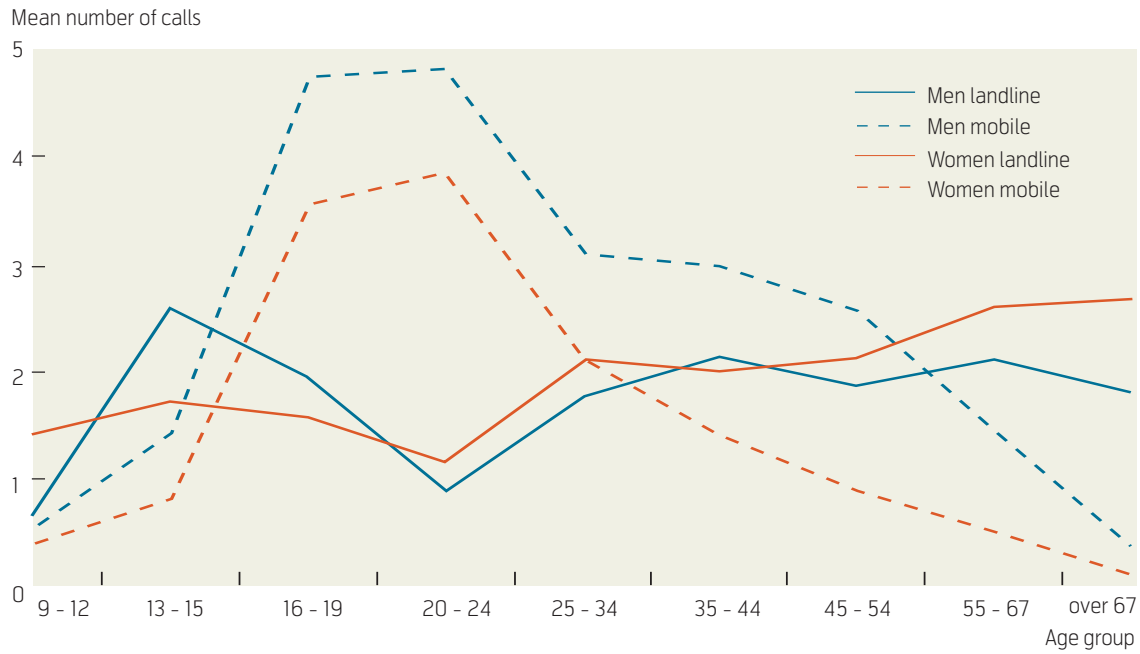


Figure 5 Number of calls per day (all platforms) and mobile

ISSN 0085-7130 © Telenor ASA 2004

the late teen years and the early 20s. It is during this decade in one's life that the telephone seems to be the most central, at least when thinking of the number of calls made. Males reported making about six calls a day during this period and women reported making about five. Roughly eight of ten calls are mobile based for men and slightly more than seven of ten are mobile based for women during this period of life.

The intensity of calling and the fact that it is largely mobile based is not surprising in many respect. It is in this period of life that one is quite nomadic. Often teens and young adults are in transition from the homes of their parents into either student life or into the early portions of their careers. In addition, partnerships are often in flux and there is not the routinized stability that one finds in the earlier portions of one's life or in those portions that follow this period. Teens and young adults do not necessarily have the convenience of having their most intimate sphere living in the same home and thus there is a need to contact others via the telephone in order to coordinate activities and maintain friendships. In addition, their nomadic life style means that mobile telephony fits well into their daily routines.

In the 9 – 15 age group one often operates within a relatively bounded range of movement, and particularly while one is in elementary school. During this period the stirrings of emancipation have only begun. The child operates within the local sphere and there is little need for telephonic contact on any broad scale. In a different way the stability of middle-aged life means that there is less need for telephonic contact with one's peer group. The most intimate sphere is often equivalent to those who live within the four walls of the home. In addition, the pressures of scheduling within various organizations (work, day care, children's free time activities etc) mean that there is a regularity to one's scheduling. There is not the same room for free renegotiation of activities that perhaps characterizes the lives of teens and young adults. This is reflected in the number of calls made. While there are many calls being made in the private sphere, the volume is approximately a third less.

Interestingly, the number of landline calls drops during the "nomadic" period. The drop for both sexes is about one call per day. However, this group is still about two calls per day over the average of others when considering the use of mobile telephone. Thus, the number of mobile calls more than makes up for the drop in landline calls. In addition, a certain percent of this group do not have access to a landline telephone.

Length of the calls

Where men call more often, women speak longer when they have first called, particularly in the case of landline telephony (see Figure 6). The other interesting aspect in terms of the time spent on the telephone is that – with some exceptions – there is not the same "golden age" during the late teens and young adulthood that one finds when looking into the number of calls. In general women report using the telephone just about 30 minutes a day where men say that use it about 20 minutes. There is a rapid rise in voice based telephony use when considering the difference between the youngest users and the late teen users. The data shows that teen girls between 16 and 19 report using the longest time on the telephone in general and also on the mobile telephone. The older female users reported somewhat lower use, particularly when considering their landline use. The other interesting issue here is that mobile telephony is used in about equal parts with landline telephony among men. However, aside from the teen respondents, women use about twice as much time on the landline telephone as the mobile for voice based communication.

Thus, while SMS has generally been the realm of teens and specifically teen girls, voice mobile telephony is in many respects the realm of middle aged men. There is a certain irony here. It is often women who have the responsibility for the coordination of familial life. This is in terms of the instrumental aspects of daily activities (the "soccer mom" as portrayed in the US experience). In addition there is often an expressive dimension to women's work in the sense that they are often the person who is in contact with the extended family and carries out various forms of care giving. Given this situation, one might suggest that the mobile telephone would be ideal as a tool to carry out at least the instrumental portions of this work. However, it seems that they have not adopted the device to the same degree as men. By way of explanation, men often have a job subsidized mobile telephone and thus there is not the same economic barrier for men. At another level, the mobile telephone is not seen as being the appropriate device for a longer telephone conversation. It is seen as being expensive, the device becomes warm during longer calls etc. Thus, care giving and the more expressive aspects of network maintenance – that is those areas of activity that are often within the purview of women – are better carried out via the landline telephone.

Clearly this means that the average length of a call increases as one goes from the younger age groups to those that are older. The elderly female callers report the longest mean call length (about 12 minutes per call) while teen boys report the lowest mean (less than 2 minutes per call).

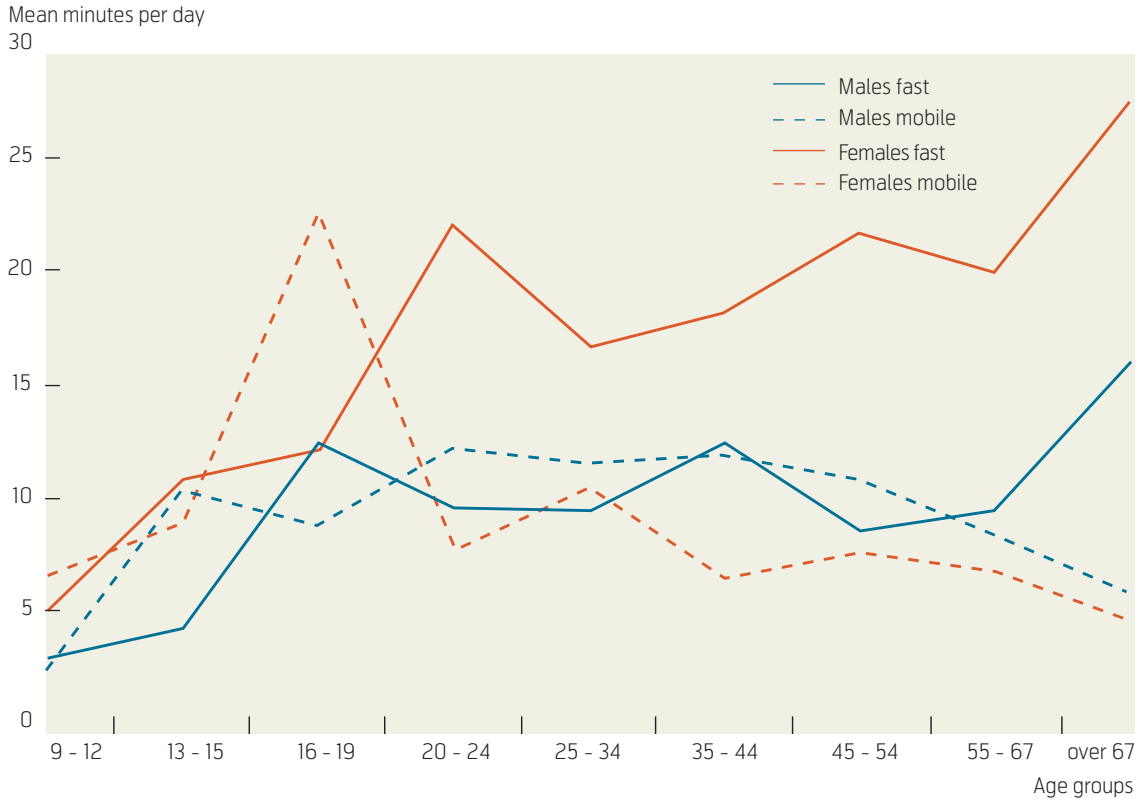


Figure 6 Minutes per day for all telephony and mobile by age and gender, Norway 2003

The use of SMS

There is a rough similarity when looking at the curves describing the daily use of voice mobile telephony and those describing the use of SMS (Figure 7 shows SMS use). In general the curves are low for the youngest pre-teens. Those who are in the teen groups and the young adult groups are at the top of the curve and finally those who are in the adult and the elderly groups report far lower use. We have seen this same general pattern when considering the use of mobile

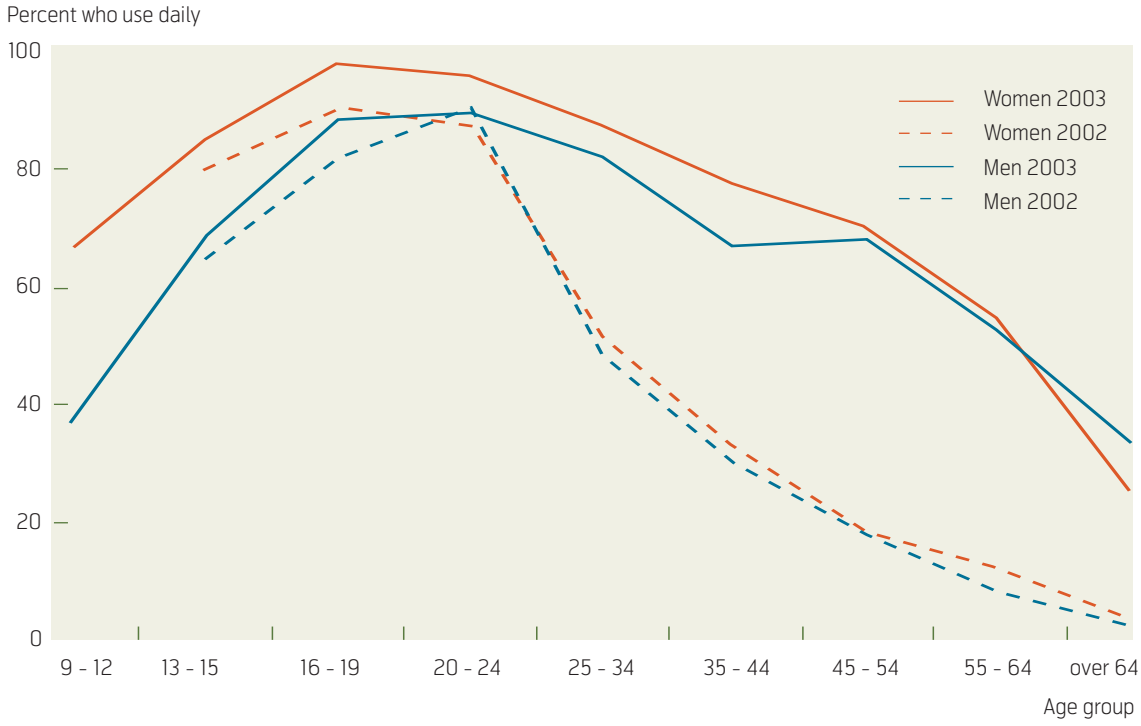


Figure 7 Percent using SMS on a daily basis by age and gender, Norway 2002 – 2003

telephony for calling. It is also apparent here. About half of the pre-teens with a mobile telephone reported sending SMS messages on a daily basis. This rose to 90 % for those in the late teen and early adult groups and then fell gradually to around 30 % of those with a mobile telephone in the oldest age group.

It is interesting to note that women report somewhat higher rates of use[2]. While 72 % of the men with mobile telephones reported using SMS on a daily basis, 81 % of the women reported the same. The data shows that for almost all age categories women were more inclined to use SMS than men. It is only among the oldest groups that there is parity.

There is also data on the daily use of SMS in 2002 contained in Figure 7. One can see that there has been a transition between 2002 and 2003 in this regard. Where in 2002 SMS was largely a teen and young adult form of interaction, older users have started to use this form of communication to a greater degree. This point will be further developed below.

What has not changed

General time used on the phone

One of the first things that seems to be quite stable since the late 1990s is the general amount of time we use for calling (see Figure 8). In 1998 the mean time respondents reported spending on the telephone was 23.8 minutes per day. In 2003 it was 24.3. The same is generally true with regard to the reported use of mobile telephony for voice calls. The respondents who had a mobile telephone in 1999 reported using 7.1 minutes per day while in 2003 it was 8.7 minutes.[3]

The age based profiles have also been rather stable (see Figure 9). The youngest respondents have not used much time on the telephone. However as they progress through the teen years their usage more than triples. For the youngest users the mean daily time on the telephone is about 10 minutes. For those in their early 20s it is over 30 minutes. This change is the most dramatic shift in use for all age groups. There is no other life transition that witnesses the same radical change in telephonic behavior. Clearly this change takes place against the backdrop of the child's emancipation from their parents. The telephone is that realm where many of the aspects of teen life find place. Gossip is exchanged; romances are established, maintained and ended. Social activities are arranged and teen identity is developed. As one moves away from their parents' home and into a more or less nomadic young adult period, the telephone continues to play a central role. During this period in one's life many of the same social affairs need to be arranged telephonically as in the case of teens.

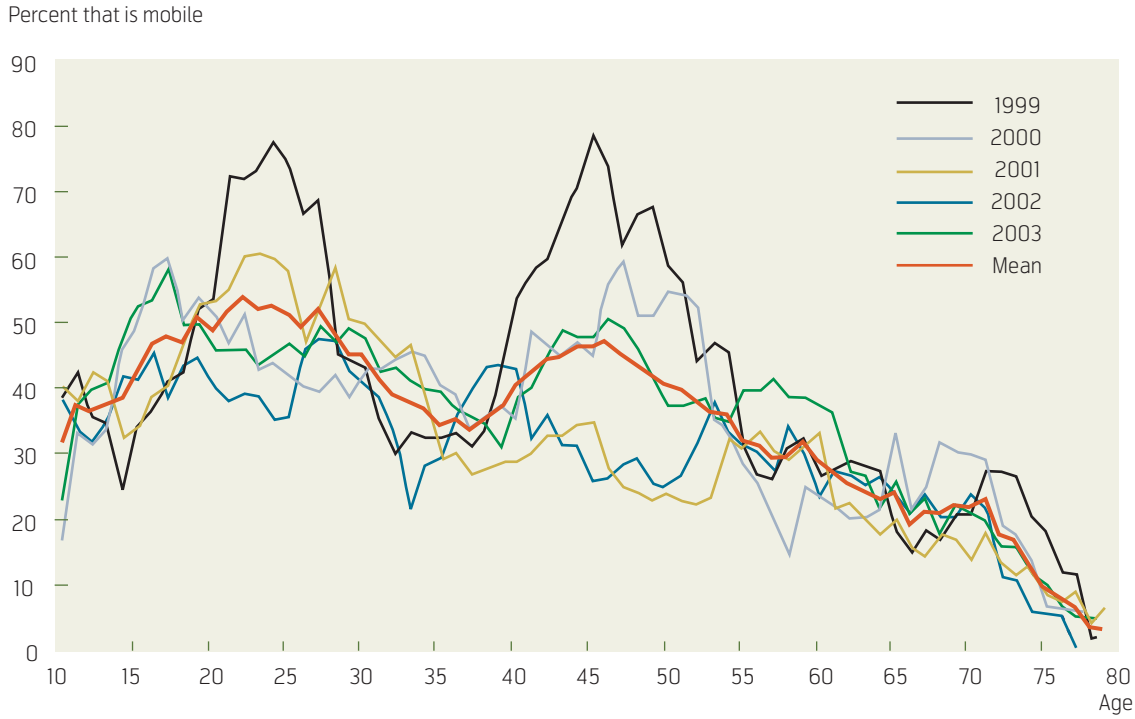


Figure 8 The mean minutes per day on the telephone

2) *Chi2 (1) = 9.9, sig. = 0.002*

3) *It needs to be noted that there were more mobile telephone users in 2003 than in 1999.*

ISSN 0085-7130

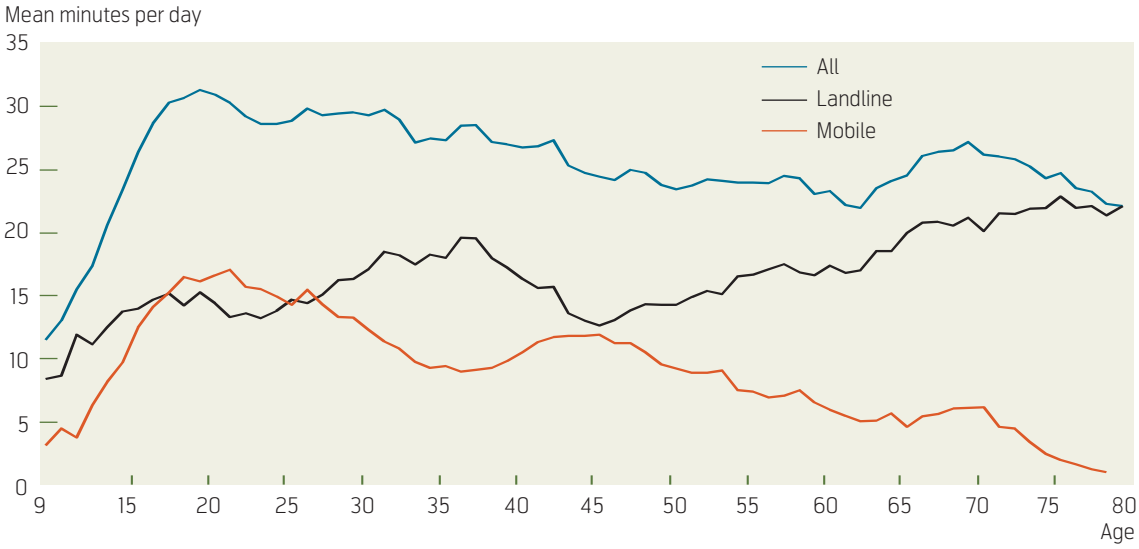


Figure 9 Percent of calls that are made on the telephone

As noted the peak time of use comes when one is in their mid 20s. From that point the respondents reported a decreasing amount of time on the telephone. This long slow decline continues until approximately the point at which the retirement period begins. The summary data from 1998 to 2003 shows a slight rise in reported use among the oldest respondents.

It is interesting to also look at the age groups who choose to use mobile and landline telephony. Young adults reported using more time with mobile than landline telephony. By way of contrast, the elderly users reported almost no time on mobile telephony, all the while maintaining a relatively high level of landline telephony use.

The percent of all telephony time that is spent on the mobile telephone also shifts for different age groups. In reality it is wrong to discuss this issue in a chapter examining stable elements in the use of telephony since it has increased with time. In 1999 approximately 30 % of all telephone time – that is all time in telephone conversations – was spent on the mobile telephone. In 2002 this had risen to 41 %. The year 2003 saw a drop to 36 %. Thus, while the total time spent on the telephone has remained about the same, it seems that mobile telephony has nudged its way into the total time budget here.

When looking at this same material across the age groups one sees that it is particular groups that have adopted the use of the mobile telephone. Here one sees that it is the young adults and middle aged individuals that are the most active voice mobile users. The young adults report that more than half the time used on the telephone is spent talking on mobile telephones. The other group is middle-aged telephone users. This is the group that most often has subsidized access to mobile telephony via their job. Thus, one might expect that the increased use is a consequence of this.

Those outside the revolution

We have been interested in following the adoption of mobile telephones. Early in the adoption process the mobile telephone was often seen as part of a family's holdings of equipment. It was used by that individual who may have had use for it at a particular moment. Children often had to, in effect, show a stronger need than adults since there is often a moral barrier to a child's use of the device. As time has gone, a more individualized relationship to the device has arisen. Instead of being a commonly held item in the home, it is an individual's. Indeed, many of the services are focused on individual consumption. As noted above, the catalog of names, sent and received SMS messages, various personalizations such as the ringing sound, logos and covers, and even the specific model – are all individualized. Even the fact that the PIN codes are commonly used underscores the sense that the mobile telephone is an individualized device.[4)]

4) *It is interesting to note that concrete ownership is often a slippery concept. In many cases the de jure and the de facto owner of the mobile telephone can be different. Many people who work in larger organizations report owning a mobile telephone that has actually been purchased by their employer.*

Thus, as the market develops it seems that one moves from a situation in which the mobile telephone is seen as a type of common property to a situation where it belongs to a single individual.

As we will see below, generally most people between 15 and early 70s either have a mobile telephone, or they have access to one. The two major groups who do not have one are the very young and the very old. Data shows that the elderly are only slowly adopting the mobile telephone. Part of this may be due to the parsimonious ethic of the pre-WWII generation. As a cohort that experienced both the depression and the depravation of the Second World War, they often carry an attitude of careful consumption. This is in contrast to the post WWII "baby boom" children who perhaps carry an ethic of freer consumption and also were often exposed to more electronic developments in their working lives. As the generation of pre-war persons moves from the scene and is replaced by those born after the WWII, one can suggest that there will be a more open attitude towards the ownership of the mobile telephone.

It is unlikely that young children will have mobile telephones unless they become some form of ambient device. The data shows that only about 10 – 15 % of the youngest children have a mobile telephone of their own.[5)] This has been quite stable for the last few years. There are some who receive one in the case of their parents' divorce. In this way, the non-resident parent has a direct line of contact with the child and does not need to go through the perhaps hostile filter of their ex in order to speak with the child. Beyond this group, there is little need for younger children to own a mobile telephone. Their social world is often quite close by and the responsibility of owning and maintaining a device may be beyond them.

Gendered use of the telephone

As noted above men make more calls per day than women, particularly when considering mobile telephony. In almost all age groups this is the case. It is particularly obvious when considering the use by young adults and in the case of middle aged persons. By way of contrast women talk longer on the telephone. Aside from the situation of teens and their use of the mobile telephone, this is most often the case for when considering the landline phone and its use by adult women. Finally SMS seems to be more of a female than a male arena.

These patterns, and in particular those associated with voice telephony seem to play on more traditional relationships to the technology. Where men seemingly have a large number of shorter interactions while ostensibly moving about, the data points to the notion that women have fewer, but longer telephone conversations from a fixed location. A simplification may be the quick, spontaneous call vs. the good friendly chat on the phone. These two approaches to voice telephony point, perhaps, to the socialization into different social roles. Where the spontaneous call may be an instrumental affair in order to quickly take care of some arising issue, the long conversation may point to the expressive maintenance of social ties. It is incorrect to say that it is all one way or another, but the data here seems to point to some of the broader issues that have been described by others with regard to the gendering of communication (Tannen 1991).

What is likely to soon change

The current situation indicates that the mobile telephone is overwhelmingly being used for voice interaction and to send SMS messages. These are the main areas of income for the mobile network operators. However, there are other areas where mobile communication may become more common. These include enhanced forms of interaction (MMS, push to talk and chatting), and other forms of data transmission.

MMS, that is the transmission of messages that can include text, pictures and sound, is a new form of interaction that is now on the market. In many respects this is a further development of SMS messaging in that it is relatively speaking asynchronous. It may come to occupy a somewhat similar position in the public imagination. It may become a way for individuals to share their daily experiences with others in their intimate sphere. Another function may be to enhance the communication between, for example, tradespersons who need to specify with relatively high degree of precision the situation at a work site. The MMS photo in Figure 11, for example, was sent by a carpenter to an architect in order to confirm that the conception and the execution of a particular plan were the same. Thus, we see that there are both expressive and, perhaps more importantly, instrumental issues that are at play in the eventual adoption and use of MMS. One can also speculate that, just as with the evolution of email, SMS and MMS will eventually merge into a single type of service. Along the same lines, one can imagine that multiparty chat rooms may develop in the mobile realm.

Another functionality that has already arrived on the scene is the so-called "push-to-talk" service. This is

5) *A much larger group has various forms of access to the device.*

Figure 11 An MMS photo taken by a carpenter and sent to an architect to check on the details of a construction situation

a form of IP telephony that is simplex, as opposed to the traditional duplex of the telephone world. That is, the "channel" is only available to one person at a time. Thus, users often have to adopt the "walkie-talkie" traditions of saying "roger" and "over" etc. The advantage is that more than two persons can simultaneously participate in a session. In addition, it is a relatively efficient service when seen from the perspective of system resources. Finally, it is not limited to only local use, as with traditional radio. Rather a push-to-talk session can include persons from broadly different geographic locations. In practical terms, groups that are in need of close coordination can use services like this. Families at shopping centers and working crews at building sites come immediately to mind as potential users.

Social consequences of mobile telephony

Up to this point I have focused on the number of phones, the length of conversations etc. Another issue is the actual impact of mobile telephony on our lives. It seemingly satisfies certain needs and at the same time can lead to difficult and unexpected situations. On the one hand it is wonderful to get a message from a child while we sit on the bus, it is good to coordinate (or perhaps micro-coordinate) a meeting, or it is good to call ahead to our family or friends and let them know that because of traffic we will be late in getting to dinner. At the same time, it is perhaps embarrassing to hear it ringing when we forgot to turn off the sound during that critical sales meeting.

We are in the process of working out our relationship to the mobile telephone. This is clearly not a static thing. As new situations arise we use the device in new and unexpected ways. New features give us new possibilities. At the same time we have to work out how and when to use the device. We are bending the device to our purposes, and we are reformulating our lives around the possibilities that it provides us. We are domesticating it (Ling 2004).

A short list of the mobile telephone's social consequences indicates that the device provides us with an enhanced ability to coordinate our activities with others, it gives us a sense of safety and it has become an element in teens' and parents' negotiations over emancipation.

Taking these in order, the greatest contribution of the mobile telephony will be its contribution to coordination. In a modern society, we have experienced the rapid growth of the cities and the adoption of rapid, individualized transportation in the form of the automobile. We know from many studies both in Norway and in other countries such as France, Germany, Japan and Korea that the bulk of traditional landline telephone calls are primarily used to for instrumental purposes, in other words for coordination. We use the telephone to make and confirm appointment, organize schedules and arrange the delivery and recovery of children at various events.

Mobile telephony expands these possibilities. With the mobile telephone we can plan – and re-plan – activities anywhere and at any time. This can be done far more conveniently than with the traditional telephone since we are, in effect, calling a person, and not a geographical position. Thus, we do not need to be at a specific place, node or location to receive information. This increases the efficiency of planning our everyday activities. If we get out of a meeting early we can call our spouse and alter who will pick up the kids at day-care or shop for groceries. When in the grocery store we can again call to find out if it was Swiss or Cheddar cheese that is needed for tonight's dinner. These are mundane issues, but they are significant in the way that they lubricate the machinations of everyday life. Indeed, the ability to quickly coordinate activities in a complex society is probably the most significant contribution of the mobile telephone.

Coordination is also a key use among tradespersons, sales people and other workers who spend their time away from an office. The ability to quickly exchange information with colleagues, order materials and check details with the office facilitates their work process.

The ability to coordinate on a mobile basis has indeed been revolutionary for some portions of society. A good example of this are deaf persons. Previous to mobile telephony, and specifically SMS, this group

was reliant on a special telephone and translation service in order to coordinate their activities and to coordinate their daily needs. If, for example, an unexpected problem arose and there was the need to re-plan a meeting it was often impossible to assemble the technology and competence to send out the message, particularly if people were en route. Text based SMS has cut through this Gordian knot with a simple and effective way to help this group coordinate their activities.

Looking now at safety the mobile telephone has found a niche in that it provides us with a sense of security. Mobile telephony offers persons with chronic conditions (and those who care for these people) a broader range of movement. The mobile telephone means that help is accessible should a problem arise. The same can be said for those suffering from an acute problem that can range from punctured tires to life threatening situations. We need to take care here however. It is important not to be lured into a false sense of sanctuary. The nearest base station can be out of range, the batteries can be empty and the electronics can be vulnerable to moisture. In addition, use of the mobile phone in certain situations (most particularly while driving) has been shown to be dangerous. It is easy to think that we can clear up a few quick tasks while driving. However, research shows us that driving and talking on the phone is a dangerous combination.

Finally, as discussed above, one of the most surprising aspect is its adoption and use by teens. The combination of pre-paid cards and easy access to mobile telephones has meant that this group is among the most enthusiastic mobile telephone users. To put this into context, one of the major tasks of teens is to develop a sense of identity and to emancipate themselves from their parents. The role of the parent is to provide their children with the ballast they need for this transition. The peer group is also an important element in that they provide the teen with a reference group and a milieu in which the individual can try out their nascent adult roles.

The mobile telephone is a perfect tool in this situation, particularly when seen from the perspective of the teen. It provides them with a communication channel over which they have control. It is free from the surveillance that parents and siblings can enforce over the traditional landline telephone. Beyond free access, the terminal itself is an icon to freedom. One need only look at the advertisements for the popular mobile telephone terminals. The terminal is also a locus for information on who is a part of the gang and a variety of SMS messages. Thus, the (relatively) unhindered access to friends as well as the ownership of a device that serves as a type of "friendship central" is a powerful philter.

Teen girls are central in this development. While teen boys were the first to really adopt the device, one can say that it is among teen girls that mobile telephony has found a robust form. Analysis shows that teen girls are more active users. They are also strong SMS users. When comparing the types of messages teen girls send to those sent by boys we see that teen girls write longer messages. They include more information in their messages. They use more complete grammar and they seem to be more nurturing in their messages.

In sum, mobile telephony has found its place in Norway. It is still working out how, where and when it will be used by various groups. That it will be used, however, is not in question.

Bioliography

Fischer, C. 1992. *America calling: a social history of the telephone to 1940*. Berkeley, CA, University of California.

Haddon, L., Ed. 1997. *Communications on the move: The Experience of Mobile Telephony in the 1990s*. Farsta, Telia.

Ling, R. 2004. *The Mobile Connection: The cell phone's impact on society*. San Francisco, Morgan Kaufmann.

Tannen, D. 1991. *You just don't understand: Men and women in conversation*. London, Virago.

Rich Ling (50) is a sociologist at Telenor R&D. He received his Ph.D. in sociology from the University of Colorado, Boulder in his native US. Upon completion of his doctorate, he taught at the University of Wyoming in Laramie before coming to Norway on a Marshall Foundation grant. Since that time he has worked at the Resource Study Group and has been a partner in the consulting firm Ressurskonsult. Since starting work at Telenor he has been active researching issues associated with new ICT and society. He has led projects in Norway and participated in projects at the European level. Ling has published numerous articles.

email: richard-seyler.ling@telenor.com

How Telektronikk changed the Web

HÅKON WIUM LIE

Håkon Wium Lie is CTO of Opera Software

Transferring *Telektronikk* onto the Web was not a straightforward task in 1993. The Mosaic browser had been released just months earlier, and procedures for preparing documents for the Web were not established. The problems experienced from this early work were directly used to propose better methods in CERN's Web project. Since then, the Web has firmly established itself as the electronic publishing platform of choice, and we should expect that the web version of *Telektronikk*'s Cyberspace edition from 1993 also will be readable in 2043.

Introduction

In 1993 I was asked to be the guest editor of an issue of *Telektronikk*. A few months earlier, the Mosaic web browser had been released and it suddenly became possible to demonstrate the concepts that the articles would describe: how to publish information in Cyberspace. My editing job, which in the beginning mostly consisted of nagging human authors (as other guest editors of *Telektronikk* will have experienced), suddenly became a very practical challenge of getting all the bits in the right order for the Web. I am proud of the fact that *Telektronikk* 4.93 was one of the first paper publications to become available on the Web[1)], and our publishing efforts received an 'Honorable Mention' at the first Web conference at CERN in 1994[2)]. Also, as I will recount below, the challenges facing us in 1993 had some long-term implications for the Web.

The challenges in 1993

Preparing a document for publication on the Web in 1994 was very similar to what it is today. You convert the textual content to HTML, adorn the text with images, add hyperlinks to other documents, and finally the document is served to browsers across the world over the Web. However, in 1993 these procedures were not established and much of our efforts (I received help from Norwegian Telecom's MultiTorg research group) were spent experimenting with different kinds of HTML markup, how to use images, and how to link the documents together.

In particular we were faced with two challenges. First, network connections at the time were much slower than they are now and browsers did not support progressive rendering as they do today. So, when we added images to the articles, they suddenly became slow to to load and the user experience suffered accordingly. Second, there was no established way to convey the graphics design of *Telektronikk* from paper to Web.

The first challenge was solved by making small versions of the large images. By placing a thumbnail graphics (as they would later be referred to) on the page instead of the large image, the page would load quickly. By clicking on the small images, the larger version would be downloaded and shown. This technique is one of the standard tricks of the trade today, and *Telektronikk* was – to the best of my knowledge – the first to use it.

The second problem turned out to have no good solution at the time. The paper version of *Telektronikk* was professionally styled and we wanted to use the same graphics design on the Web. However, HTML was designed to represent the *content* of the document and not its *style*. HTML was developed in a scientific environment where the content is more important than its presentation, and authors are encouraged to declare that some text is, say, a heading rather than what font to use. Therefore, in order to capture the fonts and colors used in the paper version, the web version resorted to using an image. That is, instead of sending browsers the title "Telektronikk 4.93 Cyberspace" as text, an image of the text was sent instead. This is similar to how fax machines work, but on the Web this practice is frowned upon. There are several reasons for this. Text in images is only accessible to those of us who can see, and not to speech synthesizers. Also, it is impossible to search for text in images, and images generally take up more capacity than text.

Solutions

In 1994, I joined Tim Berners-Lee's web project at CERN. Berners-Lee had written three of the specifications that formed the columns of the Web: HTML, URL and HTTP. They describe, respectively, the content of the document, how to link one document to another, and how to transfer the document from the server to the client. What was lacking from the Web

1) *See: http://people.opera.com/howcome/1993/telektronikk-4-93*

2) *See: http://botw.org/1994/awards/design.html*

ISSN 0085-7130 © Telenor ASA 2004

in 1993 was a way to describe the presentation style of documents – fonts, colors and layout. Could we design a solution that would solve the problem for *Telektronikk* and other web publications? In October 1994 I proposed a solution: Cascading Style Sheets (CSS). Using CSS, authors could describe the presentation of their documents. The style sheet could be placed in a separate file and many documents could refer to it. For example, one style sheet could describe the look and feel of a web site, or perhaps all electronic editions of *Telektronikk*.

CSS was not an immediate success. Even if authors liked the idea of using style sheets, the popular browsers at the time did not support style sheets. Three days after CSS had been published, a company called Netscape Communications announced their first browser. The new program was a significant improvement over Mosaic and users flocked to it. Naturally, the new browser did not support CSS. I spent the next few years of my life trying to convince browser vendors to add support for CSS. Netscape finally added support in version 4. Microsoft showed eagerness when they started competing with Netscape, but the quality of their work was – to use a polite term – uneven. Only consistent pressure from web authors and users forced the big software makers to fix problems and make CSS usable.

Meanwhile, a small Norwegian browser company had started to make headlines on the web. Opera Software. Two other members of the MultiTorg group, Geir Ivarsøy and Jon S von Tetzchner, had seen the opportunity for a better browser and founded a company to turn ideas into products. I admit to not having much faith in the venture in the beginning. Competing with Mosaic, Netscape and Microsoft – or Americans in general – is a challenge. Thankfully, they did not listen to me and went ahead with their ambitious plans. In 1998 it was time for Opera to add support for CSS. In three months, Geir Ivarsøy had implemented better support for CSS than Microsoft and Netscape had spent three years on. At that point I was convinced that Opera would make it and joined the company.

Web future

More than a decade has passed since *Telektronikk* 4.93 was published. In that time, the Web has changed the way humans access information. URLs are printed on all kinds of products and "google" has become a verb, much like "xerox" did in a paper-based world. The browser has become the window to a world of electronic information where electronic equivalents of newspapers, brochures, dictionaries, laws, application forms, and bulletin boards abound.

At a technical level, however, the Web has not changed that much in ten years. *Telektronikk* 4.93 was written in HTML and the web version is still readable in browsers of today. HTML, along with HTTP and URLs, are still basic building blocks of the Web. CSS, JavaScript and a few other specifications have been added to the list, but today the Web has stabilized as a platform for publishing.

There are several reasons why the technical foundation of the Web is not developing quickly any more. First, there are hundreds of millions of browsers out there that use the basic building blocks. Adding support for a new specification requires that browsers are replaced and this is a major undertaking in 2004, much more so than in 1994. Second, the Web has a solid foundation which does not necessarily need much more functionality and which has performed remarkably well under the pressure of growth.

As a rule of thumb, in order for a technology to replace an existing one it has to offer something better by an order of magnitude. I do not foresee any new publishing technology stepping forward to replace the Web in the foreseeable future. How long is the foreseeable future? I bet that the web version of *Telektronikk* 4.93 will live for at least 50 years. That is, common computers in the year 2043 will still be able to read those web pages we authored in 2003. Anyone against?

Håkon Wium Lie (39) is a Web pioneer, having worked on the WWW project at CERN, the cradle of the Web. He first suggested the concept of Cascading Style Sheets in 1994 and he later joined W3C (the World Wide Web Consortium) to further strengthen the standards. In 1999, he was listed among Technology Review's Top 100 Innovators of the Next Century and has also been invited to the World Economic Forum as a Technology Pioneer. He is currently a member of the W3C's Advisory Board, Technology Review's "TR 100", and World Economic Forum's "Technology Pioneers".

Wium Lie holds a master's degree in visual studies from MIT's Media Laboratory, as well as undergraduate degrees in computer science from West Georgia College and Østfold College, Norway.

email: howcome@opera.com

 ISSN 0085-7130

The cover

In 1993 it was not possible to represent *Telektronikk*'s cover (seen on the right) without resorting to an image. Today, with the addition of CSS, it is possible. The code fragment below encodes the *Telektronikk* cover from 4.93 and works in common browsers:


```
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 3.2//EN">
<html>
<head>
<title>Telektronikk</title>
<style type="text/css">
BODY {
  background: rgb(87, 116, 139);
  font-family: "Arial", "Helvetica", sans-serif;
}

div, h1, h2, h3 {
  margin: 0;
  padding: 0;
}

div {
  width: 20em;
  text-align: right;
  float: left;
}

div h1 {
  color: rgb(200, 200, 230);
  float: right;
  background: black;
  padding: 0.3em 0.3em 0.7em;
}

div h1 span {
  color: rgb(255, 250, 255);
  background: black;
  padding: 0 0 0.1em 0;
  font-size: 1.3em;
  border-bottom: 0.15em solid rgb(200, 200, 230);
}

h1 {
  float: left;
  padding: 0.6em 0.3em 0.7em;
  font-size: 2em;
}

h1 span {
  color:  rgb(200, 200, 230);
}

h2 {
  color: rgb(186, 255, 201);
  margin-right: 0.4em;
  font-size: 1.5em;
}
</style>
</head>
<body>
<div>
<h1><span>Telektronikk</span></h1>
<h2>Cyberspace</h2>
</div>
<h1><span>4.93</span></h1>
</body>
</html>
```

From radiotelegraphy to fibre technology

Telenor's history and development on Svalbard

VIGGO BJARNE KRISTIANSEN

Viggo Bjarne Kristiansen, now retired, was Managing Director for Telenor Svalbard

In a few years Telenor[1)] will celebrate the centenary of their presence on Svalbard. Of course it is not possible to actually quote the people who were central in the pioneering work carried out in this arctic area in 1911. However there is some material, though very little written material exists, which can provide us with insight into the background for this brave project which was not only important for Telenor, but also had national and international political ramifications.

The discovery of Svalbard

To understand Telenor's interest in Svalbard, one has to have some insight into the prevailing conditions on Svalbard during the first two decades of the last century. Svalbard, or Spitzbergen as it was then called, was a terra nullius (no man's land). Nations who had an interest in the area made a stake for themselves, annexing areas where they anticipated immediate or future financial gain.

Svalbard was discovered in 1596 and activity in the area began only 15–20 years later. A Dutch expedition consisting of two ships prepared for an expedition north. The objective of the voyage was to find a sailing route through the Northeast Passage. Willem Barentsz (1549–97) was on board one of the ships. Barentsz was not the captain, but the navigator, perhaps the most important position on the ship for such an undertaking.

On 17 June 1596 the crew saw land at 80 degrees north. They saw a landscape of tall, pointed mountains. For this reason they called it Spitzbergen, which means pointed mountain. A few days before that, around June 9 or 10, they also visited Bjørnøya (Bear Island).

Looking at a map of this northern area today, the natural question is: "What on earth was Barentsz doing near Spitzbergen when his goal was to find the Northeast Passage?" The answer is an easy one. There was no GPS or other navigational tools that were accurate enough to allow the crew to "blindly" sail to their destination. Navigation in this era was complicated and depended on trial and error.

Nonetheless we can be pleased that Barentsz got lost. Because of his inaccurate navigation, we now know when Svalbard was discovered and who the discoverer was. Unfortunately, Barentsz did not make it back to the Netherlands, his home country. He died on this expedition after his ship was packed down by ice at Novaya Semlya later that year. The crew managed to get ashore, but several crewmembers died from the exertion.

Many historians believe they have reason to argue that Svalbard was discovered long before Barentsz' arrival in 1596. Some argue that Icelandic Vikings arrived on Svalbard as early as the end of the 11th century. Others believe that Russian Pomors hunted and fished on and around Svalbard in the 14th century. Still others maintain that there were settlements on Svalbard as early as the Stone Age. Because there is so little evidence, these theories are difficult to confirm. Therefore, until proved wrong, it is correct to assume 1596 and Willem Barentsz. In other words, we can state that the discovery of Svalbard was relatively recent, only approximately 100 years after the discovery of America.

What were the consequences of the discovery of Svalbard?

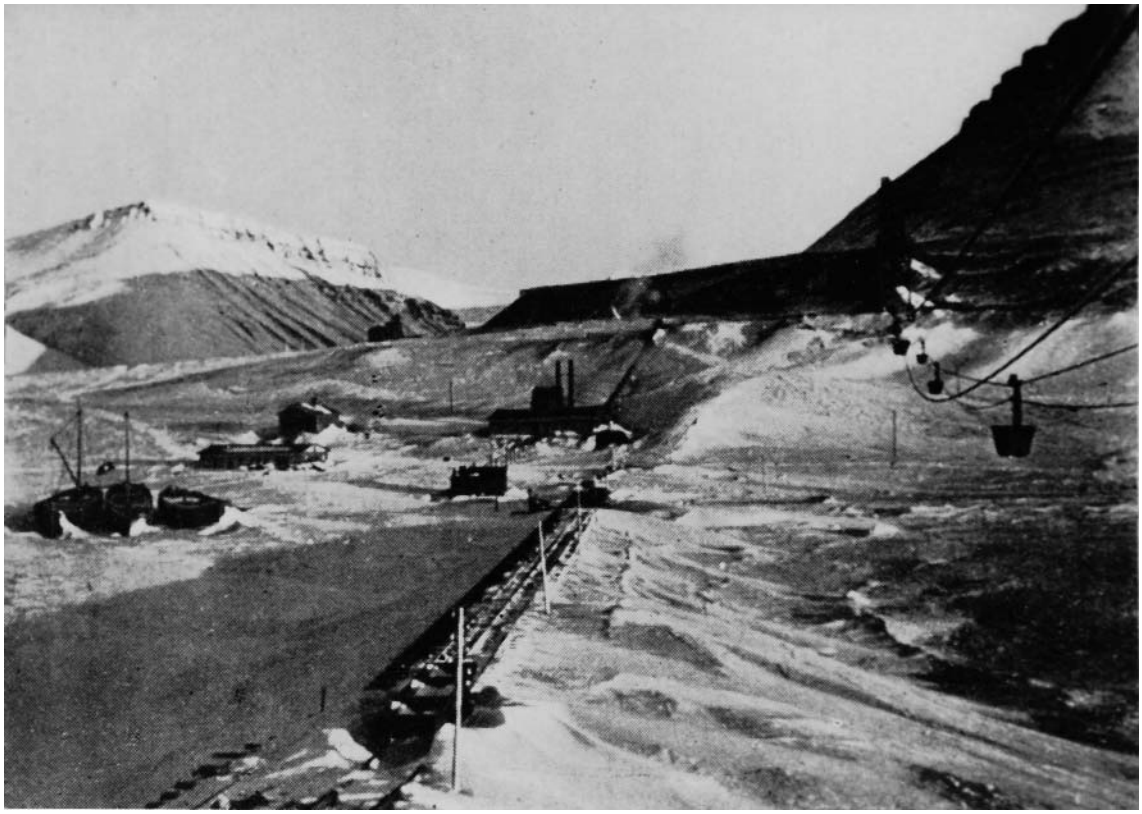
Barentsz and his crew did not merely see ice, snow and pointed mountains in this arctic area. They also saw a large number of sea mammals such as whales, walrus and seals; in other words, considerable resources that could be exploited. It was not long before interest in these resources was awakened in Europe.

Englishmen began hunting for walrus in the area around Svalbard as early as 1605. Later, Dutch, French, Danish and Basques came as well, and gradually the hunt also included polar bears and reindeer. When whaling became popular these activities really took off.

1) *Telenor has changed its name several times. Up until 1969 it was called Telegrafverket, then Televerket, and from 1995 the name has been Telenor. In order to avoid unduly confusing the readers of this article, the name Telenor will be used regardless of which name was actually used at the specific dates in the article.*

 ISSN 0085-7130

Svalbard is the name of an archipelago lying between 74 and 81 degrees north and between 10 and 35 degrees east. The land area is approximately 63,000 square kilometres. Two-thirds of this is covered by ice and snow. The largest island of Svalbard is Spitzbergen. Other islands are Bjørnøya, Hopen, Egdeøya, Nordaustlandet, Prins Karls forland, Barentsøya, as well as some smaller islands



Longyear City (now Longyearbyen) as it used to look when The Arctic Coal Company ran its operations here between 1906 and 1916

The hunt for whales and walrus went on until the mid 16th century. In spite of the enormous number of animals, the hunt resulted in these species being almost wiped out and the vessels had to sail further out to sea along the icy region near Greenland.

After the heyday of whaling, Greenland whales were still hunted, but the seal hunt became the main occupation for the next two hundred years. However, the Greenland whales were also practically extinct, and hunting for land animals such as polar bears, reindeer and foxes increased. It was at this time that hunters began to stay on Svalbard through the winter. The hunt for polar bears continued unabated up until this species became protected as late as 1972.

Basically the hunt was uncritical and went for 'everything that moved'. There was no regulation, and as a consequence, several species were at the point of becoming extinct on Svalbard. The only reason that this greedy practice could continue was that no country was responsible for regulating the exploitation of resources. After all, Svalbard was no man's land.

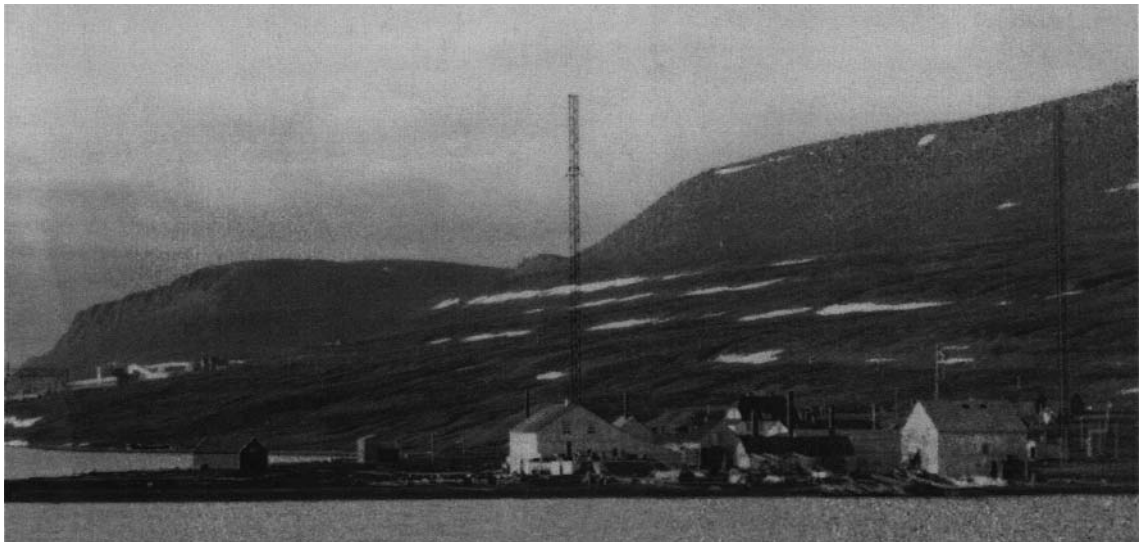
Around 1900, industrialisation began on Svalbard. This era commenced when large coal deposits were found, enticing many people to travel to Svalbard to extract this sought-after resource. Various minerals were also found, among them gypsum, lead, iron, marble and asbestos. These discoveries led to many companies and individuals travelling to Svalbard to extract and exploit these resources, often with the motive of making a quick profit.

Since the area did not formally belong to another country, those sufficiently interested simply annexed the areas they were interested in. Few of those coming to Svalbard planning on earning money by exploiting the island's resources had much luck. Many started, but only a few actually succeeded.

Among the first to start serious commercial coal operations from Svalbard was John Munroe Longyear, a rich American businessman. He began coal operations in what is today called Longyearbyen. The place bearing his name was then called Longyear City ('byen' is the Norwegian word for city [translator's remark]). The name of his company was The Arctic Coal Company (ACC). The company started operations in 1906, and it would turn out that initiatives by this company provided the basis for Telenor's first establishment on Svalbard.

Telecommunications in no man's land

Coal was a product which was sought after all over the world, and if ACC were to sell their coal, they were dependent on contact with the world beyond Svalbard. The only contact with the mainland was by



Spitzbergen radio at Finneset in Grønfjorden was built and put into operation in 1911. The station was active until 1930 when it was moved to Longyearbyen

boat. A trip to Svalbard from the mainland usually took more than two days each way.

Radiotelegraphy had come into use many places around the world, even some places in Norway. John Munroe Longyear was certain that this had to be the solution. So in 1910 he contacted the Norwegian authorities and launched a proposal that ACC would build a radiotelegraphy station in Longyearbyen. But in order for it to operate, a station had to be built in Norway that could communicate with ACC's station on Svalbard. Of course, Munroe Longyear believed that constructing this station was unproblematic, as the area was in the possession of ACC. What remained was only to convince the Norwegian authorities to build the station in Norway.

An alternative to Norway funding a station on the mainland was that ACC be permitted to build this station as well. For Munroe Longyear the matter was simple and straightforward. But it turned out to be more complicated than he had anticipated.

The Norwegian authorities used all possible means to consolidate their position on Svalbard, and they now regarded this as an opportunity to strengthen their presence. This meant however that Norway would build both stations; the one on Svalbard and the one on the mainland. This gradually became the Norwegian consensus, but only after the matter was politically evaluated nationally and internationally.

Thomas Th. Heftye, the telegraph director at that time, was a keen politician with an interest in radiotelegraphy, which was the new telecommunication technology at the time. Heftye's attitude and role in this matter was of crucial importance for the resolution of the Storting, the Norwegian National Assembly, to grant NOK 300,000 to the construction of a station on Svalbard and a station on the coast of Finnmark. The decision, which was made at the beginning of May 1911, aroused some irritation among the leaders of ACC who had taken it for granted that the station on Svalbard would be built by them. In addition, several nations were interested in Svalbard and expressed their dissatisfaction on Norway's position on this matter.

Nor was the chosen location on Svalbard that which ACC had expected, it was not to be Longyearbyen, but Finneset in Green Harbour (today called Grønfjorden, close to Barentsburg). Here, Norway already had interests in a whaling station, and it was clear that the Norwegian authorities wanted to build up a Norwegian "colony" in this area, despite the fact that the site, due to its location, was not particularly suitable for radio communication. The station on the mainland was located at Ingøy on the coast of Finnmark.

Construction of the Svalbard station started at late June 1911 and the station was formally opened on 22 November the same year. An impressive amount of work was carried out during this short time span: several buildings were built, mast foundations installed and antennae were mounted and tested on the station. Hermod Petersen, department engineer, was in charge of its execution and was also the station's first station manager with responsibility for the station from the autumn of 1911 to the summer of 1912. In its first year of operation, the station had a crew of six men, including the station manager. The station was named Spitzbergen radio, a name that was

kept until 1925 when Norway assumed sovereignty over Svalbard. The name was then changed to Svalbard radio, which it still has today.

The Arctic Coal Company was forced to use Spitzbergen radio until 1912, when they were able to construct their own station in Longyearbyen. ACC's station communicated with Spitzbergen radio which imparted ACC's traffic to be dispatched out of Svalbard. Until ACC had their own station in operation, telegrams were delivered to and picked up from Spitzbergen radio by couriers from ACC.

After Spitzbergen radio came on the air, several small private stations opened around the islands of Svalbard. One of the most important stations was Kings Bay radio, later called Ny-Ålesund radio. This station was made operational in 1918 and became a vital link during Roald Amundsen's attempt to reach the North Pole by seaplane in May 1925, and during his expedition on the airship "Norge" from Ny-Ålesund, via the North Pole to Alaska in May 1926. During Umberto Nobile's attempt to reach the North Pole on the airship "Italia" in 1928, the Ny-Ålesund station assisted with the radio traffic to and from the airship. As most people know, this voyage ended tragically, and Ny-Ålesund and Svalbard radio stations were important parts in the search for the airship and her crew. Roald Amundsen also participated in the search party, which ended tragically as his airplane lost control and everyone aboard died, including himself. Neither the airplane nor the crew were ever found. It is thought that the airplane crashed into the sea somewhere between Tromsø and Bjørnøya.

Photo: Norwegian Telecom Museum

Thomas Thomassen Heftye was the telegraph director from 1905 until he died in a train accident in 1921

Radiotelegraphy – the solution of the time

Radiotelegraphy was first used around 1900. The first station for traffic between Europe and USA was opened in 1907, but already a year earlier Norway began using radiotelegraphy between the stations at Sørvågen and Røst. This was the second connection in the world that was used in the public telecommunication network. Interest in this communications solution was huge all over Europe, just as in Norway where radiotelegraphy was seen as a good solution for covering the long coastline that otherwise would be very expensive to cover by use of physical communications. It was only natural that people inhabiting the coastal areas demanded that this solution be implemented. It was stressed that building land stations would provide greater safety for the fishing fleet when they could acquire radio equipment on board their boats.

As already mentioned Heftye, the telegraph director, was strongly in favour of this form of telecommunications, but he knew that the financial situation, which he was responsible for, would not permit the construction of radio stations to the extent he felt was necessary. The state coffers were also practically empty. So it was seen as somewhat of a sensation when the Storting permitted the allocation of NOK 300,000 to construct these two stations on Svalbard and Ingøy respectively. There was no way these investments could be justified by anticipated revenues. For this reason it was apparent that there were other motives than economic ones which were behind the decision of the National Assembly. It was clear that this was a Norwegian move towards acquiring a stronger foothold on Svalbard.

Spitzbergen radio – a Norwegian meeting place on Svalbard

Telenor was the first public enterprise to set up permanent operations on Svalbard. Actually, the Post Office was there first, but the postal service was a more seasonal operation. Incidentally, the Post Office was located at Spitzbergen radio when this was opened.

Spitzbergen radio soon became a meeting place for Norwegians. This was not merely to search for communications with the mainland, but also to hear news from all over the world. And besides, all the visits helped the radio station acquire a reputation as something of a social gathering place.

Many people visited the radio station at Finneset during its operative years. Among the most famous was Fridtjof Nansen, who visited the station twice during his oceanic expedition to Svalbard in 1912. In his book *A journey to Spitzbergen* he wrote about his visit to Spitzbergen radio. He gave the station crew a fine testimony and particularly mentioned the hospitality and assistance he received. Among other things he was able to adjust his chronometer with the aid of a signal which was sent from Paris every evening at midnight.

Thomas Heftye, the telegraph director, had already visited the station in July 1911, while the station was under construction. During his visit he insisted that the Norwegian flag should be flying at the station. He had his way and the flag was hoisted on 23 July 1911. There is reason to believe that this was the first time that the Norwegian swallow-tailed flag was flying on Svalbard. Svalbard was still no man's land, and this event did not go unnoticed, but brought reactions within Norway and from countries with an interest in the events on Svalbard.

When the first Governor was appointed in 1925, there was no office or house for him on Svalbard. Instead he carried out his duties from Spitzbergen radio, where he lived and worked during this initial period.

Hunters and fishermen with hunting stations around Svalbard could now make their way to Spitzbergen radio to contact their employers on the mainland. From the radio station they could provide information about hunting and fishing conditions and not least make arrangements for when they wanted to be brought home.

In many ways Spitzbergen radio contributed to creating a new and better existence for the people living on Svalbard.

On the move

The Arctic Coal Company sold its operation in Longyearbyen to the 'Store Norske Spitsbergen Kulkompani' in 1916. In this way Longyearbyen was "on Norwegian hands". International work was underway to determine who would have sovereignty over Svalbard.

On 9 February 1920 the Svalbard treaty was signed in Paris. Nine countries were behind the treaty, although it did permit the inclusion of others, and today more than 40 countries have joined. The treaty gives Norway full and unrestricted sovereignty over Svalbard, but gives all citizens of the signatory states equal rights for hunting, fishing and business enterprise on Svalbard. Norway took over the formal sovereignty of Svalbard on 14 August 1925. This date is celebrated as Svalbard's "national holiday" and is an official flag-flying day.

Following this, in 1928, it was resolved that Spitsbergen radio would be moved from Finneset to Longyearbyen. The station was now named Svalbard radio. The decision about the location was later changed, and it was decided that the station would be moved to Ny-Ålesund.

Ny-Ålesund was then to be the new location of the station and work to move the station began in 1929, but following a major mining accident in Ny-Ålesund on 16 August 1929, the Kings Bay Kulkompani decided to shut down mining operations at Ny-Ålesund. As a result of this, the resolution to move the station was once again changed, and the new location would be Longyearbyen. Svalbard radio moved to Longyearbyen in 1930 and is still there today. That same year, the station at Finneset was shut down after 19 years of service. Radiotelegraphy was still the ruling communications solution and would remain so for many years to come.

Isfjord radio is built

The location of Svalbard radio between high mountains at Longyearbyen did not provide optimal sending and reception conditions, and in particular, boat traffic complained about the conditions. It was therefore decided that a radio station and a lighthouse would be built at Cape Linné, at the mouth of the Isfjord estuary. The Norwegian Polar Institute was responsible for building the station which was completed in 1933. The topographical conditions of this area with an open south-facing horizon were considerably better and connection quality was noticeably higher with this station. In particular boat traffic had a considerably improved coverage between Svalbard and the mainland, as well as in the area around the islands.

Svalbard radio was still active, however, and traffic from this station now went via Isfjord radio as the station was called. During its first year Isfjord radio was run by a crew of three: the station manager, the telegraph operating assistant and a cook. Later on, an engineer and a handyman joined their ranks.

Isfjord radio was built in 1933. It was destroyed during World War II, but reconstructed in 1946. The picture shows how the station looked before the new station building came in 1957

Svalbard radio in Longyearbyen and Isfjord radio at Cape Linné were actively operating radiotelegraphy until after the outbreak of World War II.

The war years at Svalbard

Spitzbergen radio was active throughout World War I, from 1914-1918. The station was instructed to keep a close eye on all shipping activities and in particular to notify in the event of anything suspicious. In addition the station was responsible for reporting weather data and data about ice conditions around Svalbard. Except for this, life went on as normal and there was no military action on the archipelago.

However, Svalbard did become involved in World War II, 1940 – 1945, which led to the evacuation of inhabitants. On 3 September 1941 evacuation was carried out but prior to this, Svalbard radio and Isfjord radio were destroyed so that they could not fall into enemy hands.

In September 1943 Barentsburg and Longyearbyen were bombarded and set on fire by a German fleet consisting of two battleships and nine torpedo boats. When the fleet drew back through Isfjorden, they fired a few volleys at Isfjord radio in spite of the fact that the station had long been destroyed.

The Germans' interest in Svalbard was primarily to get a foothold there in order to be better equipped to cut off allied sea traffic to and from Arkhangelsk. In addition they also wanted to get their hands on the coal reserves on Svalbard. Industries in Germany, particularly the weapons industry, were running at full capacity and coal was a sought-after source of energy.

An efficient local home guard on Svalbard helped to prevent the Germans from succeeding. Several people lost their lives on Svalbard as a consequence of the fighting. There is a monument not far from the telegraph building in Longyearbyen commemorating this.

Back after World War II

When the war ended in May 1945, work started to rebuild Svalbard. No one was opposed to resuming coal mining activities, nor to rebuilding and improving the infrastructure, including telecommunications. Svalbard radio started up again in Longyearbyen and was in operation from the autumn of 1945. Isfjord radio was rebuilt on Cape Linné and was operative from the autumn of 1946. The station building at Isfjord radio was built identical to the original building.

Development – slow but sure

Radiotelegraphy was still the ruling communications solution, but in March 1949 Svalbard received its first telephony connection to the mainland. A connection was established between Longyearbyen and Harstad radio. Thus a small step had been taken in a new era in telecommunications history on Svalbard.

In addition to the tasks of operating the station and attending to the radio station, the crew at Isfjord radio were assigned the task of making meteorological observations for the Norwegian Meteorological Institute.

In 1954 a new telecommunications building with a technical and administrative wing and a wing for employees was built and put into use in Longyearbyen. Svalbard radio, which had been co-located with the District Governor, now moved into its own premises in the new telecommunications building. In 1957 a new station building was erected at Isfjord radio. The old building was kept and is still standing in the same spot next to the "new station".

Radio and TV conditions

Reception and broadcasting conditions on Svalbard were rather poor. After the war an antennae several kilometres long was built, running from Longyearbyen up to Platåfjellet, which improved conditions somewhat. Telenor had the technical responsibility for radio and TV transmissions to Svalbard and in 1969 a local 'studio' was set up in the telecommunications building in Longyearbyen. From this studio, TV programmes were relayed to receivers in Longyearbyen. The unusual thing about this was that the programmes were recorded on the mainland and the tapes were sent to Longyearbyen by plane, and then they were aired two weeks after being shown on the mainland. The same programme was sent twice per day, first in the morning and then again in the evening, so that shift workers in the mines would be

ISSN 0085-7130

able to see the programmes. This remained so until 1980 when programmes were sent with "only" one week's delay. It was only several years after satellite connections were established between Svalbard and the mainland that viewers in Longyearbyen could see live programmes. This took place on 22 December 1984 – a nice Christmas present for the inhabitants of Longyearbyen, in other words.

New possibilities with satellite

In 1974 work began to introduce satellite communication with Svalbard. Telenor's research and development department initiated testing, while there were many sceptics as to this being at all possible. In theory this was on the borderline of what is possible because Svalbard is so far north that elevation would be too low.

First of all it was important to find a suitable place for receiving signals from the satellite. Of all the places that were investigated, it turned out that Isfjord radio, where Telenor already was established, was the most suitable place. Isfjord radio lies on the mouth of the Isfjorden and has an open horizon to the south. A higher elevation of the site would have been desirable, but nonetheless this was the place chosen for testing. Another advantage with Isfjord radio was that the place already had the infrastructure required, such as buildings and electrical power. Testing took place over the course of several years, but in 1977, based on the results so far achieved, it was decided to build a satellite station at Isfjord radio. Construction started, and to make a long story short, satellite connection with the mainland was put into operation on 12 December 1978. This revolutionised telecommunications on Svalbard.

The satellite antennae at Isfjord radio has a diameter of 13 metres. Notice that the antennae is nearly pointing parallel with the terrain. This is because the elevation angle is only approximately 3 degrees

Developments speed up

From this point on many possibilities lay open and the first suggestions for automating telecommunications were launched. Only a little over two years were to pass before this became a reality. The fact that it nonetheless took as long as two years was due to the local subscription network in Longyearbyen being owned by Store Norske Spitsbergen Kulkompani and had to be taken over by Telenor in order for the automation to take place. In addition the subscription network was in such poor condition that it needed a considerable upgrade. It was also necessary to expand the technical part of the telecommunications building in Longyearbyen.

After the formalities concerning the takeover of the subscription network from Store Norske Spitsbergen Kulkompani were arranged, the subscription network had been improved and the technical part of the telecommunications building in Longyearbyen was extended, the foundation was in place for automating telephones on Svalbard. At the same time radio connections were established between Longyearbyen and Ny-Ålesund via a radio link station at Kongsvegpasset. In this way the telephones in Ny-Ålesund could be automated at the same time as in Longyearbyen.

On 20 May 1981 Svalbard was connected to the international long-distance network. At around this time Svalbard became a separate telecommunications area, the 28th such area. By establishing the telecommunications area, Svalbard had its own telecommunications director with a separate administration in keeping with other telecommunications areas on the mainland.

So a big step had been taken into the world of modern telecommunications, but many unsolved tasks remained. Internal communications on Svalbard still had some defects. Sveagruva (the Svea mine), where Store Norske Spitsbergen Kulkompani had considerable activities, still did not have acceptable telephone connections. The connection to the mine was a result of an analogue radio link being built between Longyearbyen and Svea, via Isfjord radio and a small station on Cape Martin. This connection opened on 28 August 1984.

In August 1998 the radio link station on Cape Martin burned to the ground and a new connection to the Svea mine had to be established. This time a radio link route was chosen from Longyearbyen via a radio link station on top of the Skolten mountain (appr. 1000 metres above sea level), to a repeater on the Deinbolt mountain, and then down to Sveagruva.

With satellite connections to the mainland it was also possible to transmit NRK's live TV-broadcasts to Svalbard. As previously mentioned this took place in December 1984. Not everyone was happy to receive live television broadcasts and this became a topic of discussion. Those who wanted to keep the old set-up complained that they would lose the opportunity to broadcast repeats, which would be negative for shift workers.

Sveagruva received live broadcast television in 1986, and Ny-Ålesund in 1987. The Norwegian settlements of Longyearbyen, Sveagruva and Ny-Ålesund now had satisfactory telecommunications. Stereo radio broadcasts came in 1988, and in 1990 a videoconference studio was set up in Longyearbyen. The pager service was implemented in 1991 and in the same year attempts were made for live broadcasts from mobile cameras via satellite. These broadcasts were sent between Longyearbyen and the North Cape Plateau.

Photo: Per Krokan

Skolten radio link station on the Longyearbyen to Sveagruva connection lies approximately 1000 meters above sea level and is normally only accessible by helicopter

In 1989 there were still two Russian settlements that only had radio connections to the rest of the world. They were Barentsburg and Pyamiden, and although their total population exceeded that of the Norwegian settlements on Svalbard, they lacked the communication technology that Norwegian settlers had. There was great concern for these places to be connected to the international telecommunications network on the same level as the Norwegian settlements. After long negotiations with the Russians this idea was realised and both locations were connected to the Norwegian telecommunications network on Svalbard during the summer of 1989.

On the mainland digitalisation was well underway and naturally the question of digitalisation on Svalbard was discussed. According to Telenor's digitalisation schedule, Svalbard would be the last area of the country to be digitalised. But as it turned out this work was brought forward considerably. The situation of the Svalbard telecommunications area at that time was that professional operations were the responsibility of Telenor and thereby the (Norwegian) Ministry of Transport and Communication, but financially the Svalbard telecommunications area was the responsibility of the (Norwegian) Ministry of Justice and the Police. In order to get the necessary funding for digitalisation, this would have to be allocated via the Svalbard budget of the Ministry of Justice and Police. Previous experience had shown, however, that getting the necessary investment capital from that authority was rather difficult. Efforts were therefore made to remove the telecommunications activities on Svalbard from the Svalbard budget and over to Telenor's regular budget. After a lot of work and case handling in Telenor and in various ministries, it was resolved that Telenor Svalbard would be removed from the Svalbard budget with effect from 1 January 1991.

This decision became known well ahead of this date, and therefore preparations were made for the construction of a digital switchboard in Longyearbyen. This was erected and made operational on 19 October 1990. The analogue KV exchange was not yet nine years old when it was replaced.

It is perhaps a curiosity that Svalbard was the first fully digitalised telecommunications area in Norway.

The telecommunications area's humble content and size was of course the reason for this, making it difficult to compare it with other telecommunications areas.

The digitalisation provided opportunities for new services, among them ISDN, a popular and sought-after solution. An upgrading and expansion of the exchange in April 1999 provided a long-awaited possibility to satisfy a greater demand for ISDN. With the digitalisation, Ny-Ålesund also acquired the same possibilities as Longyearbyen. Until 1996 the local telecommunications network in Ny-Ålesund was owned by Kings Bay AS until Telenor Svalbard took over in 1996.

In connection with the 400-year anniversary of the discovery of Svalbard, Telenor Svalbard issued their own phonecards. They were released on 17 June 1996, and in the same year the mobile telephone was introduced. A GSM base station covering Longyearbyen and its environs was put into operation and later expanded by setting up base stations on Skolten and in Sveagruva. This year Barentsburg also received mobile phone coverage. Internet became accessible on Svalbard when an Internet node was put into operation in 1997.

Organisation of the operations on Svalbard

In the course of the 1990s Telenor underwent several restructuring processes, including the cessation of the geographical organisation structure with telecommunications areas and regions. To start with Telenor became more division-oriented and it later became a limited company, as is well known.

The dissolution of the telecommunications areas also had an effect on operations on Svalbard. With a modest staff of less than 20 employees, the operations were now split up into several areas (network, private, installation, administration, etc.), and each of these was managed by its respective overarching unit on the mainland. This form of organisation was not suitable, as it did not allow for local coordination of the operation. After a few years experience, it was decided to consolidate the operations on Svalbard and to give the coordinating responsibility to the on-site manager. In 1996 it was decided that Telenor's operations on Svalbard would be united in one private company, a subsidiary of the Telenor group. Telenor Svalbard AS became operational on 1 January 1997. The company had its local management on Svalbard and since then has had its headquarters in Longyearbyen.

The company today has 10-12 employees and manages to a great extent all of the tasks which the operation is assigned. In some cases expertise is hired in from other Telenor units. Svalbard radio, with its crew of six employees, does not belong to Telenor Svalbard AS, but is connected to Telenor Networks maritime department.

Photo: Herta Grøndal

The present-day main building at Isfjord radio was erected in 1957. Before the radio office was moved to Longyearbyen, 12–15 people worked and lived here. Now the station is no longer staffed but is remotely monitored from Longyearbyen

Development in the satellite area requires new solutions

As far back as the late 1980s several players showed an interest in Svalbard's unique location in terms of satellites in polar orbits. In simple terms, the further north you are, the more orbits you can register from a polar orbiting satellite. So many establishments were looking for a location in the far north where one could read and download data from such satellites. Svalbard, with its extensive infrastructure and easy access to an airport, was clearly a natural and well-suited place. The only question was where on Svalbard. Several places were considered: Ny-Ålesund, Isfjord radio and Longyearbyen. When the decision was finally made, Platåberget at Longyearbyen was chosen as its location and the establishment is named SvalSat. Here data are downloaded from several satellites in polar orbits. Because the downloaded data were going to be conveyed further, this development also had significant consequences for Telenor Svalbard. It became necessary to increase capacity on the radio link network between Longyearbyen and Isfjord radio for redirecting the data via satellite to the desired destination in the world. In addition a connection had to be established between SvalSat and Longyearbyen by use of fibre optics, which were later supplemented with a radio link. The transmission capacity of the satellite connection from Isfjord radio had to be expanded considerably. The number of SvalSat users increased, which required more and more capacity. The users felt that leasing of satellite capacity was so expensive that it was necessary to consider the possibility of other communications solutions from Svalbard.

For this reason the idea of a fibre optic cable between Svalbard and the mainland was launched in 2002. Telenor was invited to participate in financing such a project, but in the end they refused, which led to Norsk Romsenter (Norwegian Space Centre) realising the project in 2003, in cooperation with the users of SvalSat. The fibre optic cable, i.e. two cables in their respective routes, was officially put into use in January 2004. Each of the cables is 1300 kilometres long, and Telenor Svalbard has now transferred its own connections from satellite to this fibre optic cable.

Even though Telenor declined to becoming owner or partial owner of the fibre optic connection, both Telenor and Telenor Svalbard were heavily involved in the construction phase. Through this solution Svalbard has capacity and high quality telecommunications connections superior to most places on the mainland.

The development of telecommunications on Svalbard, from its inception with radiotelegraphy to fibre optic technology being implemented on the mainland took over 92 years. It would however be fair to say that the changes in the past 20 years have transformed Svalbard from a "developing country" in terms of communications to being a state-of-the-art teletechnological society.

Photo: Herta Grøndal

Telenor Svalbard's administration building in Longyearbyen

Changes as a consequence of new technology

For the Telenor employees on Svalbard the work situation and work tasks have changed in keeping with the technological developments. For the employees responsible for the technical side of operations, these developments have presented new challenges demanding increased expertise. The staffing level has been under continuous assessment and has always been kept at a minimum level. Isfjord radio, which had a staff of 12-15 persons before the radio office was moved to Longyearbyen, had a staff of 4-5 persons for many years. In 1999 all staff were removed, and it is now totally remotely operated/monitored from Longyearbyen. After the implementation of the fibre optic cable to the mainland, Isfjord radio is no longer as important as it once was.

Flexibility and enthusiasm

In many ways a lot of the work on Svalbard can be considered as pioneering work. Expertise and experience in the construction of telecommunications in arctic areas have been difficult to come by. On-site learning by doing has been necessary. In total, an impressive piece of work has been done in this area. When as a rule results have been successful, this is not least because of the enthusiasm and tenacity that employees have always shown. No task has been too big or too small to be realised. With the minimum levels of staff which the operation has always had, the flexibility of the employees regarding execution

of tasks in progress has been especially important. As a matter of course everyone would make an effort and cooperate for tasks to be solved. It was never a question of what one knew or was able to, but of what one had to do. This way the job of the employees has been interesting and varied. Without this flexibility and enthusiasm exhibited by employees, there is good reason to believe that the evolution to present-day telecommunications on Svalbard would have taken considerably longer.

Viggo Bjarne Kristiansen (63) started his career in Televerket (Telenor) in 1958 as a telecommunications installer. Through internally sponsored education he later advanced to work on telephone automation. He has held various management positions in Telenor, and also in the Norwegian labour union of telecom engineers (TMLF/DNTO). He attended the Norwegian school of Administration in 1984 and the Norwegian Defence Academy in 1990–91. In 1994 he became Department manager on Svalbard and advanced to Managing director for Telenor Svalbard from 1997. He retired from this position in 2004.

email: vbkristi@online.no

INMARSAT – a success story!

How it was established. Later developments. The role of Telenor – former NTA[1)]

OLE JOHAN HAGA

Ole Johan Haga retired in 2000 after working in various positions in Telenor for 40 years

It is well known that the establishment of the International Maritime Satellite Organisation – INMARSAT– has been a great success. Norway, and notably NTA, played a substantial, some believe a determining, role in its creation. Today, the INMARSAT system supports links for telephone and data communications to more than 287,000 ship, vehicle, aircraft and other mobile users. For the customers, notably the ship owners, it has allowed the introduction of modern management of their fleets. For Telenor the provision of mobile satellite services has been, is, and should continue to be, an outstanding commercial success. For the world, INMARSAT has been a glorious example of peaceful cooperation between all parts of the world, in a grand project of great practical value for communication and commerce, already at the time of the "cold war". The main emphasis of this paper is on the establishment of INMARSAT. Part A describes the activities which led to the formation of INMARSAT in July 1979. The contentious aspects and their resolution are described in some detail. Later developments are briefly described in part B. Part C covers the participation of Telenor in the organization, as well as Telenor's utilization of the space segment.

Part A The establishment

1 What INMARSAT is

INMARSAT operates a constellation of geostationary satellites which extend mobile telephone, fax and data communications to every part of the world, except the polar regions.

INMARSAT was conceived and established as an international and fully global organization. The purpose of INMARSAT was to make provision for the space segment necessary for improved maritime communications and, in particular, for improved safety of life at sea communications and the Global Maritime Distress and Safety System (GMDSS). This purpose was later extended through amendments to the constitution instruments to provide space segments also for land mobile and aeronautical communications, and the name of the organization was changed to the International Mobile Satellite Organization (IMSO) to reflect the extended purpose.

INMARSAT owns (or leases), and operates, communication satellites – the space segment – for mobile communications. It sells capacity on the satellites to earth station operators, later termed distribution partners, who use the capacity to produce and sell telecommunication services to end users[2)]. End users are mobile stations on ship, aircraft, and mobile users on land, and "fixed" land users. In addition to operating the space segment, INMARSAT specifies the communication standards required for using the space segment, and controls and approves all stations which seek access to the system.

The establishment of INMARSAT was based on two international public law instruments developed under the auspices of the International Maritime Organization (IMO). These are:

a Convention on the International Maritime Satellite Organization (INMARSAT) between State Parties to the Convention; and

b Operating Agreement between telecommunications entities public or private (one per Party) called "Signatories" designated by a State.

INMARSAT was structured with three principal organs:

a The Assembly of Parties (one State, one vote), which dealt with general policy matters and the long term objectives of the Organization;

b The Council, composed of 22 Signatories, or groups of Signatories. It decided on all financial, operational, technical and administrative matters, and made provision for the space segment for carrying out the purposes of the Organization. Signatories' voting rights were linked to their utilization of the system via investment shares;

1) *NTA = The Norwegian Telecommunications Administration.*

2) *The term "end user" is used to designate the (ultimate) user – the customer – of the service, to distinguish him from the user of the space segment, who is the earth station operator.*

 ISSN 0085-7130

Figure 1 INMARSAT Headquarters in London. It accommodates the administration, as well as the operation centre

c The Directorate, which was the executive body of the Organization headed by a Director General, who was the Chief Executive Officer and legal representative of the Organization.

Investment in the space segment was originally financed by the owners of the organization. The owners were signatories to the Operating Agreement. INMARSAT pays compensation for use of the invested capital, and repays the capital contributions when revenues allow repayment.

INMARSAT receives revenues from the sale of capacity on the space segment. The revenues shall cover all administrative and operational costs of operating the space segment, as well as financial costs related to the investment in the space segment.

It is important to underline that the provision of share capital to INMARSAT is a financial undertaking, quite separate from the business of producing and selling communication services to end users.

The economy of the organization has been sound. Utilization of the capacity has been good and, due to lack of competition, it was for a long period possible to set the tariffs at such a level that the owners received a comfortable compensation for invested capital. In fact, it is specified in the (original) Operating Agreement that compensation for capital provided by signatories should be 17 per cent per annum.

INMARSAT has not been unaffected by the changes which have taken place in the telecommunications industry, namely liberalization and fierce competition. To meet the new situation and new challenges, INMARSAT was changed into a private corporate structure in 1999.

2 National initiatives

For thousands of years ships plying the high seas sailed without any communication with land, and their business was adapted to that situation. About a hundred years ago radio entered the maritime scene. Radio communications on short and medium wavelengths (HF and MF) radically changed the sailing conditions. A comprehensive global safety at sea system was gradually developed, which over the years has saved innumerable lives and large material values; ship owners could have contacts with their ships; and seamen could talk to their families while at sea.

Even if HF radio was of tremendous importance to shipping, it suffered from certain defects, notably poor regularity in many ocean areas. With the advent of satellite technology and telecommunications via satellite in the sixties, it was quite natural that the idea of utilizing satellites for communication also with ships at sea was considered. The problems, however, seemed formidable. In order to enable antennas of reasonable size on board ships, the output power from the satellite towards the ship has to be high, in particular in the case of geostationary satellites. This

implies large and expensive satellites. Economy then dictates extensive, preferably global, cooperation to establish the space segment. Intricate systems for accessing the satellite on a shared basis had to be developed. Development of steerable antennas for the ship required considerable R&D effort. On the basis of early experiments Comsat of USA concluded that maritime satellite communications "could be implemented with available technology, but the economics of the project did not justify a commercial venture".

With such challenges research institutions in Norway were skeptical. Reflecting the general attitude, Finn Lied, in the capacity of chairman of the physics committee of NTNF[3)], was skeptical. Bjørn Rørholt, at the time chairman of NTNF's committee for space activity, advised strongly against committing funds and efforts in this area. "I think we will do Norwegian-owned industry a disservice by leading them into such a demanding development as a ship terminal at this stage", he warned Finn Lied. Even the potential users, the ship owners, did not show any interest. NTA took quite an opposite view. Early 1968 NTA (responsible Per Mortensen) provided some funding for ELAB[4)] for experiments with ship equipment working in 1.5 GHz frequency band. In 1969 NTA's research institute TF initiated a study on maritime satellite communication, where A/S NERA contributed greatly. TF with Nic Knudtzon as its head, was a driving force in these efforts. A tri-party constellation of NTA, EB/NERA, and ELAB largely dominated further development. The result of these early studies enabled Norway to table concrete and well founded proposals to the ITU WARC (Space) in 1971, which allocated frequency bands to maritime satellite communications.

The 1971 WARC was exclusively devoted to space communication, and several new problems had to be resolved. Both radio frequencies and the geostationary orbit are limited natural resources. To mention one interesting and important issue, the Conference decided on an equal right for all countries to use these resources. Some equatorial countries had claimed sovereignty of its part – its arc – of the equatorial orbit.

The Norwegian contribution at WARC 71 was substantial – in my opinion material – to the satisfactory result, with regard to maritime satellite frequencies. There was strong opposition to allocating frequencies to such a service, even from the United States. Arne Bøe of the Norwegian delegation, should in particular be mentioned as being very competent and effective in the tough negotiations.

It is in my opinion reasonable to conclude that Norway with NTA took a leading role in these early stages of developing maritime satellite communications.

3 Establishment of a new inter-governmental organization – The basic problems

Norway with NTA was convinced that a global satellite system should be developed. In 1972 IMO set up a Panel of Experts on Maritime Satellites to study the technical, operational, administrative, and institutional aspects of a maritime satellite system.

The creation of a new inter-governmental organization for the provision of the space segment was first proposed by the Soviet Union, who tabled a draft convention to this effect between the first and the second meeting of the Panel of Experts. At Norway's proposal the Panel decided on its second meeting in April-May 1973 to start immediately to work out a Convention based on the Soviet proposal. A complete draft was developed in the course of the Panel's work.

The Panel of Experts also considered other alternatives for organizational arrangements for a maritime satellite system, viz. to create a consortium or to make use of an existing organization such as IMO or INTELSAT. For various reasons these other solutions attracted little support. A consortium was rejected because almost all delegations felt that policy control over the global maritime satellite system should be exercised by an inter-governmental organization. Telecommunication administrations strongly opposed the idea of IMO running a communication system – as IMO's primary concern was and should remain that of safety at sea, so IMO was rejected. INTELSAT was acceptable to the USA and some other countries. Other important maritime countries, however, with little or no interest in INTELSAT's fixed traffic, strongly believed that they would get too little control over the maritime service facilities if INTELSAT were chosen. The absence of some important maritime countries in INTELSAT – notably the USSR – was also regarded as a serious disadvantage. The USA, on the other hand, was critical to the entire idea of creating a new inter-governmental organization. However, at the first session of the INMARSAT Conference the United States had decided to support the creation of INMARSAT, but stated that it was

3) *The Royal Norwegian Council for Scientific and Industrial Research.*

4) *ELAB: Research Institute, at the time organized and located as an extension of the Norwegian Institute for Technology (NTH).*

ISSN 0085-7130

impossible for the US Government to participate in a commercial organization and that the US Government could not undertake economic responsibilities or even give any guarantees of an economic nature. US participation would therefore have to be through a private commercial company. The simplest arrangement would be to follow the INTELSAT pattern, i.e. one agreement between governments on basic principles and organizational questions and a second agreement between the private or public telecommunication entities which were to finance and operate the system. The US view was supported by some countries, but the majority rejected the complicated arrangement proposed by the US. These delegations argued that since telecommunication services were operated by public corporations in most countries, an organization along the lines of the US proposal with all its inherent complications and delays would be unreasonable.

4 The package deal

Due to the difference of opinion on this and some other basic questions the first session of the Conference was not able to make much progress. However, in the last days – or rather nights – of that session, informal talks were held between the delegations of the USA, Japan, the Soviet Union, the United Kingdom, Norway, the Federal Republic of Germany, France and the German Democratic Republic. The informal negotiations led to the result that the Western Europeans accepted the fundamental US demand that the member States could transfer all financial and operational functions, responsibilities and liabilities in the organization to so-called "designated entities". The Western European delegations also accepted that this principle be put into effect by splitting the draft Convention into two agreements, one between governments ("Parties") and one between "designated entities" ("Signatories"). The acceptance of this concept was, however, conditioned upon the text of the Panel of Experts being used as a starting point. Furthermore, the USA and the Western European delegations agreed on a compromise solution for distribution of functions between the Assembly and the Council, and also on the procurement policy. These compromises in the first session of the INMARSAT Conference were referred to as "the package deal".

5 The distribution of functions between the Assembly and the Council

There was agreement that the Organization should have three organs: An Assembly where all member governments (Parties) would be represented, having one vote each; a Council, where the Signatories having large investment shares would be represented and vote in proportion to their shares; and a Directorate. The distribution of functions between the organs as agreed in the "package" mainly followed the US point of view that the main power of the Organization should be vested in the Council and not in the Assembly.

The large investors in the European camp did not have much difficulty in accepting the US view in this matter – in fact they preferred the US solution – since they would have a strong influence in the Council, where voting is weighted by the investment shares they hold, and would have relatively much less influence in the Assembly, where all members are represented, each having one (un-weighted) vote. Conversely, the small investors would for the same reasons, which for them have the inverse effect, want the Assembly to have at least some real power. The smallest investors would not even have a seat in the Council. The original compromise on the division of power caused considerable difficulty at a later stage for small investors, in particular those from the Third World. The division of power agreed in the "package" was not changed, but small investors were accommodated to a certain extent by the adoption of a large Council (22 Representatives) – against the wish of the large investors – and by specifying that four of the 22 representatives be elected to ensure just geographical representation, with due regard to the interests of the developing countries. A Signatory elected to represent a geographical area would be obliged to represent each Signatory in that area which is not among the 18 countries represented on the Council by virtue of their investment shares. This will give also developing countries with small investment shares some voice in INMARSAT's decision making.

Another and related problem was the voting procedure in the Council. The positions were that most Western European delegations argued that INMARSAT was a commercial undertaking which had to make decisions in an efficient and timely manner. They proposed that the minimum requirement for a positive decision should be one third of the representatives representing a majority of the total investment shares. Other delegations, including many small investors and the US, argued that it was essential that the important decisions of INMARSAT had broad support. The small investors were clearly afraid that a few large investors could force through decisions against a majority. The USA and probably many other countries were seriously concerned about the strong position Western Europe would get in the Council. If decisions could be made by only a majority, Western Europe might be able to force through e.g. procurement decisions with little or no extra support. These delegations therefore proposed that the requirement for decisions should be the majority of the representatives representing two-thirds of the

investment shares. None of the alternatives got sufficient support to be adopted by the Conference – until the United Kingdom suddenly changed its position and voted for the two-thirds alternative – to great disappointment to fellow Europeans.

6 The procurement policy

The negotiations on procurement policy repeated some of the confrontations and arguments from the negotiations of the INTELSAT definitive arrangements. The INMARSAT negotiations were, however, less lengthy and less complex – probably because the ground had been covered before. The situation was essentially as follows:

The USA held the view that the sole consideration in awarding procurement contracts should be best quality, price and delivery time, since only uninhibited competition would provide the cheapest and best service to the users. The Europeans, supported by most other countries, accepted that best quality, price and delivery time should be the dominant consideration, but wanted the idea of developing and maintaining world-wide competition to be included as an additional consideration.

The main European motive was clearly a desire to break the de facto American monopoly in space technology. They further argued that in the long run a broad base of suppliers would increase competition, thus reducing costs to INMARSAT and maritime users.

The "package deal" text states that the procurement policy shall be such as to encourage worldwide competition and to this end award contracts, based on responses to open international tender, to bidders offering the best combination of quality, price and delivery time. The idea of competition is given prominence in the Convention text and obliges the Organization to observe this aspect actively through all the stages of the procurement procedure. The competition aspect is more 'active' in the INMARSAT text than in the corresponding INTELSAT provision.

The Soviet Union participated in the informal negotiations and accepted two of the compromises of the "package deal", viz. the distribution of power between the Assembly and the Council and the procurement policy, but could not accept that member States could transfer all the economic responsibilities and liabilities to private entities nor that the draft Convention be split into two agreements. The first session of the INMARSAT Conference therefore ended without a result. The Conference appointed an Intersessional Working Group, which was charged with the task of trying to resolve the fundamental problems, and thereafter develop complete draft agreements based on the draft of the Panel of Experts and on any contributions submitted later. The Intersessional Work-

Photo: O.J. Haga

Figure 2 A tricky problem is being resolved through intense informal discussions. Mr. Freeman of US State Department surrounded by Professor Seyersted, Norway, upper right hand corner; Dr. Kolodkin, USSR, left, Secretary General of IMO, wearing glasses; Mr. Kolossov, USSR, upper middle with red tie

ing Group met three times in the summer and autumn of 1975. Before the first meeting of the Working Group, the delegation of the Federal Republic of Germany undertook to split the draft of the Panel of Experts into two agreements. A group of Western European – including Norwegian – delegates also held informal talks with the Soviet delegation in Moscow and the US delegation in Washington during the summer of 1975.

With some issues still unresolved a draft Convention and a draft Operating Agreement were submitted to the second session of the INMARSAT Conference.

7 The limitation of voting power in the Council

The important outstanding point at the end of the second session of the Conference was the upper limit of the voting power of a single Signatory in the Council.

The USA wanted no such limitation – in their view each Signatory should be entitled to vote its full investment share. The majority, however, wanted to prevent any Signatory from falling anywhere near to a veto power in the Council. After intense negotiations the Conference came very close to an acceptable compromise. The US accepted the principle of limitation and could – with considerable difficulty – accept a 25 per cent ceiling, a figure Western Europe could accept without much difficulty. The instructions did not, however, permit the Soviet delegation to go above 20 per cent, and since the Conference focused on this problem at a late stage in the session, it was impossible in the short time available for the Soviet delegation to get new instructions. A third session of the Conference therefore had to be called to solve this problem. A few other minor points were also left for the third session.

After a certain amount of informal contact between the most interested delegations in the interim period, agreement was reached on a 25 per cent voting limit with some qualifications, and the Conference adopted unanimously the Convention and the Operating Agreement at a short third session in September 1976.

In the foregoing I have dealt with the basic problems. In such a formidable undertaking – to arrange for a commercial venture in which all parts of the world were to participate – there were obviously also other areas of dispute. The most difficult additional problem was that of agreeing on investment shares – both initial and long term.

8 Long term investment shares

The Conference accepted without difficulty the general principle, that the investment share of a Signatory shall be based on that Signatory's percentage utilization of the space segment. However, the application of this principle to the maritime situation creates problems which do not arise in the case of fixed satellite service. Maritime satellite traffic must be handled on the basis of demand assignment of a channel for the duration of a call. How is then the utilization of the space segment by a Signatory to be measured?

The USA, supported by a few delegations, proposed that calls originating and terminating in the land territory of a Signatory should be attributed to that Signatory for the purpose of determining the investment share. The majority proposed that the origin of a call should determine its attribution in all cases. Although logical and in accordance with fixed service practice, the majority proposal would give the USA in particular an unreasonably low share, and a fairly complicated compromise formula was finally adopted. Utilization in both directions is divided into two equal parts, a ship part and a land part. The part associated with the ship where the traffic originates or terminates is attributed to the Signatory of the Party under whose authority the ship is operating. The part associated with the land territory where the traffic originates or terminates is attributed to the Signatory of the Party in whose territory the traffic originates and terminates. Signatories of flag of convenience countries were granted exception from the formula: their share may upon application to the Council be reduced to twice the land part, but not below 0.1 per cent. Flags of convenience in this context are countries where the ratio of the ship part to the land part exceeds 20 : 1. The considerations behind this exception were that flag of convenience countries could not afford to become members under the adopted general formula and that the majority of the delegations preferred to have these important maritime countries in as members even if they could not take their fair share of the investment burden.

9 Initial investment shares

Initial investment shares had to be determined through negotiations at the Conference, since Signatories' utilization of the space segment could not be known in advance. It was agreed that initial shares should reflect expected usage. HF radio traffic, number of ships and tonnage of the merchant fleet of the Signatories were the relevant factors on which the assessment was to be made. The delegations made bids based on their own estimates. It proved to be extremely difficult to reach a total of 100 per cent of initial shares. The main reasons were that some large maritime countries – notably flag of convenience

countries – were unable to accept what was thought to be natural shares and that others were unwilling to take up the resulting shortfall, probably because it was expected that INMARSAT would suffer economic losses for a fairly long period. There was also some disagreement as to what weight the various factors – HF traffic, number of ships and tonnage – should be given when assessing expected satellite traffic. It took long hard hours of negotiation in Plenary sessions, working groups and informal meetings, and considerable good will, before the 100 per cent target finally was reached in the Plenary in the last but one night of the second session of the Conference.

It is ironic to note that in the months immediately prior to entry into force of the agreements, a decision by the US to increase its share from 17 to 30 per cent caused a struggle to avoid a consequent reduction of the initial shares of other Signatories. An increase is permitted under a provision which was included in order to secure entry into force despite the high threshold figure of 95 per cent. The substantial unilateral increase by the US was feared by many to upset the relative balance of power in the Council to an unacceptable degree. The only way to restore the balance would be for all or most of the other Signatories to increase their shares by a similar factor as that of the US. The results of such actions would be a substantial oversubscription of shares at entry into force. Upon proportional reduction of all shares at entry into force, to get a total of 100 per cent, the original balance would be restored. The struggle culminated in a race in the evening of July 15, 1976. Each increase by the US was met with corresponding increases by the others. It was the US against the rest of the world. This went on until minutes before entry into force of the agreements at midnight. The latest increase by US was unknown (it was later known to be 65 percent), so it was met with a joint declaration by the others, that we increase by an amount proportional to the latest increase of the US. Thus the balance was kept, but the US protested against the method, so the formal existence of INMARSAT started with a conflict between the US and the rest of the world. This was of course very unfortunate but, on a positive note, the incident also illustrates the fact that the economic prospects for INMARSAT were considered then to be much more promising than they were three years earlier. Table 1 shows the resulting initial investment shares (in rounded figures).

Despite the conflict, the Council commenced work immediately, with its first meeting starting the following day, on 16 July 1979. This first meeting was chaired by the author of this paper.

Luckily, the conflict was resolved at the second meeting of the Council.

10 Other problems

- Legal questions such as liability, arbitration and privileges and immunities were difficult to resolve and required lengthy negotiations. Some new ground was covered which could be of value also in other contexts.

- Formulation of financial principles created some difficulty. Some delegations wanted to stress that the only basis for INMARSAT's operations should be "accepted commercial principles". Others disagreed with this being the only basis. The resulting compromise was that INMARSAT shall operate on a sound economic and financial basis "having regard to" accepted commercial principles.

Country	Per cent	Country	Per cent	Country	Per cent
USA	23.5	Canada	2.6	P.R. of China	1.2
USSR	14.2	Kuwait	2.0	Belgium	0.6
UK	9.9	Spain	2.0	Finland	0.6
Norway	7.9	Sweden	1.9	Argentina	0.6
Japan	7.0	Denmark	1.7	New Zealand	0.25
Italy	3.4	Australia	1.7	Bulgaria	0.1
France	2.9	India	1.7	Portugal	0.1
Western Germany	2.9	Brazil	1.7	Algeria	0.05
Greece	2.9	Poland	1.7	Egypt	0.05
The Netherlands	2.9	Singapore	1.7		

Table 1 Initial investment shares

ISSN 0085-7130

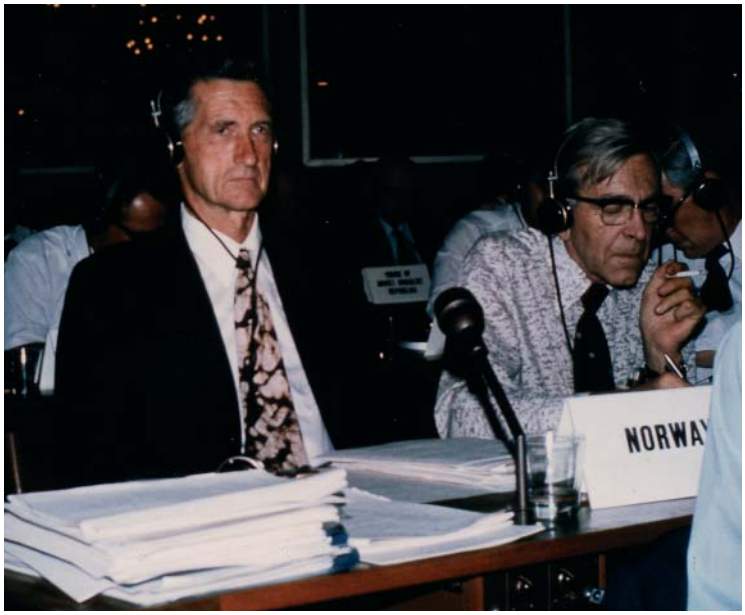


Figure 3 Professor Finn Seyersted of the Norwegian delegation – in typical attentive stance (Photo: O.J. Haga)

- The relation to other (competing) space segments which Parties might wish to establish also required a compromise text: INMARSAT shall express views in the form of a "recommendation of a non-binding nature" with respect to technical compatibility and possible economic harm.

- The language question brought about a collision between practical/economic and prestige/political considerations. Many official languages[5)] and especially many working languages would be very costly. A requirement to translate all documents into several languages slows down the work at conferences, and may even be used tactically as a means to delay proceedings. Norway tabled a very controversial document in which it was demonstrated that, on the basis of investment shares and number of countries, English, 'Scandinavian' and Russian would be the most important languages in INMARSAT. On this background, and for economic reasons, Norway proposed that INMARSAT should work only in the English language. It was politically impossible for the US and UK delegations to support this position, even if they agreed. The French and Spanish delegations were furious at the demonstrated statistics and the proposal. Since it was not possible to attract sufficient support for an article as proposed by Norway, the tactics was to block the adoption of any language article. This tactics succeeded, so the result was that no language article was adopted, leaving it to the Assembly and the Council to decide which languages to use. This matter was important only for the Council. With the very different voting power distribution in the Council, it was decided to work with only English documentation, but with interpretation between the four languages English, French, Spanish, and Russian in plenary sessions of the Council. Very few international organizations have managed to adopt such language arrangement, which is very timesaving and therefore economical.

Norway was very active and contributed very much to the work of the INMARSAT Conference. The core of the Norwegian delegation consisted of Ole Johan Haga of NTA, Professor Finn Seyersted, representing the Ministry for Foreign Affairs, and Kjell Skar, representing the Ministry of Communications. It is worth mentioning that Professor Seyersted was elected chairman of the drafting committee for the two agreements. Seyersted was a widely respected expert on international law, and it was largely due to his merits that the language of the agreements is more concise and easier to understand by laymen, than is usual in English and American law instruments. It is also quite exceptional that a delegate with non-English native tongue is chosen as chairman of a drafting committee for law instruments written in the English language. In addition to numerous concrete factual contributions during the work of the conference, Norway was able to play a mediating and bridging role between the more extreme positions held by various delegations, notably between the US and USSR. Serious dissensions were due to differences in business culture. Norwegian delegates spent long hours in hotel rooms with certain delegations in the evenings after the formal meetings, trying to explain why certain representatives held such and such views.

11 The Preparatory Committee

The Preparatory Committee and its Panels were active in the period following the last session of the INMARSAT Conference in 1976 and the formal coming into existence of INMARSAT in 1979. 22 countries participated in its work. The Committee provided an Interim Report to Governments and presented all its findings to INMARSAT. It studied technical specifications, operational matters, economic and marketing matters, and organizational matters. It reached conclusions and made recommendations in most of these areas, although some minor points of contention remained to be resolved. INMARSAT thus had a well prepared basis on which to start its work, once the ratification period was over.

5) *In ITU terminology: "Official language" implies interpretation to and from that language in conferences; "Working language" implies, in addition, that all documentation shall appear in that language.*

12 The INMARSAT space segment and the Joint Venture discussions

The potential members of INMARSAT had a general desire to ensure that there would be an operational maritime satellite system to follow on from the existing US-owned MARISAT system, which would reach the end of its design life in 1981. Coupled with this desire was the realization that, although the lifetime of some of the MARISAT satellites might extend beyond 1981, it would be necessary to prepare for the procurement of satellites for the follow-on system before INMARSAT itself could come into being. Since the Preparatory Committee was specifically precluded from committing INMARSAT, discussions were started between interested administrations with a view to establish a 'Joint Venture', which could conclude procurement contracts for follow-on satellites. It was intended to offer these contracts to INMARSAT, when formed, and then dissolve the Joint Venture. The Joint Venture forum – later the Joint Venture Conference – studied various space segment options and developed these through de facto negotiations with potential suppliers. The result of these efforts was that two suppliers – INTELSAT and the European Space Agency (ESA) – had work in hand to manufacture complementary parts of a composite satellite system consisting of maritime subsystems included in three INTELSAT V satellites and three dedicated satellites – MARECS – from ESA. Draft procurement contracts had also been prepared. Thus INMARSAT was in a position to take procurement action quickly based on a recommended space segment solution from the Joint Venture Conference. Since manufacture of these satellites was in progress and offers remained valid for INMARSAT, it was not necessary to institute the Joint Venture formally.

As a result of the Joint Venture activity INMARSAT was in a position to have its own space segment in operation from 1981. The Organization further had the benefit of the traffic base developed and the experience gained by the MARISAT system.

Figure 4 The third generation INMARSAT satellite. Essential new feature: seven spot beams (five active) in addition to global coverage. The spot beams gave extra capacity in areas where demand from users was high, and furthermore allowed the introduction of the mini-M standard mobile station – see Figure 6

13 Concluding remarks on Part A

It may appear from the preceding chapters that the creation of INMARSAT had been nothing but a constant struggle to resolve controversies and problems. It is true that it had taken much time and effort to find workable compromises where there were differences which, incidentally, is quite normal in international contexts. But this was fortunately not the whole story. Despite the differences, the areas of common interest and agreement dominated by far, and it is this fact, in addition to the good will to compromise which prevailed among all, which made it possible to find workable and acceptable solutions. Although it took considerable effort to solve the various problems, more efforts were devoted to analyze requirements and develop technical specifications and operational procedures, to carry out market studies and economic analyses, and to prepare institutional matters such as organizational structure, rules of procedure for the organs, financial regulations and many other detailed matters. INMARSAT, when finally established, was well prepared to become a fully working and thriving organization.

INMARSAT generation	I-1			I-2	I-3	I-4
	MARISAT	Intelsat V	MAREX			
Launch from year	1968	1980	1980	1991	1996	2005
Number of satellites	2	3	2	4 + 1	5	
Capacity in equivalent telephone channels	30	50	80	150	4 times that of I-2	10 times that of I-3

Table 2 Generations of INMARSAT space segment

ISSN 0085-7130

It is the first time in history that parties from so diverse parts of the world joined in a large commercial undertaking. In addition to providing general public maritime satellite communications in an efficient and economic way, the INMARSAT system provided a basis for developing an immensely advanced future distress system on a fully global basis. Furthermore, I strongly believe that the very constructive and positive cooperation we had seen in preparing the ground for INMARSAT, and the cooperation which developed within INMARSAT, contributed to bring East and West, North and South a little closer together.

The lengthy maritime satellite negotiations have required a large number of international meetings. NTA had its fair share in hosting meetings. One Preparatory Committee meeting, three Joint Venture meetings and several working group meetings were held in Norway.

Part B Later developments

14 Development of service standards. Extension to aeronautical and land mobile

The original service standard, called INMARSAT A, provided analogue telephony and telex services. The antenna of the ship station was mounted on a gyro-stabilized platform. The ship station was expensive and the market comprised mostly large vessels. Over the time the technology was refined to a large extent, greatly facilitated by the introduction of digital technology. Standard B, operative from 1993, is a digitized and improved version of Standard A; it has a wider spectrum of services and require half as much power from the satellite, thereby allowing lower tariffs. Standard C is the "lilliputian" station, which provides data transmission at bit rate 600 bit/s. It supports data transmission, telex and distress messages. Standard M is also based on digital technology. It provides telephony at a low bit rate, in addition to data communication. Standard M and C are suitable for use on land, being much smaller than Standard A and B. Telenor developed in partnership with UK's BT a mini-M lap-top terminal. Under the name Mobiq it was launched into service in 1997. The most recent development for the maritime market is the *Fleet* range of service, with the largest antenna solution offering voice and data services with a capacity of up to 64 kbit/s.

Figure 5 Ship earth station Nera F77, which belongs to the so-called Fleet *range of services; capacity 64 kbit/s for voice and data – global coverage. World market share 41 per cent*

As in all other areas of telecommunications, the Inmarsat[6)] industry has seen a continued shift from analogue to digital services, as well as a shift from voice services to data and internet driven services. Inmarsat's revenues from data services are now higher than for voice services, even in the maritime market.

With the development of new standards, satellite communication could be made available for use on land as well as on aircraft. INMARSAT therefore decided to extend its services to land and aeronautical applications and amended its constitutional instruments to this effect. Amendments for land mobile was adopted on 19 January 1989 and entered into force on 26 June 1997. Amendments for aeronautical applications were adopted on 16 October 1985 and entered into force on 13 October 1989.

The change of name to International Mobile Satellite Organization (IMSO) was adopted on 9 December 1994 and entered provisionally into force immediately.

15 Introduction of the "Global Maritime Distress and Safety System" (GMDSS)

Satellite communication readily lends itself to developing a vastly improved maritime distress and safety system. Consequently, IMO decided that the new GMDSS should be obligatory for all new oceangoing ships as from 1992, and for all ships from 1999. These decisions gave a strong boast to the development of the market for satellite communications.

6) *Lower case letters are used after privatization of the Organization.*

NTA and ELAB contributed significantly to the specification of the system.

16 Privatisation of INMARSAT

After twenty years of successful operation, Member States and Signatories of INMARSAT decided to challenge rapidly growing competition from private providers of satellite communications services and pioneered the first ever privatisation of all assets and business carried on by the intergovernmental organization while adhering to the continuous provision of the public service obligations and governmental oversight as a pre-requisite of the privatisation.

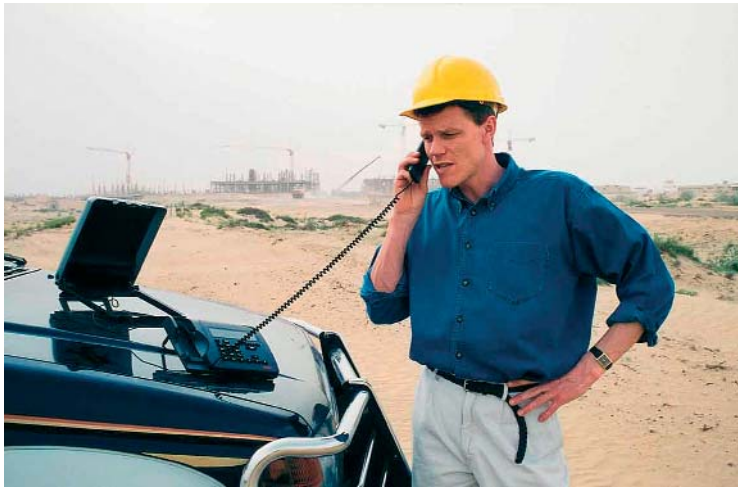


Figure 6 Application of INMARSAT M

At its Twelfth Session in April 1998, the INMARSAT Assembly adopted amendments to the INMARSAT Convention and Operating Agreement which were intended to transform the Organization's business into a privatised corporate structure, while retaining intergovernmental oversight of certain public service obligations and, in particular, the Global Maritime Distress and Safety System (GMDSS). These amendments were implemented as from 15 April 1999, pending their formal entry into force. In doing so, it was recognised that early implementation of the new structure was needed to maintain the commercial viability of the system in a rapidly changing satellite communications environment, and thereby ensure continuity of GMDSS services and other public service obligations; namely: peaceful uses of the system, non-discrimination, service to all geographical regions and fair competition.

The restructuring amendments entered into force on 31 July 2001 and became binding upon all Parties, including those which had not accepted them, and the Operating Agreement terminated on the same date.

The restructuring of INMARSAT involved the incorporation of holding and operating companies, located in England and registered under British law on 15 April 1999. On the same day, the Headquarters Agreement between the UK Government and the IMSO was signed. A Public Services Agreement between IMSO and the privatised Inmarsat was also executed with immediate effect. The Operating Agreement was terminated and the Signatories received ordinary shares in the privatised Inmarsat Ltd in exchange for their investment shares. Future capital requirements will be met from existing shareholders, strategic investors and public investment through a listing of the shares on a stock exchange (IPO). The INMARSAT satellites and all other assets of the former intergovernmental organization have been transferred to the privatised operating company which continues to manage the global mobile satellite communications system, including maritime distress and safety services for GMDSS at either no cost or at a special rate.

The residual intergovernmental organization (IMSO) continues with 88 Parties, operating through the Assembly of Parties, its Advisory Committee and a small Secretariat, headed by the Director who is the Chief Executive Officer and legal representative of the Organization. Under the relevant provisions of the Convention, as amended, the Public Services Agreement and the Articles of Association of the Company, IMSO is charged with overseeing, and under some circumstances may enforce fulfilment of the Company's public service obligations and, in particular, GMDSS services. In performing this role, IMSO acts as the natural ally of IMO and watchdog of proper provisions and implementation of IMO's requirements in respect of GMDSS by Inmarsat Ltd. To facilitate these functions, an Agreement of Cooperation has been concluded between IMSO and IMO. Under a similar Agreement with the International Civil Aviation Organization (ICAO), IMSO ensures that Inmarsat Ltd. takes into account the applicable ICAO Standards and Recommended Practices in line with the Public Services Agreement and regularly informs ICAO accordingly.

Administrative Arrangements have also been signed between the Secretary-General of the International Telecommunication Union (ITU) and the IMSO Director. These provide the Organization with direct access to the relevant bodies of the ITU, enabling IMSO to play an active role in the development of international telecommunication policies.

In recent years, the horizons of mobile satellite communications are expanding with ever-increasing speed, and there are several different options for the design and capability of new services. The adoption by the IMO Assembly of Resolution A.21/Res.888 –

Criteria for the Provision of Mobile Satellite Communication Systems in the Global Maritime Distress and Safety System (GMDSS), has provided a clear indication of IMO's intention to consider opening up provision of GMDSS services in the future to any satellite operator whose system fits these Criteria. However, that remains for the future. At present, Inmarsat Ltd., with the satellite communications system that it operates, is the sole global provider of these services, although, after the restructuring of INMARSAT, the process of liberalization and privatization of global and regional satellite communications services is fast developing.

Quoting the Director of IMSO, Mr. Jerzy W. Vonau, it is encouraging to note, in this context, that IMSO has not been able to detect any reduction or deterioration in the level and quality in the provision of GMDSS services by Inmarsat Ltd. under the new regime, compared with the situation prior to privatization. All other public service obligations were also observed or due attention has been given by the Company. It may therefore be concluded that, as a result of right political and legal decisions of the Member States, followed by a distinct but workable interface between IMSO and Inmarsat Ltd, the restructuring and privatisation of the business and assets of the former inter-governmental organization have paid off and the principles under which the process of restructuring took place have proved to be effective.

The formal structure of the new company is somewhat complex, but the essential organs are Inmarsat Ltd, which executes all the activities of the Company, and its parent company Inmarsat Group Holding Ltd, which is a financial holding company owned by the shareholders. The main shareholders today are, with share holdings in rounded figures:

- The venture capital companies Apax and Permira, who together hold 52 %
- Telenor 15 %
- Lockheed Martin 14 %
- KDDI (Japan) 8 %

In late 2003 most of the Signatories of "old" INMARSAT sold their shares in Inmarsat Group Holdings Ltd, although some have retained a minute share in order to receive information. They remain, however, Parties to the residual intergovernmental organization IMSO. In addition to the ordinary shareholders, IMSO holds a "Golden Share", which gives it special rights in questions related to Inmarsat's obligations with respect to certain public services, notably GMDSS. The Board of Directors of Inmarsat Group Holdings Ltd consists of seven directors, three being executive officers in Inmarsat Ltd, and four being non-executives representing major shareholders. The executive officers on the Board are CEO, CFO, and COO[7]. The shareholders Apax, Permira, Telenor, and Lockheed Martin each have one Board Director. Bjarne Aamodt, Senior Vice President of Telenor, is the Board Director from Telenor.

The executing company Inmarsat Ltd performs the same functions as the Directorate in the original INMARSAT. The change from the INMARSAT Council with 22 members, to the Board of Inmarsat Group Holdings Ltd, with seven directors, is but one illustration of the profound change which has taken place. It is comfortable to observe that the business of Inmarsat Ltd has continued to be financially very sound in the new situation of fierce competition.

Part C Telenor as investor in the Organization, and as exploiter of the space segment

17 The Nordic Earth Station at Eik in Rogaland

The purpose of NTA's efforts in the creation of INMARSAT was obviously to provide its allocated share of necessary funds to INMARSAT for establishing the space segment, and to utilize the space segment to provide high quality telecommunication services to the maritime industry around the clock in all seas and oceans of the world.

Norway through NTA has dutifully fulfilled all its obligations under the Convention and the Operating Agreement, including payment of its investment share. The first investment called for in 1981–82 amounted to some NOK 100 million, which was the highest investment ever by Norway in space technology.

In order to produce services an earth station "on land"[8] is required. The earth station is very costly, mainly due to the complicated arrangement for accessing the space segment on a demand basis. The total traffic volume in question was rather low, so routing the traffic over a remotely located station implied insignificant expenses compared to the cost

---

7) *CEO: Chief Executive Officer; CFO: Chief Financial Officer; COO: Chief Operating Officer.*

8) *"Earth station", as opposed to space segment, is the term used to designate stations located on the surface of the earth; i.e. also ship stations. So we have "earth stations on land" and "ship earth stations". The last term may sound strange, but so it has been defined.*

Figure 7 Eik earth station (Photo: Ole Gunnar Mosvold)

of constructing and operating an earth station. This again called for international cooperation. A study conducted by STSK[9)] concluded that it was not economical to construct more than one station to serve all the Nordic countries. For economic reasons the station should be located at one of the existing satellite earth stations, viz. Tanum in Sweden or Eik in Norway. Since Norway by far would have the highest traffic volume over the station, it was decided to locate a Nordic station at Eik. An agreement to this effect was signed late 1979. NTA should invest in, and operate, the station, whereas the other Nordic countries should pay parts of the operating expenses, including capital cost, proportional to their respective use. The net profits of the maritime communications services utilising Eik were split in a similar proportionate way. For a period of many years, the Eik earth station was instrumental in the Nordic countries' ability to turn the provision of mobile satellite services into a commercial success.

The Eik Nordic operation was terminated at the end of 2001, at which point Telenor remained sole owner and operator of the station.

Procurement of the earth station was based on international competition. Only two bids were received: from NEC (Japan) and EB/NERA. Both were technically acceptable. EB/NERA was chosen, having a slightly lower price. A Norwegian supplier was considered very advantageous. The tri-party cooperation of NTA/EB NERA/ELAB described in chapter 2 helped EB/NERA to be able to come up with a competitive bid for this large and complicated undertaking. The Nordic earth station at Eik was put in operation early 1982, as the first one in the INMARSAT system in Europe.

In the years which followed, NERA captured about 25 per cent of the world market for maritime land earth stations as well as for ship stations. This represented a successful industrial spin-off of the early tri-party studies initiated by NTA.

18 The Tri-party Earth Station Cooperation between Norway, UK, and Singapore, with stations Eik, Goonhilly, and Sentosa

International cooperation in maritime satellite communications required yet another dimension. The space segment consisted at the outset of three satellites located over the Atlantic (AOR), the Pacific (POR) and the Indian (IOR) oceans, respectively, which provided coverage of all important ocean areas. To provide near world coverage (polar regions cannot be reached with geostationary satellites), land earth stations obviously need to "see" all three satellites. Eik can see AOR and IOR, so cooperation with an earth station owner in the Pacific area was necessary. It made very good economic sense to cover only the IOR from Eik, and let Goonhilly in the UK cover

9) *STSK = the Scandinavian Committee on Telecommunication Satellites*

AOR for the Nordic countries, whereas Eik handled British traffic destined for the IOR. With Sentosa of Singapore included in the cooperation, we all achieved world coverage at minimum cost. A similar model was later adopted by other groups of countries.

We may observe that this kind of cooperation was traditionally unheard of within the maritime community. It was a serious prestige matter for maritime countries to have their own HF coast stations. The hard economic facts in the case of maritime satellite communication forced countries to cooperate also in this matter.

Eik was one of the most successful stations in the INMARSAT system – at times the best – with high traffic volume, excellent quality and regularity. The provision of telecommunication services over the station became a lucrative business for Telenor, with annual revenues in the order of NOK 7–800 mill.

With the Inmarsat industry becoming increasingly deregulated and commercialised, there has been a high degree of consolidation amongst the major players. Some of the these players have through acquisitions or strategic alliances achieved global coverage; i.e. they now "see" all satellites in the INMARSAT system from their own or controlled ground infrastructure. This is one of the reasons for the termination of the cooperation with the UK and Singapore in the late 1990s.

19 Acquisitions

In 2001 Telenor took another bold step in mobile satellite communications with the acquisition of Comsat Mobile Communications (CMC). Comsat Mobile, with revenues in excess of US$ 100 mill, on 27 May 2001 became a subsidiary of Telenor, under the name Telenor Satellite Services Inc. The transaction transferred CMC's operations, including all employees as well as the two earth station facilities in Southbury, Connecticut, and Santa Paula, California, to Telenor.

Figure 8 The headquarters of Telenor Satellite Services (TSS) US operation at Rockville, Maryland

With the CMC acquisition Telenor has become the largest provider in the world of mobile satellite communications. Telenor now offers satellite communications customers worldwide a broad portfolio of truly seamless global services. In February 2001 Telenor also acquired the Brussels, Belgium, based Accounting Authority SAIT Communications (SAIT). With the acquisition of SAIT, Telenor strengthened its position in the retail, or end user, segment of the Inmarsat industry.

Annex 1 Historical summary

- National studies on how to make use of satellite techniques for maritime communications started in the second half of the 1960s.

- International studies were also started around this time – in the International Radio Consultative Committee of ITU (CCIR) and the Inter-Governmental Maritime Consultative Organization (IMCO, later IMO).

- The World Administrative Radio Conference WARC (Space) allocated frequency bands to the Maritime Mobile Satellite Service in 1971.

- IMCO set up the Panel of Experts on Maritime Satellites in 1972 to study the technical, operational, administrative, and institutional aspects of a maritime satellite system. The Panel held six meetings over the 1972 to 1974 period. It recommended, inter alia, that INMARSAT be created as an intergovernmental organization and that a conference be convened to effect this. The Panel's report contained a complete draft convention for such an organization.

- IMCO convened The International Conference on the Maritime Satellite System (the 'INMARSAT Conference'), which held three sessions between April 1975 and September 1976.

- The Conference adopted and opened for signature the Convention and the Operating Agreement on the International Maritime Satellite Organization (INMARSAT). The Final Acts of the Conference were signed by representatives of 46 States. A summary of the INMARSAT instruments is given in Annex 2. Both instruments entered into force in July 1979, 60 days after States representing more than 95 per cent of the initial investment shares had become Parties to the Convention.

- The INMARSAT Conference established a Preparatory Committee, which was active from January 1977 to the summer of 1979.

- The US consortium MARISAT established a maritime satellite system in 1976. The two MARISAT satellites later became part of the first generation INMARSAT system.

- Considerable preparatory work for INMARSAT's space segment had been carried out over a period of some two years under what was termed the Joint Venture Conference.

- A Nordic earth station was opened for service early 1982 at Eik in Rogaland. This was the first European earth station in the INMARSAT system.

- A new Global Maritime Distress and Safety System (GMDSS), based on satellite communications over the INMARSAT system, was introduced by IMO in 1992.

- The intergovernmental organization INMARSAT was privatized in 1999.

- Telenor purchased Comsat Mobile Communications (USA) in 2001 and became the largest provider in the world of mobile satellite communications.

Annex 2 Summary of the (original) INMARSAT agreements

The organization INMARSAT was based on two agreements: a Convention between participating States ("Parties") and an Operating Agreement between telecommunication organizations of the member States ("Signatories"). Article 2 of the Convention provides that each member State or Party shall either itself sign – or shall designate a competent entity to sign – the Operating Agreement.

Article 4 of the Convention states that the Party is not liable for obligations arising under the Operating Agreement. In other words, the State is not responsible for financial, economic and operational matters, but the Party shall guide and instruct the Signatory to ensure that it fulfils its responsibilities and does not violate obligations which the Party has undertaken under the Convention.

The purpose of INMARSAT, as given in Article 3 of the Convention, is "to make provision for the space segment necessary for improving maritime communications, thereby assisting in improving distress and safety of life at sea communications, efficiency and management of ships, maritime public correspondence services and radiodetermination capabilities". Also, INMARSAT shall act exclusively for peaceful purposes.

Operational and financial principles of the organization are stated in Article 5 of the Convention and some of the articles of the Operating Agreement. The Organization shall operate on a sound economic and financial basis having regard to accepted commercial principles. Furthermore, the Organization shall be financed by contributions of Signatories. Each Signatory shall contribute capital in proportion to its use of the space segment and shall receive capital repayment and compensation for use of capital when revenues allow such repayment.

Article 6 of the Convention provides that the Organization may own or lease the space segment.

INMARSAT is open for membership by all States. Also ships of non-member countries may use the space segment on conditions determined by the Organization.

The Organization will have three main organs: the Assembly, where all member States are represented and have one vote each: the Council, with twenty-two Signatories and voting-power in relation to investment shares; and the Directorate headed by a Director General. Substantive decisions are made with a two-thirds majority both in the Assembly and in the Council. The functions of the Assembly are to consider and review the general policy and long-term objectives of the Organization and express views and make recommendations thereon to the Council. In addition, the Assembly has some specific listed functions. All other decisions will be made by the Council. The Council will thus decide on all financial, operational and administrative questions.

The headquarters of INMARSAT will be located in London.

The procurement policy shall be such as to encourage, in the interest of the Organization, worldwide competition in the supply of goods and services. To this end the Organization shall award contracts to bidders offering the best combination of quality, price and delivery time.

The charges for use of the space segment shall have the objective of earning sufficient revenues for the Organization to cover its operating, maintenance, and administrative costs, the provision of such operating funds as are necessary, the amortization of investments made by Signatories, and compensation for use of capital.

The sum of the net capital contributions of Signatories and of outstanding contractual capital commit-

ments of the Organizations is subject to a ceiling of 200 mill US dollars.

In the initial phase, before the Signatories' use of the space segment is known, the negotiated investment shares are related to expected use of the space segment and are specified in the Annex to the Operating Agreement. The Annex to the Operating Agreement also contains a provision which makes it possible for Signatories to increase the initial shares before entry into force of the Convention. This provision was adopted by the INMARSAT Conference in order to ensure that the instruments would enter into force even if some countries due to lengthy domestic procedures were not able to become Parties within the time limit specified in Article 33 of the Convention.

Those who are familiar with the INTELSAT Agreements will observe substantial similarities between INMARSAT and INTELSAT. Indeed, INTELSAT was to a fairly large extent used as model, although many delegates believed that the INTELSAT Agreements are unnecessarily complicated. The texts of the INMARSAT Agreements are more concise and are believed to be clearer and easier to read for non-experts.

Bibliography

1 Collett, J P. *Making Sense of Space – a History of Norwegian Space Activities*. Oslo, Scandinavian University Press, 1995.

2 Gallagher, B. *Never Beyond Reach – The World of Mobile Satellite Communications*. London, INMARSAT, 1989.

3 Haga, O J. INMARSAT – an example of Global International Cooperation in the Field of Telecommunications. *Telektronikk*, 75 (4), 373–381, 1979.

4 Solbakken, K. INMARSAT Nordisk maritime jordstasjon på Eik. *Telektronikk*, 77 (1), 24–27, 1981.

5 Solbakken, K. INMARSAT – Systembeskrivelse. *Telektronikk*, 77 (1), 21–23, 1981.

6 Grimsmo, N, Bøe, A. INMARSAT – International Maritime Satellite Organization. *Telektronikk*, 77 (1), 16–20, 1981.

7 Knudtzon, N. Telesatellitter – et helhetsperspektiv. *Telektronikk*, 86 (2/3), 81–92, 1990.

8 Haga, O J. Mobil satellittkommunikasjon – Perspektiver for 1990-årene. *Telektronikk*, 87 (2/3), 231–236, 1991.

9 *Telenor completes acquisition of Comsat Mobile Communications from Lockheed Martin*. Press release by Telenor/Lockheed Martin, December 2001.

10 Vonau, J W. *Restructured INMARSAT and public service obligations*. London, IMSO, 2004. (Internal document of IMSO.)

11 *Inmarsat Group Holdings Ltd*. October 6, 2004 [online] – URL: www.inmarsat.com.

12 Interview with Bjarne Aamodt, Senior Vice President of Telenor, 27.09.2004.

13 Interview with Knut Grødem, Telenor Satellite Services, on 29.09.2004. (Assisted with paragraphs 14, 17, 18 and 19.)

Ole Johan Haga (71) holds a BSc (Honours) degree in Electronic Engineering from University of St. Andrews, Scotland, 1960. He has worked with Telenor for 40 years in various positions. After his retirement in 2000 he has been available as senior consultant. His main responsibilities and positions have comprised:

- *10 years of duty in the construction of the national radio relay link network*
- *various responsibilities in the Technical Department of NTA Headquarters*
- *international activities in various Nordic committees, ITU, CEPT, IMO, and INMARSAT. Chairman of INMARSAT Council 1984–86*
- *regional director of Oslo 1985–95*
- *international activity, mainly on investment projects in mobile networks in South East Asia and Eastern Europe*

Ole Johan Haga's managing positions include: Head of Radio Link Department of NTA Headquarters, Assisting Technical Director, Vice President of Telenor International, Managing Director of Telenor Satellite Services AS, Managing Director of Telenor Invest AS.

email: ojo-ha@online.no

Features of the Internet history
The Norwegian contribution to the development

PAAL SPILLING AND YNGVAR LUNDH

Paal Spilling is professor at the Department of informatics, Univ. of Oslo and University Graduate Center at Kjeller

Yngvar Lundh is an independent consultant

This article provides a short historical and personal view on the development of packet-switching, computer communications and Internet technology, from its inception around 1969 until the full-fledged Internet became operational in 1983. In the early 1990s, the internet backbone at that time, the National Science Foundation network – NSFNET, was opened up for commercial purposes. At that time there were already several operators providing commercial services outside the internet. This presentation is based on the authors' participation during parts of the development and on literature studies. This provides a setting in which the Norwegian participation and contribution may be better understood.

1 Introduction

The concept of computer networking started in the early 1960s at the Massachusetts Institute of Technology (MIT) with the vision of an "On-line community of people". Computers should facilitate communications between people and be a support for human decision processes. In 1961 an MIT PhD thesis by Leonard Kleinrock introduced some of the earliest theoretical results on queuing networks. Around the same time a series of Rand Corporation papers, mainly authored by Paul Baran, sketched a hypothetical system for communication while under attack that used "message blocks" each of which contained an address to identify the destination. In the latter half of the 60s, these ideas had got enough momentum for the U.S. Defense Advanced Research Project Agency – ARPA (later renamed DARPA) – to initiate development and fund research on this new promising communications technology now known as packet switching. A contract was awarded in December 1968 to Bolt, Beranek and Newman (BBN) to develop and deploy a four node network based on this technology. Called the ARPANET, the initial four nodes were fielded one a month from September to December 1969. The network quickly grew to comprise academic research groups across the United States, including Hawaii, and in 1973 also extended to Norway and England. During the early 1970s, DARPA developed two alternate implementations of packet switched networks – over satellite and ground radio. The protocols to link these three networks and the computers connected to them, known as TCP/IP, were integral to the development of the Internet. The initial nascent Internet, consisting of those three networks, was first demonstrated in 1977, although earlier two network tests had undoubtedly been carried out. Through independent implementations, extensive testing, and refinements, a sufficiently mature and stable internet technology was developed (with international participation) and in 1980 TCP/IP was adopted as a standard for the US Department of Defense (DOD). It is uncertain when DoD really standardized on the entire protocol suite built around TCP/IP, since for several years they also followed the ISO standards track.

The development of the Internet, as we know it today, went through three phases. The first one was the research and development phase, sponsored and supervised by ARPA. Research groups that actively contributed to the development process and many who explored its potential for resource sharing were permitted to connect to and use the network. This phase culminated in 1983 with the conversion of the ARPANET from use of its initial host protocol, known as NCP, to the newly standardized TCP/IP protocol. Then we had the start of the interim phase. All hosts on the ARPANET were required to convert to TCP/IP during early January 1983, but in reality the conversion lasted until June 1983, during which time both the old protocols and the new protocols were run simultaneously. ARPANET was divided into two parts interconnected by a set of filtering gateways. Most defense institutions were attached to one part called MILNET, which was to be integrated with the Defense Data Network and operated by the Defense Communications Agency (DCA). The other – open – part, still called ARPANET, contained university institutions, non-defense research establishments and a few defense organizations including DARPA. The newly reconstituted ARPANET remained in operation until 1990, when it was decommissioned. By that time, responsibility for the open part was taken over by National Science Foundation (NSF). NSF had created a small experimental network which was replace in 1988 by a higher speed network called NSFNET. The NSFNET was the result of efforts by IBM, MCI and MERIT, the latter having their contract with NSF. Other organizations also provided funding for relevant parts of the Internet. And gradually many of the regional parts of the network were privatized. The network was now, in

 ISSN 0085-7130

principle, open to anyone doing computer science research and international extensions were soon put in place. The number of attached institutions and users grew rapidly. The ARPANET technology had served its purpose, was now being phased out and replaced by higher-capacity lines and commercial routers. This coincided approximately with the appearance of the very first Web-browser. The World Wide Web was invented at CERN in 1989 and has proved to be a major contributor to the usefulness of the Internet. A few years later, in 1993, the restrictions on commercial activity in the NSFNET were lifted, and the Mosaic browser was introduced by the University of Illinois. This was the start of the third phase, the commercial phase, resulting in an explosive growth in geographic coverage, number of users, and traffic volume. Of course, email and file transfer had already been in place for two decades, but the use of web browsers made it easier to use and opened up a larger world of information access on a scale never before seen.

For many years the Internet and its concepts were neglected by the telecom operators and the other European research communities. In the last chapter we attempt to shed light on some of the important factors contributing to this effect.

2 The prelude; the inception of packet switching

There has been a debate for some time about who invented packet switching; was it Kleinrock at MIT, Paul Baran at Rand Corporation, or Donald Davies at the National Physics Laboratory in England? Donald Davies is recognized as the person who coined the term packet. We do not take a stand here. We believe all three studied, from a conceptual viewpoint, different aspects of the store-and-forward technology, a key concept behind packet switching. We provide a brief description of their research relevant to packet switching.

Leonard (Len) Kleinrock, a PhD student at MIT, published his first paper on digital network communications titled *"Information Flow in Large Communication Nets"*, in July 1961 [1]. This was the first paper describing queuing networks and analyzing message switching. He developed his ideas further in his 1962 PhD thesis, and then published a comprehensive analytical treatment of digital networks in his book *"Communication Nets"* in 1964 [2].

After completing his PhD in 1962, Kleinrock became Professor at UCLA. There he later established and led the Network Measurement Center (NMC), consisting of a group of graduate students working in the area of digital networks. In October 1968, ARPA awarded a contract to Kleinrock's NMC to perform ARPANET performance measurements and identify areas for network improvement.

Paul Baran, an electrical engineer, joined RAND in 1959. The US Air Force had recently established one of the first wide area computer networks for the SAGE radar defense system, and had an increasing interest in robust and survivable wide area communications networks. Baran began an investigation into development of survivable communications networks. The results were first presented in a briefing to the Air Force in the summer of 1961, and later, in 1964, as a series of eleven comprehensive papers titled *"On Distributed Communications"* [3, 4].

The series of reports described in remarkable detail an architecture for a distributed, survivable communications network for transport of speech and data. It was based on store-and-forward of message units of 1024 bits, dynamically adaptive routing, and could withstand serious destruction to individual nodes or links without loss of end-to-end communications. At the time, the technology to implement this architecture cost effectively did not exist. Apparently the Air Force did not see the value of this new concept at that time and did not follow up the recommendations in the report.

Donald W. Davies at the National Physical Laboratory (NPL) in England, apparently unaware of Baran's ideas, developed similar concepts. He got his original idea in 1965: to achieve communication between computers by utilizing a fast message-switching communication service [5]. Long messages had to be split into chunks, called *packets*, and sent separately so as to minimize the risk of congestion. This was the same approach taken by ARPA, but the ARPANET initially used the term "message switching" and later adopted Davies' terminology. *The store-and-forward of packets became known as packet-switching.*

Davies proposed in 1967 a plan [6] for a communications system between a set of terminals and a set of computers. It was based on store-and-forward of packets in a mesh of switching nodes interconnected by high-speed lines. Terminals were to be served by one or more interface computers. These interface computers acted as packet assembly/disassembly between the network and the terminals. The practical outcome of the NPL activity was a local packet-switched communication network that grew in the coming years, to serve about 200 terminals and a dozen or so computers.

3 ARPANET; the start of a new era

In the latter half of the 1960s the packet-switching concept was mature enough to be realized in practice. We introduce the four persons most influential to the creation of ARPANET and subsequently the internet. Dr. J.C.R. Licklider, around 1960, had the vision of a "Galactic network" and provided the main inspiration. Lawrence Roberts published the network plan in 1967 and led the early development of ARPANET. When Roberts left ARPA in 1973, Robert Kahn took over the responsibility for the development process and brought it to its full fruition over a period of more than ten years. He was assisted by Vinton Cerf in developing the TCP/IP protocols, the true heart of the internet. In our opinion these two people are the main inventors of the internet, but assisted by many individuals and research groups in a great collaborative effort.

Dr. **J.C.R. Licklider** did research on psychoacoustics at MIT in the late 1950s. He had the unusual educational background as engineer and psychologist, and saw early the need for computers in the analysis of his research results. Licklider joined Bolt, Beranek and Newman in Cambridge, Massachusetts in 1957 to pursue psychoacoustic research. Here he was given access to one of the first minicomputers, a PDP-1 from Digital Equipment Corporation (DEC). He developed the vision of an "On-line community of people", expressed in a seminal paper in 1960 called *"Man-Computer Symbiosis"* [7, 8], in which he described an interactive computer assistant that could answer questions, perform simulation, display results graphically, and extrapolate solutions for new situations from past experience. In this way, computers could facilitate communications between people and be a support for human decision processes. These were quite futuristic ideas at that time, and was the true start of the project that later got named ARPANET.

Dr. J.C.R. Licklider was employed by ARPA in 1962 as leader of the division that was later named "Information Processing Techniques Office" or IPTO. This office is tasked with initiating and financing advanced research and development in information processing of vital importance to the American Defense. The military had a long tradition of partnership with university research. Most of the basic research for DOD was performed in the academic arena. In Licklider's days the office funded top-level academic scientists called "Principal Investigators". Licklider left the IPTO office in 1964, but had a second term from January 1974 through August 1975.

Lawrence (Larry) Roberts, after finishing his PhD at MIT in 1958, joined MIT Lincoln Laboratories and started research on computer networks. He was inspired by Licklider's visions. In February 1965, ARPA awarded a contract to Lincoln Laboratory to experiment with computer networking. In October 1965, the Lincoln Labs TX-2 computer talked to the Q32 computer at System Development Corporation (SDC) in Santa Monica California, via a dial-up 1200 bit/s phone line. This was one of the world's first digital network communications between computers. The results of this research were presented by Merill and Roberts, at the AFIPS Conference in October 1966, in a paper titled "Toward a Cooperative Network of Time-Shared Computers" [9]. In December 1966 Lawrence Roberts was asked to join ARPA to lead the IPTO effort in developing a wide area digital communications network, which was later named the ARPANET.

Larry Roberts based his network plans on the MIT research performed by Kleinrock, the BBN research of Licklider, and also on his own experience. He presented his networking plan at the ACM Gatlinburg conference in October 1967 [10]. It contained plans for inter-computer communications and the interconnection of networks. Roberts met with the NPL researcher Roger Scantlebury during the conference and learned about their work and the work of Baran. Later Roberts studied the Baran reports and met with him. The Baran work did not have any significant impact on Roberts' plan, according to Roberts' "Internet Chronology" (http://www.ziplink.net/~lroberts/InternetChronology.html). The NPL paper [6] convinced Larry Roberts to use higher speed lines (50 kbit/s) between the nodes and use the word packet.

Larry Roberts left ARPA in October 1973 to become the second President of Telenet, providing a commercial data communication service based on the X.25 standard. He was followed for a brief period by Dr. J.C.R. Licklider as director of the IPTO office. In 1979 Telenet was sold to GTE, to become the data division of SPRINT.

Larry Roberts has been the recipient of numerous awards. He shared the Charles Stark Draper Prize for 2001 with Robert Kahn, Vinton Cerf, and Leonard Kleinrock for their work on the ARPANET and Internet.

Robert (Bob) Kahn obtained his PhD degree from Princeton University in 1964. He then worked with the Technical Staff at Bell Laboratories and subsequently became Assistant Professor of Electrical Engineering at MIT. Then he joined Bolt, Beranek and Newman (1966 – 1972), where he was responsible for the system design of the ARPANET, the first packet-switched network, and wrote the technical

proposal to ARPA that won the contract for BBN. The research team, led by Frank Heart, proposed to use mini-computers as the switching element in the network. The team consisted of Bob Kahn, Severo Ornstein, Dave Walden and others. After awarding the contract to BBN, Bob Kahn wrote the Host – IMP technical specification.

In 1972 Bob Kahn was asked by ARPA to organize a demonstration of an ARPANET network, connecting 40 different computers, at the International Computer Communication Conference in October 1972, making the network widely known for the first time to technical people from around the world [18]. Organizing the demonstration was a major undertaking, to push and coordinate all involved parties to get everything to work reliably in time for the conference. The demonstration was a great success.

Bob Kahn moved to ARPA immediately after the conference, initially as program manager with responsibility for managing the Packet Radio, Packet Voice and SATNET projects. He later became chief scientist, deputy director and subsequently director of the IPTO office. While Director of IPTO, he initiated the United States government's billion dollar Strategic Computing Program, the largest computer research and development program ever undertaken by the federal government. Dr. Kahn conceived the idea of open-architecture networking. He is co-inventor of the TCP/IP protocols and was responsible for originating ARPA's Internet Program. Dr. Kahn also coined the term "National Information Infrastructure" (NII) in the mid 1980s, which later became more widely known as the "Information Super Highway". Kahn left ARPA late 1985, after thirteen years. In 1986 he founded the Corporation for National Research Initiatives (CNRI). CNRI was created as a not-for-profit organization to provide leadership and funding for research and development of the National Information Infrastructure. He has been the recipient of numerous awards and received several honorary university degrees for his outstanding achievements. In 1997 president Clinton presented the US National Medal of Technology to Kahn and Cerf.

As President of CNRI, Kahn has continued to nurture the development of the Internet over the years through shepherding the standards process and related activities.

Vinton (Vint) Cerf did graduate work at UCLA from 1967 until he got his PhD in 1972. ARPA released the "Request for Proposals" in August 1968. As a result, the UCLA people proposed to ARPA to organize and run a Network Measurement Center for the ARPANET project. The team included among others: Len Kleinrock, Stephen Crocker, Jon Postel, Robert Braden, Michael Wingfield, David Crocker, and Vint Cerf.

Vint Cerf's interest in networking was strongly influenced by the work he did at the Network Measurement Center at UCLA. Bob Kahn, then at BBN, came out to UCLA to participate in stress-testing the initial four-node network, and had a very productive collaboration with Vint Cerf. Vint did the necessary programming overnight, and together they did the experiments during the day.

In November 1972, Cerf took up an assistant professorship post in computer science and electrical engineering at Stanford, and was one of the first people there who had an interest in computer networking. The very earliest work on the TCP protocols was done at Stanford, BBN and University College London (UCL). The initial design work was done in Vint Cerf's group of PhD students at Stanford. One of the members of the group was Dag Belsnes from the University of Oslo. He did work on the correctness of protocol design. The first draft of TCP came out in the fall of 1973. A paper by Bob Kahn and Vint Cerf on internetting appeared in May 1974 in IEEE Transactions on Communications [20] and the first specification of the TCP protocol was published as an Internet Experiment Note in December 1974. Then the three groups began concurrent implementations of the TCP protocol. So the effort at developing the Internet protocols was international from the beginning.

Vint Cerf worked for ARPA from 1976 till 1982, having a leading role in the development of the Internet and internet-related technologies. In 1982 he became vice president for MCI Digital Information Service, leading the engineering of MCI Mail System, the first commercial mail service to be connected to the internet. Then, in 1986, he became Vice President of the Corporation for National Research Initiatives (CNRI), a position he held until 1994. Then he joined MCI, and is now senior vice president of Technology Strategy.

Vint Cerf has been the recipient of numerous awards, both nationally and internationally. In 1997, President Clinton presented the US National Medal of Technology to Cerf and Kahn.

4 ARPANET; the research and development process

Here we go briefly through the research and development process, from the point in time where the requirements were formulated and the first research contracts were awarded, till the network was operational and covered the continental USA and with two "tentacles" – to Hawaii and to Norway, and from there onwards to England.

The planned network consisted of two main components:

- The packet-switches or nodes (implemented on minicomputers), should be interconnected in a mesh network by means of high-speed telephone lines and 50 kbit/s modems, to permit alternative routes between any sender-receiver pair. The interface between a node and a host was standardized to enable hosts of different makes and operating systems to connect to the network.

- Each host computer was connected to its dedicated node (communications front-end). The software in the hosts should permit resource sharing and support person-to-person communications.

To implement the plan, ARPA awarded contracts to the following institutions in the last quarter of 1968:

- Bolt, Beranek and Newman (BBN) in Boston; Frank Heart led the group with responsibility to develop the packet-switching nodes (called Interface Message Processors, or IMPs), deploy them, and to monitor and maintain the network;

- University of California Los Angeles (UCLA); Professor Len Kleinrock led the group with responsibility for performance studies of the network by means of simulation and real measurements;

- Network Analysis Corporation (NAC); Howard Frank and his team with responsibility for developing the network topology subject to cost and reliability constraints, and for analyzing the network economics;

- Stanford Research Institute (SRI); to establish a Network Information Center (NIC) as part of Doug Engelbart's group.

The initial four IMPs were fielded at the end of 1969. The first one was installed at UCLA in September. Due to Len Kleinrock's early theoretical work on packet switching and his focus on network analysis, design and measurements, his Network Measurement Center (NMC) was selected to be the first host on the network. The next node was installed at SRI in October. Doug Engelbart led a project called "Augmentation of the Human Intellect" at SRI, which included the early hypertext system NLS. NLS was the first advanced collaborative oN-Line text System [11], and included many of the modern text editing functions like mouse, cut-and-paste, hypertext, and a window-based user interface. In fact, many of the visionary concepts demonstrated by Engelbart in 1968 really became practical many years later when personal workstations interconnected in networks became economically feasible. Doug Engelbart also had a journaling system under development. It was intended to be the basis for the Network Information Center (NIC). Dick Watson was the leader of this task initially.

Soon after SRI was on the network, the first host-to-host message was sent from Kleinrock's group at UCLA to SRI.

Subsequently, in 1972, NIC was established as a separate project at SRI, with Elisabeth (Jake) Feinler as leader. NIC should maintain hostname-to-address mapping tables (in use in 1970) and be a repository for network-related reports and documentation (Requests for Comments, etc).

Two other IMPs were then installed, one at UC Santa Barbara and one at the University of Utah. These two groups did research in visualization; visualization of mathematical functions at Santa Barbara and 3-D visualization at Utah.

A key feature of the network was to use dedicated computers, called packet switches, interconnected in a mesh network and responsible for the transport of data in the form of packets. The network should be robust against link and/or node failures. Hence the mesh network should provide alternative routes when forwarding packets, to circumvent failures in the network.

Another key feature of ARPANET was the use of a network control center – NCC. NCC had the ability to monitor each node in the network, start and stop node interfaces, start and stop nodes, perform diagnostic tests of individual nodes and lines, and download new software into nodes from NCC. This made ARPANET a very powerful laboratory for studying and developing networking technology, all without requiring personnel to travel to all the greatly separated sites. It should be noted that NCC served two purposes, to be an efficient tool in the development process and to manage and maintain the network. It was always an important goal in the development

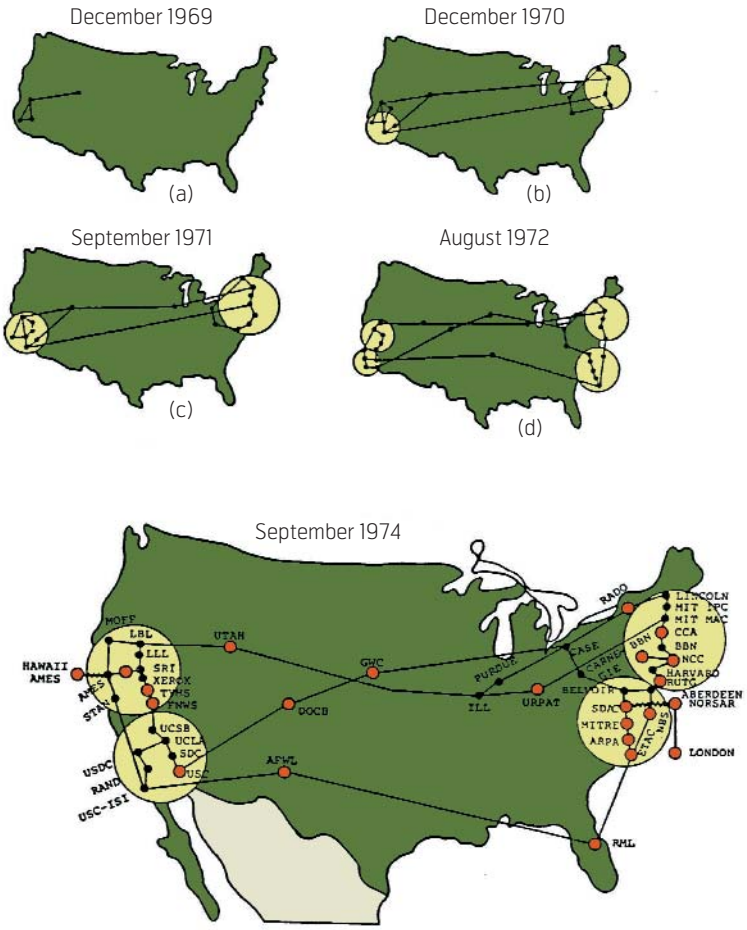


Figure 1 Some of the early stages of ARPANET

process to arrive at a technology that would *not* need centralized control.

While the performance of the network was analyzed and measured, the Network Working Group led by Steve Crocker at UCLA worked intensely to finish designing the host-to-host protocol called "Network Control Program" (NCP) [12]. NCP was part of the node software and provided a standardized interface to the attached hosts. The design was completed in December 1970, and was then implemented and installed in the increasing number of nodes during the 1971/72 period. This enabled the development of the long-awaited user services (host applications) like "Telnet", File transfer (FTP), and electronic mail. Ray Tomlinson at BBN implemented the initial electronic mail system on the ARPANET, using the now well-known address notation user@host [13]. An improved mail management program was written using TECO macros by Larry Roberts in 1972, for reading, storing, forwarding and sending mail.

After the initial test period, the ARPANET started to grow [14], see Figure 1. It spanned across to the East Coast in December 1970, with a total of 13 packet switches with numerous attached hosts. By 1975 it consisted of about 50 IMPs and between 150 and 200 hosts, permitting up to four hosts per IMP. The network interconnected defense establishments, and research institutions and universities with defense contracts scattered all over the USA. The continental part of the network had two "tentacles", one to Hawaii and one to Kjeller/Norway and onwards to London/England. As can be seen from Figure 1, all IMPs, excluding those in Hawaii, Norway and London, were interconnected with at least two neighbor IMPs for reliability purposes. In 1975, the operations of the continental part of ARPANET was transferred by ARPA to the Defense Communications Agency, while the responsibility for policy, further research and the international extensions continued to remain with ARPA.

5 Internetting was part of the vision from the very beginning

In parallel with the ARPANET development in the 70s, ARPA also initiated and funded development of mobile communications for tactical purposes based on portable radio units and packet-switching technologies, and packet switched satellite communications for wide area coverage.

Before he left the ARPA office, Larry Roberts had initiated the development of a packet switched satellite network (SATNET) by funding BBN to build the ground station nodes, the satellite IMP. When Bob Kahn moved to ARPA in late 1972, he initiated the development of a communicatitons network based on mobile radio-based units, called the Packet Radio Network (PRNET). Later, when Larry Roberts left ARPA, Bob Kahn took over the management of the ARPANET project. He also significantly changed the nature of the packet satellite effort – the original satellite node was split into three units: the IMP, the SIMP and a gateway (now router) in between. Although this was relatively straightforward technically, it was a non-trivial accomplishment politically and required the internet architecture to guide it.

The intention was to interconnect PRNET and SATNET with ARPANET. The ARPANET was viewed as a terrestrial backbone network. The PRNET was a broadcast radio network interconnecting geographically distributed clusters of packet-radio units, while SATNET was believed to be a means to interconnect widely dispersed ARPANET-like networks, for example located on different continents.

The PRNET development was mainly done as a collaborative effort among many parties and led by Bob Kahn. The packet radios were built by Collins Radio

in the Dallas area, the packet radio stations were built by BBN, and the whole system was assembled and tested in the San Francisco Bay Area, with SRI International as the leading party. The first field tests were started in 1975 [15]. A forerunner to the PRNET project was the radio-based Aloha system, conceived by Professor Abramson and developed at the University of Hawaii by Norman Abramson, Frank Kuo and Richard Binder [16] with funding from ARPA. The purpose was to provide access for user terminals, initially within range of the university, but later scattered over the Hawaiian archipelago, to a central computing facility at the University of Hawaii. It made use of a common inwards radio channel, shared on a packet basis between users in a random access manner. Another common radio channel was used for the outwards direction to broadcast the response from the central computer to the various user terminals. The system became operational in 1970. The SATNET project was delayed till 1975/76 due to tariff and regulatory problems with INTELSAT that were later solved by Bob Kahn [17]. The project had eight collaborating partners, including Norway and England, and made use of one 64 kbit/s channel in the Intelsat IV satellite system shared between three ground stations – one in USA, one in England, and one in Norway (actually located at Tanum in Sweden). Sweden took no part in the collaboration.

6 How did Norway get involved?

There were several contributing factors leading up to inviting Norway to participate in the collaborative effort. The main initiating factor was probably Larry Roberts' idea to connect ARPANET with the network at NPL in England. We provide a brief description of these factors, subjectively presented in order of importance, as we saw it. Thereafter we mention the key persons and the work involved, in the USA (ARPA), England, and in Norway, in getting the collaboration established and operational. In addition, we also had the NORSAR project which made working with Norway desirable, since it already had a 2.4 kbit/s line to the US that could be upgraded.

In late 1970 ARPANET covered a main part of continental US with about 13 nodes. ARPA now showed interest in linking ARPANET with the network at NPL. Larry Roberts made a proposal to Donald Davies regarding the linking [18]. The proposal from Roberts suggested that the UK's share in the collaboration should be to provide a line from NPL to NORSAR at Kjeller. This was impossible for NPL to handle. England had just applied for membership in the EU. And NPL, as a governmental institution, had to turn its focus on European research issues. The result was that Donald Davies had to turn down the proposal from Larry Roberts. It is also worth mentioning that in a memo from Vint Cerf in 1973, a plan to link up with CYCLADE in France was discussed. But it was never realized.

Professor Peter Kirstein of University College London (UCL) had expressed interest in joining ARPANET. Since Donald Davies was unable to accept Larry Roberts' proposal, Kirstein came up with a research plan that included the attachment of a large mainframe computer to ARPANET, to monitor and measure the academic traffic over the link to USA, via Tanum, and to participate in the planned satellite project [18]. The network topology now required a line from London to NORSAR, and Roberts suggested that the UK's share in the collaboration should be to provide a line from UCL to Kjeller. Donald Davies supported this plan. ARPA accepted it, and was prepared to install a TIP at UCL. Peter Kirstein was able to persuade British Telecom (BT) to offer free of charge a 9.6 kbit/s line to Kjeller, initially for one year. This was sufficient for Peter Kirstein to tell Roberts to proceed with the plan, and in September 1973 the UCL-TIP became operational on the ARPANET. In 1974, the British Ministry of Defence (MoD) took over the cost of the line to NORSAR; somewhat later BT also offered free of charge a 48 kbit/s line from UCL to the British ground station at Goonhilly for the packet satellite connection and the British part of the uplink to the satellite. Bob Kahn worked on the procurement of the packet satellite connection and made all the arrangements for the UK participation with John Taylor, then of the British MoD.

In 1965 contacts were established between ARPA's Nuclear Monitoring Research Office (NMRO) and the Norwegian Defence Research Establishment (NDRE). A seismic detection facility was under installation at Billings in Montana, USA, and later similar installations were made at other places (Iran, Alaska and Korea). The close proximity to USSR made Norway an attractive location for a seismic detection facility, in connection with "The Comprehensive Nuclear-Test-Ban Treaty". Director Finn Lied, research supervisor Karl Holberg, and research scientist Yngvar Lundh participated on behalf of NDRE. The result was the establishment of *NORSAR* (the NORwegian Seismic ARray), that became operational in 1970.

NORSAR was funded by the Research Council of Norway (NTNF at that time) with additional financial support from ARPA. The main processing center, the Seismic Data Analysis Center (SDAC) was located in Virginia. There were leased lines to all detection facilities from SDAC. The line from NORSAR to

SDAC went originally via the British satellite ground station at Goonhilly and an underwater cable to Norway. When the Nordic satellite station at Tanum became operational in 1971, the line from SDAC was relocated to go via Tanum. The line was paid for by ARPA. Until 1973, the line had a capacity of only 2.4 kbit/s but was upgraded to 9.6 kbit/s thereafter.

So the main two reasons for inviting NDRE and the Norwegian Telecommunications Administration (NTA) to participate in the further development of packet switching were:

- The development of packet-switched satellite communications would profit from Norwegian participation. The NORSAR array was seen to be a major potential user of the network. It was also assumed that this work would be of interest to Norway as a large shipping nation.

- The collaboration between ARPA, UCL and NDRE would make it substantially cheaper to link up both UCL and NDRE to ARPANET, when making use of the NORSAR – SDAC line.

Some other minor arguments may also have contributed to inviting NDRE to join:

- Previous contacts and the establishment of NORSAR in 1970 had contributed to a good relationship between the ARPA office and NDRE.

- Yngvar Lundh of NDRE had been on a sabbatical at MIT in 1958 in the same laboratory and at the same time as Larry Roberts completed his PhD. They got to know each other. Years later Larry Roberts started to work for ARPA.

Larry Roberts, at that time the current director of the IPTO office at ARPA, and Bob Kahn visited NDRE in the early fall of 1972 to discuss a possible participation by NDRE in the further development of the packet switching technology. Many aspects of this new form of communication were discussed at the meeting, with relations to a possible future Norwegian participation. Among other things, Roberts and Kahn pointed out wireless communications, and more specifically satellite communications, as important for Norway as a large shipping nation. They recommended that NDRE should attend the upcoming ICCC meeting later in 1972, where a presentation and demonstration of ARPANET were to take place. Prior to the Norway meeting ARPA had contacted NTA – The Norwegian Telecommunications Administration (now Telenor), but they declined to participate.

Yngvar Lundh attended the ICCC meeting in Washington DC [19] and was convinced of the great potential in the applications of this technology. He decided to join the development and participate in the planned multiple-access packet-switched satellite project. He had moral support from the director Finn Lied and the research supervisor Karl Holberg. Lundh established a small research group at NDRE, consisting of himself and a few master students. He started to participate in the regular ARPA project meetings. On 15 June 1973 a terminal-IMP (TIP), on loan from ARPA, was installed on the premises of NORSAR. This location was selected for two reasons. Firstly, it was important to have the TIP outside the restricted area of NDRE to permit other Norwegian groups to participate in the project. Secondly, to have easy access to the NORSAR-SDAC line for multiplexing the TIP and NORSAR traffic over that line. The line capacity was upgraded from 2.4 kbit/s to 9.6 kbit/s.

Paal Spilling started to work for NDRE in 1972 as a research scientist. He was a former nuclear physicist and joined full time with Lundh's efforts in 1975. He had no knowledge in data communications and protocols and was given time to educate himself – partly by following university courses and partly by practical trials and errors. Paal Spilling became a highly needed addition of qualified manpower to Lundh's group. One of his first assignments was to work in Kirstein's group at UCL for two months, for a flying start. UCL was at that time ready to start testing their implementation of the early version of the Internet Protocol, TCP. Since then, Paal has been a major contributor both to the development of Internet itself and to other aspects of computer communications in Norway.

Bob Kahn, while at ARPA's IPTO office, was eager to get the satellite project started. The idea was to use a fixed 64 kbit/s SPADE channel in the INTELSAT IV satellite, with one ground station at Tanum in Sweden, one at Goonhilly in England, and one at ETAM in West Virginia in USA. The SPADE channel would be used in a time-shared modus between the three ground stations in a modus called "Multi-destination half Duplex". It was assumed that each ground station operator would pay for its part of the uplink to the satellite. This was a modus operandi the INTELSAT organization could not handle at that time. As other telecom operators, they were used to the mode "Single Carrier per Channel", which meant that the two end-points of a channel or line had to be owned by the same customer/operator. It took Bob Kahn between one and two years to convince INTELSAT to change their policy, to permit the new way of operating the channel and the ground stations; this included also a new tariff for this operational modus

[17]. The satellite project – SATNET – therefore was delayed until 1975/76.

In parallel with Bob Kahn's effort to convince INTELSAT, Yngvar Lundh had discussions with the Research Department of NTA (NTA-R&D – now Telenor R&D) about possible participation. Finally he was able to persuade them to participate in the planned satellite project, but only as observer. Bob Kahn had argued strongly that the line from Kjeller to Tanum should have a capacity of at least 50 kbit/s, since it would transport traffic between USA and London, NORSAR and NDRE. He was afraid that a low capacity of that line might constrain the performance too much and thus bring the packet switching technology in discredit. NTA-R&D was now willing to provide free of charge two lines from Kjeller to Tanum, a 48 kbit/s line and a 9.6 kbit/s line. The higher-capacity line should be used for the SATNET experiments, while the 9.6 kbit/s line would replace the Norwegian part of the existing line between NORSAR-TIP and the SDAC-IMP in Virginia. The installation order went out on December 23, 1976. In addition, NTA permitted a satellite-IMP (SIMP) to be installed inside the Tanum ground station, and to provide free of charge the 64 kbit/s satellite uplink. This agreement was initially for one year, but was later prolonged till the end of 1980.

In connection with the Norwegian participation there were plans to attach NORSAR's two IBM-360 systems and RBK's (Regneanlegget Kjeller-Blindern) Cyber-74 to NORSAR-TIP in addition to NDRE's computer laboratory. NORSAR's two systems went on the air in 1977, about four years after NORSAR-TIP was installed, while the Cyber-system was never attached.

In 1975 Yngvar Lundh started planning the attachment of NDRE's computer laboratory to NORSAR-TIP. When Paal Spilling was back from the two-month stay at UCL he started the detailed planning on how to connect NDRE's SM-3 computer to NORSAR-TIP. Towards the summer of 1976 the connection was working. This effort will be described in Chapter 8.

Some further details of the developments in Computers and Communications were reported in [20].

7 From ARPANET to INTERNET

The initial internet concepts were published in May 1974 [21] by Vint Cerf and Bob Kahn. Based on the technology behind the three networks, ARPANET, PRNET and SATNET, all funded by ARPA, Bob Kahn got the vision of an open architecture network model: any network should be able to communicate with any other network, independent of individual hardware and software requirements. It included a new protocol, replacing the NCP used in the ARPANET, and gateways (later to be termed routers) for the interconnection of networks.

The design goals for the interconnection of networks, as specified by Kahn, were:

- Any network should be able to connect to any other network via a gateway;
- There should be no central network administration or control;
- Lost packets should be retransmitted;
- No internal changes in the networks should be needed in order to enable their interconnection.

The original paper described one protocol, called Transmission Control Program (TCP). It was responsible both for the forwarding and the end-to-end reliable transport. In the following we will describe the main events converting ARPANET into being part of the INTERNET.

The initial three contracts to develop TCP were awarded by ARPA to Stanford University (SU), where Vinton Cerf was a new assistant professor, to BBN (Ray Tomlinson) and to Peter Kirstein's group at University College London (UCL). As mentioned previously, the Stanford group was responsible for developing the initial specifications for TCP. The early implementations at SU and UCL were field-tested between one another in 1975 to support the work on the specifications. As a result of extensive testing, the TCP specifications went through several iterations. It also turned out that the TCP protocol was not modular enough to support certain protocol requirements, such as those needed for packet voice (real-time requirements). Speech traffic has sufficient redundancy, so it is far less serious to lose a speech packet now and then, than to retransmit lost packets by the TCP protocol and thereby increase the play-out delay at the receiving side. It was decided in March 1978 [22] to split TCP into two parts. One part (IP – the Internet Protocol) was responsible for the networking aspects such as addressing, routing and forwarding, while the other part (TCP – Transmission Control Protocol) was responsible for end-to-end requirements – mainly reliability and flow control. The splitting permitted the development of a simple transport protocol, called the User Datagram Protocol (UDP), interfacing with IP and living side-by-side with TCP.

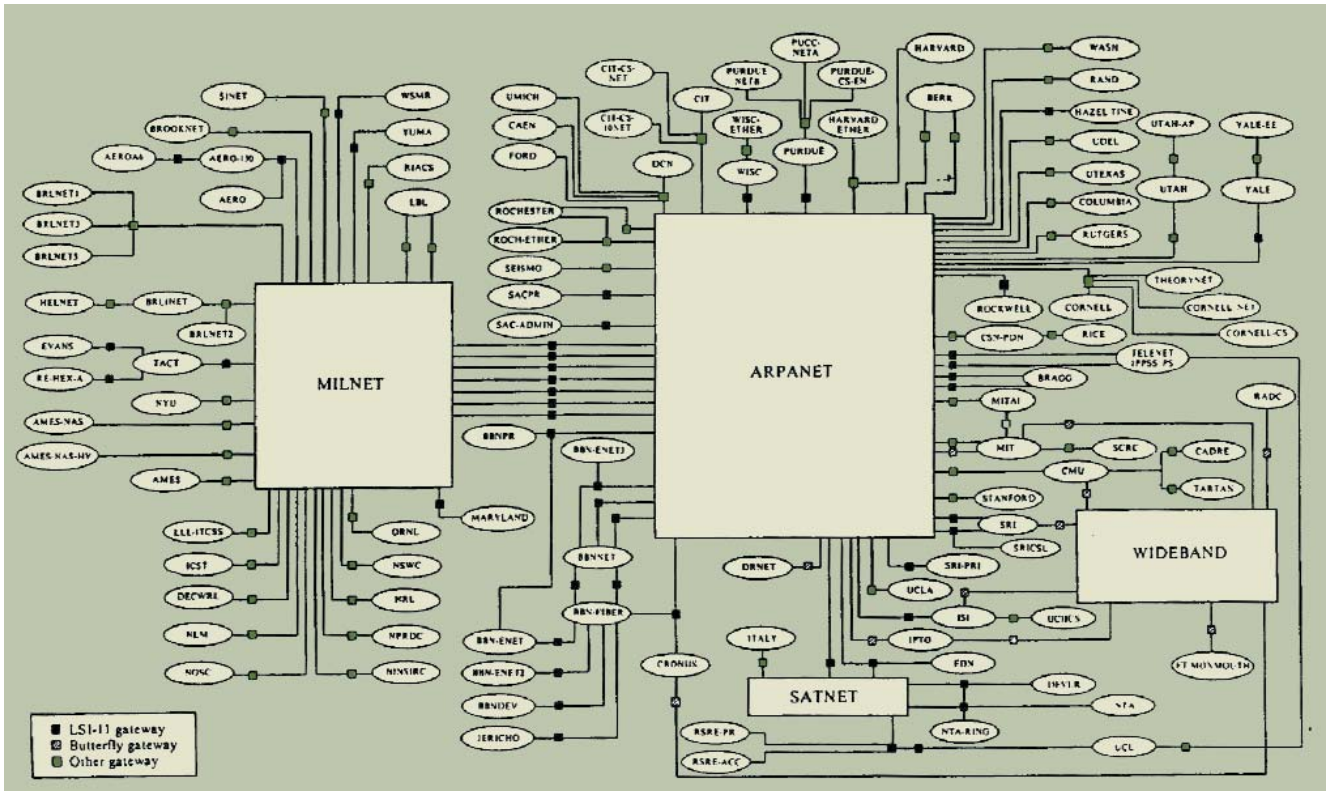


Figure 2 A map of the INTERNET around 1986

After the IP-TCP split, other parties started implementing TCP/IP under popular operating systems such as TENEX, TOPS20, and to integrate it with communication services like TELNET, FTP and email. BBN's version of TCP/IP was integrated into Berkeley's ARPA-supported work on UNIX. BBN also did the initial work to develop internet gateways.

So early in the 1980s a mature suite of protocols had been developed. Simultaneously several high-speed local area networks (LAN), graphical workstations and IP routers had been developed within the research community. All the necessary ingredients were now available to make the internet technology proliferate. In addition the SATNET experiment was completed, and the system was now being used on a semi-operational basis to interconnect ARPANET with LANs at UCL and NDRE. Two new European partners were attached to SATNET. DFVLR (German Air and Space Research Institute) was attached to SATNET in the summer of 1980, via the satellite ground station in Raisting. A little later the research institute CNUCE in Pisa, Italy, was also attached via a ground station in Italy.

ARPA now initiated a transition plan to gradually convert the communications software in all important ARPANET hosts over to the internet suite of protocols. This transition was to be completed and the transition in place by January 1, 1983. This actually happened over a period of several months in early 1983. Many of the hosts were then moved off ARPANET and reconnected to LANs, as well as many new workstations. ARPANET then served as a backbone, inter-connecting the collection of LANs, see Figure 2. This we call the *INTERNET* with capital letters. The backbone was still managed by BBN under contract with ARPA, and institutions wanting to connect to the INTERNET needed permission from ARPA. This effectively limited the growth of the INTERNET for some time. In 1983 the internet technology was sufficiently mature, and the whole community had converted to TCP/IP. This enabled the Defense Communication Agency (DCA) to split off all defense-related organizations into MILNET and integrate it with the Defense Data Network.

This was done to support non-classified defense operational requirements. The other part of ARPANET, still called ARPANET, supported the needs of the general research community. This open part was eventually taken over by National Science Foundation (NSF) (once their NSFNET had been established around 1985) and other funding parties, and regional internets were gradually privatized or new ones established.

A first step in the process was to split ARPANET into approximately two halves that were interconnected by a set of filtering routers, see Figure 2. The

filtering routers would prevent unauthorized communications initiated in the open part to penetrate into the closed MILNET. Electronic mail was considered harmless, and was permitted to flow freely both ways between MILNET and ARPANET. This was the first example of what later became known as "firewalls".

NSF agreed to fund part of the infrastructure needed in the academic environment for the interconnection of the Super-computer sites and to provide access from universities to these facilities. This infrastructure consisted of leased lines and routers. This was a more flexible, higher-speed solution than ARPANET, and the ARPANET became obsolete after a short while. It was completely phased out in 1990. The Internet's structure as we know it today started then.

This transition took away the strict control performed by ARPA regarding permission to connect to the Internet. *So from now on we spell the Internet with small letters*. Academic institutions, research organizations and even research departments belonging to industrial organizations, not only in the US but also in Europe, were permitted to connect to the Internet. NSF and European research funding shared the cost of several leased lines between key centers in Europe and USA. The initial main Internet sites in Europe were The Center for Mathematics and Computer Science (CWI) in Amsterdam and CERN in Geneva, Switzerland. "Internets" grew up in most European countries. In Scandinavia the four Nordic countries Denmark, Finland, Norway and Sweden joined forces and interconnected their growing national academic internets via NORDUnet, with the hub in Stockholm and from there leased lines to CWI in Amsterdam and to CERN. This increased the growth-rate, with the result that the whole Internet approximately doubled in size every 12–18 months, and in 5–6 years time covered 50 plus countries and more than 20 million users. This development is important, however, but not part of our presentation.

The current Internet is not owned or operated by one organization. The many pieces of it are owned by different organizations and operated in a distributed fashion. It is therefore surprising how well it does function. This must primarily be ascribed to the robustness of its protocols and routing mechanisms.

One of the motivations behind the introduction of packet switching was the need to share expensive and scarce computer resources among a large and geographically distributed set of users. Such resources could for example be editors, program compilers and debuggers, programs for scientific calculations and various database applications like document archiving and retrieval. Hence two of the early services being provided in ARPANET were TELNET and FTP. TELNET provided means to interface a local terminal with a program/application in a remote host computer and make use of the concept of communicating virtual terminal entities in the local and remote computer and a negotiation procedure to select the right terminal options. FTP provided the ability to transfer files to and from a remote computer, either in binary or character-oriented mode. In addition to these two services, electronic mail was also provided, and was soon to become the dominating service in ARPANET, and later in the Internet. During the 70s, these services were steadily refined as experience was gained in using them. These services have essentially been the same since they were standardized in 1982.

As the Internet grew in size and more stored information/documents became available online, one needed help in locating documents – searching for titles and/or keywords. Services such as "archie" and "gopher" were developed, and later "The Wide-Area Information Servers" (WAIS) permitting natural language searches among standardized database servers.

A very important boost to the Internet community was provided by the "World-Wide Web". Originally it was developed in 1989–91 by Tim Berners-Lee at CERN, the European Laboratory for Particle Physics, and released in 1991. The purpose was to provide the research staff with a universal means to disseminate text, graphical information, figures and database information. It makes use of the concept of hypertext, i.e. non-sequential text. This concept in the form of the MEMEX machine, was first described by Vannevar Bush in 1945 in the Atlantic Monthly article "As we may think", and later refined as a concept by Ted Nelson in the 1960s. It was Ted Nelson who coined the term 'hypertext'. Another early contributor to the development of hypertext systems was Douglas Engelbart, who had demonstrated working hypertext-links in a prototype application (NLS) at Stanford University in 1968.

"Clicking" on one such link to a reference emphasized in a document automatically opens a new connection to that reference. It may be anywhere in the net, possibly in another computer in a different country. The referenced document is retrieved and presented on the user's terminal, all in a seamless fashion. The "language" used is the "HyperText Markup Language" – HTML – a subset of the "Standard Generalized Markup Language" – SGML. It is used to tag the various pieces of a document, so as to specify how it should be presented on the screen.

Two popular front-end clients – "browsers" – emerged. One called Mosaic, distributed free of

charge from the National Center for Supercomputing Applications – NCSA, in 1993/94. The other, Netscape, was the commercial version of Mosaic. These Web browsers have turned out to provide the most viable service on the Internet, and really set off an explosion in the traffic volume.

8 The Norwegian contributions to the Internet development

This chapter provides a short summary of the Norwegian contributions to the collaboration with the ARPA community.

The first Norwegian "host" on ARPANET

The part of the computer laboratory at NDRE relevant to our packet switching activities consisted of a computer named SM-3, produced by Kongsberg Våpenfabrikk. It had 64 kB of memory, one punch-card reader, a paper tape reader, and a fast line printer. No hard disk, no network interface, and no operating system. Programs were written on IBM punch cards, then assembled and linked, and the debugged and loadable version punched out on paper tapes to ease later loading of working programs. Assembler, linker and loader also had to be read in from paper tapes.

The first task for Paal Spilling was to make a multi-tasking operating system for SM-3. Via the collaboration with ARPA, he got hold of a technical report describing the ELF operating system, developed at SRI for the PDP-11/45 [23]. This was a good guideline and a good help in the understanding of a multi-tasking system. The PDP-11/45 was a much more advanced system than the SM-3. Hence only the rudiments of the ELF functional constructs were applicable. After a substantial period of trials and errors, Spilling finally had a robust and reliable multi-tasking system, with process scheduling, process-to-process communications, buffer management, and interrupt handling.

The work to get the SM-3 computer connected to the ARPANET node at NORSAR started around summer 1975, but was interrupted for two months by Spilling's visit to UCL. The physical distance between the computer laboratory at NDRE and NORSAR-TP required us to make use of the "Very Distant Host Interface" (VDH-interface) on NORSAR-TIP [24]. An SM-3 VDH-interface was built. Paal Spilling had to design and program the corresponding driver. The driver was integrated with the newly developed multi-tasking system. After an intensive debugging phase, the interface was operating correctly and reliably. As a final test of the VDH-interface and the multi-tasking system, Spilling performed a set of round-trip time measurements between SM-3 and a number of nodes in the ARPANET. These results were consistent with similar results performed elsewhere. Spilling also measured some specific features of the NCP protocol. The NCP protocol was running in all ARPANET nodes at that time, and was responsible for fragmentation of messages into packets at the entry node and reassembly of the packets into the original message at the destination node before presentation of the message to the attached host.

TCP implementation and testing

As mentioned previously, Paal Spilling stayed at University College London (UCL) in September and October of 1975. There he had the pleasure of participating in the first transatlantic tests of TCP. The tests were conducted between two independent implementations, one done at Stanford University (Professor Cerf's group) and one at UCL (Professor Kirstein's group). It was very exciting to observe that these two implementations were able to establish connections and exchange data, after some hectic debugging. Later, to demonstrate the robustness of the TCP protocol, the line between LONDON-TIP and NORSAR-TIP was taken down for 10 – 15 minutes in the midst of transferring data between UCL and Stanford. Then the line was brought up again, and the two ends of the TCP connection continued happily the transfer from where they had stopped, without losing data.

Back at NDRE, Spilling and a colleague started the implementation of the early version of TCP [21, 25] on the SM-3 computer. After about half a year of work, it was decided to stop the implementation and move over to a more modern system – the Norwegian NORD-10 computer. Looking back, this was probably not a good decision. It would have been better to complete our implementation to get the satisfaction and experience in fulfilling this task. Starting to work with the NORD-10 computer, it turned out to be more difficult than expected. We had to get acquainted with a new operating system, design and build a new VDH-interface, and implement the driver under the new operating system (SINTRAN). This was not a trivial task. Then, starting on the design and implementation of TCP in the SINTRAN operating system turned out to be difficult too. SINTRAN had a very primitive process-to-process communications (signaling) system. If a process received two signals, one after the other, the first one was overwritten and lost. Hence this was useless, and we had to invent some hacks to circumvent the problem. It was also next to impossible to convince the software group at Norsk Data, responsible for the SINTRAN operating system, that this was an important deficiency of their operating system. It took a few years before that was appreciated and corrected. But then it was too late for

us. In addition SINTRAN was a complicated system, and made the implementation of TCP cumbersome. And after some time the work had to be stopped, unfortunately.

The SATNET project

In the time period 1976 through 1979, NDRE was heavily involved in the development of SATNET. The purpose of the SATNET project was to explore the feasibility of operating a 64 kbit/s SPADE channel in the INTELSAT system, common to a set of ground stations, in a packet switched modus. As mentioned previously three ground stations were involved in the project, one at Etam in West Virginia on the US East Coast, one at Goonhilly at the English West Coast, and the third one at the Nordic satellite ground station at Tanum in Sweden, see Figure 3. To enable the packet-switched operation of the satellite channel, so-called Satellite-IMPs (SIMPs) were installed in the ground stations – interfacing with the SPADE channel equipment [26]. Each SIMP was then interconnected, via a leased line, with a gateway computer – the ETAM-SIMP with a gateway at BBN in Boston, the Goonhilly-SIMP with a gateway at UCL, and the Tanum-SIMP with a gateway at NDRE. The other interface of each gateway was connected to an ARPANET node. As mentioned previously, the line from NDRE to Tanum and the satellite uplink were kindly offered free of charge by the Norwegian Telecommunications Administration (NTA) for the duration of the project. The capacities of the lines between the SIMPs and the gateways were in the order of 50 kbit/s.

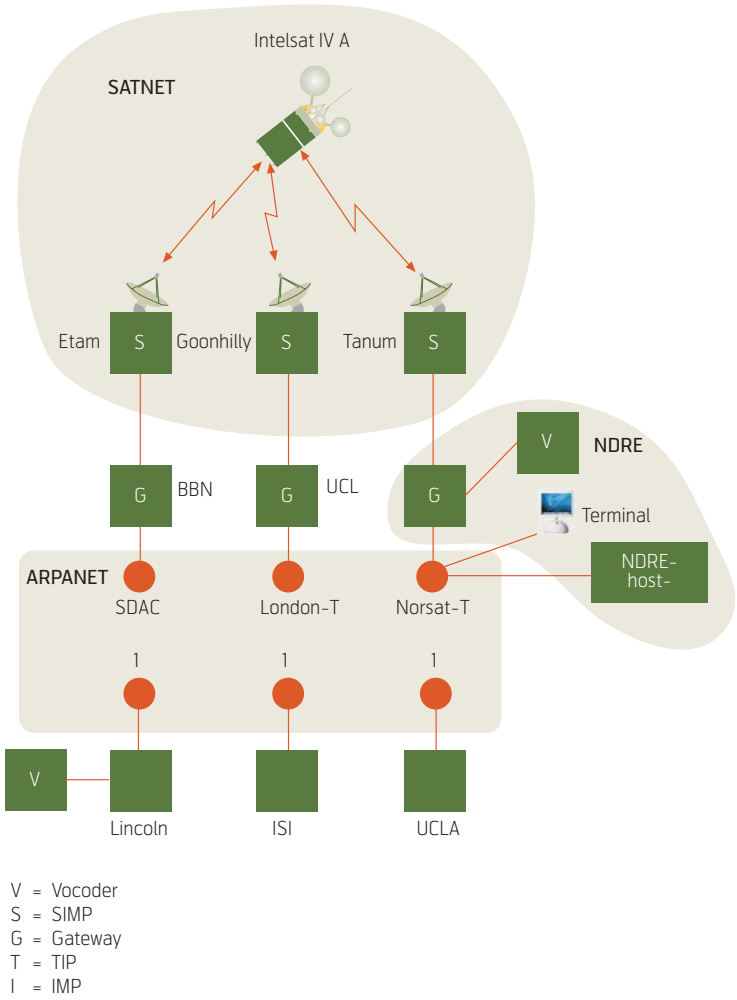


Figure 3 Network configuration for the SATNET and packet speech experiments

The SATNET research program, organized by Bob Kahn much as Larry Roberts had done for ARPANET, was performed as a joint effort between Linkabit Corporation in San Diego, University of California in Los Angeles (UCLA), Bolt, Beranek and Newman (BBN) in Boston, Communications Satellite Corporation (COMSAT) in Gaithersburg, Maryland, University College London (UCL), and NDRE in Norway [20]. Linkabit had the project's technical leadership. BBN was responsible for the development of the SIMPs, including the various channel-access algorithms the participants wanted to test out. The project participants met about four times a year, with the meeting location circulating among the participating institutions.

The Norwegian contingent was headed by Yngvar Lundh, with **Finn-Arve Aagesen** and Paal Spilling as work force. Aagesen was responsible for performing simulation studies of the most promising channel access algorithm, the "Contention-based, Priority-Oriented Demand Access" algorithm (CPODA). Paal Spilling developed management software on the SM-3 computer, to control artificial traffic generation in the SIMPs, and to fetch data collected in the SIMPs during measurements. Since the SM-3 computer did not have any storage medium, the easiest solution was to dump the measurement data out on the fast line printer. The analysis of the measurements then had to be performed by hand. Several access algorithms were studied experimentally, among others TDMA, Reservation-TDMA, and C-PODA [27, 28]. Measurements and simulations were also performed by Kleinrock's Network Measurements Group at UCLA. Mario Gerla was a key person here.

Packet speech experiments and demonstrations

NDRE participated in packet-speech experiments performed in 1978 – 1979 in collaboration with, among others, MIT Lincoln Laboratories just outside Boston. The packet speech activity was part of the SATNET project. Lincoln Lab had developed a speech vocoder (voice coder and decoder), under contract with ARPA, providing a stream of 2.4 kbit/s digitized speech. The

vocoder was interfaced with the PDP-11/45 at NDRE, with a similar arrangement at MIT Lincoln Lab, see Figure 3, and later also at UCL. The PDP-11/45 acted then not as gateways, but as speech packet assembly at the sending side and packet disassembly and play-out via the vocoder at the receiving side. In addition the PDP-11/45 contained a conference control program, developed by Lincoln Lab, that handed over the "floor", in a FI-FO queue manner, to the parties indicating their whish to talk.

Paal Spilling performed a set of measurements to examine the profile of packet-speech traffic [29]. The programming of the PDP-11/45, performed in connection with the experiments, is a good example of resource sharing, one of the driving forces behind the early development of packet switching. The computer was located close to Spilling's office. The programming tools and the source code for the conference control program was located at a TOPS-20 machine at Information Sciences Institute (ISI) in Los Angeles. Using a terminal in his office connected to NORSAR-TIP (see Figure 3), Spilling could log on to the computer in Los Angeles, write the necessary modifications to the control program, and have it compiled and loaded across the network into the PDP-11/45 next door. The downloading was facilitated by a cross-network debugger (X-NET), also in the TOPS-20 machine, and enabled Spilling to debug the modified control program loaded into the PDP-11/45. This was an exciting and fascinating experience.

NDRE participated in several packet-speech demonstrations. At one of the regular project meetings, held at UCL, Yngvar Lundh made use of the conference facility and could participate in the meeting from Norway simultaneously with someone at Lincoln Lab, i.e. three-way Internet speech conference. The quality of the speech when compressed to 2.4 kbit/s was noticeably impaired, but packet transmission through this early Internet connection worked fine in real time.

A large packet-speech demonstration in 1983 is worth mentioning. Paal Spilling had then moved over to NTA-R&D. Speech traffic was exchanged between an airplane-carrier in the Pacific Ocean off the Californian coast and NTA-R&D at Kjeller. The communications went via a PR-network, involving the airplane-carrier, an airplane flying in the vicinity of the ship, and a PR-node at SRI attached to the INTERNET, then across the INTERNET to the East Coast, then across SATNET via the Tanum ground station to NTA-R&D. The purpose of the demonstration was to show high-ranking military personnel onboard the airplane-carrier the feasibility of communicating through an interconnection of different networks (subnets), the plethora of sophisticated techniques involved, and the usability of the Internet for communicating data and digitized speech.

In 1979/80 Paal Spilling had leave of absence from NDRE, and stayed with SRI International in Menlo Park, California working on the ARPA-funded Packet Radio Network (PRNET). There he made a proposal to improve the software architecture in the PR-nodes [30] to have a better logical layering of the program structure, performed experiments on packet-speech performance with QoS-control [31], and suggested a "Busy Tone" method to overcome the "hidden terminal" problem in wireless communications [32].

Spilling was back at NDRE in the last quarter of 1980. SATNET was now considered operational, and used to interconnect local networks at UCL and NDRE with ARPANET. UCL was also using it for the total academic service traffic between the UK-SRCnet and ARPANET. Spilling made a proposal to the Research Council of Norway, and obtained funding for purchasing an LSI-11/23 computer. Through his ARPA connections he got all necessary software from Dave Mills [33] at the University of Delaware. This software package had the nickname "Fuzzball", and contained the TCP/IP suite of protocols. The LSI-11/23 was interfaced with a Proteon Token-Ring network, together with the PDP-11/45. Hence this computer was the first real Norwegian Internet host. Spilling had obtained a class B network address (128.39.0.0) for this network, from the Network Information Center (NIC).

Spilling and Lundh had no luck in convincing the management of NDRE to continue the packet-switching effort. Apparently the internet technology, including Packet Radios, was too premature both for the management of NDRE and for the Norwegian Defense.

9 The first Norwegian Internet

Paal Spilling left NDRE in the summer of 1982 to start working for the Research Department of NTA. Being inside NTA, this enabled Spilling to create the first Norwegian internet and make that a part of the Internet.

NDRE showed no interests in exploiting the knowledge and experience obtained in the collaboration with the ARPA community. Paal Spilling therefore attempted to create interests among the research people at NTA-R&D, in spite of Yngvar Lundh's previous attempts. As a result of this attempt, Spilling was invited to move over to the R&D-department by one of its research supervisors. He did so at the end of the

summer of 1982, with the hope to be able to strengthen the Norwegian effort. Unfortunately this did not happen – with one exception. We will come back to this below.

In agreement with ARPA and NDRE, all internet-related equipment was moved over to NTA-R&D, so that the collaboration could continue from this new location, but now without participation from NDRE. NTA was at that time a state-owned monopoly. Being inside NTA gave Spilling several opportunities. The first action was to purchase a VAX-750. Through his ARPA connection Spilling got permission to acquire the Berkeley version of the UNIX operating system (4.1 bsd), although NTA-R&D had to pay a significant fee to ATT. The VAX was connected to a Proteon Ring network, the same with the PDP-11/45 gateway to SATNET. In getting the UNIX system up and running, Spilling had very good help from Helge Skrivervik, and from Tor Sverre Lande at the Department of Informatics at the University of Oslo.

From 1983/84 there was a growing interest at NTA-R&D in getting access to the Internet services. The paradox was that this interest did not result in any interest in collaborating with the Internet community. There was also a desire to get access to the ATT-version of UNIX, hence a Pyramid machine running both versions of the UNIX operating system was purchased and installed. The Ringnet was now replaced by an Ethernet. In addition the PDP-11/45 gateway to SATNET was replaced by a Butterfly machine from BBN (both hardware and software), on loan from ARPA. We now provided terminal access to these two UNIX machines, so everyone in the research staff could get access to the standard Internet services. We also bought a few SUN and PERQ workstations, but they were at that time too expensive to be for everyone. In a few years' time the whole lab was Ethernet-cabled and most people had their own workstations.

There was also a growing interest at the universities in Oslo, Bergen and Trondheim, to get access to Internet services. Being inside NTA-R&D, this enabled Spilling to set up a 9.6 kbit/s line to the Department of Informatics at the University of Oslo. The line was terminated at the VAX at NTA-R&D and at the VAX at Department of Informatics. The SLIP protocol was used here to interconnect the two machines. Later, in 1984/85, we installed own-developed gateways (routers), based on equipment from Bridge Communications Inc, where Judy Estrin was technical director – a former student of Vint Cerf at Stanford and with whom Paal had carried out the first TCP tests a decade earlier. Through this connection we were able to get the software development tools

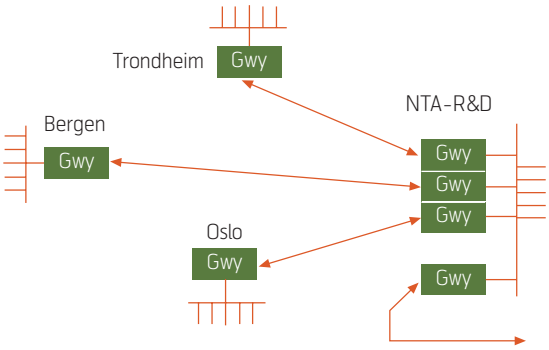


Figure 4 The Norwegian Internet around 1986

used by Bridge Communications. This enabled **Kjell Hermansen**, the only person besides Spilling at NTA-R&D interested in cooperating with the ARPA community, to develop an IP-router package for the Bridge boxes. When the router was operating reliably, Spilling requested NTA-R&D to set up lines to the universities in Oslo, Bergen and Trondheim – and with our home-built routers at each end of the lines. A few years later the Bridge routers were replaced by commercial ones from Cisco. This was the first Norwegian academic internet, see Figure 4. It was managed for some years by Spilling. UNINETT, the academic network operator funded by the Department of Education, took over the network in 1987/88, after a substantial pressure from a strong group of academic users requiring Internet access.

In the beginning there was no Domain Name Server. The mapping between host names and addresses was done via a local host file in each computer. The original host file was maintained by SRI-NIC. Every new host on the Internet had to be reported to NIC. And at regular intervals (daily or so), one had to pull over the host file from NIC and install it at the right place in the Unix file system. The development of DNS in Norway started in 1985. Jens Thomassen at the Department of Informatics at the University of Oslo was responsible for maintaining the Norwegian domain (.no), on behalf of NTA-R&D.

It is worth mentioning that there was an ongoing debate for some time in the UNINETT community whether the transport network should be based on the IP protocol, international standards like X.25, or even DECNET. This continued for probably a few years, until it was obvious that internet communications was the salient technology to focus on.

10 Other technology drivers

In addition to the ARPA-sponsored research, other very important research activities gave a significant collective momentum to the development of the Internet. In the following we will mention those which we feel are the most important ones.

Bob Metcalfe outlined in his PhD thesis at Harvard the basic idea for Ethernet. It was an extension of the Aloha concepts developed by Norman Abramson at the University of Hawaii. After obtaining his degree, Metcalfe joined *XEROX-PARC* in 1973, the research center of XEROX in Palo Alto. He had the vision of "small" powerful personal workstations (Altos) interconnected in high-speed local networks called *Ethernets*. It is highly likely that this vision was inspired by Doug Engelbart's work on the NLS-system. The initial work on Ethernet was demonstrated in 1973 [34]. A set of such Ethernets would later be interconnected into a company-wide network, spanning XEROX offices scattered all over continental US. The personal workstations should have the capability to be moved from one network and plugged into another one, following its owner on his/her journey around in XEROX. Therefore each workstation was provided with a unique Ethernet address, which could be used to locate and address its owner. When ARPA initiated the work on TCP/IP, XEROX-PARC joined in the experimentation and participated in the development process for some years until the internet concepts were mature enough for them to specify and implement their own internet protocol suite. In 1977/78 XEROX had implemented a company-wide network interconnecting about 25 Ethernets via a set of gateways and leased lines. Amazingly XEROX, as a commercial company, did not see any business potential in this technology at that time.

A little later, in 1979, IEEE started the standardization work on local-area networks (LAN). XEROX, DEC and INTEL (*DIX*) joined forces and proposed a 10 Mbit/s Ethernet standard. Due to the large amount of fielded Ethernet interfaces and the significant weight of the DIX effort, the hardware/physical layer part of the international standard was made compatible with the DIX proposal [35].

Another very important contribution to the usability and popularity of the TCP/IP suite of protocols was the further developments and refinements of the *UNIX* operating system done initially by Bill Joy at *Berkeley*, also under contract with ARPA. The hardware platforms used were VAX-750 and VAX-780. Part of this work included the integration of the TCP/IP suite of protocols, developed by BBN, into the UNIX system [36]. To make the protocol package user-friendly, flexible and efficient, it was found necessary to develop a set of support functions surrounding the protocol suite. It was also relatively easy to make device drivers for a variety of popular network interfaces, such as Ethernet, Proteon-Ring, ARPANET and X.25.

This work was very successful and has been an example for very many other implementations of the protocol suite in other operating systems. Since only a small part of the UNIX system depends on the computer's hardware, it was relatively straightforward to port the system to other hardware configurations. The Berkeley version of UNIX was made available to universities free of charge and, for a fee to AT&T and approval from ARPA, to non-educational research organization, not only in the US but also in Europe.

The flourishing of TCP/IP implementations at many places in the US and Europe provided a large basis for field tests and experimentation. Hence TCP and IP and the accompanying services like TELNET, FTP and email were steadily refined and improved. It was therefore a mature, efficient and user-friendly set of protocols that ARPA proposed as a standard for the American Defense in the middle of 1980 and approved as a standard in mid 1982.

In the late 70s and early 1980s *Stanford University* had an ARPA supported project to develop a means of supporting VLSI design. It was to be connected to a nascent *Stanford University Network*. Here, under the leadership of Forest Baskett and Andres Bechtolsheim, they developed a Motorola-68000-based CPU card and a frame buffer based display system. As a spin-off from this project, a few people left to start a small company called *SUN Microsystems* which also involved Bill Joy from Berkeley. They ported an early version of the Berkeley version of UNIX onto this hardware and developed a graphical window user-interface to go with it, and offered this as a commercial product under the name *SUN-1* in 1982. This was supported by ARPA and managed by Bob Kahn. A little later another group formed a small company called *CISCO*, offering IP routers based on the same CPU card from SU, also with support from Bob Kahn at ARPA.

11 From Resource Sharing to Information Sharing

One driving force behind the development of packet-switching networks was to share expensive resources like computers, software, and communication facilities, as efficiently as possible among a large group of users. With the proliferation of PCs and services like the web, the resource sharing got another meaning – the sharing of information.

At the time when this evolution started, computers were expensive, software programs were expensive, and communication lines were expensive. Much of the initial research focused on bridging the geographical distances between the users at their terminals, and the computers with the valuable programs they wanted to use. Hence those expensive resources needed to be shared among as many users as possible, and in such a way that the users should not need to physically move to the computers to have their jobs done. When email came into practical use, this became the dominating service – still rather centralized, but accessible from geographically distributed terminals. It is worth mentioning that Yngvar Lundh and Paal Spilling had their mailboxes on a host computer at Information Sciences Institute (ISI) in Los Angeles, and accessed from terminals connected to NORSAR-TIP. Keeping the resources centralized made the administration and maintenance of the resources straightforward and easy.

Then came the period with affordable minicomputers, where sets of terminal users were clustered around geographically distributed but networked computers. Computers were still too expensive to be affordable as single-user workstations.

Later, with the proliferation of PCs, each user got enough computing power for the daily work. The resources were now distributed, which resulted in a more cumbersome administration and maintenance of these resources. In essence it was only the networks that were shared among the users.

When the Web was developed and commercially available (1991/92), coinciding in time with the lifting of the restriction on commercial usage of the Internet, we saw an explosion in geographical coverage, in number of users, and in traffic volume. Now the Internet gradually turned into an information sharing network. Information could now be stored at any location in the network. With the use of powerful search engines and the Web with the hyper-links, the Internet is acting as a gigantic repository of information transparently accessed by the user via point-and-click in the Web browser.

12 Epilog

Why was it so difficult to create interest among Norwegian (and European) computer scientists for this new ARPANET/Internet technology? And did Norway benefit from its participation?

In December of 1973, half a year after NORSAR-TIP was installed, a Norwegian ARPANET committee was established to promote and coordinate possible Norwegian participations in ARPA activities. It consisted of members from the Research Council of Norway, NORSAR, RBK, NTH (later NTNU), and NTA-R&D. A condition for connecting to NORSAR-TIP, or making use of its terminal service, was that this should contribute to the furthering of the technology and be beneficial to both ARPA and NDRE. The committee encountered two main problems. One was the uncertainty about the future of ARPANET. Larry Roberts had left the IPTO office in 1973, and Dr. J.C.R. Licklider took over as director. He needed time to be informed and make a decision regarding the future of ARPANET. Hence it took a while before the future was known. In July 1975 the continental part of ARPANET was transferred to the Defense Communications Agency (DCA), while IPTO continued to be responsible for the international engagements. The other problem for the committee was the various upcoming European competing communication initiatives, similar to what Donald Davies experienced in England. The committee did not come up with any constructive proposals, and dissolved itself in 1975.

This is just one example of the difficulty in creating research interests for the internet technology. For many years the Internet and its concepts were neglected by the telecom operators and the European research community. Due to the tremendous popularity, growth and global coverage of the Internet, the traditional telecom operators had more or less unwillingly been forced to accept reality and offer Internet access. We will attempt to shed light on some of the factors contributing to this effect.

Competitions between alternatives

Simultaneously with the growth of the ARPANET in the 1970s, we saw the emergence of other similar competing communication concepts, like CYCLADE [37] in France presented by Pouzin in 1973, the European Informatics Network (EIN) [38] presented in 1976, and the CCITT's Orange Books [39] containing the X.25 standards published 1977. In 1983 the International Standards activities presented the "Reference Model for Open Systems" [40] and then in succession a set of layered communication protocols. The dominating feature of X.25, and the ISO standards in general, was the virtual circuit principle, in contrast to the flexible packet and datagram modes of operation in the ARPANET and later the Internet. A virtual circuit in the packet switched arena is the equivalent to end-to-end circuit established in the telephone network. The dominant part of the Norwegian research community, including NTA-R&D was for a long time convinced that packet communications had to be based on virtual circuits.

The TCP/IP family of protocols was expected to be replaced by international standards

The management at NDRE and NTA-R&D, and the Norwegian communication research community at large did not believe in the Internet technology before the end of the 80s and the beginning of the 90s. In general most communication experts believed that the TCP/IP suite of protocols eventually would be replaced by internationally agreed standards. So when we attempted to create interests for participation in the further development of this technology, the responses were negative.

Skepticism to defense matters

At the time Norway entered the development activities, there was a general antipathy against everything connected with defense matters, and especially with US defense. This antipathy was of course not reduced by the fact that participation in the activities and usage of the communications network had to be approved by ARPA.

International standards development

The national authorities and the academic communities believed strongly in international standards. It was relatively easy to obtain funding for participation in standards activities. This was less committing/ demanding than the hard practical work that went on in the ARPA community. Standards were worked out on paper within a study period of four years, and when ready accepted more or less without practical experience. Later, when standards were to be implemented and tested out, deficiencies were surely detected and the standards had to be revised, and then re-proposed as standards in the next study period. In contrast the ARPA research went via implementations, testing, modifications, more testing and refinements, and when mature enough and sufficiently stable, finally adopted as standard. This included also a set of support functions like “name to address mapping”, “service type to service-access-point mapping”, and “IP address to MAC address mapping”, to make the combined protocol layers work efficiently and user friendly.

When the ISO standards came out in the middle and latter half of the 1980s, after a substantial work effort, a set of standards had been defined for each layer in the reference model. These standards included many options. Before the standards could be implemented, one had to make use of workshops to prepare and agree on the options to use in practice.

This took quite a while. It is worth mentioning that the options agreed upon made the ISO standards, for all practical purposes, functionally equal to the Internet protocols.

Agreed international standards were not openly available. They had to be purchased for a certain fee. In contrast, all Internet protocol specifications and related documentations were freely available.

The American dominance

A very important contribution to the usability and popularity of the TCP/IP suite of protocols was the refinements of the *UNIX* operating system done at *Berkeley*. This work included the integration of the TCP/IP suite of protocols into the UNIX system. This work was very successful, and has been an example for very many other implementations of the protocol suite under other operating systems. The Berkeley version of UNIX was made available to universities free of charge and, for a fee, to AT&T and approval from ARPA, to non-educational research organization, not only in the US but also in Europe. Unix was for years the dominating system used in higher education and in research. Hence a whole new generation of higher educated people were exposed to and influenced by Unix and internet technology and services.

The development of computers and their software was dominated by American companies. The TCP/IP suite of protocols was part of the software systems delivered with the computers. There was no incentive by the software companies to spend money on implementing other standards, unless someone paid for it. So this was also a major factor gradually making the TCP/IP suite of protocols a de facto communications standard.

Did Norway benefit from the Internet cooperation with the ARPA community?

The opportunities provided to Spilling when he moved over to NTA-R&D in mid 1982 enabled him (1984/85) to establish a small Norwegian Internet, interconnecting informatics departments at the universities in Oslo, Bergen and Trondheim, and NTA-R&D. The network at NTA-R&D was interconnected, via the Nordic satellite ground station at Tanum and SATNET, with ARPANET in the US. This implied that influential academic research people could make use of Internet services and communicate with colleagues in the US, and experienced that this form of communications worked well and provided a set of reliable, effective, and attractive services.

A few years later (1987/88) the academic communications network UNINETT was established, interconnecting, in the first instant, informatics departments at the main Norwegian universities. As mentioned pre-

viously, there was a debate in the UNINETT community regarding transport network technology – X.25, DECNET, and IP. But gradually the experience with and the increasing desire to use Internet services paved the way for UNINETT to provide this kind of communications. In the beginning as some sort of hybrid network, but later converted to a pure IP-based network when the situation had ripened sufficiently. UNINETT gradually evolved to encompass all higher educational institutions in Norway.

The knowledge and experiences gained in participating in the ARPA projects led to the establishment of a computer communications research group and an early curriculum in computer communications at the Department of informatics at the University of Oslo. This effort were initiated by Yngvar Lundh and Paal Spilling, and gradually led to the establishment of similar activities at all universities in Norway.

In the late 80s, all Nordic countries had installed academic Internets. The operating organizations combined forces and created NORDUnet. It interconnected the academic networks in Norway, Sweden, Finland, and Denmark, with the main hub in Stockholm. From there leased lines were installed to CERN and CWI in the Netherlands and later directly to the US.

The Nordic countries have the highest percentage of Internet users in the world. This is in part due to the early exposure to this technology, first in the academic world and later in the public sector. NTA (Telenor) was among the first telecom operators, around 1994/95, in Europe to be convinced to offer Internet access to customers. This is certainly due to the close exposure to this technology over a long period of time at NTA-R&D.

Acknowledgements

The authors are very grateful to Bob Kahn and Peter Kirstein for providing valuable comments on various parts of this article.

References

1 Kleinrock, L. Information Flow in Large Communication Nets. *RLE Quarterly Report*, July 1961.

2 Kleinrock, L. *Communication Nets*. NJ, USA, McGraw-Hill, 1964.

3 Baran, P. On Distributed Communication Networks. *IEEE Trans. On Communication Systems*, CS-12, March 1964.

4 Baran, P et al. *On Distributed Communication networks*. RAND Corp., 1960–1962. (Series of 11 internal reports.) URL: http://www.rand.org/MR/baran.list.html

5 Davies, D W et al. A digital communication network for computers giving rapid response at remote terminals. *ACM Gatlinburg Conf.*, Gatlinburg, TN, USA, 2.1–2.17, Oct. 1967.

6 Davies, D W. Communication networks to serve rapid response computers. *IFIP Congress*, The Hague, Netherlands, August 1968.

7 Licklider, J C R. Man-Computer Symbiosis. In: *IRE Transactions on Human Factors in Electronics*, HFE-1, March, 1960. URL: http://memex.org/licklider.pdf

8 Licklider, J C R, Clark, W. On-Line Man Computer Communication. *AFIPS Conference Proceedings*, 21, 113–128. New York, Spartan Books, 1962.

9 Marill, T, Roberts, L. Towards a Cooperative Network of Time-shared Computers. *AFIPS Fall Conf.*, 29, 425–432, Oct. 1966. New York, Spartan Books, 1966.

10 Roberts, L. Multiple Computer Networks and Intercomputer Communication. *ACM Gatlinburg Conf.*, Gatlinburg, TN, USA, Oct. 1967. (Original ARPANET Design Paper)

11 Engelbart, D C, English, W K. A Research Center for augmenting human intellect. *AFIPS Conference Proceedings of the 1968 Fall Joint Computer Conference*, San Francisco, CA, Dec 1968.

12 Carr, C S, Crocker, S, Cerf, V. HOST-HOST Communication Protocol in the ARPA Network. *AFIPS Proc. of SJCC*, 36, 589–597. Montvale, NJ, AFIPS Press, 1970.

13 Tomlinson, R. The first network email. October 21, 2004 [online] – URL: http://openmap.bbn.com/~tomlinso/ray/firstemailframe.html

14 Roberts, L G, Wessler, B D. Computer network development to achieve resource sharing. *AFIPS Conf. Proc.*, 36, 543–549. Montvale, NJ, AFIPS Press, 1970.

15 Kahn, R E. The Organization of Computer Resources into a Packet Radio Network. *IEEE Transaction on Communications*, COM-25, 169–178, January 1977.

16 Abramson, N. The Throughput of Packet Broadcasting Channels. *IEEE Trans. On Comm.*, COM-25 (10), 1977.

17 Kahn, R E. The Introduction of Packet Satellite Communications. *IEEE National Telecommunications Conference*, Washington, DC, 45.1, December 1979.

18 Private correspondence with Professor Peter Kirstein, June 2004.

19 First ARPANET public demonstration, organized by Robert Kahn of BBN, at the ICCC Conf in Oct, 1972.

20 Lundh, Y. Computers and Communication – Early Development of Computing and Internet technology – a Groundbreaking Part of Technical History. *Telektronikk*, 97 (2/3), 3–19, 2001.

21 Cerf, V, Kahn, R. A Protocol for Packet Network Interconnection. *IEEE Trans. Comm. Tech.*, 647–648, May 1974.

22 Cerf, V G, Postel, J, Cohen, D. IP – TCP split in 1978. Private communication.

23 Retz, D. *Structure of the ELF Operating System*. Paper presented at the National Computer Conference, AFIPS, 1974.

24 BBN. *Specifications for the Interconnection of a Host and an IMP*. Cambridge, MA, USA, Bolt, Beranek and Newman, May 1978. (Report N 1822)

25 Spilling, P, Cerf, V G, Kirstein, P. A Study of the Transmission Control Program, a Novel Program for Internetwork Computer Communications. *Nato Grant*, 1149, 1976.

26 Jacobs, I M, Binder, R, Hoversten, E V. General Purpose Packet Satellite Network. *Proc. IEEE Spec. Issue on Packet Communications Network*, 66, 1448–1467, 1978.

27 Spilling, P, Lundh, Y, Aagesen, F A. *Final Report on the Packet Satellite Program*. Kjeller, NDRE, 1978. (Internal Report E-290, NDRE-E.)

28 Chu, WW et al. Experimental Results on the Packet Satellite Network. *NTC-79*, Washington DC, November 1979.

29 Spilling, P, McElwain, C, Forgie, J W. *Packet Speech and Network Performance*. Kjeller, NDRE, 1979. (Internal Report E-295, NDRE-E.)

30 Spilling, P. *Low-Cost Packet Radio Protocol Architecture and Functions – Preliminary Requirements*. Menlo Park, CA, SRI International, October 1980. (Technical Note 1080-150-1.)

31 Spilling, P, Craighill, E. Digital Voice Communications in the Packet Radio Network. *ICC-80*, Seattle, WA, June 1980. Also available: SRI International, 1980. (Packet Radio Technical Note 286.)

32 Spilling, P, Tobagi, F. *Activity Signaling and Improved Acknowledgements in Packet Radio Systems*. Menlo Park, CA, SRI International, 1980. (Packet Radio Technical Note 283.)

33 Mills, D L. The Fuzzball. *Proc. ACM SIGCOMM 88 Symposium*, Palo Alto, CA, August 1988, 115–122.

34 Metcalfe, R M, Boggs, D R. Ethernet: Distributed Packet Switching for Local Computer Networks. *Comm. ACM*, 395–404, 1976.

35 IEEE. *Standard for Information technology – Telecommunications and information exchange between systems – Local and metropolitan area networks – Specific requirements, Part 3*. New York, IEEE Press, 1985. (ANSI/IEEE Std 802.3-1985.)

36 *Berkeley UNIX 4.2 bsd, a complete rewrite of the original AT&T UNIX version.*

37 Pouzin, L. Presentation and major design aspects of the Cyclades computer network. *3rd IEE/ACM Data Comm. Symp.*, St Petersburg, Nov. 1973.

38 Porcet, F, Repton, C S. The EIN communication subnetwork principles and practice. *Proc. ICCC*, Toronto, 523, Aug 3–6, 1976.

39 *Series X recommendations*. The Orange Book, ITU, Geneva, 1977.

40 ISO. *Basic Reference Model*. Geneva, ISO, 1983. (ISO 7498-1.)

Paal Spilling obtained his cand.real. degree in physics from the University of Oslo in 1963. In January 1964 he started on a PhD program at the University of Utrecht, The Netherlands, and finished his degree in summer 1968 in experimental low-energy neutron physics. Subsequently he joined the nuclear physics group at the Technical University in Eindhoven, The Netherlands. In January 1972 Spilling started to work for the Norwegian Defence Research Establishment (NDRE) at Kjeller, and from end 1974 he got involved with the ARPA-funded research program in packet switching and internet technology under the leadership of Yngvar Lundh. He was in 1979/80 visiting scientist with SRI International in Menlo Park, California, where he worked on the Packet Radio program. Spilling left NDRE in August 1982 to join the Research Department of the Norwegian Telecommunications Administration (NTA-RD), also at Kjeller. Subsequently he established the first internet outside USA, interconnected with the US internet. At NTA-RD Spilling worked with communications security and the combination of internet technology and fiber-optic transmission network. In 1993 Spilling became professor at Department of informatics, University of Oslo and University Graduate Center at Kjeller.

email: paal@unik.no

Yngvar Lundh is Electrical Engineer from the Norwegian Institute of Technology, 1956. He worked as research engineer at Norwegian Defence Research Establishment until 1984, then as engineer in chief at NTA/Telenor. From 1996 he has been running his own consulting company. He was guest researcher at MIT 1958–59, with Bell Labs 1970–71, and part time professor of computer science at the University of Oslo from 1980. Lundh concentrated on development of new technologies in electronics, computing, automation and telecom. He initiated and headed research groups for that and was dedicated to the possibilities for new industry in Norway. Computer production (Norsk Data) and digital switching ("Node technique" STK/ Alcatel), electronic mail in Norway and other enterprises were among the results.

email: Yngvar@joker.no

Why and how Svalbard got the fibre[1)]

ROLF SKÅR

Rolf Skår is Director General of the Norwegian Space Centre

The Norwegian ground station for polar orbiting satellites at Svalbard, SvalSat, has the best location in the world for serving owners of such satellites. Satellite communication was a prerequisite for establishing the station in 1997. However, in 2002, the future of SvalSat was threatened due to the fact that it was not connected to the global fibre network. In record-breaking short time, and in cooperation with Telenor, the Norwegian Space Centre managed to finance and finalize the project connecting Svalbard to the mainland by fibre-optic cable for the benefit of SvalSat, scientific institutions, and the society at large in Longyearbyen.

Introduction

With all due respect to coal mining, tourism and all other activities at Svalbard, it was Svalbard's role in space activities that got Svalbard the fibre.

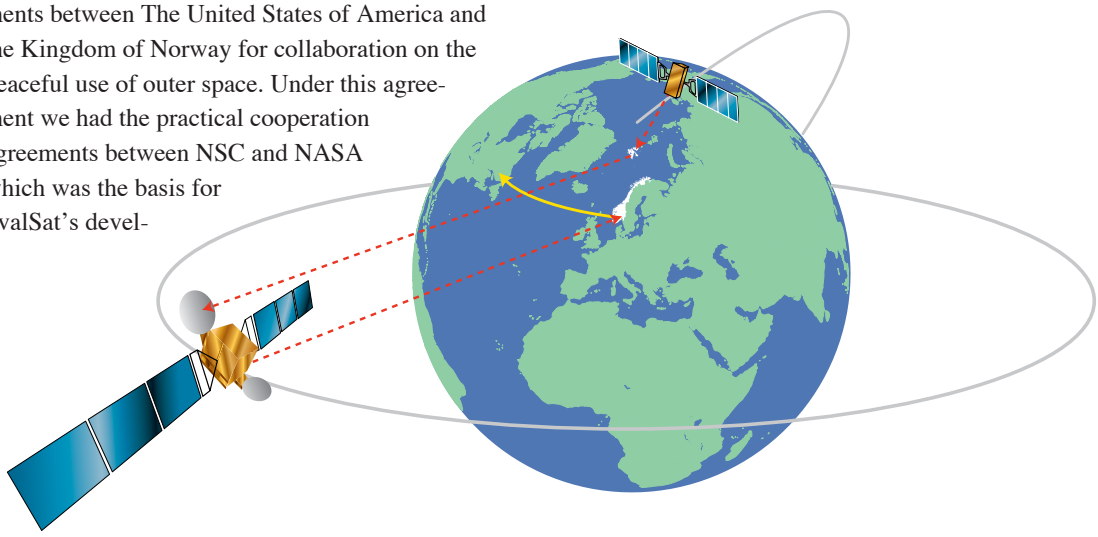
It is a simple geographical fact, supported by an airport and an enjoyable community with a mild climate that makes SvalSat the best location in the world for supporting polar orbiting satellites. SvalSat can see all the orbits.

The short version of why Svalbard got the fibre is that the Norwegian Space Centre (NSC) felt that SvalSat's future was threatened when Northrup Grumman/Raytheon for NPOESS did not select Svalbard for its global network of 15 ground stations. They instead selected Helsinki because Helsinki was connected to the global fibre network. To put it mildly, we were not happy.

We were fortunate to have very high-level agreements between The United States of America and the Kingdom of Norway for collaboration on the peaceful use of outer space. Under this agreement we had the practical cooperation agreements between NSC and NASA which was the basis for SvalSat's development, and 2002 saw the signing of the first agreement between ourselves and IPO/NOAA for SvalSat's role in the future next generation US polar orbiting weather satellites, the NPP and NPOESS.

Abbreviations used in the text

NSC	Norwegian Space Centre
NPP	NPOESS Preparatory Project
NPOESS	National Polar-orbiting Operational Environmental Satellite System
NASA	National Aeronautics and Space Administration
IPO	Integrated Programme Office
EUMETSAT	European Organisastion for the Exploitation of Meteorological Satellites
ESA	European Space Agency
UNIS	University Studies at Svalbard



Until 2004 the data from polar orbiting satellites were acquired by SvalSat, then sent via geostationary satellites directly from Svalbard to SvalSat customers on the East Coast of the US

1) *This article is based on a speech given by Rolf Skår at the inauguration ceremony of the fibre cable in Longyearbyen on January 31, 2003.*

ISSN 0085-7130 © Telenor ASA 2004

Even more importantly, over many years the traditional good relations and friendship between USA and Norway had developed into a friendship and mutual trust between individuals who were in a position to shape the future.

I have been criticized for working against the interests of space, or specifically against satellite communication, by promoting fibre instead of communication via geostationary satellites. Let me simply say that it was because of satellite communication that SvalSat was developed; it was very reliable, we never lost a pass due to satellite communication problems, and it has served the Svalbard community very well since 1979.

Communication costs were high

However, SvalSat's future was in need of a more competitive solution, and I would simply like everybody to hear me say that I will promote and fight for space based solutions when they provide the best solution for the users, and only then.

The story of fibre to Svalbard began by both Telenor and NSC doing feasibility studies in early 2002. Both studies had the same conclusion: A sub sea fibre was feasible; the recommended route was from Tromsø via Bear Island to Spitzbergen.

Both studies concluded that a system would cost around 50 mill USD or 400–500 mill NOK for one cable with satellite communication as backup. This led to the formal process of Telenor Svalbard and NSC informing all parties with relevant space interests of this possibility and inviting them to a stakeholders' meeting in Longyearbyen 24–25 July 2002. Here the Telenor Group met with decision makers from NASA and IPO, plus EUMETSAT and ESA.

As a result of this meeting we had several telephone conferences with the stakeholders and another meeting was planned for November 1 at IPO in Washington DC.

Then the project collapsed, first NPOESS selected Helsinki and then the Telenor Group withdrew.

Was the game over?

No, this is when the real game started, and this is the more interesting story of how Svalbard not only got the fibre, but how it got 32 fibres in record time.

We went to America and during the three days from October 30 to November 1, we laid the groundwork. Based upon the interest from the feasibility studies and stakeholders' meetings, we were approached by a very hungry telecom fibre sub sea cable industry.

On October 30 we visited one such company – it was not the one finally chosen – who convinced us that they could do a turnkey installation of one cable from Tromsø to Longyearbyen for 40 mill USD. Some other suppliers even indicated much lower prices, below 30 mill USD.

If there is one date to mark how Svalbard got the fibre it is October 31, 2002. At 10 a.m. we met with NASA's Bob Spearing at NASA Headquarters. He was the architect of the implemented solution. Bob told me that NASA could not invest in a fibre to Svalbard; however they could pay for a data transmission service from Svalbard to the US. Then I gave Bob an offer NASA could not refuse, and it truly reflects the mutual trust, the friendship and the common interests which had developed since 1995 when SvalSat was planned as a joint effort between NASA and NSC. I offered NASA that they should pay its current cost for satellite communication from SvalSat to the US, around 6 mill USD per year, for a few years until 20 mill USD in Net Present Value had been accumulated, and we would provide a 25 year transmission service with six times the current bandwidth. Among friends and partners you can be generous, therefore I offered 25 years – not 10 or 15, and six times the bandwidth – not double.

When the meeting was over, Bob arranged for me to meet Sean O'Keefe, the NASA administrator, and we talked about Svalbard and about Norway, and I invited him to come to Norway.

The same afternoon I met with John D. Cunningham from IPO, we met alone, and that was when Svalbard got the fibres. John was enthusiastic and eager to replace Helsinki with Svalbard. He promised to work with NASA and provide another 20 mill USD paid over a few more years than NASA because their bandwidth requirements were lower, but the value of Svalbard compared to Helsinki was worth the 20 mill USD he agreed to pay.

On November 1, the stakeholders' meeting did not develop as planned. First Telenor (the satcom people) announced that Telenor did not wish to invest in a Svalbard cable because of its negative economic impact for the Telenor Group.

Back in business

Then I announced that NSC would like to invest in a cable based upon an understanding between NASA, IPO and ourselves. We had some indications that

there was indeed a buyer's market for sub-sea fibre cables, which we wished to benefit from.

We then started to rush the project and went into a higher gear. On November 12, the Board of NSC gave me the go-ahead based on the NASA and IPO – NSC understanding. On November 15, I met with Jan Edvard Thygesen, Senior Vice President of Telenor and responsible for all Telenor networks.

A very important 'yes' to the project was his answer. And more than that, he was enthusiastic and promised full support from Telenor Networks and Telenor Svalbard. On November 18, the main terms of Telenor's role were agreed to.

Cable at low price

We rushed out the Invitation To Tender (ITT) on December 21. It had been prepared by us and a large group of Telenor experts and was for a turnkey end-to-end transmission system from Tromsø to Longyearbyen, with 40 mill USD available and vendor to choose route and technology, and an option was left whether 40 mill USD was sufficient for two independent cables forming a ring.

The vendors were asked to do the so-called Desk-Top-Study, select the route and then do the detailed survey of the selected route. The ITT deadline was February 3. Early January we discovered that to go from Tromsø to Longyearbyen was too risky due to the lack of protection from trawlers.

Another rush job: we extended the tender due date to February 25, and we did the Desk Top Survey with good help from the Harstad company Seaworks together with Telenor. We also decided to do the detailed survey for the new route going from Harstad over Andøya to Longyearbyen. The area outside Andøya is the only trawler free zone between Kirkenes and Trondheim and therefore the ideal place to get from land into deep water above 1600 m water depth.

When we opened the tenders on February 25 we had a huge surprise. We could get two completely independent cables forming a ring for around 40 mill USD.

However, there was also a major disappointment: None of the companies accepted payment from NASA and IPO over seven years. So we decided to recompete with only the four vendors that we believed could do a turnkey solution, preferably a ring, for 40 mill USD.

The recompetition resulted in a clear winner, Tyco, and on March 7, we announced to all bidders that Tyco had been selected for contract negotiation, and we asked Tyco for their Best and Final Offer for a ring with two independent cables.

During contract negotiations early April we had some 25 telecom experts, including seven lawyers on our side, we worked very long hours to complete the very detailed turnkey contract to mirror our ITT and Tyco's Best and Final Offer.

We signed the contract on April 14; however, it would only come into force when financing was in place and we had all the necessary permissions.

We committed 300,000 USD to Tyco's early planning and ordered the detailed Survey, without which we could not get the permissions, and together with Tyco we set ourselves a deadline on May 15 to get financing in place and another deadline on July 1 to get all the permissions.

Financing was to become more difficult than anticipated and the week starting with my birthday, May 13, proved to be very dramatic. Tyco had proposed a specialist US financing company, Hannon Armstrong, to arrange the financing both for the construction period and also for them to purchase the yearly revenue stream from NASA and IPO as payment for their transmission service.

One challenge was that NRSE (Norsk Romsenter Eiendom), the legal customer, a 100 % owned subsidiary of Norwegian Space Centre, a foundation, did not have any meaningful equity, nor any cash left after paying for the Detailed Survey.

I used all my tricks; I invited the decision makers for the very best of the 'Huset' dinners; Huset being the famous restaurant at Longyearbyen. I invited them on the best of snowscooter safaris to Barentsburg and to Isfjord Radio. I did not succeed.

The most dramatic week

We left Svalbard on May 15 without an agreement, and Stan Kramer and myself signed an extension of the deadline until May 19, which we also missed.

Hannon Armstrong wanted a government guarantee. I told them that only the National Assembly could authorize such. Finally, by involving the Director of Public Prosecutions[2)] and the Minister for Trade and Industry, a compromise was accepted. Mr. Ansgar

2) *In Norwegian: Riksadvokaten*

Representatives from IPO, NASA and Raytheon (antenna provider) are leaving the foundation of the IPO antenna at SvalSat on June 24, 2003. They are clearly satisfied with what they have seen

Gabrielsen wrote a letter to Hannon Armstrong whose text they had agreed to beforehand. From then on the project was unstoppable. We only needed the final permissions.

Did we take too high risks? I believe I understood the risks involved, and certainly the rewards. The offer from Tyco depended upon doing the project during 2003. It was probably our only chance to get two cables instead of one. We had invested around 10 mill NOK; the majority of this was the Desk-Top-Study and the order for the Detailed Survey which started May 19.

I was so convinced that we would get the necessary permits that Tyco also believed this to be the case, even more so when on May 21 we committed 2 mill USD in a more serious downpayment to Tyco. So after May 21, the serious work started.

To give you a flair for Norwegian bureaucracy, I will give you two real examples of how fast it can be: At Andøya we needed a building of some 140 m^2 to house the power-feed equipment, a total investment of around 5 mill NOK. We contacted the owner of the land on May 7, on May 13 we agreed on the price and the same day we sent an application to the local authorities for permission to use farmland for our purpose and for building permits. We got all the approvals needed by May 22, construction work started the next day and the building was ready to receive its first cable on July 25.

On June 17, Sean O'Keefe took Helle Hammer, the State Secretary of the Ministry of Trade and Industry, together with a few people from NSC on his NASA plane to Svalbard, accepting our invitation to visit Norway and Svalbard. During the visit I learned that the Detailed Survey necessary for our application for the main Government permit was ready. I signed the application for permit and personally handed it in to the Governor's office. Four days later we had the critical permissions and the following week we returned to Svalbard to celebrate and sign the final contract documents. The trust that Tyco, Telenor and ourselves had put into the project was truly rewarded.

The project was well under way to fully benefit from the Arctic summer with midnight sun and continual daylight. Look at these milestones: The complete wet plant, 2,700 km of cables and 40 repeaters were ploughed and laid at the bottom of the sea from July 21 until August 15; 25 days using two of the world's most advanced specialist ships and setting a world record for the deepest water depth for ploughing, 1671 meters.

The good planning was rewarded

The project team that completed this project in record time did a fantastic job. Torbjørn Dyb from Telenor

ISSN 0085-7130

The final metres of the fibre are put in place by the cable ship on the sea floor reaching Spitzbergen in August 2003

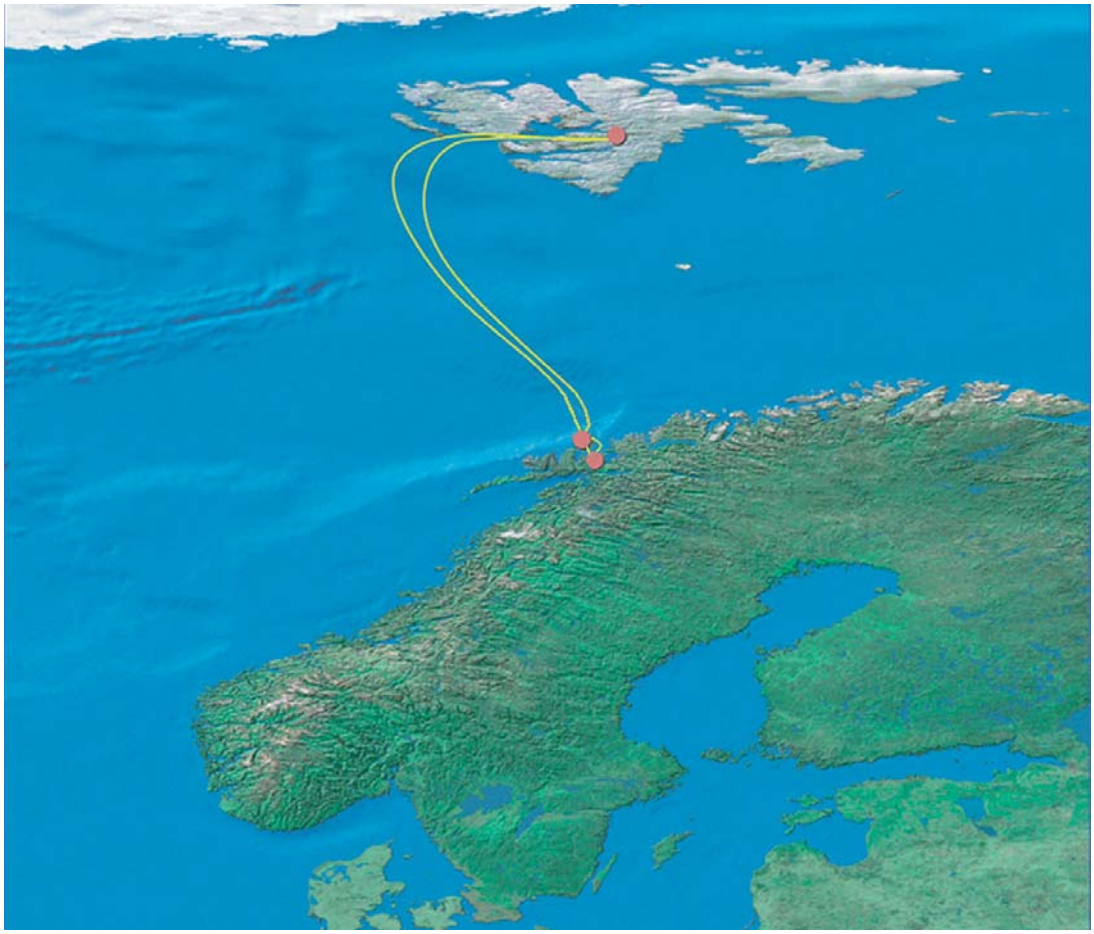
was our Project Manager leading a team of Telenor engineers, and from Tyco were Debbi Brask and Dave Willoughby.

There were no surprises, no cost overruns. On April 14, we signed the conditions of turnkey contract with Tyco. The agreed price was the final price, not a cent in additional charges. And we got a little more than we asked for.

Our plan was one cable with two fibre pairs, four fibres, and 10 Gbit/s. Tyco offered one cable with eight pairs and one cable with two pairs in contract negotiations, and with some gentle arm twisting and some goodwill what is now installed are two fully independent cables, each with eight pairs fully repeatered with 40 repeaters, and with double initial capacity, 20 Gbit/s on each cable. The system may be expanded to 2500 Gbit/s on each cable. We believe this to be sufficient for any foreseen or even unforeseen future need.

I am convinced that it is the local community that will benefit most from having two cables instead of one with satellite communication as back-up. Satellite communication would be a very expensive solution to be paid for by Telenor Svalbard and thereby by its customers.

It was Telenor Svalbard through their agreement to cover the costs exceeding 40 mill USD that made it possible to get two cables. These additional costs will have to be paid by the local users, and we have agreed that this will be done over the first six years, so after 2009 there should be a significant decrease in the cost of using the fibre cables.



Two independent cables, each with very large capacity, now connect Svalbard with mainland Norway and the international fibre network

I am pleased that we have been able to find a new pricing model for selling really broad broadband. There is 20 Gbit/s in each cable, so by using the pricing model in mainland Norway, almost all this capacity would be unused. This new model will benefit the science community most. They will have a 1 Gbit Ethernet connection to the science community in the rest of the world, and Svalbard will have the best communication system of any higher learning or research establishment in Norway. I promised this as our gift to UNIS on its 10 year anniversary.

What we are really looking forward to at the Norwegian Space Centre is for all the data that will now be downloaded from Earth science and weather satellites at SvalSat to actually be used by the scientists at UNIS and in Norway.

There is a unique opportunity that with unlimited bandwidth, raw data at full sensor fidelity may be downloaded from the sensor with 400 Mbit/s or more, and through the fibre network made available to scientists.

Rolf Skår (63) is Director General at the Norwegian Space Centre (NSC) in Oslo. He holds an MSc in Cybernetics from the Norwegian University of Science and Technology (NTNU) in Trondheim from 1966. He has been engaged at the Norwegian Defence Research Establishment (NDRE) at Kjeller, and in 1967 he was one of the co-founders of the Norwegian Computer Company Norsk Data A.S in which he held several management positions and CEO from 1978 to 1989. From 1990 to 1992 he was General Director of the Norwegian Council for Scientific and Industrial Research (NTNF) after which he entered the position as President of the consulting company Norconsult International. From 1994, he has been engaged in space research. He is the Norwegian delegate to the European Space Agency (ESA) council and from 1998 he holds the position as Director General of NSC.

Rolf Skår has extensive experience in management of large and complex information technology projects, as well as from international sales and marketing management, including sales of large information technology systems. He is an active participant in various government policy studies, in particular related to space, science and technology policy.

 ISSN 0085-7130

Technical solution and implementation of the Svalbard fibre cable

EIRIK GJESTELAND

Eirik Gjesteland is an Engineer at the Operational Centre of Telenor Networks

Longyearbyen in Svalbard at 78° North is an ideal location to support polar orbiting satellites. All 14 daily passes of a typical sun synchronous orbit are visible with satellite contact every 7 and 17 minutes.

SvalSat was developed in the period 1996–1999 as a joint effort between NASA and the Norwegian Space Centre. It has been supporting NASA earth observation satellites through an Intelsat satellite at 1° W. The available transmission rate is around 55 Mb/s and in addition there are some ISDN lines and one 2 Mb/s line. SvalSat has now grown to be the world's largest polar ground station to serve earth orbiting and polar weather satellites.

A fibre optic telecommunications cable between the Norwegian mainland and Svalbard now replaces the Intelsat connection. It was deployed in 2003, and put into operation in January 2004. The article provides an overview of the technology used together with explanations of the different building blocks of the complete transmission system. The cable and additional equipment has an expected lifetime of 25 years and the current capacity utilisation is 10 Gb/s. Telenor Networks is responsible for the technical implementation of the project, as well as the operation and maintenance.

Introduction

This article provides a technical overview of the fibre optic cable and the necessary components to make up a complete transmission system. The items covered are the cable technology itself, the repeater systems including the power feeding, as well as the monitoring and maintenance systems.

The cable system

Overview

The fibre optic cable system goes between Harstad and Longyearbyen via Andøya. The US Company Tyco has delivered the system. Tyco has its roots in the well-known corporation Bell. There are two separate cables for redundancy; only one is carrying traf-

Figure 1 The SvalSat Earth station is placed at Platåberget, outside Longyearbyen, Svalbard

ISSN 0085-7130 © Telenor ASA 2004

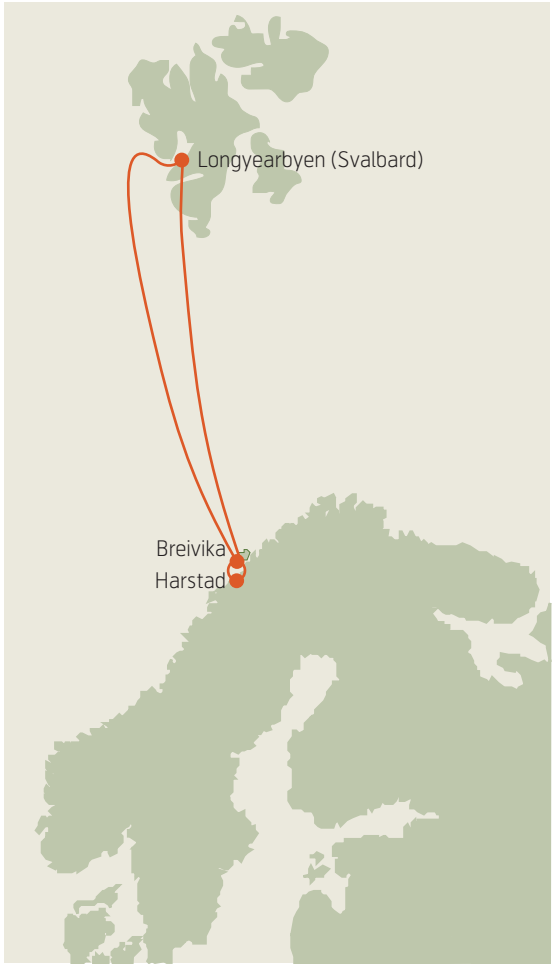


Figure 2 Two separate fibre optic cables are laid between Harstad on the Norwegian mainland and Longyearbyen on Svalbard. It goes via Breivika on the island of Andøya, before leaving the North Norwegian coast

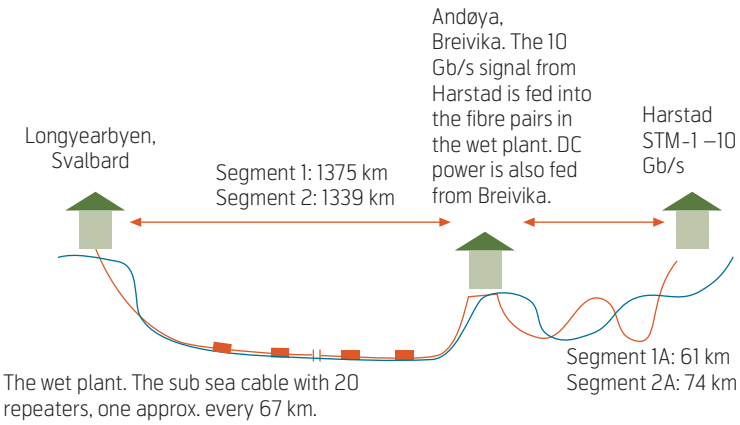


Figure 3 The two separate cables; Segment 1 and Segment 2, are 1375 and 1339 km, respectively. They are partly buried under the sea floor. The cables between Breivika and Harstad (1A and 2A) are both sub sea and land cables with a length of 61 and 74 km, respectively

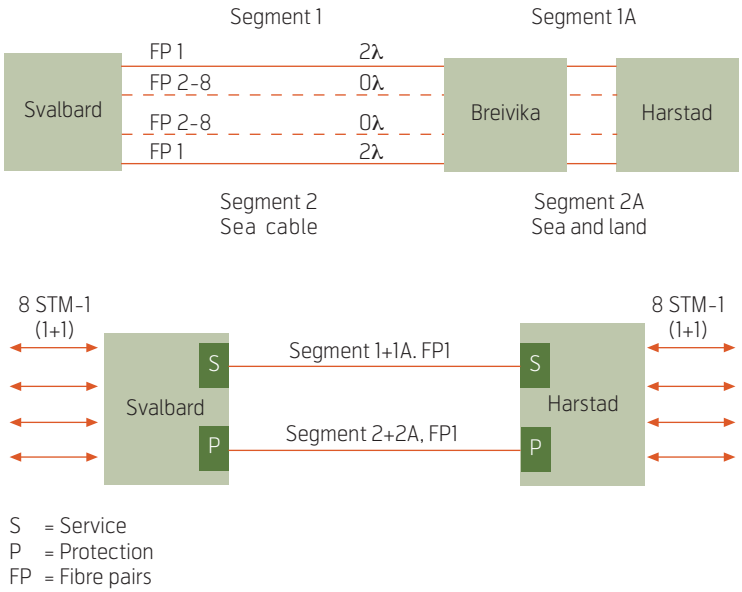


Figure 4 The fibre cable segments contains eight fibre pairs, of which only one is currently used

fic at a time. Each cable has 20 optical repeaters, spaced approximately 67 km apart, to keep the signal power up on the more than 1300 kilometre travel between Andøya and Svalbard. The repeaters are actually amplifiers, which do not regenerate the signal but amplify it. DC power to the repeaters is fed from Andøya. Each repeater has eight amplifier pairs (one per fibre pair) and a 'high loss loop back' for monitoring.

The two separate cables are called Segment 1 and Segment 2. Segment 1 is 1375 km with 172 km buried at an average depth of 2 m below the ocean floor to keep the cable safe from external hazards. Segment 2 is 1339 km and is buried for 173 km at the same average depth as Segment 1. Between Breivika and Harstad, the cables are called Segment 1A and 2A, both sea and land cable is used. Segment 1A is 61 km and Segment 2A is 74 km long. These sections are so short that optical repeaters are not needed.

System configuration

Of the eight available fibre pairs in each cable segment, only one is currently used. To use the remaining pairs, additional equipment must be installed in Svalbard and Harstad. One fibre pair has more than enough capacity for today's traffic load. If Segment 1 should fail, the traffic will automatically be switched to Segment 2.

The fibre optic cable

Different types of cable are used on the way to Svalbard. The type of cable used depends on the risk of damage. Close to shore where the risk of damage from fishing equipment and ships' anchors is significant, the cable has a thicker armour to protect the

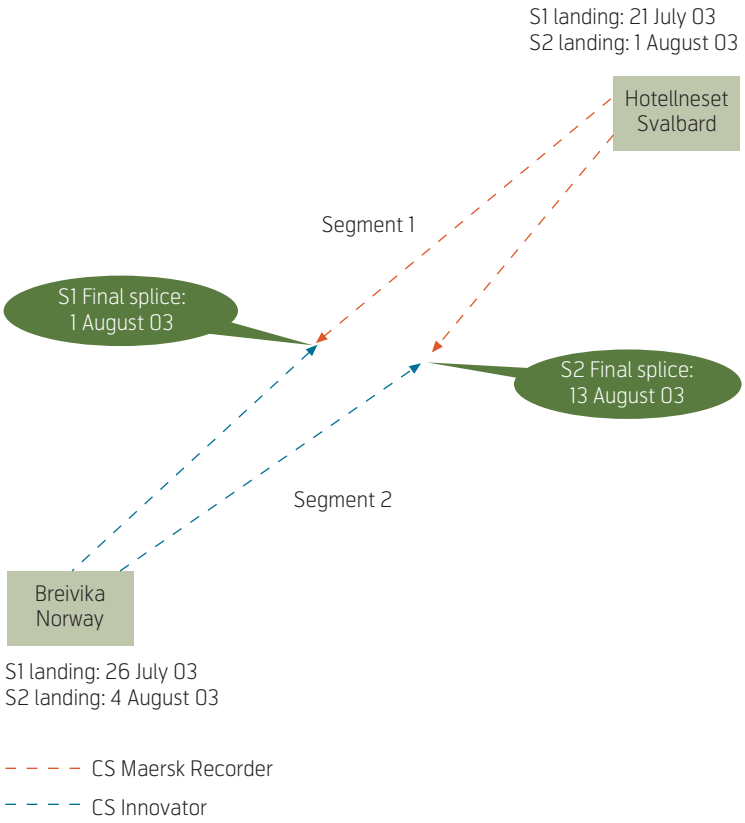


Figure 5 The two cable ships deployed the cable from each end and met midway where the ends were spliced

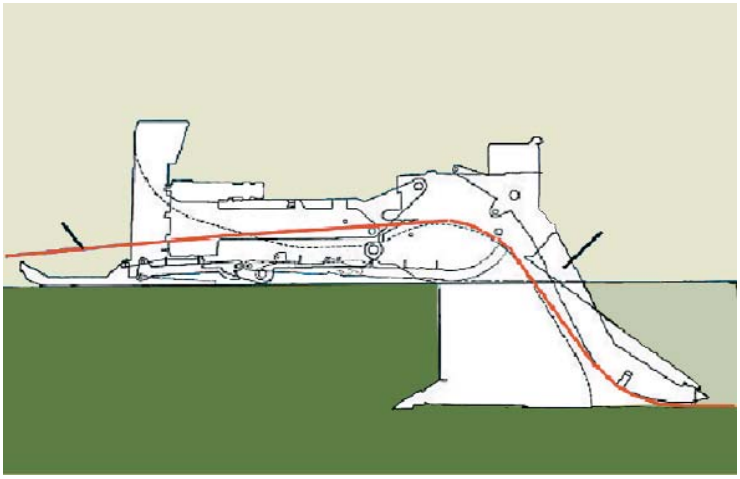


Figure 6 Example of the type of plough used to bury the cable

fragile fibre pairs. The cables are also buried in the ocean floor close to shore where the water is relatively shallow, and a special plough is used to do so. Two cable ships were used to lay the cables because one ship alone could not hold the amount of cable necessary to cover the distance. One ship started from Longyearbyen and one from Breivika, and they met midway to splice the two ends. The two ships then returned to lay the second cable, see Figure 5.

The different cable types used are from the SL 21 family shown in Table 1. The SL 21 provides the

Figure 7 The cable ship 'Cable Innovator' covered the distance from Breivika to the point where it met the other cable ship, 'Maersk Recorder', to splice the two ends

Figure 8 The cable ship 'Maersk Recorder' covered the distance from Longyearbyen to the point where it met 'Cable Innovator', to splice the two ends

optical path for the undersea digital transmission, supervision signals and a power path for the optical amplifiers.

It is not only the armour that varies, but also the type of optical fibre. In this system three different fibre types are used:

- Large mode field fibre (LMF)
- High dispersion shifted fibre (HDF)
- Non dispersion shifted fibre (NDSF)

The different fibre types affect the signal in different ways, so to minimise distortion the design includes different fibre with different characteristics.

The optical repeater

There are 20 optical repeaters on each segment, 40 in total. There is a repeater approximately every 67 kilometres. The repeaters have eight optical amplifier pairs each, one per fibre pair, and operate on DC power fed from Breivika. The system is what we call a 'single end feed system'. Each repeater also contains a 'high loss loop back' (HLLB) used for moni-

toring purposes. The Tyco repeaters use state-of-the-art optical technology to achieve high performance and reliability. The repeater amplifies multiple channel signals on multiple fibre pairs and is designed for a lifetime of 25 years on the ocean floor and on board cable ships during laying, burying and repair operations.

The repeaters are designed using the minimum number of components necessary. In this way, the risk of failure is minimised. Each repeater contains eight amplifier pairs to accommodate the eight fibre pairs. Each amplifier pair consists of the components necessary to support all wavelength division multiplexed channels on a fibre pair. The maximum number of wavelengths that may be used is 32. Each amplifier pair contains two erbium doped fibre amplifiers (EDFAs) that provide broadband amplification on two transmission paths in opposite direction. Each amplifier also has its own power supply and operates independently so a failure on one amplifier pair will not affect the amplifier pairs on the other fibre pairs.

Figure 13 shows one amplifier pair, including the laser pump unit (LPU) and the loop back coupler module (LCM). The LPU includes four 980 nm laser pump modules. The output is combined using wave-

Cable type	Application	Features
LW	Benign, sandy bottom. Depth to 8000 m	Core cable, light protection. Medium-density polyethylene white jacket, providing improved visibility when submerged
SPA	Somewhat rocky bottom; Risk of shark attack Depth to 6500 m	Metallic tape and high-density polyethylene white outer jacket applied over core provide additional abrasion protection and hydrogen sulfide protection
LWA	Rocky terrain; moderate risk of trawler damage. Depth to 1500 m Typically used for burial (depth to 1200 m buried)	Light armor wire layer applied to core cable[a]
SA	Rocky terrain; high risk of trawler damage. Depth to 1000 m (depth to 800 m buried)	Heavy armor wire layer applied to core cable[a]
DA	Very rocky terrain; high risk of trawler damage; moderate risk of abrasion. Depth to 400 m	One armor wire layer applied to LWA cable[a]

[a] Tar-soaked nylon yarn is used for the outer jacket of armored cable.

Table 1 SL21 cable types

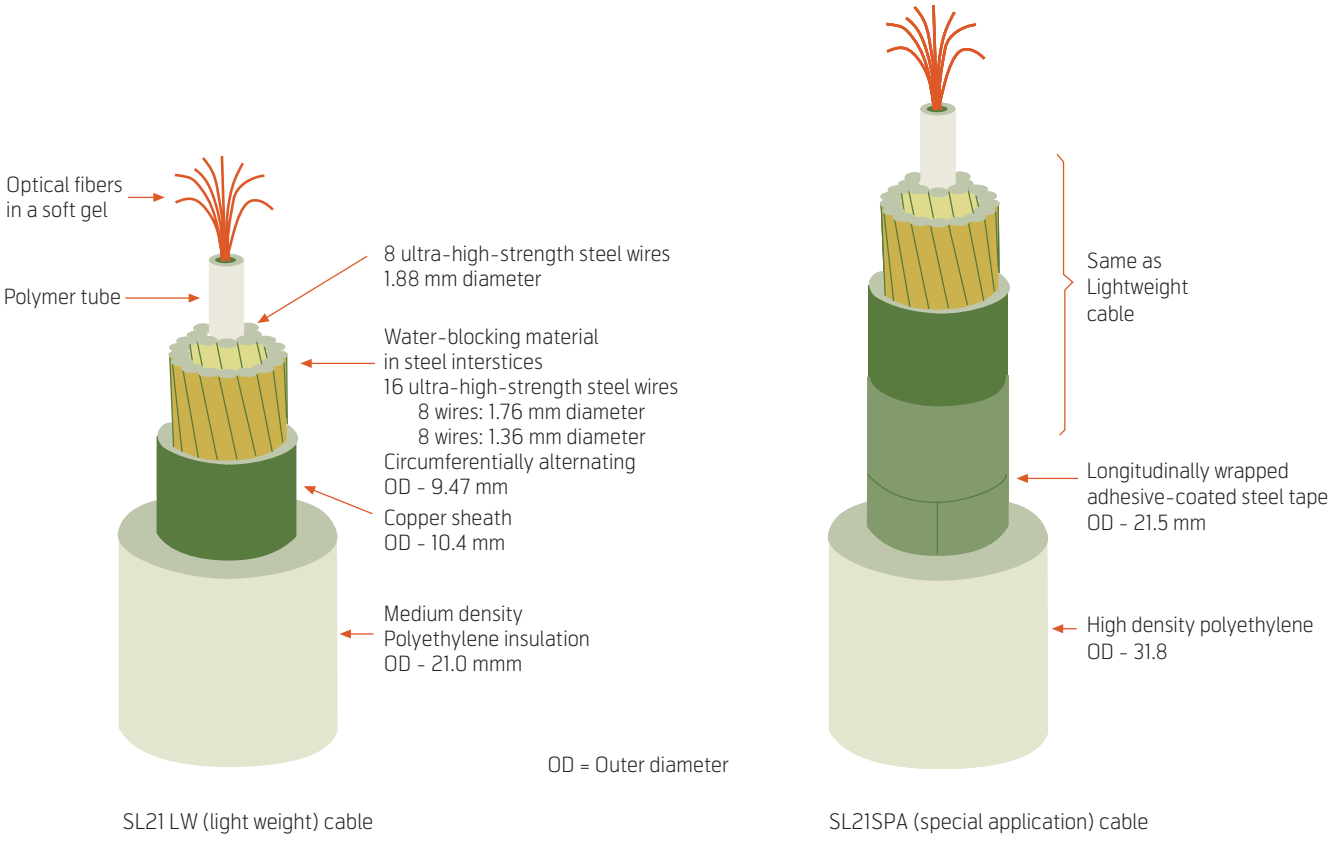


Figure 9a The cables used are from the SL21 family of cables

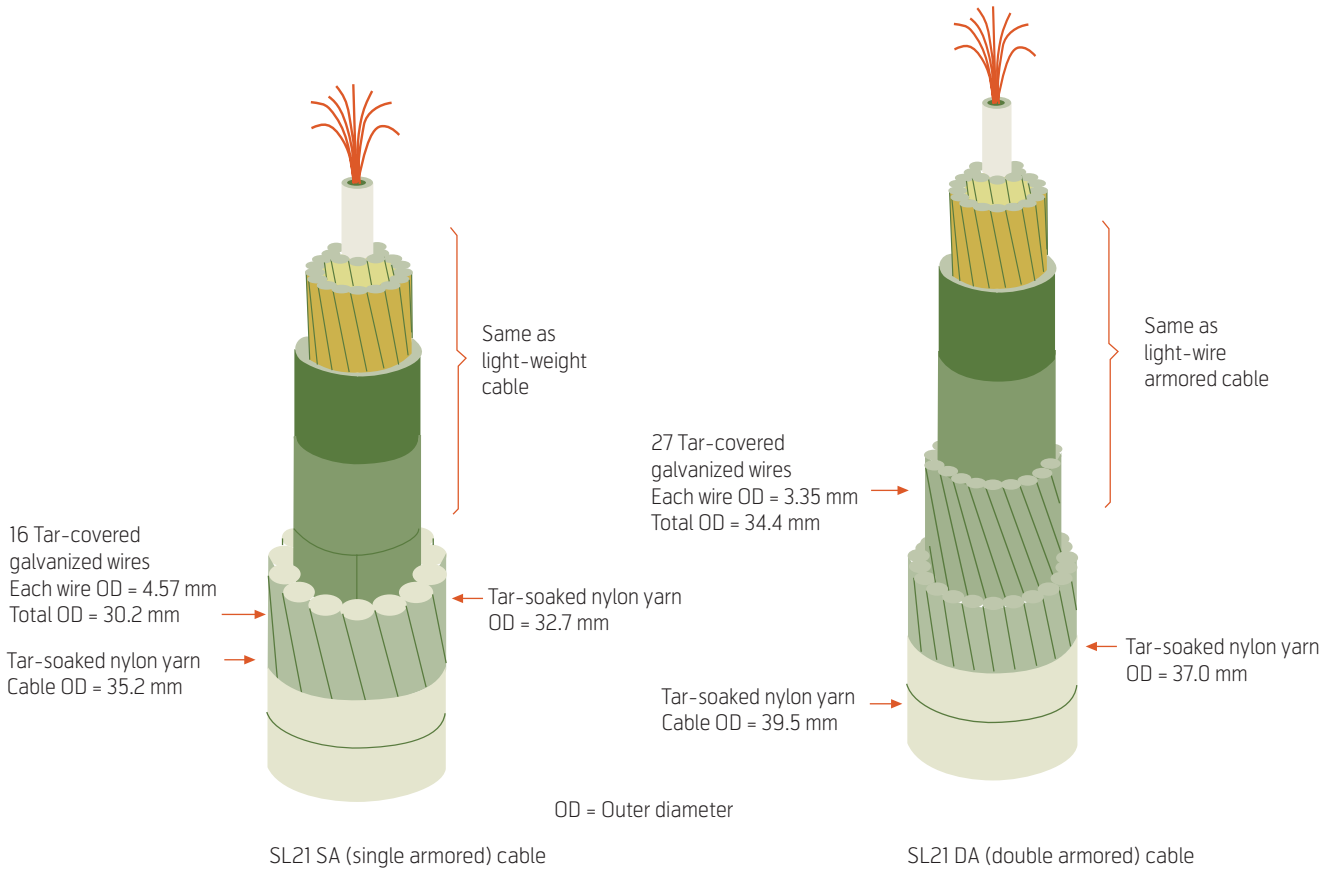
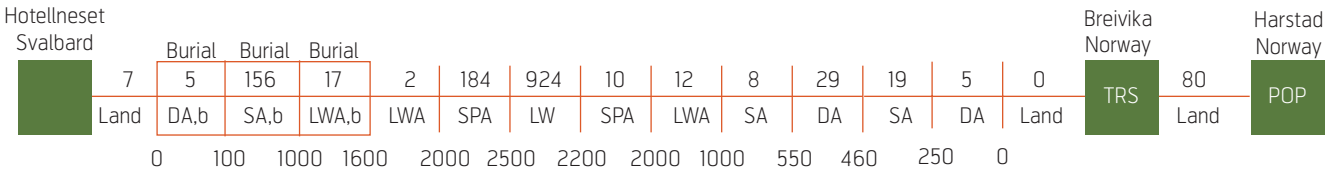


Figure 9b The cables used are from the SL21 family of cables (continued)



Summary of cable types, quantities & spares		
Type	Length	Spares
Land	87	
DA	34	2
DA,b	5	

Figure 10 The Straight Line Diagram (SLD) shows where the different fibre types are used. Segment 1 is used as an example

length division multiplexing or polarisation beam combiner as the figure shows. The 3dB-coupler splits the output of the LPU so that the LPU pumps the EDF of each fibre of a fibre pair. This provides redundancy as well as uniform amplification in each direction through the amplifier.

The LCM provides a wavelength selective, high loss loop back (HLLB). The LCM also provides a broad-band loop back for the optical time domain reflectometer (OTDR) and the coherent time domain reflectometer (COTDR). These are both used to monitor the state of the sub sea cable system. They can

Figure 11 The cable ship installing the cable off the coast of Breivika

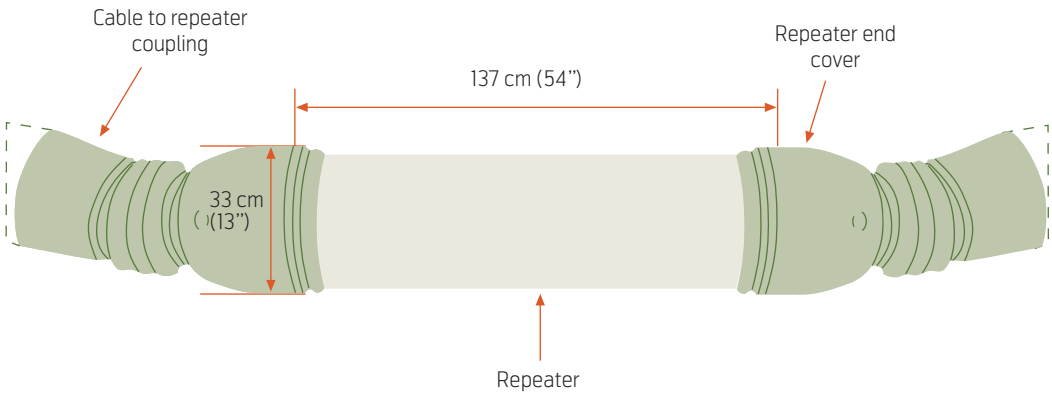
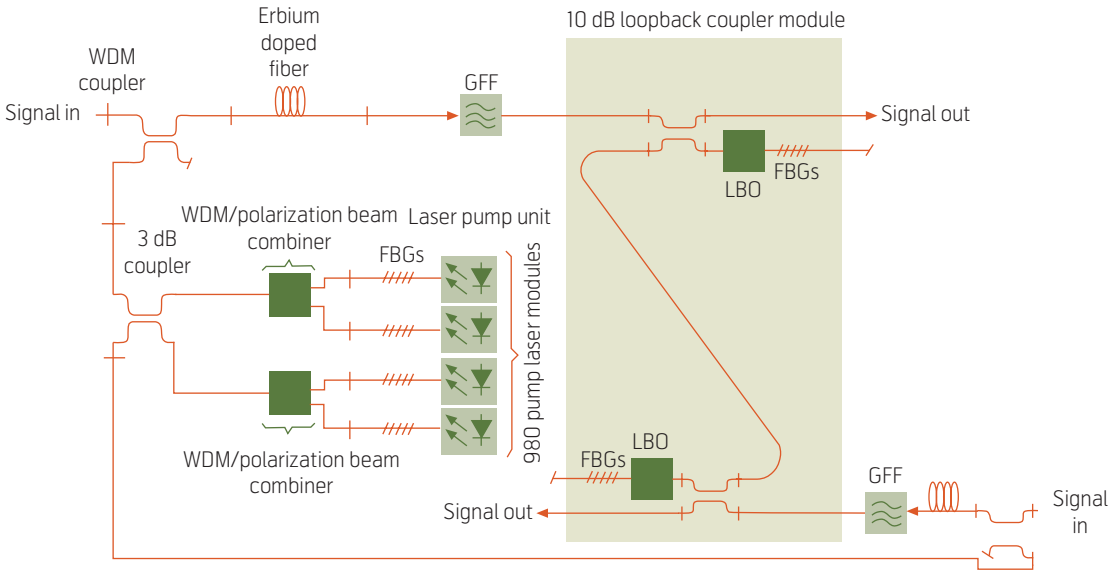


Figure 12 The optical repeater is quite large and weighs approximately 750 kg in total, including the couplings on each side. A total of 40 repeaters are installed



WDM = wavelength division multiplexing
GFF = gain flattening filter
LBO = line buildout
| = fusion splice
= optical fiber termination
= fiber bragg grating (FBG)

Figure 13 One amplifier pair includes a laser pump unit (LPU) and a loop back coupler module (LCM)

also be used to determine the location of a fault that has occurred.

The optical amplifiers are pumped with a 980 nm pump laser source. Pumping at this wavelength results in an optical bandwidth of the amplifier that is greater than 20 nm centred around 1551 nm. The noise figure for such a 980 nm pumped EDFA is less than 5 dB. Figure 14 shows a typical EDFA gain and noise characteristics. As we can see, the noise is way below the signal at the operating range of the amplifier.

Gain compensation

The length of the EDF (Erbium-Doped Fibre) and the pump level determines the gain and noise characteristics of the amplifier. At very low input the gain and noise characteristics are constant. Then when the power is increased the gain of the repeater will eventually start to decrease. When the input power level is raised sufficiently the gain of the amplifier will be 0 dB as shown in Figure 14. This gain compression characteristics of optical amplifiers make them especially suitable for transmission system applications because the power level along a chain of amplifiers in a transmission system is self-correcting and self-limiting. This means that if an amplifier for one reason

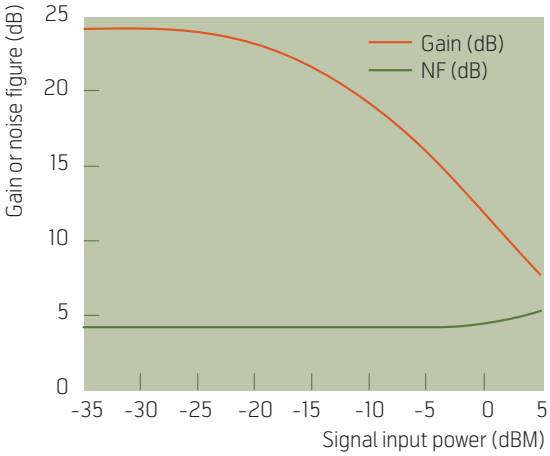


Figure 14 Gain and noise figure as a function of input power for the erbium doped fibre amplifier (EDFA)

or another gets a drop in input power, the gain in this amplifier will increase to compensate for that low input power. There can be different reasons for a drop in input power. Some examples could be a degraded pump in the previous amplifier, component aging, a degraded splice or a repair. This results in a lower input power to the next amplifier, which has to compensate for the low input by increasing its gain. The following amplifiers in line will compensate for the power loss until the signal is back at its equilibrium level.

Mechanical design

The eight amplifier pairs are mounted on assemblies inside a high-pressure container. The material used is copper-based alloy that withstands corrosion and supports hydrostatic pressure to a depth of 8000 m. Cathodic protection of the housing is not necessary. The high-pressure container protects the electrical and optical elements in the repeater from the ocean environment and provides a high strength connection between the cable terminals. The container consists of a cylindrical body and a repeater cone in each end connecting the body to the coupling hardware for the cable. The cable is connected to the cones using a special coupling. The coupling provides a path for both dc-power and an optical connection via splicing of cable and repeater signal from the cable to the repeater.

The lives of electronic components increase with decreasing temperature, therefore the amplifiers inside the repeater are mounted on a heat transfer plate to keep the temperature down. The seawater outside the repeater keeps the repeater temperature down.

To prevent the repeater from corrosion, the parts that are exposed to seawater are made of a tested copper based alloy. The parts carrying the electrical power are insulated from the seawater using a thick layer of extruded polyethylene.

The Power Feed Equipment

The Power Feed Equipment (PFE) provides power to the optical repeaters. The PFE is located in Breivika. In this system single end feeding is used. Systems that are fed from both ends are safer, but the power supply in Breivika should provide enough redundancy to make the system reliable. The primary return current path is the seawater and the seabed.

The PFE converts the –48 V battery supply at the cable station in Breivika to the regulated voltage needed for the system. The PFE feeds the system with a constant current of 1.1 A at 2500 V. A generator back-up maintains the 230 V mains voltage used to feed the –48 V battery system in case of a mains power failure.

In normal operation the two converters are serially connected, each supplying half of the load to the cable. Each converter can supply the cable system on its own, so if a problem occurs at one converter or it has to be stopped for maintenance reasons the other converter will automatically take over the load, providing active redundancy. The monitor functions are also redundant. The switches allow converters and monitors to be reconfigured for in-service testing and repair. In the Svalbard system the cable is a single-end feed system, but for systems that are dual-end feed the system will automatically switch to single-end feed if one PFE fails completely.

The output from the PFE can be in constant current mode or constant voltage mode. At normal operation, constant current mode is used.

The parameters monitored on the PFE are output voltage and current, station ground current, seabed voltage and converter output voltage and current. The PFE is monitored using the Tyco Element Manager System (TEMS). There are three installed, one in Harstad, one in Longyearbyen and one at Telenor's Network Operation Centre, which provides 24 hours surveillance of the system.

To protect the undersea system the PFE will shut down if the voltages or currents exceed certain threshold values.

The PFE can operate on either Ocean Ground or Building Ground. In normal operation ocean ground

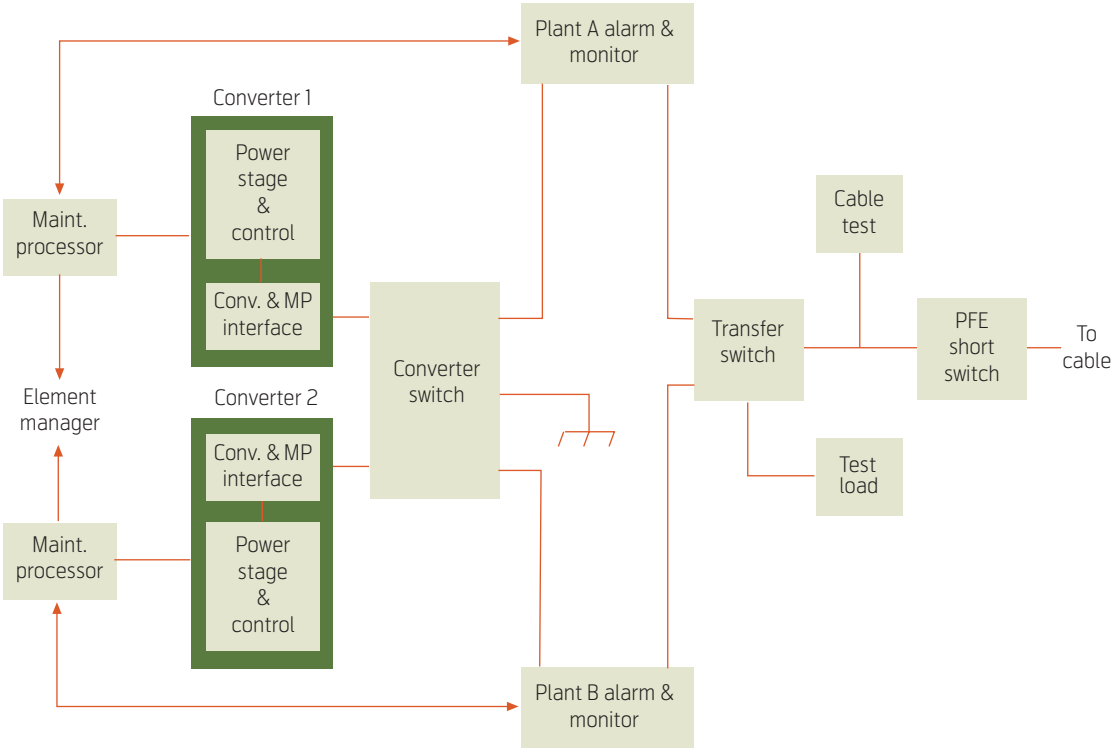


Figure 15 Power feed equipment (PFE) architecture

is used. To limit the risk of damage due to lightning strikes an Ocean Ground Protection Panel (OGPP) is installed. The OGPP provides isolation between ocean ground and building ground.

If there is a problem serious enough for picking up a part of the cable for repairs being necessary, the PFE can provide a 4 to 5 Hz sine wave on its high voltage output. This signal can be detected by the cable ship to locate the cable on the ocean floor.

Line Terminating Equipment

The LTE terminal equipment platform provides an interface between the terrestrial network and the undersea system.

The LTE provides powerful Forward Error Correction (FEC), precision wavelength control and high stability receivers and transmitters.

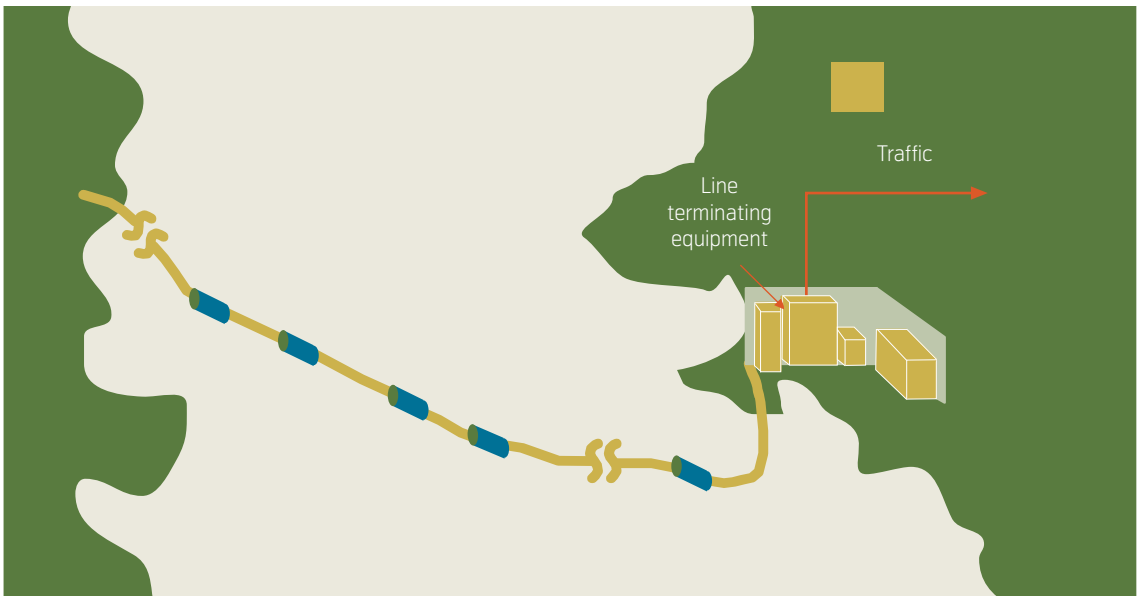


Figure 16 The Line Terminating Equipment (LTE) converts the STM-1 signal from the Telenor network into a 10 Gb/s signal suitable for long distance sub sea transmission

ISSN 0085-7130

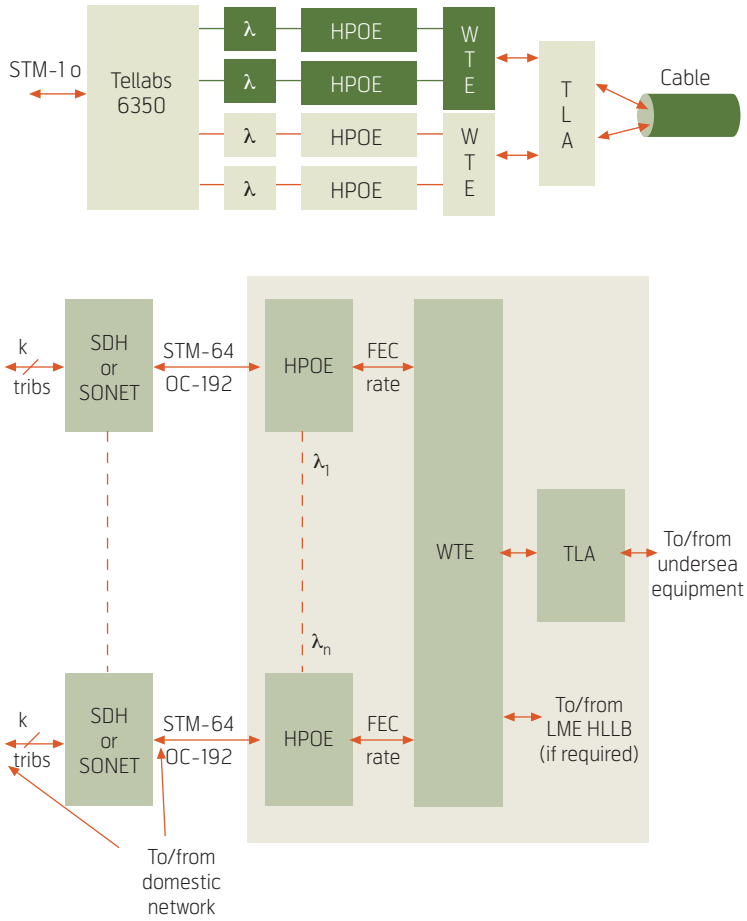


Figure 17 The Line Terminating Equipment (LTE) used in the Svalbard cable system

The LTE uses 'clear-channel transmission'; i.e. it is independent of the protocol used. In our case, the input is an SDH signal, but any protocol may be used.

The LTE processes the 10 Gb/s signal for ultra long haul:

- Up to 10,000 km without regeneration
- Dynamic pre-emphasis
- Wavelength multiplexing and de-multiplexing
- Dispersion compensation
- Pre- and post-amplification
- Automatic wavelength adjustments
- Chirped return to zero (RZ) modulation

The LTE consists of the following building blocks:

- High Performance Optical Equipment (HPOE)
- Wavelength Terminating Equipment (WTE)
- Terminal Line Amplifier (TLA).

The High Performance Optical Equipment provides special transmit and receive line signal processing, the Wavelength Terminating Equipment provides optical wavelength processing, and the Terminal Line Amplifier amplifies the signal before sending it into the sub sea system. One HPOE is needed per channel, but two are actually used per wavelength to provide redundancy. In the Svalbard project two wavelengths are used.

The input from the Telenor network is an SDH STM-1 (155 Mb/s) signal. At both ends, Harstad and Longyearbyen, the STM-1 signal is fed into Tellabs STM-64 equipment. The output from the Tellabs STM-64 is sent to the LTE equipment. The STM-64 signal has a bit-rate of 9,953 Gb/s. The first module in the LTE is the High Performance Optical Equipment (HPOE). Each LTE has two HPOEs for redundancy. In the HPOE extra Forward Error Correction (FEC) bits are added to the STM-64 signal making the new bit-rate 10,664 Gb/s.

For the receive operation the opposite sequence is engaged.

Functions:
Transmit direction:

- Reception and optical/electrical (O/E) conversion of the STM-64 input signal
- FEC encoding
- Insertion and encoding of overhead bytes for FEC
- Insertion of user communication channels into the FEC overhead data
- Optical line signal generation
- Post amplification

Receive direction:

- Reception and O/E conversion of the 10,664 Gb/s signal
- FEC decoding and bit error correction
- Generation of line performance parameters
- Extraction of the user channel from the FEC overhead for inter station communication
- Generation of the STM-64 optical output signal

High Performance Optical Equipment (HPOE)

The HPOE accepts any signal having a one-zero density consistent with STM-64/OC-192 protocols, the bit-rate being 9,953 Gb/s. The signal is converted to electrical, and FEC overhead bytes are added making the bit-rate 10,664 Gb/s.

In addition to the Forward Error Correction, the HPOE provides phase modulation, automatic transmit and receive direction channel optimisation, FEC performance parameters and a unit alarm interface. The output power is 9 dBm regardless of the input power.

Forward Error Correction (FEC)

FEC is a technique used to detect and correct digital transmission errors. The forward error correction code improves the signal to noise ratio (S/N) with 4 to 4.7 dB. The HPOE adds overhead bytes used to detect and correct errors that may occur in the undersea system. In other words, redundancy is added to the transmitted signal.

The performance of the FEC is shown in Table 2.

Wavelength Termination Equipment (WTE)

The Wavelength Termination Equipment (WTE) is used to provide optical wavelength processing. The WTE does automatic wavelength pre-emphasis, wavelength multiplexing and demultiplexing, side-tone LME wavelength combining (the Line Monitoring Equipment is explained later), and wavelength dispersion compensation. Figure 20 shows the WTE's location, while Figure 21 shows a more detailed view.

The Optical Pre-Emphasis Equipment (OPEE) provides channel equalization by adjusting the optical power of each channel prior to transmission such that all have the same end-to-end line error performance. The OPEE maintenance circuit automatically determines the appropriate channel pre-emphasis based on detected FEC error.

Because the response of the sub sea system is known the output signal can be adjusted for this before it reaches the undersea system. This is called pre-emphasis. Figure 22 shows the principle of this.

The optical filters, splitters and combiners are passive components as shown in Figure 23.

Terminal Line Amplifier (TLA)

The Terminal Line Amplifier (TLA) provides high-availability optical, wideband gain. The TLA may be used in constant output power mode or constant gain mode. The TLA can be used both in transmit and receive directions for post- and pre-amplification, respectively. The TLA contains two laser pumps, one is a 980 nm at 150 mW constant output and the other is a 1480 nm with a maximum output of 160 mW. The optical signal amplification is done by using erbium-doped fibre amplifier (EDFA) technology. This is a well-developed method for amplification of optical signals without optical to electrical conversion and regeneration of the signal. The TLA has two

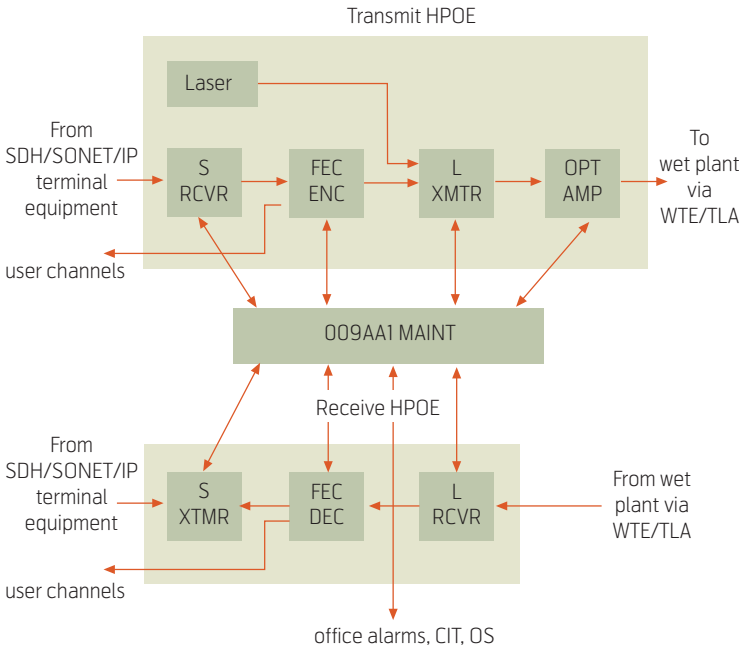


Figure 18 The High Performance Optical Equipment (HPOE)

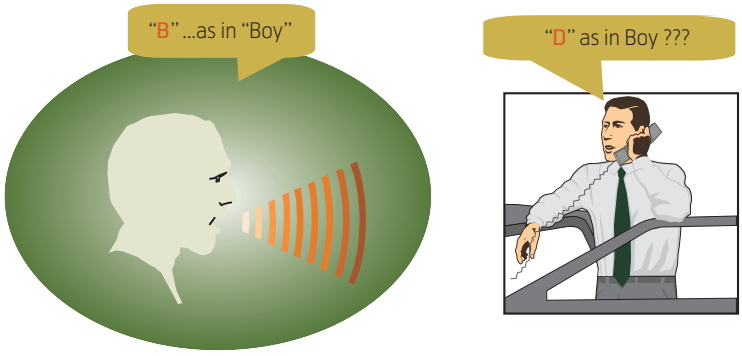


Figure 19 By using FEC you get increased system capacity, longer transmission distances are possible and repeater spacing may be increased. The optical power can be lower, the pump power gets lower and there are less non-linear impairments

Feature	Specification or applicable standard
Line bit error rate	Delivered bit error rate
10^{-4}	$\leq 10^{-9}$
10^{-5}	$\leq 10^{-13}$
10^{-6}	$\leq 10^{-17}$
Error correction protocol	Reed-Solomon 14/15, per ITU-T G.975

Table 2 Specifications and performance of the Forward Error Correction (FEC) scheme

ISSN 0085-7130

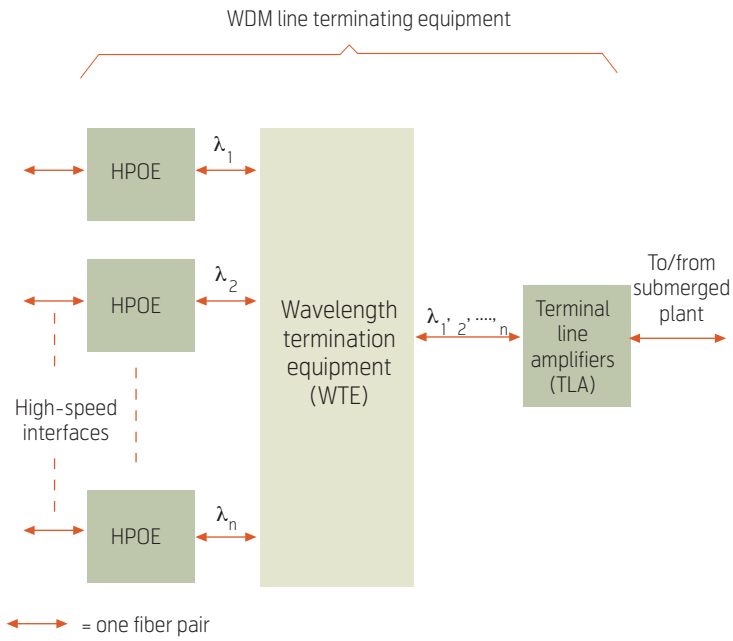


Figure 20 The Wavelength Terminating Equipment (WTE)

stages. The first stage holds the 980 nm laser pump and gain-flattening filter, and the second stage holds the 1480 nm laser pump. Passive components are used in the TLA because they are usually more reliable than active ones. The TLA also has a maintenance circuit pack, which monitors the components, records the output power, bias current and semiconductor chip temperature. Any value outside the limits of normal operation will trigger an alarm.

Multi Side-Tone Line Monitoring Equipment (MST LME)

The LME is used to monitor the condition of the undersea system. The LME consists of an equipment rack, a computer and a workstation. The LME operates together with the loop-back coupler modules in the repeaters and in the LTE (Line Terminating Equipment). The whole system is called the Line Monitoring System (LMS). At start-up of the sub sea system a delay calibration has to be performed. The Multi Side-Tone (MST) LME injects an optical 2.5 MHz, three level (–1, 0 and 1) pseudo random square wave signal directly into the outgoing fibre through the LTE into the sub sea system. The injected signal is called the LMS signal. The injected signal comprises two wavelengths which are different from the ones used for the transmission itself, hence the name MST LME. The LTE and all the repeaters contain a 'high loss loop-back' coupler module, which sends a fraction of the LME signal back in the receive direction and back to the LME. The location and loop-gain of each repeater and LTE is measured during the time the signal takes to reach a particular repeater. The results are stored in a database as a ref-

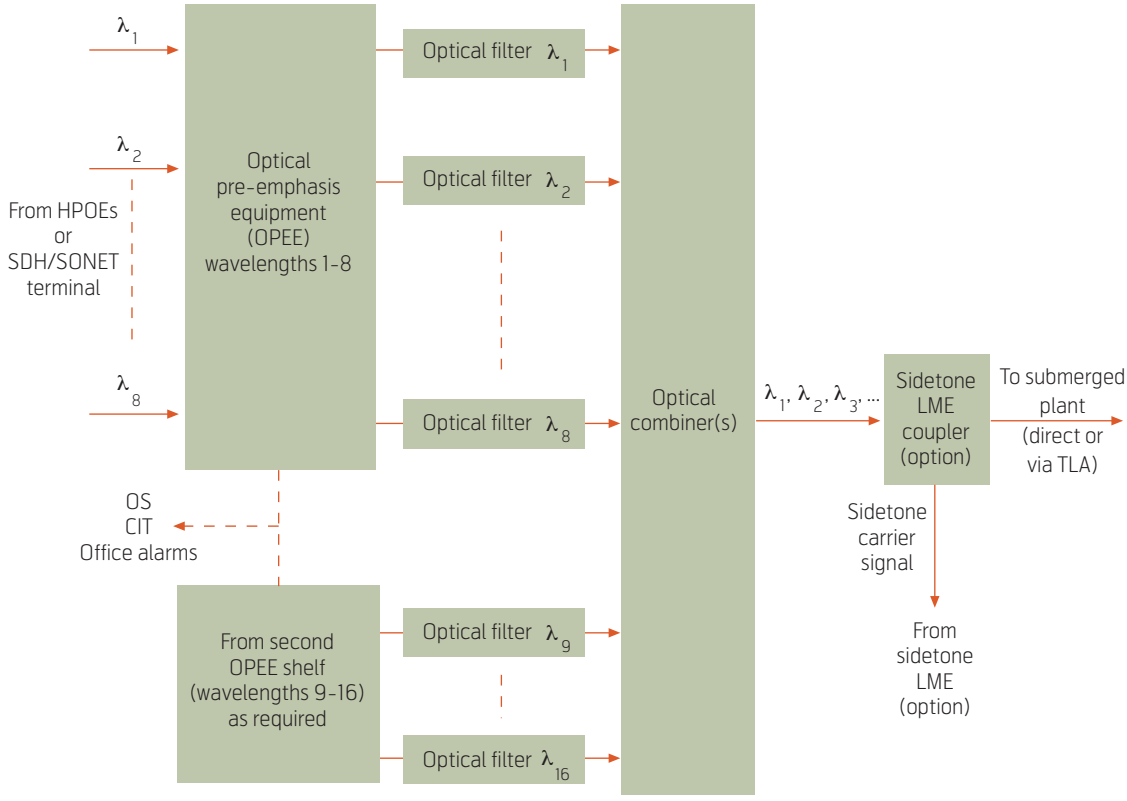


Figure 21 Detailed diagram of the WTE

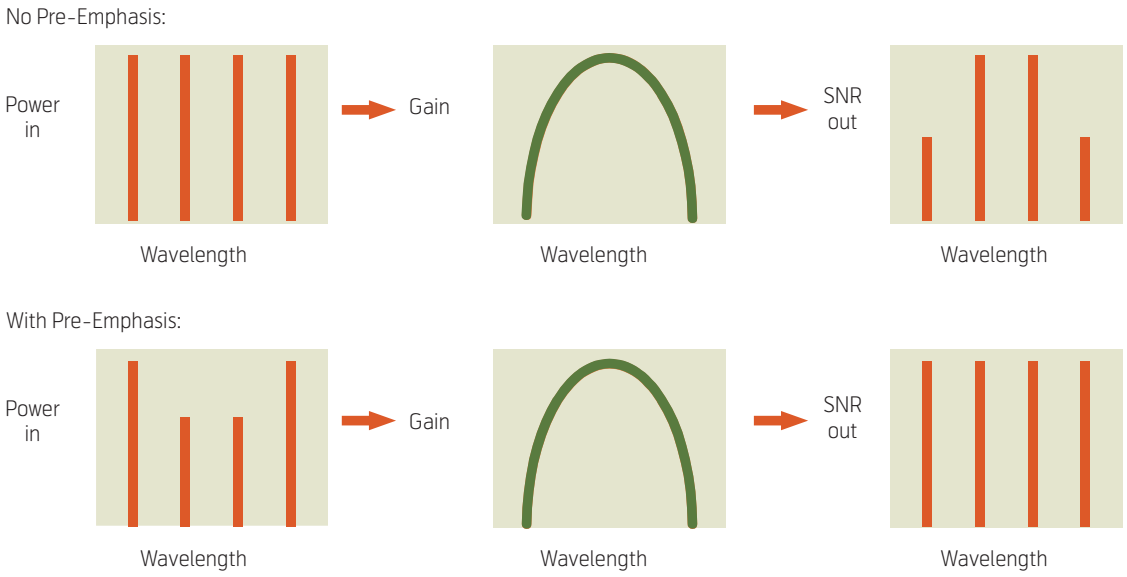


Figure 22 Principle of pre-emphasis

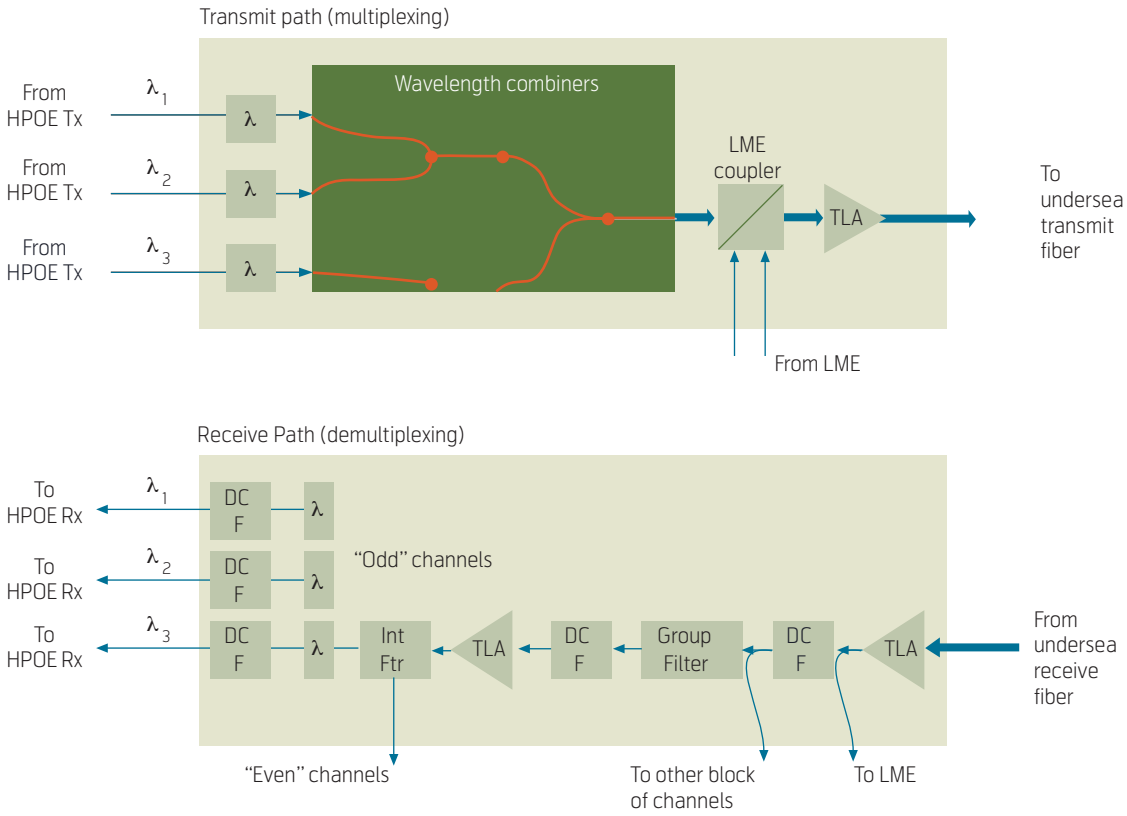


Figure 23 Optical filters, splitters and combiners

erence to future monitoring runs. In this way we can look for changes that may indicate a problem with the undersea system. If the system is altered in any way, a new baseline has to be established. Periodic monitoring runs are performed. If a fault has occurred in the undersea system the monitoring run will differ from the baseline. The test-run results can be analysed manually by comparing the signature from the test-run with a library of fault scenarios. This will give us the opportunity to determine what the fault may be and where it may be located. The LME also has an automatic signature analysis function. The faults that may be determined are laser-pump failures, fibre breaks, cable breaks or changes in gain and loss. One of the wavelengths from the MST LME is placed near or below the lowest traffic wavelength and one

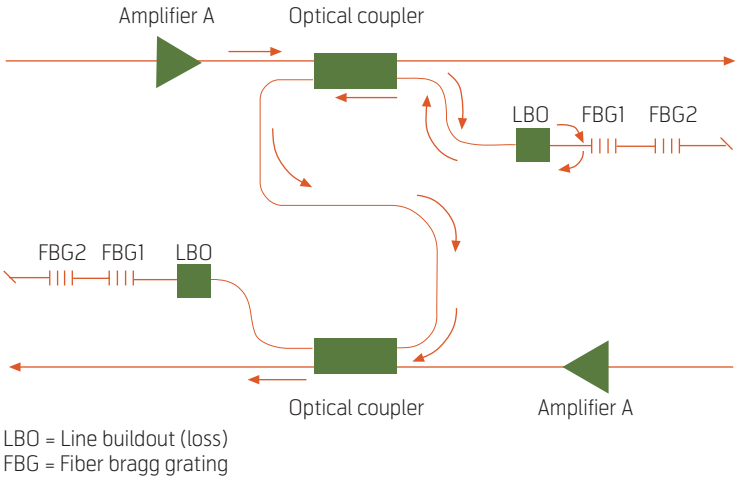


Figure 24 The repeater High Loss Loop Back (HLLB) path

near or above the highest traffic wavelength. In our system only one wavelength is used for traffic, but in other sub sea systems engaging more wavelengths, this method is used.

Repeater High Loss Loop Back (HLLB) paths

Each repeater has a High Loss Loop Back (HLLB) path used for the LMS signal as shown in Figure 24.

Using optical couplers, the signal is looped back after the amplifier in the repeater. In case one of the optical pumps of each amplifier fails the loop gain will be changed and the LMS signal will have a different shape baseline telling us that we have a problem with a repeater. The laser pumps are very reliable and one failed pump does not require the repeater to be picked up for repair. The repeaters after the one with a fault will restore the signal strength up to the normal value. The MST LME may be used to monitor the undersea system both in service and out of service. For out of service monitoring a more powerful LMS signal may be used.

The management system

The management system used for the Tyco equipment is called TEMS (Tyco Element Manager System). The TEMS provides network management functions for the line monitoring equipment (LME), the line terminating equipment (LTE) and the power feed equipment (PFE). The 24-hour surveillance is done from Telenor's Operation Centre at Fornebu, while monitoring runs and other management functions are done from Longyearbyen and Harstad, where the two other TEMS workstations are located. All functions cannot be performed from the TEMS; some have to be done using a laptop computer directly connected to the equipment in the cable station.

Summary

A fibre optic telecommunications cable was deployed between the Norwegian mainland and Svalbard in 2003, and put into operation in January 2004. The article has provided an overview of the technology used together with explanations of the different building blocks of the complete transmission system. The cable and additional equipment have an expected lifespan of 25 years and the current capacity utilisation is 10 Gb/s.

Eirik Gjesteland (29) graduated from Kongsberg Technical College in 1998 and has a BA with honours in Electrical and Electronic Engineering from Heriot-Watt University, Edinburgh, Scotland from 2000. He has since then been working at the Operational Centre of Telenor Networks.

email: eirik.gjesteland@telenor.com

The engagement of Televerket in the specification of GSM

A SHORT INTRODUCTION BY BJØRN LØKEN

Bjørn Løken is Director in Telenor Nordic Mobile

Mobile communications have expanded its grounds from covering pure communication needs to become a lifestyle product, and Telenor has become one of the business' major players, even on a global scale. In Norway mobile phone is a natural first choice to many people. But it was not written in the stars that the future of wireless communications should become that bright. I am therefore pleased that the Editorial Board invited Telenor to highlight its contributions to the development of the GSM standard and reflect upon why GSM became such a success.

Telenor (formerly Televerket) has a long record in operating public mobile telephone systems. A manual system in the 160 MHz band opened in 1967 (OLT). The service became quite popular, and the system soon ran out of capacity. A few years later a supplementary system was introduced in the 450 MHz band in some heavy traffic areas in the southern part of Norway. The introduction was coordinated with neighbouring countries which experienced similar capacity strains.

The request for more capacity was uniform in the Nordic countries. Hence the idea to develop and establish a common automatic mobile telephony system was born. The system was called NMT. An important requirement was that the system should offer international roaming throughout the Nordic countries.

The appreciation of OLT and NMT in the Norwegian market revealed some basic knowledge that became crucial for Televerket in the years to come.

- The market potential of mobile communications seemed substantial, but unpredictable.
- Coverage is crucial, both in terms of area coverage and in terms of services offered, i.e. seamless handover, international roaming, and voice mail services.
- The steady miniaturisation of electronic components and the ever lasting increase in processor speed imposed some inherent market drivers that are particularly valuable in mobile communications, for success so much relies upon the ability to create mobile applications.
- The operator's best salesmen are the mobile users themselves and the importance of forefront users eager to exploit new technology should not be underestimated. Equally important is the need to identify user groups inclined to generate traffic volume.

Two years ahead of opening the analogue NMT system to the public in 1981, Televerket launched some initiatives to study what would be the impact of digital treatment of speech, modulation and coding on mobile communications. Those activities were tightly connected to an advanced expert community at the Norwegian Institute of Technology (NTH).

At the same time Televerket once more joined forces with its Nordic sister operators to study what might be the best strategy for the Nordic countries for mobile communications after the NMT era (FMK, "Fremtidens Mobilkommunikasjon"). In parallel, Televerket took some initiatives within ITU CCIR and CCITT to engage those bodies in the specification of global standards of vital importance for the smooth operation of automatic international mobile communications systems. I myself was appointed Special Rapporteur on mobile questions in CCITT Study Group XI (ISDN and telephony swiching and signalling), a position I held from 1980 to 1992.

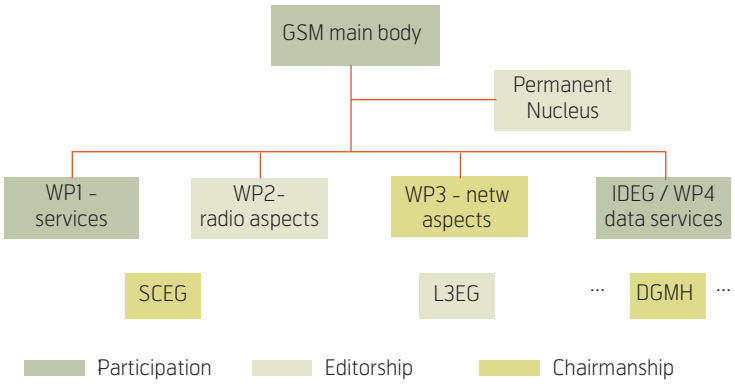


Figure 1 An extract of the organizational structure of the GSM project around 1988. GSM Main Body was the group to approve or reject the work items or specifications proposed by the four working parties. Below the working parties, there would be ad hoc groups with special assignments. Three of those groups are depicted in the figure; SCEG (Speech Coding Experts Group), L3EG (Layer 3 Expert Group), and DGMH (Draft Group on Message Handling). The colour codes indicate which groups in which Televerket was a) represented, b) had editor responsibilities, or c) had the chairmanship

ISSN 0085-7130

When the GSM initiative was launched Televerket found itself in the happy situation of possessing experts in many different fields that appeared to be crucial in the GSM work. These resources where offered to the GSM group. In my opinion Televerket's efforts made a significant contribution to the success of GSM. Indeed, without operational experience and convincing expertise we would have had a much harder time advocating for a system equally well suited for both urban and rural areas like those we find in the Nordic countries.

The organisation of GSM is depicted in Figure 1.

From Televerket Jan Audestad, Petter Bliksrud and myself participated in the main body. When the working parties were established, Jan became chairman of WP3. Several experts joined WP2 and took part in the endeavours therein, in particular Torleiv Maseng and Rune Rækken. Helene Sandberg and Hans Myhre from the operational division of Televerket's mobile activities joined WP1. Jon Emil Natvig, being at that time a distinguished expert on speech coding techniques, became chairman of SCEG. Knut Erik Walter entered WP3 and became heavily involved in its work. Finn Trosby entered IDEG (later WP4), and became chairman of DGMH, one of its four drafting groups. Stein Hansen entered the Permanent Nucleus in Paris (PN) and became editor of several of the specifications that became assignments of PN.

Many of the individuals mentioned above are among the authors of the articles to follow. In addition, we have been so lucky as to get the chairman of the GSM project itself, Thomas Haug from the Swedish Televerket (now TeliaSonera) to write about how GSM came about.

It is for the historians to judge what the future had held for mobile communications in Europe without heavy involvement of Televerket and the other telecom operators in the GSM specification work and their firm commitment to launch services based upon the emerging standards (MoU). Having reached such an overwhelming penetration worldwide, it is in my opinion obvious that the efforts have paid off. Telenor – being the 12th biggest mobile operation company in the world and entirely basing its business on GSM – has indeed benefited from it.

Bjørn Løken (57) graduated from the University of Oslo in 1972. After graduation he joined Norwegian Computing Center as a researcher on teletraffic simulation models. In 1975 he was employed by the Research Department of Televerket, and in the following five years he worked on specification and testing of the NMT system. In the 1980s he was engaged in the CEPT/ETSI specification work on GSM. From 1980 to 1992 he also held a position as Special Rapporteur on mobile questions in CCITT Study Group XI. From 1987 to 1992 Bjørn Løken was Chief of Research with responsibility for network development for both fixed, mobile and data communications. From 1992 to 1995 he was Director of Strategy in Televerket. Since 1995 Bjørn Løken has been with Telenor Nordic Mobile, primarily concerned with questions related to telecom regulations.

email: bjorn.loken@telenor.com

How it all began

THOMAS HAUG

Thomas Haug, formerly of the Swedish Telecommunications Administration, now retired

This article describes the background for the European Telecom Administrations' decision to start the work on the GSM mobile communication system and the fundamental issues the committee was faced with in the first five years, as well as how they were dealt with.

The basis for the work

The European telecommunications market was for a long time badly fragmented and in many areas lacking universally applied standards despite efforts by i.a. ITU. As a result, economy of scale could frequently not be achieved. In an attempt to remedy that situation, the organisation of 26 Western European Telecom Administrations (CEPT) in 1959 started standardisation activities, but the progress was often hampered by differences in policy in the member countries. As time went by, the fragmentation in the rapidly growing mobile communications field became particularly annoying because the very idea of mobile communication is of course to communicate while on the move, and the incompatibility made this impossible while visiting a foreign country. (One exception was the NMT system.) The spectrum available for new mobile communication systems was very limited in frequency ranges suitable for this purpose, among other things because of the large amount of spectrum reserved for radio and television broadcasting.

At the meeting of the World Administrative Radio Committee (WARC) in June 1979, it was realized that it was desirable to allocate a part of the 900 MHz band for land mobile communication in Zone 1 (which in the terminology of the Radio Regulations means Europe). The reason for this decision was that the 900 MHz band, which up to that time had not been extensively used, now was becoming attractive because of the technical development . A large block of spectrum was therefore reserved for land mobile use, but nothing was said about the character of the systems, whether they should be private or public, automatic or manual, analogue or digital.

CEPT discussed for a while the opportunity thus created. At the CEPT Telecommission meeting in June 1982, the Netherlands and the Nordic Telecom Administrations presented a proposal, stating that because of the rapidly increasing demand for mobile communication, the frequency band allocated by WARC'79 was in danger of being put to use for national and probably incompatible systems. This would almost certainly remove the only opportunity for the coming decades to achieve a form of harmonisation in Europe in the field of public land mobile services, since frequencies above 1 GHz were not at that time considered suitable.

The Telecommunication Commission decided to entrust its Harmonisation Committee (CCH) with the task of coordinating the activities of CEPT in the field of mobile services. A special study group should be set up, reporting to CCH and working out an action plan with the aim of ensuring that compatible mobile systems could be implemented by the early 1990s. The new study group was given the name *Groupe Spécial Mobile* (GSM). The work was to be completed by 1986, and the Commission accepted an offer from the Swedish Administration to supply the author of this article as Chairman. (I was in a meeting somewhere else and was quite surprised when I heard this, since nobody had asked me in advance.)

The decision made by the Commission was quite vague, and in effect left it to the new group to propose its own terms of reference. Representatives of the Netherlands and the Nordic Administrations therefore met during the summer of 1982 in order to write a proposal for an action plan, which was subsequently approved by CCH in November that year and used (Doc. GSM 2/82) as the basis for the work of GSM for many years.

The decision only mentioned "harmonisation", which indicates that the compatibilty aspect was the dominating factor behind the decision, and few delegates at the T-Commission meeting sincerely believed that free circulation of radio users across international borders could be achieved within the foreseeable future, given the serious political and military obstacles that existed at the time. In fact, later comments have indicated that in view of the many failures in European standardisation, many doubted that a common system could be agreed at all.

ISSN 0085-7130

The start

From the start, it was clear that many more aspects than harmonisation had to be addressed if the goal of a modern Pan-European service was to be reached. Clearly, the goal envisaged by the CEPT Telecommission and the CCH of completing the work by the end of 1986 could not be met if important new developments were to be taken into consideration. It was therefore decided that by the end of 1986 only an outline specification should be available, comprising the main system parameters for the various parts of the system and their interfaces, including the air interface.

The list of basic requirements for the new system was to a large extent patterned on the corresponding list for NMT, but with a number of modifications due to the fact that there would be a huge number of technological advances that had to be exploited, and of course also the necessity for the new system to satisfy the needs of a far larger community than NMT all over Western Europe. The need for services other than speech was stressed, but since nobody knew exactly what those services would be, the system structure had to be modular and flexible. The same philosophy as for ISDN and OSI should be applied in order to achieve this, and standards for protocols etc. should as far as practicable be compatible with such developments. Furthermore, the system must provide the same facilities as those offered in the public telephone and data networks, and the need for security must be taken into account, all of this without significant modifications of the existing fixed networks.

Hardly anyone among us expected that an analogue system would be preferable since the technology was clearly moving toward digitalisation. It had to be proved, however, that a digital system would meet the needs better than an analogue system. The digital technology had never before been tried in terrestrial public mobile communications, and it was necessary to convince the operators and manufacturers that a digital system was a realistic goal. Obviously, if "harmonisation" had been the only goal, an analogue system could be based on well established techniques and could therefore be specified and built in a short time. This may have been in the minds of some of the CEPT T-com delegates, whose main concern was to make sure that the allocated spectrum would be used for a harmonised system without paying much attention to the technical solution. The choice of analogue techniques would mean, however, that many desirable objectives could not be met, such as ease of introducing modern services, compatibility with the emerging digital fixed networks, etc., and it was clear that the digital technique had a far greater potential than the analogue one for mass manufacturing processes. To establish the feasibility of digital technique in the mobile network was therefore among the first steps of the work in order to establish a truly modern system. The greatest problem would be caused by the need to achieve a satisfactory speech quality and at the same time conserve spectrum.

The first meeting of GSM was held in Stockholm in December, 1982. One of the most important issues at that meeting was the spectrum situation in the CEPT countries, and there was great concern that it might prove impossible to keep the 900 MHz band "on ice" for GSM for such a long time as envisaged in the work plan for GSM. The feasibility of a standardised system was thus in great danger, since it was unrealistic to expect the manufacturers to invest in a system with a very uncertain market potential. This point was a constant concern for several years until the EU Directive reserved the band for GSM.

Another important issue was the question of free circulation of users. The existing regulations in many European countries regarding the use of radio equipment by foreign visitors would clearly be a serious obstacle to the system since free circulation would be one of the great advantages in a common system. To eliminate such obstacles would therefore be among the most important tasks of the committee, but it would be a very difficult task since those obstacles were usually outside the authority of the national telecom administrations.

Some assumptions had to be made for testing the solution, regardless of the choice of analogue or digital techniques, and one such assumption must be a traffic model. This was a very important issue during the first year of the group. In retrospect, of course, those traffic figures look almost ridiculous today.

Mode of operation

The written and unwritten rules of CEPT were in many ways very different from those generally used in purely national bodies. As usual in CEPT, we had to work on the basis of consensus, so there would be no voting. Earlier experience in CEPT and other standards bodies had shown that making recommendations by majority decision could easily lead to a situation where some countries chose to disregard the recommendations, and the compatibility was then lost. (It is correct to say that the possibility of voting in a few special cases was introduced towards the end of the 1980s, but it was never used by CEPT/GSM.) Consensus is not the quickest way to reach a decision, but on the other hand, a consensus makes it almost certain that everybody is going to stick to the decision made.

The choice of working language was of great importance for the efficiency of the work. In CEPT, both English, French and German were established as working languages, and the committees and many large working groups therefore worked through translators. In GSM, we agreed that it would be far better to work in English only, since just about everybody in the technical field had a sufficient command of the English language, and since working through translators in such a highly specialised field would actually hamper the work. Furthermore, making arrangements for translators and interpreters for every meeting would involve practical problems and major costs. The language question turned out to be a politically sensitive issue in certain countries, however, and the respective delegations had to do some arguing at home in order to convince their managements. They succeeded, and English was accepted somewhat reluctantly as the only working language. The issue resurfaced after a few years but the rules were then firmly established and nothing happened.Without this decision, we would never have been able to reach consensus within the time frame agreed.

The GSM committee had to utilise other bodies in- and outside CEPT as much as possible. Within CEPT, there were many working groups with experts in the fields of radio issues, signalling, data transmission, commercial issues, etc. In addition, the GSM committee liaised with several bodies outside CEPT, such as various COST groups and industrial bodies like EUCATEL. The expertise in those groups was very useful for the work on the new system. Another body with which we had frequent contacts was the European Commission, which on the political level proved to be very helpful. Above all, through their activities, the GSM frequency bands (890–915 and 935–960 MHz) were reserved for the coming Pan-European system.

It soon became clear that it would not be possible to deal with all issues in the Plenary, so GSM decided to set up three ad-hoc sub-working parties, reporting to the plenary, to deal with the three major areas of services and facilities, radio aspects and network. As is usually the case in most organisations, the increasing workload made it necessary to increase the number of working parties and to make them more independent, and in 1986 we had to set up a Permanent Nucleus in Paris in order to support the work, both in the Plenary and in the working parties.

The question of IPRs, often a difficult issue in standardisation bodies, arose early in the work. In line with the usual CEPT strategy, it was decided to avoid patented inventions unless they could be made available royalty-free for the CEPT standards, and we succeeded in acquiring solutions along those lines for two areas, i.e. speech encoding and voice activity detector. After 1988, the question of IPRs was taken over by the MoU and it assumed a much greater importance.

Crucial and controversial Issues

Two crucial issues during the first years of the GSM work, i.e. analogue versus digital techniques and the access and modulation method in a digital system required a lot of work. Proposals were invited and no less than eight were submitted for testing, which was done in Paris in late 1986 by CNET assisted by the GSM Permanent Nucleus. The results were discussed in a meeting in Funchal, Madeira in February 1987. The group found unanimously that the digital technique offered great advantages over the analogue one, and the measurements showed that the requirement for speech quality could be met with a good spectrum economy. Controversy arose around the second question, however, since France and Germany had proposals which did not agree with those of the other CEPT countries. This was due to political decisions made in France and Germany, not to any disagreement on the part of the delegates present, who were privately in agreement with the rest of CEPT/GSM, but they had to follow the line taken on the highest political level at home. Despite endless discussions, the issue could not be resolved in the meeting, so on this fundamental question there was a serious risk of failure of the entire GSM effort. During the following months, however, a considerable amount of pressure was put on the governments in the two countries, above all by the European Commission, which was very anxious to get a Pan-European system. As a result, in the next meeting in June 1987, France and Germany informed us that they had agreed to the solution accepted by the others. Thus ended the only really serious controversy that we had in CEPT/GSM, and we were now certain that there would be just one European system. Without agreement on that point, the Pan-European system of today would not exist.

Conclusion of the first phase

The 1987 decision was a milestone in the work on the system, since it eliminated the only major controversial issues and the road towards the final goal now lay open. There were many changes to our work arrangements, such as the joining of ETSI in 1989, but basically the established organisation with working parties, Permanent Nucleus, etc. stayed although the names were changed. As examples of those changes one should note the signing of the Memorandum of Understanding in September, the opening of CEPT groups to the manufacturing industry and later the

creation of ETSI. Without the commitment of those bodies, hardly any manufacturer would have dared to invest the huge sums required for the development of the system. Still, five years of detailed work remained to be done. There is no doubt that in terms of man-years, they represent a far greater part of the work than Phase 1 and was basically a detailed design project, which cannot be said of Phase 1. The support by the operators, national regulating authorities and manufacturers grew, and a huge number of people were gradually enlisted. However, that part of the work is well known and it would lead too far to go into it here.

The interested reader will find it in [1].

Reference

1 Hillebrand, F et al. *GSM and UMTS, The creation of Global Mobile Communications*. London, John Wiley, 2002.

Thomas Haug (77) was employed by the Swedish Telecommunications Administration (now Telia) from 1966 until his retirement in 1992. He was Chairman of Groupe Spécial Mobile in CEPT, later ETSI TC/GSM, from its inception in 1982 until the start of the system in 1992, and in that capacity he led the work on the specifications for the GSM system. From 1970 to 1978 he was Secretary of the joint Nordic NMT committee which worked out the specifications for the NMT system, and from 1978 until 1982 he was Chairman of the NMT committee.

email: thomas.haug@wtnord.net

My work in the GSM services area

AN INTERVIEW WITH HELENE SANDBERG BY FINN TROSBY

Helene Sandberg is Vice President in Telenor Nordic Mobile

FT: In which time period did you take part in the GSM work on services, and what became your main area of objectives?

HS: I entered the Working Part 1 of the GSM activities in CEPT and later ETSI in 1986 and stayed with the work items of WP1 – later GSM1 – until 1990. One of the major objectives in WP1 at the time of my entrance was a 'modification' of the current drafts for supplementary services, and I was fortunate to become heavily involved in the work with those.

At approximately the same time, I also took part in activities in the MoU; mostly in the subgroup for billing and accounting (BARG).

FT: Why was it necessary with a 'modification'?

HS: The work on specifying the supplementary services had commenced some time before I entered WP1 and just before the specifications of supplementary services for ISDN were stabilized. Unfortunately, WP1 had gone on quite rapidly with the Stage 1 definition of the supplementary services for GSM without paying attention to the fact that there would soon be a complete and approved set of such for ISDN. When that was the case, a decision was made to review the whole portfolio of supplementary services for GSM to align them with the ones of ISDN as much as possible. In fact, 'redoing' might be a better word than 'modifying', because almost all of the old stuff was replaced by new ISDN 'alignments'.

FT: How easy was it to find replicas of each of the supplementary services of ISDN for GSM?

HS: For every supplementary service that did not have mobility or roaming implications, the task was fairly easy. For those that did have mobility or roaming implications, the objectives were harder. In some of those cases – e.g. with the Advice of Charge – the service got a somewhat different interpretation in GSM than it had in ISDN. And in some other cases, a supplementary service of ISDN did simply not apply in GSM. However, I remember that for some of those – e.g. Call Transfer – it took quite some time and effort to remove them from the list of supplementary services!

FT: Retrospectively, how would you assess the success of the service portfolio of GSM?

HS: This is a question that is difficult to answer, but let me try to comment on the GSM services from different angles.

In my mind, the success of GSM was due to two basic factors –

- The scheme of agreements and functionality to provide for a smooth way of international roaming;
- SMS, turning out to be an extremely popular service to supplement telephony.

Defining the fundamentals of the GSM system coincided with one of the last hopes of substantial innovations within the domain of circuit switching: ISDN. I do not think that ISDN deserves the designation of a regional or global 'success' – far from it. However, one has to admit that the only sensible and viable strategy of the GSM design in the last half of the 1980s was to align the service repertoire of GSM with what was anticipated to be the future bandwagon of fixed network communications.

In terms of supplementary services, I think that the forwarding services, the barring services and the Calling Line Identification Presentation (CLIP) have also been vital elements of the GSM success.

FT: When taking into account what has happened afterwards, which is the one additional task or service you would have appreciated that the WP1 crew of those days had incorporated?

HS: What I'm immediately thinking of is the task to prepare prepaid solutions for roaming. Even if that perhaps is a MoU/GSMA matter more than a GSM/WP1 matter and even if prepaid was not at all a mature concept in those days, the prepaid business domain has proven to be so extremely important that this task is the one that springs to mind. Having been able to hasten the launch of those services on a broad scale would have meant a lot to the economy of the operators.

ISSN 0085-7130

FT: How would you assess the way of working in WP1; how were the discussions and the ability to resolve its objectives? How was the collaboration with other bodies inside and outside GSM? Who were the most profiled members of WP1 in those days?

HS: I think that the working procedures of WP1 were fairly good and efficient. I recall discussions in some of the bodies outside the body hierarchy underneath GSM, e.g. MoU, in which the ratio between the urge to keep on discussing and the importance of the subject were quite stunning. For instance, there was a discussion on the shape of the arrow of the SIM cards that raised very strong opinions. In WP1, however, the economy of the man-hours of the attendants was far better.

Another risk of a working group structure like the one of GSM is that the technical bodies will be left idle waiting for Stage 1 descriptions and will start producing those themselves. I did not recognize any such during my time in WP1. I would say that WP1 had a fairly good grip on things and that the collaboration with other groups went on fairly well. It should be mentioned however that WP1 deliberately did give the technical groups quite some latitude in terms of implementation. For instance, I do not think that there was very much contact between WP1 and the speech coding guys in SCEG.

In a group of this nature, the chairman will normally be the most or one of the most profiled persons. In WP1 this was also the case with our chairman Madame Martine Alvehrne from France Telecom. Among the other participants who caught your eye and ear, I would like to mention Alan Cox from Vodafone (at that time Racal/Vodafone). I think he was on full time catching up with anything within GSM that contained service aspects, and he combined his competence in this field with the clear business strategy that has characterized Vodafone ever since.

FT: Finally, would you like to add some more to the overall characteristics of WP1?

HS: For people who may have an interest in comparing the standardization of those days with the one of today, I think WP1 emphasized a quality of the whole GSM: It was all very much the mobile operator's ballgame. In WP1 of the years of my attendance, only operators were allowed to participate. In the technical groups, vendor representatives gradually seeped in. At first, they were only allowed to speak, after a while they were given the same rights as the representatives from the operator industry. But the foundation walls of the GSM, i.e. the basic specifications with version number 3.0.0 and slightly above were produced under strong influence of the mobile operators. Without speculating any further, I would say that is worth some considering.

Helene Sandberg has been employed at Telenor (formerly 'Televerket') since 1966 and has more than 20 years experience from mobile operations both domestically and abroad. She participated in the launch of Esat Digifone in Ireland and Connect Austria (One) in Austria. She returned 18 months ago from Thailand where she had been COO for two years. She is now Vice President in Telenor Nordic Mobile with a responsibility for Service & Delivery.

Wideband or narrow band? World championships in mobile radio in Paris 1986

TORLEIV MASENG

Torleiv Maseng is Director of Research at the Norwegian Defence Research Establishment

The story I will tell started in 1981 when I came back to ELAB/SINTEF in Trondheim after exile in the Netherlands working for SHAPE Technical Center in The Hague. My assignment was to propose a new digital mobile system replacing the successful analogue Nordic Mobile Telephone system. The new digital system which was to be defined was coined "Fremtidig Mobilkommunikasjonssystem" (FMK). Jan Audestad at Televerkets Forskningsinstitutt at Kjeller outside Oslo had formulated the task and given the project to ELAB.

In the beginning, this was a lonely assignment since I worked alone and few others took part in our project. Audestad did his best in linking us with other activities in the Nordic countries. The work towards FMK was arranged through Nordic meetings in various subcommittees.

Why not wideband?

At that time the common thinking was to transmit symbols so slowly that radio propagation reflections from mountains and buildings did not matter. This meant that the rate should not exceed 25 kbit/s. If this rate was exceeded, the symbols from various paths would be sufficiently shifted in time to overlap with the next symbol and make it difficult to decode[1)]. The problem with this idea was that in such a narrow band system, the reflecting paths would sometimes cancel each other and cause the signal to disappear.

Since we did not know which bit rate that was the best choice, Odd Trandem at ELAB designed an adaptive modem shown in Figure 1. This system could transmit at a bit rate of 256, 512, 1024 and 2048 kbit/s and could display the instant channel response on an oscilloscope. It could cope with up to 4 bits propagation delays difference between all paths. If this was exceeded, that part of the signal would appear as interference. Fortunately the longer the signal propagates, the weaker it becomes. Therefore this interference is normally small.

Our idea

If the bandwidth of the channel was sufficiently wide, however, the various paths would not overlap and could be distinguished. Only when there was no path would the signal disappear. My idea was to measure the channel continuously using radar principles and calculate how the signal would have been received, given the channel, for various data sequence alternatives. The receiver would continuously choose the sequence which most closely resembled the received signal. This task is called equalization. There is a simple and efficient implementation of this operation called the Viterbi algorithm. The idea was to merge demodulation, equalization and error correcting decoding in a joint Viterbi decoder. [1–4] The smart thing was that the modem would not make instant decisions about received bits, but defer decisions and observe the received signal until typically 4 bits had been observed and then make a decision on the bit value. As the number of bits which could be considered increased, the modem could handle longer and longer echoes or higher bit rates, but at the same time the complexity of the modem grew exponentially.

Figure 1 Odd Trandem: The brain behind the prototype development. Here he is skiing with what eventually ended up becoming a GSM terminal. I am behind him outside the picture with car batteries

1) *This is the same reason why psalms intended for churches with a lot of echoes are slow.*

ISSN 0085-7130

Committee works with no result?

Around 1984 we started developing hardware to prove our ideas. At the same time Audestad introduced us to some very enthusiastic French scientists from CNET, a PTT lab outside Paris. At that time they were working with Germany towards a German/French system. The idea was that when Germany and France finally agreed, the rest of Europe would follow. They were keen to start working with us. Their principal idea was "Slow Frequency Hopping" to randomize interference and to provide diversity to avoid deep signal fades. Shortly after, the "Groupe Spécial Mobile" (GSM) organization was born and all the Nordic FMK staff put into a similar GSM organization. The work was assigned to three Working Groups: 1 Services, 2 Modulation and coding and access system, 3 Protocols. I participated in WG2. Around 18 European monopoly PTT operators from different countries eventually participated under the auspices of CEPT. The work in WG2 was slow, but many ideas from nations were presented. Even if we were supposed to propose a modulation and coding scheme for radio access, everybody had their own favorite ideas and it was hard to reach an agreement. The slow progress in WG2 delayed the work in the other groups. The European Commission threatened to take over the control of the work in GSM unless progress was made. Something needed to be done.

Overview of GSM experimental systems

Developers	ATR, SAT, SEL, AEG Italtel	ANT, Bosch, Telettra	LCT	Philips/TRT Mats-D		Ericsson	Televerket Sweden	Mobira	Elab, Trondheim
System	CD-900	S-900-D	SFH-900	(MS-BS)	(BS-MS)	DMS-90	MAX		ADPM
Multiple access	wideband TD/CDMA	narrowband TDMA	nb./wb. CD/TDMA	narrowband FDMA	wideband CD/TDMA	narrowband TDMA	narrowband TDMA	narrowband TDMA	TDMA
Traffic rate (kbps)	12.8+3.2	9.6+1.4	16+25.6	16+2	16+2	16+8	16+4.6	16+8	16
Signallling rate	0.8+1.6	0.25	10.4	0.75	0.75	1+0.67	1.6+0.5	0.5+0.5	–
Multiplexed rate	18.4	11.25	52	18.75	18.75	25.67	22.7	25	–
Channels per carrier	63	10	3	1	64	10	11	9	10-160
TDMA-factor per carrier	63	10	3	1	4	10	11	9	10-160
CDMA-factor per carrier	1	1	1	1	8	1	1	1	1
Multi access rate (ksymb/s)	1496.25	128	201	19.5	39	340	302	252	256-4096
Modulation rate (kbaud)	3990	128	201	19.5	1248	340	302	252	256-4096
Gross bit rate (kbps)	7980	256	201	19.5	19968	340	302	252	256-4096
Carrier spacing (kHz)	6000	250	150	25	1250	300	300	250	200-4000
Modulation	QPSK	4-CPFSK	GMSK	GTFM	QAM	GMSK	GMSK	GMSK	DPM
Pulse type/filtering	1SRC	2AC	0.30 gss	gss	–	0.25 gss	0.50 gss	0.25 gss	SRC
FM/PM	PM	FM	FM	FM	PM	FM	FM	FM	PM
Symbol alphabet	4	4	2	2	4	2	2	2	2
Modulation index	0.50	0.33	0.50	0.50	0.50	0.50	0.50	0.50	1.20
Double sided bandwidth (kHz)									
• -20 dB	–	130	200	20	2000	370	300	270	–
• -70 dB	–	260	600	50	2500	810	1000	600	–
Gross symbol duration (µs)	0.251	15.6	4.98	51.3	0.801	2.94	3.31	3.96	–
Time slot duration (ms)	0.4762	3.1875	1.3333	16	4	0.8	1	1.77	–
Speech frame duration (ms)	15	32	12	16	16	–	–	–	–
M acc frame duration (ms)	30	32	240	16	16	8	11	16	
M acc frame length (gr symb)	119700	2048	48240	312	19968	2720	3322	4032	–
Frequency hopping rate (s^{-1})	–	–	250	–	–	125	–	–	–
Diversity BS	–	yes	–	yes	–	–	yes	yes	–
Diversity MS	–	–	–	–	–	–	yes	yes	–
Speech coding	SBC	RELP	SBC	RELP	RELP	–	–	–	–
Channel coding	RS	–	RS	RS	comp	RS	–	RS	–

Figure 2 Eight different prototype modems were offered for the Paris trials. Among these were Philips, Ericsson, Mobira (Nokia) [From Communications System Worldwide, Sept 1986]

The need for a championship to get further

After two years and no real firm decisions between various proposals, WG2 decided to invite nations to implement prototype modems and to conduct comparative measurements. Initially The Netherlands offered to host a measurement campaign, but very soon CNET in Paris offered superior laboratory facilities for the measurements. We agreed upon how to choose the best modem: *The modem which is able to handle most traffic within a limited frequency band, considering interference from surrounding base station reusing the frequencies. For this purpose a set of propagation channels were selected.* The neat thing about this criterion was that it considered both the modem's ability to withstand interference and noise and made it important to use little bandwidth. The most important channel was called "Typically Urban" and was selected after a lot of propagation experiments throughout Europe under the auspices of the EU in the group called COST 207.

Help from Sweden

Even if the Swedish Televerket had their competing modem candidate, they realized the threat from the competing consortia in ending up with something which would not be advantageous to Sweden. They therefore decided to help us in Norway with field measurements. They provided a chauffeur and a car with measurement facilities in Stockholm. The measurements were a great success since they confirmed my calculation that there must be an optimum bit rate for a certain terrain [1]. It turned out that 512 kbit/s outperformed 256 and 1024 kbit/s. For low bit rates, the likelihood of losing the signal was greater. If the bit rate was too high, a practical modem would not be able to handle the complexity. In Figure 3 this would correspond to a too small window. In between there was an optimum!

The Paris field trials, preparing for the championships

The Paris field trials were subject to a lot of attention. In the journal *Communications System Worldwide* from September 1986, Figure 2 was presented along with a description of the various proposals. About the Norwegian proposal was written:

> "Finally in this brief overview, some mention should be made of the system proposed by Elab of Trondheim, Norway. This was apparently designed in the behest of the Norwegian PTT, which wanted to be seen to be making some contribution to GSM.
>
> Since Elab is attached to a Technical Institute and has no manufacturing capability, it would appear to have little chance of success. Nevertheless, Hans Werner Lawrenz of the Deutche Bundespost describes it as a 'very interesting approach'.
>
> It uses a system called 'adaptive digital phase modulation', the basic idea of which is to cater for the difference between operating in large cities and in rural areas, and thereby increase the capacity of the system. Thus in cities, a lot of channels per carrier with a short coverage range and small cells are used. 'It's a good combination of coverage and bandwidth,' comments Lawrenz."

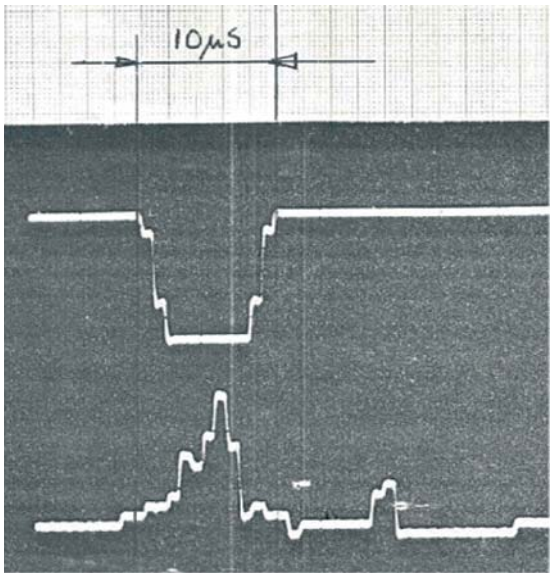


Figure 3 The upper curve is the window inside which the channel response (lower curve) must fit since that part outside will disturb. The modem is constantly trying to minimize the energy outside. The picture was taken at Skeppsbron in Stockholm in 1986. The window is around 4 bits at a rate of 512 kbit/s. When the bit rate is incresed, the window becomes shorter and there will be more disturbance from multipath propagation. When the bit rate is reduced, the window will be longer, but the components in the channel response less resolvable, which may cause the signal to disappear (fading)

During the Nordic Seminar on Digital Communication 14–16 October 1986 in Stockholm, William Gossling at The Plessey Company Plc gave a keynote talk:

> "What we are faced with now is a technology breakpoint in mobile radio. The true significance of second generation digital cellular is that it gives Europe not only a technically superior solution to the mobile telephony problem, but also the oppor-

ISSN 0085-7130 © Telenor ASA 2004

Figure 4 The Paris field trials [From New Scientist, Feb 1987]

tunity to create a world competitive supporting industrial structure.

It is time for our men of affairs to make the most positive outcome for Europe to be achieved. Let us be in no doubt that we already have a distinguished record of success. European achievements with Airbus Industrie, EUTELSAT, Arienne, Jaguar and Tornado is beyond debate. Now the time has come again.

Results from the championships in Paris

By Christmas of 1986 one of Europe's largest-ever multi-company research and development programmes successfully passed a critical phase of evaluation in Paris. This was accomplished with the usual drama of cancelled leave, overtime working and ruined fingernails that is common to such endeavour. I refer, of course, to the Paris field trials, designed to evaluate the various proposals for a European digital cellular radio standard."

In February 1987 the *New Scientist* (Figure 4) wrote:

"Cars equipped with the various technologies on offer were driven around Paris. The winner was a surprise- it was a system called ELAB, that was developed not by a large company but by Trondheim University in Norway."

The best modem proposed by Germany and France from Standard Electric Lorenz (SEL) coined "Wideband" became the candidate of France and Germany while all the other countries wanted the Norwegian proposal which was called "Narrow band". After some negotiation, our system was chosen for further optimization. Soon after, GSM decided on a common standard based upon our ideas described above, and the bit rate was set to 270 kbit/s. This choice of bit rate would ensure that the modem could work well in cities as well as in rural areas with long echoes.

After this, efforts to make a Norwegian mobile industry were started, but this is another story [5,6].

References

1 Maseng, T. On Selection of Systems Bit-rate in the Mobile Multipath Channel. *IEEE Trans. on Vehicular Techn.*, VT-36 (2), 51–54, 1987.

2 Maseng, T. Digitally Phase Modulated (DPM) Signals. *IEEE Trans. on Comm.*, 911–918, Sept. 1985.

3 Maseng, T. *The Power Spectrum of digital FM as produced by digital circuits*. Signal Processing, Amsterdam, Elsevier-North Holland, 253–261, Dec. 1985.

4 Maseng, T, Trandem, O. Adaptive Digital Phase Modulation. *Second Nordic Seminar on Digital Land Mobile Radio Communication*, 14–16 Oct. 1986, Paper no. 19.

5 Sand, G, Skretting, K. *Fortellinger om Forskning*. Trondheim, Tapir, 2002.

6 Gulowsen, J. *Bro mellom vitenskap og teknologi*. Trondheim, Tapir, 2002.

Professor Torleiv Maseng is Director of Research at the Norwegian Defence Research Establishment where he is responsible for communications and information systems. He worked as a scientist at SINTEF in Trondheim for ten years where he was involved in design and standardization of GSM. For seven years he was a scientist at the NC3A NATO research center in The Hague. During 1992-94 he was involved in the start-up of the new private mobile operator NetCom GSM in Norway where he had technical responsibility. Since 1994 he holds a chair in radio communications at the University of Lund in Sweden. In 1996 he took up his employment at the Norwegian Defence Research Establishment (FFI) located at Kjeller outside Oslo. He is the author of more than 150 papers, holds patents and is a Technical Editor of the IEEE Communications Magazine. He has received an award for outstanding research and has arranged large international conferences.
email: Torleiv.maseng@ffi.no

GSM Working Party 2 — Towards a radio sub-system for GSM

RUNE HARALD RÆKKEN

Rune Harald Rækken is Senior Research Scientist in Telenor R&D

The task of designing GSM was a complex issue requiring extended co-operation by several players. The overall design criteria for GSM was, amongst others, that the system should be at least as spectrum efficient as the analogue systems, and that the speech quality should be comparable to or better than that of NMT 900. The work on the radio sub-system involved computer simulations to evaluate the advantages and drawbacks of the different candidates, and the goal was always to optimise the solutions. However, the final solutions often turned out to be sub-optimal due to the system complexity. Another lesson learned is also that standardisation is sometimes about finding the best technical solution, and sometimes it is just about politics.

My entry into the GSM work

During my studies at the Technical University in Trondheim in the mid 1980s I got in touch with Torleiv Maseng, who was highly involved in working out a radio system candidate for the new pan-European mobile system that was being developed within CEPT. Maseng convinced me that mobile communications was a thrilling area where lots of development would be ongoing for years.

A guest lecture given at the University by Jan Audestad further convinced me. So, thanks to these two gentlemen, Torleiv Maseng became my tutor and Telenor R&D became my future employer. Since then my work has been dedicated to research and development of mobile communications systems and services.

During the spring of 1986 I simulated the behaviour of a digital mobile radio system exposed to various radio propagation conditions as modelled by the COST 207 group (European Co-operation in the field of Scientific and Technical Research, COST action 207: "Digital Land Mobile Radio Communications"), trying to figure out an optimum bit rate for a future mobile system that had to combat or exploit multipath propagation. The simulation work was really to predict the performance of the radio modem that Maseng and his companions were designing for the forthcoming "world championship in mobile radio", which is more comprehensively described by Maseng in another article in this issue of *Telektronikk*.

In my diploma thesis carried out during the autumn of 1986, my task was to analytically calculate the frequency reuse distance in a cellular mobile radio system. Mobile communications since the introduction of the NMT system in 1981 had become so popular and widespread that it was quite clear that a coming mobile system would be limited by interference from other users rather than by the link budget. Hence, tight re-use of radio frequencies was a critical issue to increase the overall capacity in the system. Luckily, my analytical calculations of frequency re-use in GSM aligned very well with the simulations carried out by people studying the GSM radio subsystem within CEPT.

My student work had introduced me to the work that was ongoing throughout Western Europe to define a common standardised system for mobile communications. Hence, the next step when I entered Televerkets Forskningsinstitutt (now Telenor R&D) was to join the standardisation group on the radio subsystem. Then I realised that what I had so far been involved in was only a microscopic fraction of the design to be undertaken to establish a new standard for mobile communications.

Organisation of the work on the GSM radio subsystem

In 1987 when I entered the GSM working party 2, Alain Maloberti from CNET in France chaired the group.

The group was further divided into three working groups: one dealing with channel organisation, one on modulation aspects, chaired by Torleiv Maseng, and one on handover and cell reselection issues. The working party consisted of a lot of skillful and also colourful people, aiming to pave the way for a digital mobile radio system to provide Western Europe with a modern system for mobile communications.

The overall design criteria for GSM was, amongst others, that the system should be at least as spectrum efficient as the analogue systems, and that the speech quality should be comparable to or better than the speech quality of NMT 900.

 ISSN 0085-7130

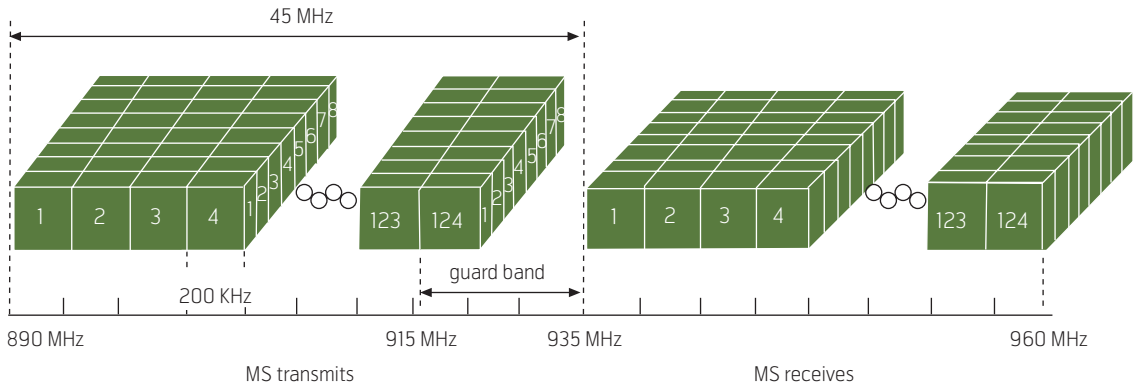


Figure 1 Radio frequencies and time slots in GSM, initial frequency band

To figure out what candidate could be most spectrum efficient and best performing was the main task during the Paris tests in the autumn of 1986. This was the most controversial issue that had to be solved with regard to the GSM radio sub-system. So – when the issue of wideband versus narrowband transmission scheme was solved, it was just "peace and harmony" in the radio subsystem work.

It is quite clear that designing GSM was a complex issue requiring extended co-operation by several players. As the development of GSM went on, huge parts of the system proposals were subject to computer simulations to simulate and evaluate the advantages and drawbacks of the different proposals. The simulations were always carried out with the goal of optimising the solutions, but often ending up with solutions that were sub-optimal. And of course, some discussions have been political rather than purely technical, as will always be the case in international co-operations.

My interest was at that time mainly in modulation, channel coding and equalisation aspects, so I joined that working group.

GSM radio sub-system

The protocols and mobility handling of GSM ("Global System for Mobile communications") are described elsewhere in this article. Here I will give a brief overview of the radio sub-system of GSM.

Frequency bands and access technique

GSM utilises TDMA (Time Division Multiple Access). Several users share the same carrier frequency, but transmit on separate time slots. Further, the mobiles are transmitting and receiving on different radio frequencies, utilising frequency division duplex (FDD).

Originally, the following frequency bands were allocated to GSM:

- 890 – 915 MHz for uplink (mobile transmitting, base station receiving);
- 935 – 960 MHz for downlink (base station transmitting, mobile station receiving);

giving a duplex separation of 45 MHz. In GSM the channel spacing is 200 kHz, giving room for 124 carriers in the available bandwidth.

Later on, there have been extensions of the GSM band, amongst others in the 1800 and 1900 MHz frequency range.

Generation of coded data blocks

GSM is designed for transmission of digitized information. As an example of how bits are organised into blocks for transmission over the radio interface, let's take a look at the full rate (13 kbit/s) speech channel.

Transmission of digitized speech is always a trade-off between maintaining a subjectively acceptable speech quality and offering the required spectrum efficiency and overall system capacity. The speech coding algorithm in GSM is called RPE-LTP (Regular Pulse Excitation – Long Term Prediction), and the work that led to this solution is described by Jon Emil Natvig in another article in this issue of *Telektronikk*.

The bit rate from the speech encoder is 13 kbit/s, resulting in a data block of 260 bits every 20 millisecond. Not all of these bits are equally vulnerable to bit errors, measured in terms of subjective speech qual-

ity. Hence, unequal error protection is used, giving strongest protection to the information bits being most vulnerable to transmission errors.

How then was this idea figured out?

Well, there were lots of simulations of error patterns that might result due to various propagation conditions and interference that GSM might be exposed to. The performance of the radio subsystem with different error protecting methods was simulated using those different error patterns, and people performed subjective listening tests to evaluate which coding scheme gave the best subjective speech performance.

Hence, the 260 bits of each speech encoder block is divided into three classes:

- Class 1A: the 50 most significant bits in each block is coded with a shortened (53,50,2) block code. Errors in these bits will significantly decrease the subjective speech quality. If errors are detected among these bits the Bad Frame Indication (BFI) is set, and the previous speech frame is repeated;
- Class 1B: 132 bits;
- The 182 class 1 bits together with 3 parity check bits and 4 tail bits are encoded with a half rate convolutional code;
- Class 2: the 78 least vulnerable bits from the speech frame is sent uncoded over the radio interface.

The encoding of a 260 bits speech frame into a 456 bits encoded data block is schematically shown in Figure 2.

For other logical channels other channel coding schemes are used. The result is always a data block of 456 bits (exception is the random access burst). These blocks of 456 encoded data bits are divided into sub-blocks of length 57 bits, being used to construct the data bursts that are transmitted over the radio channel.

The GSM transmission system

In GSM three different methods are used for forward error correction (FEC) to protect the transmitted data against transmission errors; that is, block encoding, convolutional encoding, and interleaving.

Utilising block codes, parity check symbols are added to the data symbols during the encoding operation. Block codes are capable of correcting data errors independent of the error position within the data

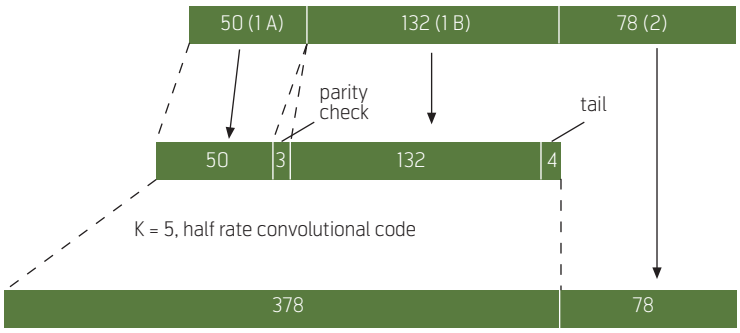


Figure 2 Forward error correction for full rate speech channel

block, meaning that block codes are well suited for correcting bursts of errors that typically occur on a mobile radio link.

Convolutional codes, on the contrary, operate on bit streams. The convolutional code applied in GSM has a constraint length K = 5, meaning that during the encoding process the bit at the input of the encoder is taken into consideration, in addition to 4 preceding bits.

What a convolutional code really does is spread the effect of one bit by introducing intersymbol interference. A convolutional code thus introduces time diversity, hence the channel noise is averaged of K data symbols. This is the mechanism giving error correction properties.

It is worth noting that this is very similar to the ideas exploited by Maseng et al. when they built their candidate modem for the "world championship in mobile telephony". It was also thought during the standardisation process that the same Viterbi decoder could be used both for resolving intersymbol interference and for decoding the convolutional code. This is probably the reason why it was clearly stated that GSM should be able to cope with intersymbol interference spread over four bit intervals.

A convolutional code performs badly when there are bursts of errors. If however the bit errors are "randomised" in position, a convolutional code can correct errors even with rather high bit error rates. It is therefore of fundamental importance to introduce mechanisms to spread bit errors, particularly due to the fact that errors on a mobile radio channel has a tendency to occur in bursts.

Interleaving is such a "whitening process" which spreads bursts of errors to optimise the decoding of

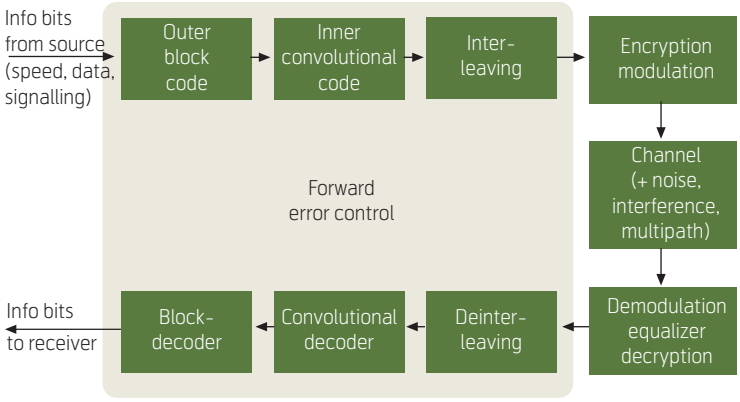


Figure 3 Structure of forward error control in GSM

the convolutional decoder. Interleaving performs better the larger the interleaving depth is (the more the bits in a block are spread) because the de-correlation between successive bits increases with increasing interleaving depths. The drawback is that increasing interleaving depth also leads to increasing decoding delay, hence there is always a trade-off between performance and delay due to the interleaving process.

For the non transparent data channels and the signalling channels, there is in addition to the FEC mechanisms also a layer 2 protocol using retransmission (ARQ) of data blocks still being erroneous after the FEC.

Multiplexing

The data bursts being transmitted over the air interface in GSM consists of information bits and a known bit pattern (midamble – which is a training sequence) that is being used for synchronisation and also to estimate the impulse response of the radio channel to utilise knowledge of the radio propagation conditions in the decoding process (soft decoding).

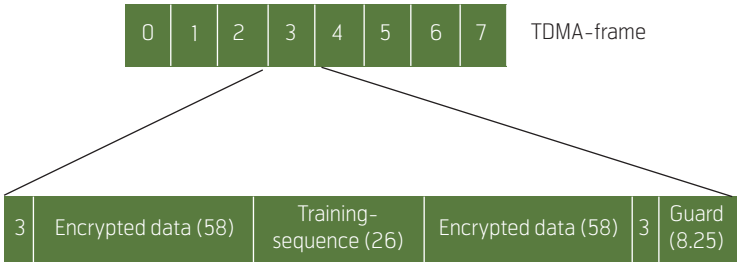


Figure 4 Normal burst and TDMA frame. The length of the components is given in number of bit intervals (3.69 μs)

The bit rate (burst bit rate) over the GSM radio interface is 270.833 kbit/s, and the burst length is hence 0.577 ms. The length of the normal burst is 148 bit intervals, in addition to a guard space of 8.25 bit intervals. 8 such data bursts are multiplexed into a TDMA frame of length 4.615 ms. The normal burst is used for most purposes in GSM.

Modulation

The modulation method used in GSM is Gaussian Minimum Shift Keying (GMSK) with a time-bandwidth product of 0.3. This is a partial response modulation with constant envelope, meaning that all the information is in the phase of the received signal. A time-bandwidth product of 0.3 means that a pulse shaping filter is used that introduces intersymbol interference in the transmitted bit stream to obtain a smaller modulation spectrum.

In the recent extension of GSM, EDGE (Enhanced Data rates for Global Evolution) the main difference from "plain" GSM is that the modulation method 8-PSK is used instead of GMSK, offering 3 bits/symbol instead of 1 bit/symbol in GMSK. This paves the way for higher bit rates in EDGE than in GSM.

The narrowband transmission scheme designed by Maseng et al. outperformed the other system candidates during the Paris tests in 1986. This modem used adaptive digital phase modulation (ADPM). ADPM however was replaced by GMSK rather quickly after the Paris tests. With today's distance to this change of modulation method, it should be revealed that the replacement of ADPM was not solely founded in technical arguments, even though it was claimed that GMSK was a better known modulation method which also allowed a smaller modulation spectrum than ADPM. The choice of GMSK was primarily a political compromise to motivate France and Germany to accept the narrowband GSM solution that the rest of the GSM players had accepted. It should also be emphasised that we are very satisfied with the solutions that the GSM radio transmission system is based on.

It was during this discussion on modulation methods that I as a newcomer to this field realised that standardisation is sometimes about finding the best technical solution, and sometimes it is just about politics.

Literature

Langewellpott, U, Rainer, M. Modulation, Coding and Performance. In: *Digital Cellular Radio Conference (DCRC)*, Hagen, Germany, 1988.

Maseng, T. Wideband or Narrowband? World championship in mobile radio in Paris 1986. *Telektronikk*, 100 (3), 161–164, 2004. (This issue)

Natvig, J E. The 1987 European speech coding championship. *Telektronikk*, 100 (3), 182–186, 2004. (This issue)

Ochsner, H. Overview of the Radio Subsystem. In: *Digital Cellular Radio Conference (DCRC)*, Hagen, Germany, 1988.

Rækken, R H. *Feilkorrigerende koding i GSM (Error correcting codes used in GSM)*. Kjeller, Telenor R&D, 1993. (R&D note U 1/93)

Rune Harald Rækken (42) received his MSc degree from the Norwegian Institute of Technology (NTH) in Trondheim in 1986 and has since then been employed by Telenor R&D. In 1987 he started to work on the GSM specifications, with focus on modulation, equalisation and channel coding. Later on, he has worked on radiowave propagation measurements and modelling. Recently he has been focusing on UMTS and IEEE 802 standards for wireless communications. He is now a senior research scientist working in the mobile systems group at Telenor R&D.

email: rune-harald.rakken@telenor.com

The Mobile Application Part (MAP) of GSM

JAN A AUDESTAD

Jan A. Audestad is Senior Adviser in Telenor

This is the story of how the Mobile Application Part was developed. MAP is the neural system interconnecting the distributed computer infrastructure of GSM (VLRs, MSCs, HLRs and other entities). The work on MAP started as a study of the general architecture of mobile systems in ITU two years before the GSM group was established. In 1985 the work was taken over and later completed by the GSM group. MAP was the first protocol of its kind in telecommunications systems. After long and difficult negotiations with the community of switching and signalling experts, MAP got its final structure in 1988. This included the use of Signalling System No. 7 as carrier and the support of services such as roaming, call handling, non-interruptive handover, remote switching management and management of security functions. MAP is one of the real technological triumphs of GSM.

Interconnecting operators: The three sausage model

This is the story of how the Mobile Application Part (MAP) was created and the problems and conflicts that this development caused between the communities of enthusiastic mobile communications experts and the conservative and preservative telephone switching engineers.

The idea to design an international land mobile communications system came up in CCITT[1)] Study Group XI (Switching) in late 1980, more than two years before GSM was established. The first document describing how mobility was achieved in the NMT system was presented to the Plenary Assembly of CCITT in 1981. The document was appended to a proposal for a new study question on land mobile systems. The authors were Bjørn Løken and myself. The Study Group found the proposal so interesting that they appointed Bjørn Løken as Interim Rapporteur so that work on land mobile systems could commence immediately without awaiting the approval of the Plenary Assembly.

From mid 1981 the work on land mobile systems in CCITT took off at a violent speed and important results were obtained early. The three most important results obtained during the early years were a simple network model nicknamed the "tree sausage" model, a method for non-interrupt handover of calls between different switching centres, and the outline of a protocol supporting mobility within and between Public Land Mobile Networks (PLMN). The latter was later christened the Mobile Application Part (MAP). These achievements were all adopted and developed further by the GSM group.

Figure 1 shows the "three sausage" model of a land mobile system. The sausages are the two PLMNs and the fixed network. The significance of this model is that it is entirely abstract; that is, independent of the particular physical design of the network. The model can be used to describe all major activities taking place in the system and it is general enough to apply to every practical configuration of a mobile system. This allowed MAP to be developed independently of a particular network architecture: MAP is thus not specific for GSM.

The PLMN represents an entity of network ownership. The interaction between PLMNs is thus a cooperation of different mobile network operators allowing mobile subscribers to roam between networks independently of ownership and subscription.

The model shows two PLMNs together with the fixed network. MAP supports mobility between the PLMNs, that is, MAP offers to mobile terminals the capability to move from one PLMN to another PLMN retaining the capability to receive and make calls. Interconnectivity between a mobile terminal and a fixed terminal or between two mobile terminals takes place across the fixed network. The signalling protocol connecting the PLMN to the fixed network is Signalling System Number 7 (SS No 7). The plan was to implement SS No 7 in the telephone network during the 1980s so when an international mobile system was ready for operation around 1990, this would be the natural choice of interconnection method. When GSM was put in operation in 1991/92, SS No 7 was in place and mobile communications could commence immediately.

1) *The Consultative Committee on International Telephony and Telegraphy of the International Telecommunications Union (ITU).*

ISSN 0085-7130 © Telenor ASA 2004

PLMN architecture: The heritage of NMT

Another development that took place prior to GSM was the specification of an internal architecture of land mobile systems as shown in Figure 2. The architecture is based on the principles first developed for NMT and were adjusted for use in more flexible configurations by the CCITT.

The architecture is simple. In addition to the radio infrastructure (base stations and cells) and the telephone exchanges switching the calls between the base stations and the fixed network (MSC), there are two types of databases: VLR (Visitor Location Register) and HLR (Home Location Register)[2]. The VLRs are in charge of a location area. Within this area the mobile stations can roam without updating their location[3]. Location updating takes place when the mobile station roams from one location area to another. The VLR contains all information about the mobile terminals currently in its location area required for establishing calls to and from these terminals. The VLR also controls the switching process in the MSC. In this respect, the VLR is an intelligent network (IN) node similar to the concept developed independently by Bellcore during the 1980s for serving telephone calls requiring centralised control (routing and payment management of free-phone and premium rate calls, and management of distributed queues and helpdesk functions). In mobile networks, identity management, location updating, routing administration and handling of supplementary services require the support of similar procedures as intelligent networks. Therefore, it is not surprising that the two groups came up with rather similar solutions to the problem of remote control of switching processes independently of each other.

The HLR is a database containing subscription information (services and capabilities) and the current location of the mobile terminal. There is usually one HLR for each GSM operator[4].

In addition, there are also other databases and entities not important for the basic architecture (equipment registers, authentication key escrows, voicemail systems and short message centres). These databases are also connected to the other network elements by MAP.

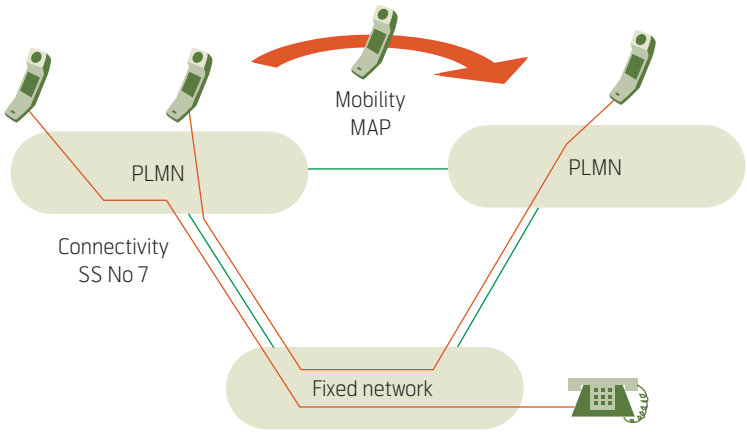


Figure 1 The three sausage model

There are permanent MAP connections between the VLR and the MSCs it is controlling. There are sporadic MAP connections between the VLR and the HLR for location updating, between two or three MSCs for handover, and between two VLRs for management of identities during location updating. Sporadic means that a relationship is established only when two such entities must exchange information and controls.

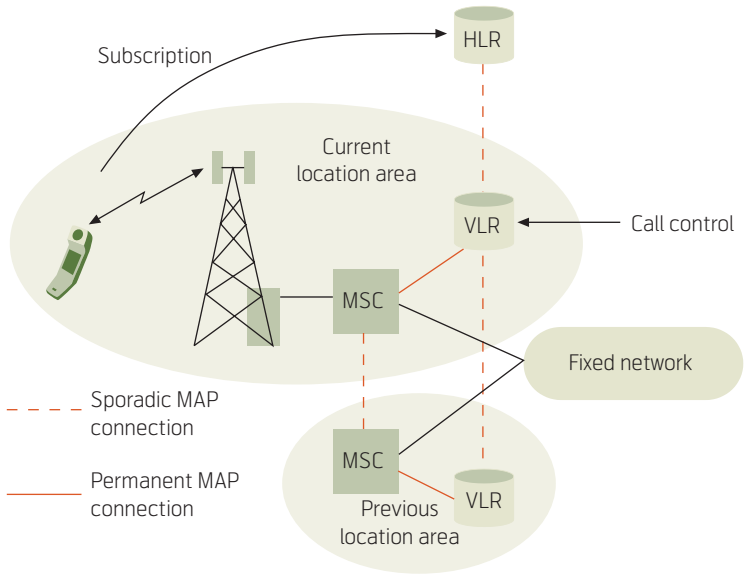


Figure 2 Architecture of GSM

2) *These names were inspired by NMT: HMX for home mobile exchange and VMX for visiting mobile exchange offering similar functions as the HLR and the VLR, respectively.*

3) *A VLR may control a number of location areas. Roaming between such location areas only require updating of the VLR and not the HLR.*

4) *If there are many mobile subscribers, two or more HLR databases may be required. Such configurations are regarded as a single, distributed HLR.*

ISSN 0085-7130

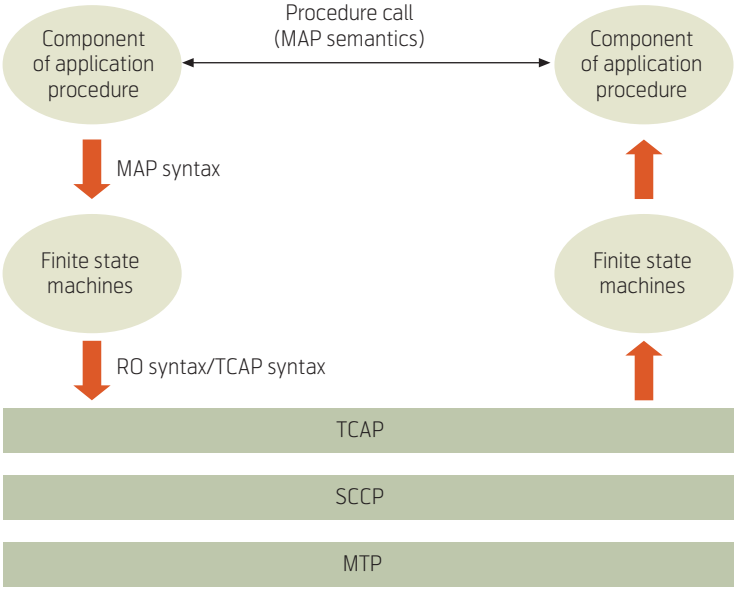


Figure 3 MAP protocol

Many of the operations and procedures supported by MAP were outlined in CCITT before the GSM group was established. The work had in fact progressed far before GSM took over the specification in 1985.

Cooperation between SPS/SIG and GSM

Until 1985, not many of us believed that Europe would ever be able to specify a common land mobile system. CEPT[5)] had this far never succeeded in its standardisation efforts so there was no reason to believe that it would succeed this time either. At the GSM meeting in Berlin in October that year, this attitude changed. The core team of the GSM group had now become tightly consolidated and it was decided unanimously that every effort should be made to make GSM a success. An outline of 110 recommendations that should comprise the complete GSM specification was drafted at the meeting so that work could commence on solving concrete problems. Amongst these recommendations was the MAP specification.

This was the kick-off the GSM group needed.

CEPT already had a working group dealing with signalling, SPS/SIG. This group was asked to take charge of the MAP specification. Christian Vernhes, who had worked on MAP right from the beginning in CCITT, became chairman of the team in charge of the specification. In addition, the team consisted of Bruno Chatras (ASN.1 specification) and myself (procedures). Alfred Karner joined the team in the autumn of 1988 cleaning up the formal aspects of the specification.

The request from the GSM group was to deliver a complete specification by January 1989. SPS/SIG met this requirement. Both the syntax and the semantics of the protocol had then been methodically tested and verified by Siemens. This included testing the ASN.1 syntax of the protocol and all procedures that were written in SDL.

Protocol structure

The final structure of the protocol is shown in Figure 3. GSM is a distributed system where most of the procedures are divided into components located in different computers, for example, in the mobile terminal, in one or two VLRs and in the HLR for location updating. A handover between cells connected to different MSCs involves simultaneous execution of software components in the mobile terminal, in two base station tranceivers, two base station controllers, two or more MSCs and one VLR. The components in the MSCs and the VLR are interconnected by MAP. The other components are interconnected by other protocols.

MAP is the core element in the network architecture of GSM, tying together MSCs, VLRs, HLRs and other specialised databases and equipment. The configuration of the protocol is shown in Figure 3.

The MAP semantics corresponds to the exchange of procedure calls between the components of the distributed application procedure as shown in the figure. The syntax and the semantics of these procedure calls are mapped onto the protocol via a set of finite state machines in order to transport the commands across the network. The machines not only map the syntax and the semantics but also take care of software synchronisation, ensure that the right software components are bound to each other during the whole transaction, and detect and respond to exception conditions. The state machines ensure that the abstract syntax of MAP can be interfaced to any application layer protocol. MAP can in fact be interfaced with any application layer protocol supporting procedure calls, for example RPC (Remote Procedure Call) of the Internet.

5) *Conférence Européenne des Postes et des Télécommunications. The part of CEPT in charge of telecommunications standardisation later became ETSI (European Telecommunications Standards Institute).*

The protocol stack on which MAP was finally implemented is shown in Figure 3. The stack consists entirely of sub-protocols of SS No 7: TCAP (Transaction Capabilities Application Part), SCCP (Signalling Connection Control Part), and MTP (Message Transfer Part).

In search of a suitable protocol stack

The reliability of MAP must be as good as in telecommunications networks at large. This means that the maximum outage of the protocol should not exceed 15 minutes per year which is the standard reliability requirement for telephone exchanges, telecommunications links and signalling systems. This corresponds to the ultrahigh availability of 99.997 % – that is, the system must function properly 99.997 % of the time. Therefore, the stringent requirement on availability puts constraints on the choice of protocol stack for supporting MAP: all layers in the stack must be at least as good as this.

Another important requirement is the time budget as shown in Figure 4. The time budget has to do with psychology. The time from a telephone call is initiated and until the called user is connected (i.e. activation of the ringing tone) must not be more than one or two seconds; otherwise the calling user may prematurely clear the call because a few seconds is a very long time when waiting for the ringing tone. The total time must be divided among a large number of processing events allowing 100 milliseconds or less for each of these processes. In the GSM system, such an event may be information exchange between the VLR and the HLR during call establishment. On average the exchange of messages, including processing time at each end of the connection, should not require more than 100 milliseconds on average and less than 200 milliseconds for 90 % of the calls. Broadly speaking, this allows for typically 30 milliseconds one-way delay (5000 km) and 20 milliseconds processing time. The delay includes processing time at any intermediate node such as router for directing the signalling information.

The availability requirement and the timing constraint put severe restrictions on the choice of protocol stack.

Only three protocol stacks suitable for transfer of MAP messages were available in the mid 1980s. These are the Internet suite of protocols, the protocols of public packet switched data networks (X.25), and Signalling System No 7. One problem was that these stacks were not even complete; that is, they did not contain all layers that were required for supporting the application protocol.

The most complete protocol stack was RPC over TCP/IP. MAP fitted nicely into the RPC format, TCP ensured end-to-end integrity and IP was able to route the messages across the network. Around 1985, the Internet was only used as a research network. The network was owned by no-one and there was no formal organisation in charge of its evolution. The reliability and quality of service of the network was unspecified except that messages were delivered on a best effort basis, that is, there were no specification concerning how much time it would take to transfer a message across the network. The Internet as such was therefore not suitable as a carrier of information requiring real time operation and ultra-high availability.

Nevertheless, one possibility we had was to apply the Internet *technology* between MSCs, VLRs and HLRs – not the Internet itself. Exploitation of the Internet technology offered us one possible, though not the wanted, solution. We kept this as the last resort solution all the time up to the summer of 1988.

Another possibility was to use the X.25 data network. A complex stack of application protocols, the Open System Interconnection (OSI) suite, was specified by CCITT and ISO during the 1980s. The problem with the X.25 network was that it did not contain secure enough mechanisms for autonomously restarting operation after a major link failure. These had to be built into the specification before X.25 could be used for high reliability transfer of information. Furthermore, the network supporting MAP had to be a dedicated network for MAP only because otherwise other traffic could cause freeze-out or excessive delays. X.25 did not offer more than what the Internet technology did for a fraction of the cost. X.25 was therefore never a serious alternative.

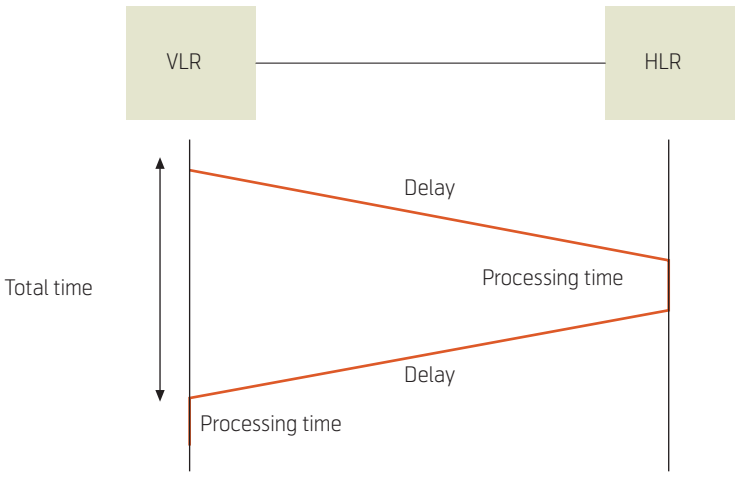


Figure 4 Time constraints

ISSN 0085-7130 © Telenor ASA 2004

The solution we wanted was Signalling System No 7 because this protocol was designed for extreme reliability and real time operation. However, the signalling system did not support higher layer protocols ensuring end-to-end integrity. The basic protocol designed for operation between telephone exchanges consisted only of a data link layer and a simple network layer that could only be used on a very simple network infrastructure. This is the Message Transfer Part (MTP). In order to convert the signalling system into a data network supporting flexible routing, the Signalling Connection Control Part (SCCP) was developed. This is essentially a data network combining the features of IP and X.25, offering both connection-oriented and connectionless operation. However, the SCCP had a separate connectionless category offering guarantee of delivery. The signalling system offered neither a transport protocol nor application layer protocols. We needed these protocol layers in order to ensure proper end-to-end control of the messages.

We requested the signalling subgroup of Study Group XI already in 1984 to specify such protocol layers but our request was ignored. It came as a surprise when we received a question from the same group in 1986 asking us to provide them with the set of requirements we needed to implement MAP. The group had started specifying an application layer protocol called the Transaction Capabilities Applications Part (TCAP). We learned much later that the reason this was done was not to satisfy our needs but to support the development of the intelligent network in the USA.

This time things also went wrong because Study Group XI specified a protocol that was adequate for the simple intelligent network protocols but was far too simple for a complex protocol like MAP. We had to make some additions to the specification before it could be used.

Transaction Capabilities (TCAP): The best we got

TCAP consists of two layers: the component sublayer supporting the Remote Operations (RO) protocol and the transaction sub-layer protocol. RO is a remote procedure call protocol standardised as part of the OSI development. It allows two application entities to exchange invoke commands requesting that a remote task is executed, return of the results of the execution and provision of error codes if things go wrong.

The transaction layer is a connection-oriented protocol where the information generated by RO is sent in Begin, Continue and End messages supporting obvious functions. The protocol offers multiplexing since several RO messages can be sent in one transaction message.

However, it turned out that TCAP was too simple for a protocol as complex as MAP. TCAP did not offer appropriate mechanisms for controlled establishment and release of the connection, explicit binding of the application modules, activity management, recovery management and software synchronisation. It never occurred to Study Group XI that there were good reasons for the complex application layer protocols defined by ISO for general data communications: association control, session management, concurrency management and reliable transfer of information.

The functions needed for proper operation of the GSM system (association control, explicit binding, session management, and recovery) were included in the MAP specification itself, making it more complex and less universal than necessary.

Alignment with the fixed network: Much hindsight, little foresight

One of the major problems we faced was the alignment of GSM and the fixed network. The general opinion of the fixed network community, including the market analysts of the telecommunications operators, was that mobile communications would remain a small service as compared to fixed telephony in all future. Statistics from the Nordic countries showing a steady and tremendous growth of mobile communications was brushed aside with the argument that this was just a transient that would soon be over.

This attitude was global, where the strongest opponents were the North-American telecommunications operators for which mobile communications represented a real competitive threat. Mobile communications was a way in which new companies could perforate the monopolies of the seven RBOCs[6)] – nicknamed the Baby Bells – that AT&T was sliced into during the divestiture of AT&T in 1984.

The small but growing community of people working on mobile systems had come to the opposite conclusion: the design of handheld mobile terminals had become feasible, moving the mobile phone out of the car and onto the pavement. The potential market for mobile communications was then just as big as the

6) *Regional Bell Operating Company.*

market for fixed telephony, or perhaps even bigger: telecommunications was about to enter a new era where flexibility and ubiquity would become the most important drivers of market growth.

One undisputable requirement that we had to accept right from the beginning was that the GSM system should not have any impact on the specifications of the fixed network. This included signalling, routing algorithms, numbering and number analysis, switching functions and network architecture.

We may regret this attitude now.

Let us look at some problems that this decision caused not just for the GSM system but for telecommunications in general.

GSM uses non-geographic numbers. This means that there is no relationship whatsoever between the telephone number of a mobile station and where that station is located. The number only identifies where in the network the subscription information and the current location of the subscriber can be found. We had to invent new methods by which we could handle the routing of calls to mobile terminals since this was the first time that the problem of non-geographic routing occurred in real systems.

For several years the MAP team and the GSM group proposed that the routing algorithms of the telephone network should be altered allowing a look-ahead procedure in MAP by which the location of the called mobile subscriber was first found and then the call was sent directly to the destination. This procedure was proposed in order to avoid "tromboning"; that is, to avoid that a call from an Italian to a Norwegian on vacation in Italy is set up all the way from Italy to Norway and then back again to Italy. The application of a look-ahead procedure would ensure that the call could be established on a short connection within Italy. At this time, the look-ahead procedure was regarded as a feature solely required in GSM. However, if implemented, the look-ahead procedure would have solved several routing problems such as number portability, allocation of personal numbers, role and time dependent routing of calls, and selection of the shortest and cheapest route for voicemail calls, free-phone calls, calls to distributed helpdesks, calls to premium rate services, and so on. It would also have enabled a complete integration of GSM and intelligent network services.

The feature was neither accepted by the CCITT nor by ETSI. This decision resulted in many problems later as the use of non-geographic numbers has become the rule rather than the exception.

Another problem was concerned with segmentation of messages in the SCCP. No such service was included in the original specification of the SCCP since the opinion was that segmentation of long messages had to be done by the application using the SCCP. It took us several years and numerous meetings to persuade the designers of the SCCP that segmentation at the application layer is impossible because there is no simple way in which the length of application messages can be determined. The reason is that the application protocol is written in an abstract syntax requiring compilation in exactly the same way as any program written on the computer. The compilation of the messages takes place at the interfaced between TCAP and the SCCP and it is only there that the exact bit pattern of the message is known. The length of a message then depends on both the syntax of the MAP message and the format of TCAP: the bit pattern is different for the same MAP message if it is sent in a Begin or a Continue format. The length and content of the bit pattern depend on such a subtle detail as the invoke number of the RO message and the transaction number of the transaction message. Segmentation must then either take place in TCAP after compilation or in the SCCP before the message is sent. The message must be reassembled before it is delivered to the de-compilation process of TCAP wrapping up the transaction messages and the remote operations in order to pick out the MAP message.

The task to make the signalling experts understand this subtle detail was difficult but in the end the SCCP was amended in order to support segmentation. Otherwise we had been forced to resort to solutions based on TCP/IP.

A final problem was concerned with addressing. SCCP offered three addressing schemes: the point codes identifying exchanges, subsystem number intended to identify a specific box or function at the signalling interface, and the global title which is nothing but a telephone number identifying the termination. The point code is not globally unique and cannot be used for MAP; MAP required so many subsystem numbers in order to identify the different functions and boxes that too many of the available codes were used up. We could therefore only use the global title. This seems trivial but it is not.

In data communications, several addresses are usually needed in order to identify the process that needs to be invoked. The global title identifies only the VLR, HLR, MSC or whatever other type of equipment is found at the termination. The global title does not identify each of the several functions that the VLR is in charge of. Noting that the VLR may not be a single

computer but a complex of many computers makes the problem even bigger. This means that the final destination of a TCAP message – the application module that shall perform the actual task – can only be identified from addressing information contained in the transaction messages, the remote operations messages or the MAP messages.

In order to understand that this is a problem, let us look at how this feature is solved in TCP/IP.

Broadly speaking, the IP number identifies the computer. The header of IP contains a field indicating which type of protocol is contained in the information field of the packet, e.g. TCP or UDP. This allows the computer to select the correct software for analysing the content of the information field. Similarly, TCP (or UDP) contains an address called the port number. The port number identifies a computer program that is capable of analysing the application message embedded in the information field of the TCP datagram. Behind each port there is then a particular software that is able to interpret the format of the message (http, email, mpeg and so on). Based on the analysis of addresses and other information included in the application message the computer can then finally start the program that can actually handle the content of the message.

When the SCCP and TCAP were specified, this way of handling computer interactions was obviously not understood properly. Otherwise, a transport layer would have been introduced just in order to handle addresses beyond simple routing addresses in complex computer systems. A protocol as simple as UDP would have done the trick.

Without the support of a transport layer protocol and supporting application layer protocols, MAP relies too much on Signalling System No 7 and became less flexible than we hoped when we started developing it.

In the end a success

MAP was the first application protocol designed for signalling purposes. The first version of it was sent for approval of the GSM group in February 1989. The work on the second version had already started a couple of months earlier. In less than one year the second version was finalised, in time for implementation in the first GSM network. Version 1 was never implemented. The first version contained more than 650 pages including appendices and supplementary material; the second and third versions contain about 750 and 1000 pages, respectively.

The first version consisted of 54 individual procedures. The newer versions are just a little bigger.

It took much time to develop the first version of MAP. The reason was that there was no similar protocol that we could use as template. In fact, we had to make all the mistakes ourselves in order to learn how to do it. The ASN.1 encoding of the protocol, for instance, was rewritten several times before we understood how to write it. I think we misinterpreted the ASN.1 language in all the ways this possibly could be done. Finally, Alfred Karner showed us how to do it.

The application protocols developed later for intelligent networks, supplementary services in the telephone network, universal personal telecommunications and UMTS are all based on the methods and principles developed in MAP. Some of the members of the MAP team were invited to various groups in CCITT and ETSI to explain how to interpret the ASN.1 and SDL specifications and apply these languages in protocol design.

The development of the MAP protocol was indeed pioneering work.

For a presentation of the author, turn to page 54.

Signalling over the radio path

KNUT ERIK WALTER

Knut Erik Walter is Senior Technical Adviser in Telenor R&D

"Mobile to network signalling" – it doesn't sound very exciting, does it? Nevertheless, it is a fact that this area is the very central part of the GSM system since it reflects the logic (or intelligence?) of the functionality that makes a GSM mobile with its SIM-card the powerful communication tool that we know. It contains a number of different protocol levels and procedures that perform the necessary housekeeping which together allows the user of a mobile station to concentrate entirely on the services. This article contains a short description of the protocol levels and main procedures, and a personal impression of the story of their invention.

How did it start?

After the original GSM group had established the three first subgroups (or "Working Parties" which were the official wording), the responsibility of all signalling was placed in WP3 "Signalling and Network Aspects". The whole idea was to define precise specifications for a set of interfaces between logical groupings of functions. One very obvious interface was the air interface, which also constitutes the border between the subscriber's equipment and the network. The usual working procedure was then, as now, to appoint an editor for a defined recommendation. The responsibility of an editor is typically to draft a first version and to include additions and changes based on written input documents and meeting discussions.

Dr. Woldemar Fuhrmann, then working for the German incumbent operator DeTeCon was appointed as the editor of the core specification of the GSM radio interface Layer 3. Dr. Fuhrmann was a very skilled young man with experience from data communications and science, and in addition he had an artistic talent for drawing, together these characteristics made him the best possible choice for the challenging task.

The first version of 04.08, which became the official GSM recommendation number for the Layer 3 specification for mobile-network signalling, was presented for WP3 early 1986. It was already then a comprehensive document with a number of SDL (System Description-Language) diagrams of the protocol behaviour included. It soon became clear that the amount of work and discussions needed to finalise this area was too large to be handled in the WP3 alone. There was simply too little available time within the meetings, and too long period between the meetings to get the work done. It was then decided to establish a subgroup chaired by the editor in order to speed up the process. The group was given the describing name: Layer 3 Expert Group, or L3EG for short. After the first meeting in Paris in August 1986 with four persons present, a very intense working period involving around 10–15 people for a period of approximately two years followed before the recommendation was considered stable.

The task and its challenges

To understand the discussions and results from WP3 and L3EG we must remember the situation in the telecom world at the time. ISDN had been the subject for definitions already for some years, but no implementation had yet been done. SSNo7, the first really digital and datacom-inspired signalling system was just in its introduction phase, and data communications was mainly handled in stand-alone dedicated networks (either circuit-switched or packet switched). The common vision for the future telecom networks was based on ISDN and SSNo7, both of them being inspired by the datacom world through the ISO OSI-model (Open System Interconnection), describing functions belonging to separate independent protocol layers. It was early decided that GSM should be a mobile extension of the ISDN network; i.e. it should encompass as much as possible of the ISDN features and functions, but still make efficient use the radio spectrum. The mobile switches (MSC) should look like normal ISDN switches on their external interfaces, and the subscriber connections (i.e. the radio path) should logically be as close as possible to the ISDN subscriber lines.

Inspired by the OSI model and the ISDN subscriber interface specifications the need for three distinct protocol layers was identified:

- Layer 1 (the physical layer), closely related to the physical medium, with the responsibility of mapping the protocol units to the physical channels, encoding/decoding, power control etc. in order to convey the information over the radio channel.

 ISSN 0085-7130

- Layer 2 (the data-link layer), with its main functions in error correction and to segment longer messages into suitable units for transfer by Layer 1 (and re-assemble them at the receiving end).

- Layer 3 (the network layer), whose main functions are to provide the actual GSM services (speech calls, data calls, SMS, roaming ...).

Just as ISDN, GSM was (and still is) mainly a circuit switched network, i.e. in order for information to be sent between two users, resources constituting a "channel" are established and maintained for the duration of the call. This also makes it meaningful to distinguish between a Signalling Plane (used for all control functions related to establish, maintain and release the information path) and a User Plane defining the information path itself. The procedures and design of the protocols described here all relate to the signalling plane.

According to the OSI model principles the protocols were defined by specifying the actual PDU (Protocol Data Units) to be sent between the protocol entities belonging to the same layers. On the mobile side the mapping between the logical protocol entity and the physical equipment was trivial, everything must reside in the same unit. On the network side it was more complicated. Layer 1 was obviously terminated in the Base Transceiver station (BTS), and the same was decided for Layer 2, since this was also specifically designed for the radio path. The Layer 3 signalling however is partly terminated in the BTS, partly in the BSC (Base Station Controller), and partly in the MSC. This reflects the fact that the concept of protocol entities of different layers is just a practical model of the system, and care was taken not to put much constraint on the actual implementation. While the actual format of the information to be sent on a physical interface needs to be standardised, the internal design should be left to the implementers. The task of the GSM standardisation group was to define interfaces and interconnection principles, not implementation of physical units.

With the OSI model and ISDN specifications present, why was it such a great task to define the equivalent version for GSM? The answer lays in two simple facts related to mobile communications:

- The physical connection is not permanently available, and its characteristics are in addition highly variable.

- The subscribers move freely around between the network access points.

The Layer 3 functions and the search for an appropriate description model

As described above, a number of additional requirements are placed on the design of a mobile system compared to a traditional circuit switched telecommunication network. Let us look a little closer at the consequences of the two areas of new requirements:

The first one is related to the fact that radio is used as a communication medium. Since radio frequencies are a scarce resource, no system for public use can allow resources to be permanently allocated to every user (contrary to the last mile of traditional telephony). A mobile system therefore must include functions to manage the frequency resources according to the user's needs. Methods for access control, mobile station identification and channel allocation are crucial. In addition, functions are needed to adjust and control the resources when the user moves during an active communication. This includes functions like power control, timing adjustments and handover between base stations. The term "Radio Resource management" (MM) was given to this group of functions.

The other area that imposed challenges is due to the core of a mobile system, the user's mobility. Since the system cannot use the access point as an implicit identification of the current user, procedures must be defined in order to verify the real identity before any access to the network resources can be granted. Similarly, there must be mechanisms for the network to locate and connect users when needed, e.g. in conjunction with incoming calls. This functionality area was named "Mobility Management" (RR), and includes functions like identity management, authentication and location updating.

In the ISDN subscriber line Layer 3 specification (CCITT Q931) is used a logical protocol model for the terminal side and network side respectively. In the model the protocol entities are defined as finite state machines, i.e. only a limited set of defined states are allowed, and moving between states is triggered by internal or external signals. (Internal signals are typically timer expiries, and external signals are protocol messages received from the peer entity). The ISDN specification contains definitions of the states related to call set-up, call modification and call release, all of them relevant to GSM. An example of the states and transitions relevant to call set-up is shown in Figure 1.

The first version of GSM Layer 3 specification tried to encompass the RR and MM functionality in the same protocol entity, i.e. the state machine had to get a number of new defined states reflecting the appropriate status of RR and MM relations (channel connected, location updated and others). It soon became

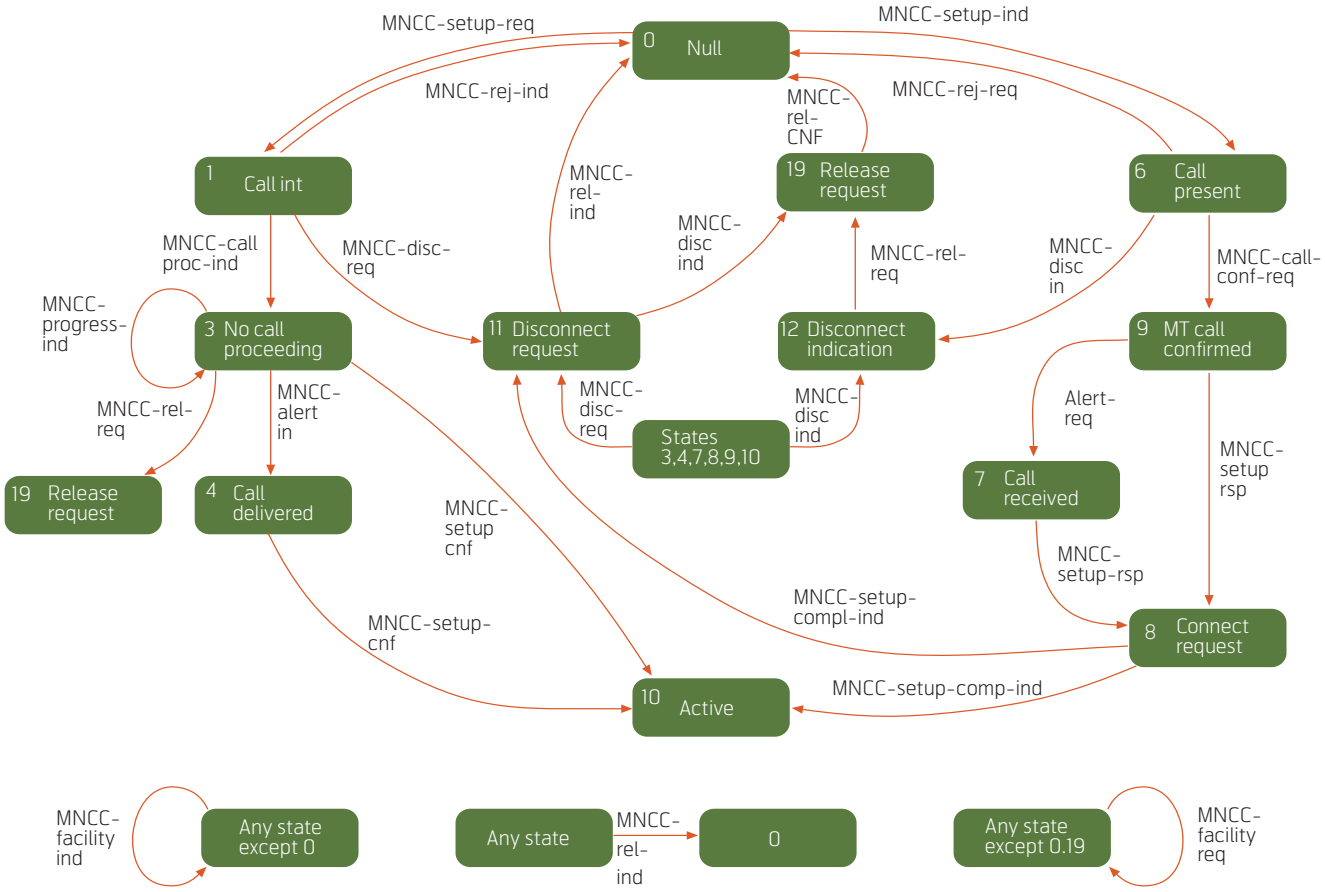


Figure 1 Service graph of Call Control entity – MS side (from 3GPP 24.007)

clear that the result was too complex; there were simply too many variables to handle and still ensure that the result was a stable protocol. The L3EG therefore decided to divide Layer 3 into three logical entities: Call Control, Mobility Management, and Radio Resource Management. Each of them could then be described with a manageable number of states and allowed transitions, and with a set of protocol procedures per unit. The challenge still remained, however, how to describe their internal relations? Long discussions both in the meetings, during lunches and in dark pubs in late evenings kept the L3EG delegates busy for months. A simple proposal was to model the three logical entities in parallel, with all of them connected to a "coordinating entity". It turned out that this did not imply any simplification, since the coordinating entity had to reflect the different states of the three logical entities, i.e. we were back to the starting point. Then a solution was proposed by the Swedish delegates, which was as ingenious as it was simple: the RR, MM and CC entities were modelled as sublayers with RR at the bottom, MM in the middle, and CC as part of a new "Connection Management" – CM entity on the top. The normal OSI principle of a higher (sub)layer governing a lower one was applied, and then suddenly the need for a coordinating function was removed. Applying this model really accelerated the progress because the definition of the three parts could be made in parallel. The model introduced the new concept "MM-connection" which is the service needed before a call (or a short message transfer) can be initiated. An "MM-connection" needs an "RR connection", which is simply a new name for channel resources allocated. The concept of MM-connection required one new Layer 3 message, but except from this the sublayering is not reflected in additional overheads with separate protocol unit headers. Messages belonging to the different logical units were all defined in the same logical space, but with different protocol identifiers. Figure 2 shows the resulting model that was decided to be included in a separate specification document (04.07: Layer 3 interface principles) for information purposes.

The Radio Resource management, the border line between radio engineering and software engineering

In addition to the protocol modelling, the Radio Resource management was the area that caused most discussion during the intensive design phase back in 1986–87. From the WP2 group (Radio aspects) a number of physical limitations were defined, their focus was naturally efficiency. On the other hand WP3 was

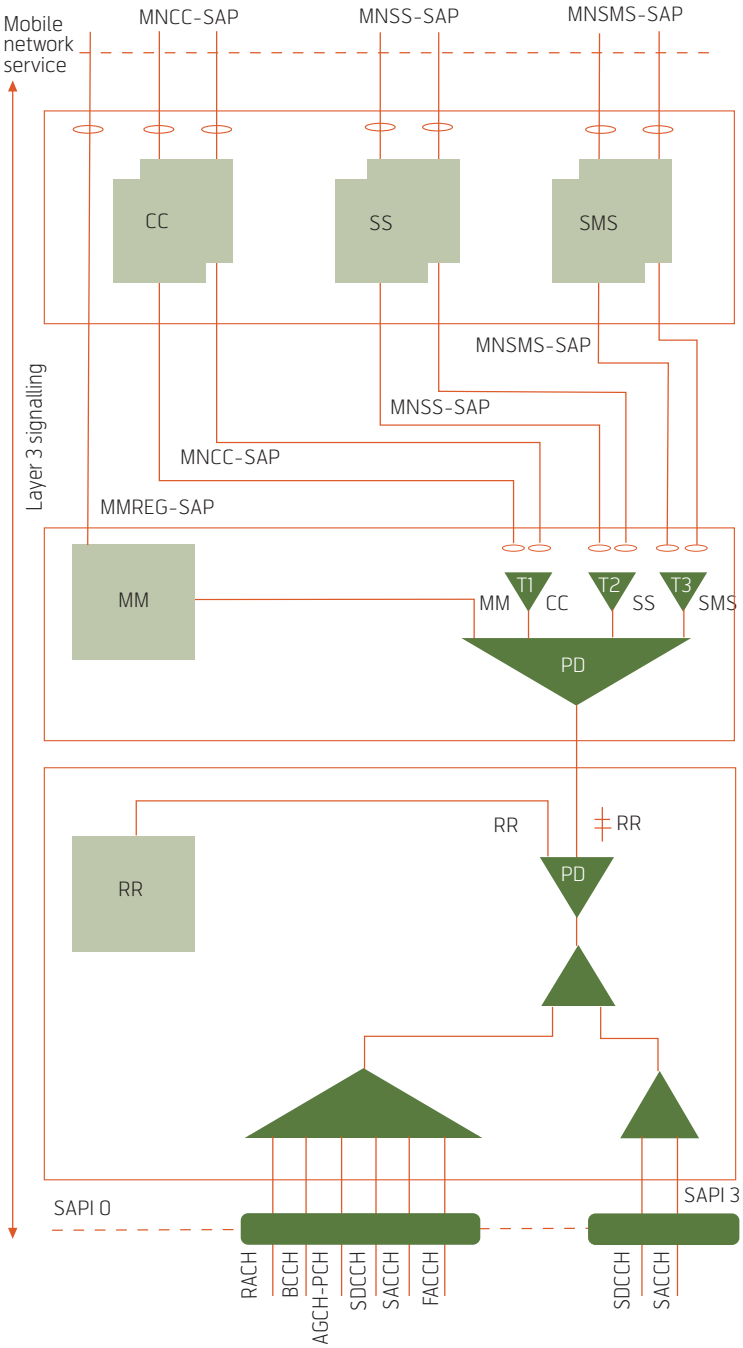


Figure 2 Mobile side protocol model (from 3GPP 24.007)

always stressing flexibility for future changes and extensions, which often meant additional protocol fields or functions. Related to the protocol hierarchy, WP2 was responsible for Layer 1, while WP3 was responsible for Layer 2 and Layer 3. It might be just a coincidence that RR was logically defined as being a part of Layer 3, since what it really does is to control Layer 1. It is therefore not surprising that a new body, the Layer 1 Expert Group, was established in order to define the details on the border of the responsibility areas of WP2 and WP3. Channel access and channel control were the main areas where discussions were made. For a long period the working assumption of the system was that signalling mainly should be performed on the so-called Common Control Channels (CCCH). Lots of proposals for the access control functions for these channels were made, and here the conflict between efficiency and flexibility was most apparent. Late 1987 the "four big ones", i.e. the German, British, French and Italian operators made a common proposal to include a concept of dedicated signalling channels. Now the use of CCCH could be reduced to a very short initial access, and the bit consumption in the later signalling phases was less critical.

The RR management protocol defines this initial access procedure as well as all transmission of the logical information from the network to the mobiles. It also contains the paging procedure and the handover procedure.

As already mentioned it is closely related to the physical layer, and as a consequence it has later been moved to a separate specification.

The Mobility Management – keeping track of and identifying the users of the system

Users of a mobile system are of course mobile, they want the services to be available to them wherever and whenever they want. Moreover, they want to be accessible when roaming both nationally and internationally. The feature of international roaming from day 1 was probably the most important reason for the tremendous success that the GSM system has experienced. Requirements were clear, and the network structure of the existing analogue networks (especially NMT) was adopted. A key principle is to distinguish between mobiles currently in use, and others. While the network knows exactly the whereabouts of a mobile active in a call, it is sufficient that it knows where to search for idle mobiles, e.g. at incoming calls. It is obviously not possible to search in real-time all over a countrywide network (or even worse, in all networks where roaming is allowed); on the other hand it is a waste of resources to keep track of every cell change for mobiles currently not in use. The concept of Location Area is the applied solution, being defined as a number of cells for which a mobile may roam without the need to update any registrations in the network. In GSM as in NMT the task of informing the network that the location area of a particular mobile has changed, is placed on the mobile itself. This core procedure is the main task of the MM part of the L3 specification. In addition to the basic procedure that is triggered when the mobile notices that a new area is entered, functions for periodic updating as well as local indication of power off and on is included. The purpose of these additions is to improve the service and network efficiency.

At the time of design most of the discussions in the MM area were related to the different security functions that have to be included. L3EG here received requirements from WP1 (Service aspects) and from the Security Expert Group defining the actual algorithms and the information that had to be transferred. Authentication of a valid subscription and encryption of the transferred information are the main functions involving MM signalling.

In short the role of the MM (and its underlying RR, L2 and L1) was defined to be provision of a secure connectivity between an identified subscriber unit so that applications can access it in a similar way as over a physical line.

The Connection Management – where the user services are handled

The sublayer model adopted identified three different protocol entities belonging to the CM layer:

- Circuit Switched Call Control
- Supplementary Services Control
- Short Message Transfer

L3EG was responsible only for the first one, the call control procedures for mobile originated and mobile terminated calls. Here the ISDN procedures were used to a very large extent, but several messages had to be reduced in size due to the limitations of the signalling channel. One example of this is the reduction from 8 to 4 bits per digit in address information fields, quite natural since no ISDN numbers so far have included anything else than numerical values. Another example is the removal of the possibility to include user-to-user information in the SETUP message. In addition all mobile-specific service elements had to be designed, i.e. identification of the different channel modes, speech codecs and so one.

One specific challenge that caused some attention was the information from the speech codec experts that DTMF tones were not possible to carry through the selected speech codec. A "workaround" consisting of signalling messages was then designed, simply based on the principle to instruct the MSC to start and stop the tones according to the user's touch on the keypad.

All in all, the design of the call control part of the Layer 3 signalling did not cause large discussions when the RR and MM peculiarities were out of the way.

Remote operations, building blocks, structured procedures

As already mentioned, care was taken to give as much freedom as possible for the implementation phase. The specification therefore nearly only consists of so-called elementary procedures, showing a signal sent and the possible allowed responses. This was inspired by the remote operations principles, which was a hot topic in fixed networks at the time. So while state machine models, SDL diagrams etc were extremely useful during the specification process, they were deliberately removed in order not to constrain the implementations. In order do implement a service, e.g. establish a circuit switched call, the elementary procedures must be combined in a structured way, but this is not specified. The mobile station was specified mainly to be a slave to the network in this respect, a principle that was needed to allow equipment from several vendors to be used simultaneously in the infrastructure provided by one manufacturer. The structuring of elementary procedures into complete signalling sequences is implemented in the different switching manufacturers' software.

The work proved to be successful and future-proof!

Looking back to the design phase of GSM one thing is clear: the flexibility that was built into the protocol was worth all its costs! It has made improvements and completely new services possible, but still allowing for early equipment to work and use their original service set. New speech codecs, higher data rates, packet switched services (GPRS), everything fits into the space provided by the original protocol design. The basic functionality is also now extended to encompass UMTS, where the simple basic MM (and extended with GPRS MM-GMM) still is the basis for the logics of the system. Moreover, the design phase gave a group of people, including myself, the chance of building a knowledge platform that has shown to be of great value over the years.

Knut Erik Walter (50) graduated as Siv.Ing. from the Norwegian Institute of Technology (NTH) in Trondheim in 1982. He was employed by Telenor R&D, where he was heavily involved in the GSM specification work. He was Research Manager for the mobile systems in Telenor R&D from 1990 to 1993. Knut Erik Walter joined Telenor Mobil when GSM was introduced in 1993 and held various positions on the technical product development and strategy side. He rejoined Telenor R&D in 2003 as a senior technical adviser.

email: knut-erik.walter@telenor.com

 ISSN 0085-7130 © Telenor ASA 2004

The 1987 European Speech Coding Championship

JON EMIL NATVIG

Jon Emil Natvig is Senior Research Scientist in Telenor R&D

From the outset the fundamental working assumption was that the GSM system would be based on digital transmission [1]. It was believed that digital technology would not only enable a more efficient management of scarce frequency resources, it would also provide high speech quality and advanced features such as security and data communications.

All this was unproven technology at the time and an important part of the preparatory work was to conduct a number of studies and tests to verify the feasibility of a digital system. These studies were a significant basis for the system design. The so-called "Paris competition" which compared different alternatives for the radio transmission – leading to the selection of the "narrow-band" TDMA radio access method – is well publicised as described in [2]. The radio access study however was itself done under the non-proven assumption that satisfactory speech transmission could be achieved at a bit rate around 16 kbit/s.

In 1985–86, the state-of-the-art in digital coding of speech had reached a level that made this assumption quite reasonable. At that time, ITU-CCITT had standardised speech coding at 32 kbit/s for telephony speech (ITU-G.721) and at 64 kbit/s for wideband (7 kHz) speech (ITU-G.722). Speech coding at 16 kbit/s and below was still a research subject even though computer simulations had demonstrated promising speech quality results. It was however unknown what quality could be obtained in the adverse bit error conditions expected in mobile radio. This fundamental issue was addressed in a less well-known but equally important study: the GSM speech coder selection process, which is the subject of this paper.

Creation of SCEG

The "Speech Coding Experts Group" (SCEG) was set up in October 1985 to deal with the speech processing issues of GSM. SCEG was not formally a subgroup of GSM but had to be set up under CEPT TM3, which was formally responsible for speech processing issues in the CEPT committee structure. In practice, however, SCEG acted as a normal GSM working group but had to report progress in both fora. The task given to SCEG was two-fold:

1) To demonstrate that a digital solution with speech coding at around 16 kbit/s would provide a speech quality at least as good as analogue systems;
2) Given that 1) could be fulfilled, SCEG should select and specify a speech coder for the GSM system.

At the plenary meeting of GSM in Madeira in February 1987, which is best known for the controversy over the radio access issue, SCEG could confirm that a digital solution would indeed offer superior speech QoS compared to analogue solutions. SCEG also proposed a compromise solution combining features from the two best codecs tested as the candidate for further testing and standardisation.

The SCEG work programme

The ultimate measure of speech quality must relate to the quality perception of live listeners. While using automated quality evaluation techniques have become popular lately, no such techniques were available in 1985–87, so speech quality had to be determined "the hard way", in subjective tests involving listeners. The standard for subjective speech quality assessment of telecommunication systems is the Mean Opinion Score test, or MOS for short. Listeners judge the speech quality by listening to recorded speech samples and assign scores on a five point scale: from (poor) to (excellent). The scores are assigned a numeric value from 1 to 5 and the scores are averaged over the listener group and the resulting value is the Mean Opinion Score, or MOS.

The initial stages of the GSM speech codec standardisation effort therefore consisted of the preparation of a set of basic requirements and the organisation of a large multi-national subjective test programme to compare candidate codecs with each other and with these basic requirements.

Basic requirements and rules

At the outset, more than 20 candidate codecs were proposed. The group therefore quickly agreed on a set of rules for the selection process:

1) Each country could only contribute one codec proposal. Several countries therefore had to carry out

ISSN 0085-7130 © Telenor ASA 2004

national "qualification" rounds to arrive at one single candidate. SCEG was not involved in this process.

2) All codec proposals had to be submitted as hardware laboratory implementations operating in real time, no simulations were allowed. This requirement was meant to ensure a certain maturity of the candidate codecs – and a real commitment from the respective proponents.

The following basic requirements for candidate codecs were evaluated (details are given in [3]):

Gross bit rate

Bit rate is a less straightforward attribute than one might think. Basically, a higher bit rate should produce better speech quality for a given algorithm. However, in a mobile environment, we have to take into account bit errors. Using some of the available bit rate for protection of sensitive information will improve the average performance over the operating conditions of the codec at the expense of a certain reduction of the maximum quality in error free conditions. In the experiments, a gross bit rate including any error protection was fixed to 16 kbit/s.

Delay

Delay is involved in several difficult trade-off considerations in codec design:

a) Delay vs conversational quality: Transmission delay causes problems in two aspects: 1) long end-to-end delays disturb the flow of conversation, and 2) reflections in the network become more audible with increasing delay.

b) Delay vs complexity: Because power consumption in CMOS VLSI is almost a linear function of the clock rate of the processing machine, a relaxation of the delay requirement could make it possible to distribute the processing over a longer period with a lower clock rate, thus reducing the power consumption.

What matters for the user is the end-to-end delay, and the maximum allowed codec delay contribution in a coder-decoder back-to-back configuration set to 65 ms.

Complexity

One basic requirement for GSM was to allow for portable handsets so complexity and power consumption of the speech codec was a major concern. This criterion was not quantified but was assessed and taken into account in the selection process. In retrospect it can be seen that this attribute was more important than we knew at the time: recent analysis [4] have documented that more than 80 % of the processing power in the GSM transmission chain is used by speech encoding/-decoding.

Transcodings

Medium to low bit rate speech coding removes subjectively redundant information. Therefore the reconstructed speech is not identical to the original speech waveform. Feeding reconstructed speech into a new coding/decoding stage could result in a large increase in the distortion introduced by the first codec.

The evaluation therefore included a transcoding condition corresponding to mobile to mobile (2 codecs in tandem).

Robustness to variations in voice spectra and speech levels

To be useful in a public service, a speech codec needs to be able to work with a wide range of speakers; men, women, children and adults over a wide range of speech levels. The candidate codec tests included male and female voices and a dynamic range of speech levels of 20 dB.

Robustness to bit errors

Obviously, robustness to errors is essential in a mobile radio application.

The selection tests were carried out with random error probabilities of 0 %, 0.1 % and 1 % representing a typical range of operation.

Digital or analogue?

The actual phrasing of the speech performance requirement in GSM was:

> *From the subscriber's point of view, the quality of voice telephony in the GSM system shall be at least as good as that achieved by the first generation 900 MHz analogue systems over the range of practical operating conditions.*

The implication of this formulation was that SCEG had to 1) consider end-to-end transmission including the public telephone network, and 2) to include an analogue 900 MHz under comparative operating conditions in the test programme.

The analogue reference system gave a substantially transparent speech quality in high C/N conditions and outperformed all codecs in error-free conditions. However, ALL digital codecs performed better on the average over the range of operating conditions than the reference FM system [3]. Thanks to error protec-

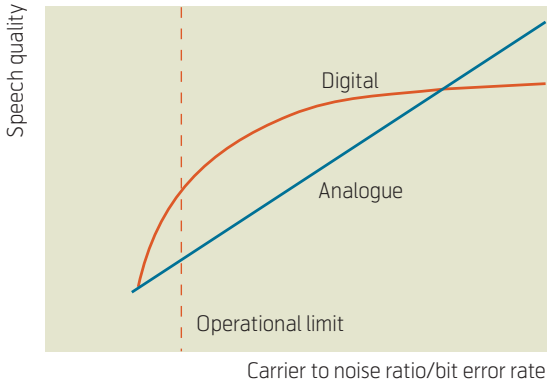


Figure 1 Analogue vs digital speech typical behavior

tion, the quality of a digital codec maintains a relatively high speech quality as transmission conditions deteriorate (until a marked breakdown point), whereas the quality of the analogue reference decreases more or less linearly. This typical behaviour is illustrated qualitatively in Figure 1.

A Solomonic decision – the RPE-LTP codec

The comparison of the codecs were organised as a large multi laboratory experiment. Running the tests in parallel in several laboratories, was essential for the credibility of subjective test results. Details of the experiments and the results are reported in [3]. Six candidate codec were presented at the first laboratory session hosted by CSELT[1)] in Torino in 1987. In this session some objective measurements were made and speech material for subjective testing was recorded. Subjective tests were carried out in seven different laboratories in the respective native languages. The overall analysis (which was undertaken by TF – Televerket's Research Institute) showed that no single codec could be declared the winner in all aspects of the "competition".

However the results showed clearly that two codecs, both based on pulse excited source-filter modelling, were superior to the remaining four sub-band coders. In comparing these two solutions the development teams from IBM/France and Philips/ Germany could see the possibility of several improvements. SCEG therefore decided to allow the French and German teams to propose a compromise solution resulting in the merging of MPE-LTP and RPE-LPC into the RPE-LTP full-rate GSM codec [6] (see box).

The specification of the codec is given in GSM recommendation 06.10, which specifies the algorithm down to the bit level. This allows verification of implementations by means of a set of digital test sequences. The specification was defined in fixed point arithmetics to be implementable on fixed-point DSPs consuming less power. A public domain bit exact C-code implementation of this coder is available [7].

Telenor/Televerket contributions

TF's strong involvement in the development of GSM in general was founded on experience and competencies built up over more than ten years of activities in several research "threads" in the mobile communication area, both conventional land mobile and satellite mobile communication.

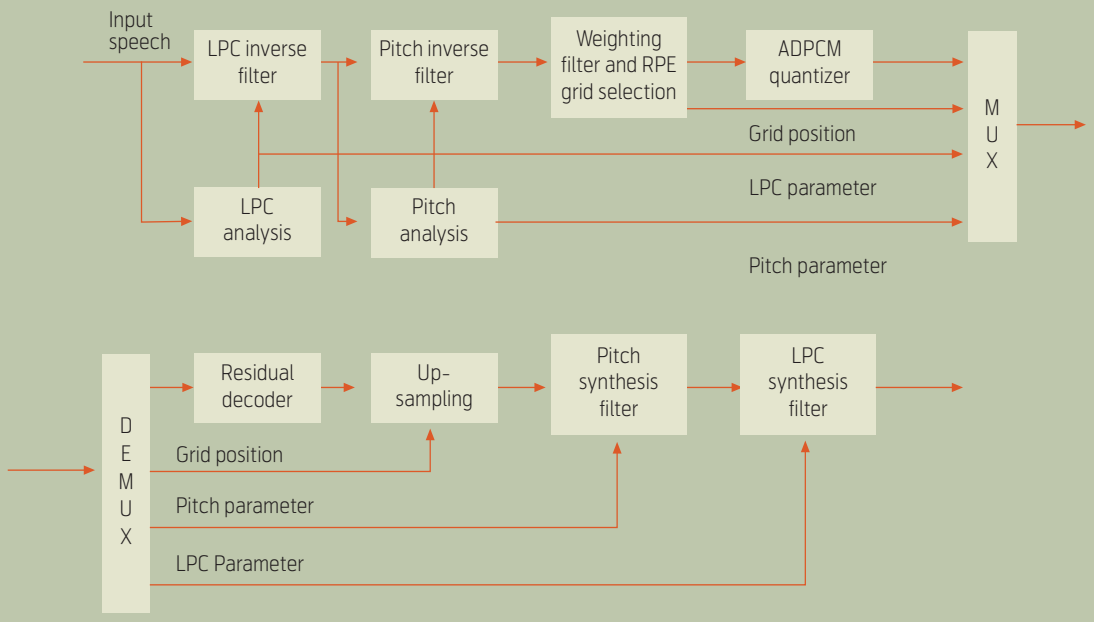
TF had been studying satellite communication and digital communication almost since it was established in 1967, and TF scientists were strongly involved when Televerket developed the NORSAT satellite system to provide telecommunications to the Ekofisk offshore oil platform in the North Sea. This was one of the first domestic satellite systems in the world and probably the first system worldwide to be based on digital transmission – a bold choice at the time. The speech encoding used 32 kbit/s adaptive delta modulation – an advanced quantisation scheme rather than a speech compression method – combined with advanced error correction. It turned out that this system provided clearly better frequency utilisation than alternative analogue solutions based on FM.

The experience from the NORSAT system convinced TF scientists that the future of wireless mobile communication lay in digital transmission. In the 1970s TF was very active in ship-to-shore satellite mobile communication in the framework of CCIR SG 8 and in the development of the MARISAT system. In this work, TF proposed a digital solution following the experiences from NORSAT, but did not succeed in convincing the committee at this time. Digital speech transmission for ship-to-shore systems was introduced later by INMARSAT (Standard B) and TF also played an active part in the studies that led to the introduction of that system.

In all this work, TF established broad know-how in system and network design, QoS aspects of speech transmission, as well as subjective speech quality, which also included laboratory facilities for subjective evaluation of speech codecs.

1) *CSELT – Centro Studi e Loboratori Telecommunicazioni.*

GSM speech coding in a nutshell



The 13 kbit/s coding scheme selected for the full-rate GSM standard is an example of a so-called analysis by synthesis system.

The sampling rate is 8 kHz and each sample is quantized at 13 bit/sample.

The GSM full-rate codec operates on time frames of 20 ms (160 samples). This corresponds to one pitch period for someone with a very low voice or ten periods for a very high-pitched voice. This is a very short time during which the speech can be assumed to be stationary. The speech frame is analysed using the source-filter model, which assumes that the speech signal is the result of a time varying linear filter excited by a source signal. For each frame, parameters are determined for a synthesis filter that models the effect of the vocal and nasal tracts. The short-term inverse filter flattens the spectrum envelope of the residual speech signal.

The next step is the long-term filter, which operates on 5 ms subframes (40 bits). In the case of voiced speech, the input signal will be a periodic impulse train where the pulses correspond to the glottal pulses. This periodicity is removed using a long-term predictive (LTP) filter, resulting in a weak signal. If the speech is part of an unvoiced phoneme, the residual is noisy and does not need to be transmitted precisely.

In the last step, the resulting residual signal is down-sampled by a factor three. This is done by selecting every third sample of each 5 ms sub-frame, starting at sample 0, 1, 2 and 3, resulting in 4 candidate sequences. The algorithm picks the sequence with maximum energy for quantisation and transmission

The *decoder* starts with the transmitted sub-sampled sequence padding the missing samples with zeroes. The resulting residual pulses are fed into the long term synthesis filter which has had its parameters updated, reconstructing the short-term residual. These impulses are then passed through the short-term synthesis filter simulating the vocal tract and thereby producing the speech signal.

More information can be found in [6] and details are given in [10].

One important basis for TF's contributions to GSM was the close collaboration between TF and ELAB. [2] gives an account of the work in the radio area. Already in 1979 – one year *before* the NMT network was launched – TF awarded its first research contract in the speech coding area to ELAB. The task given to ELAB was to study low bit rate speech coding for mobile radio applications. In the following years, a series of contracts were awarded and a fruitful collaboration evolved between TF and ELAB which established a competence environment in digital speech coding. The division of responsibilities was such that ELAB covered the work on speech algorithms and hardware implementations while TF dealt with speech system aspects, QoS and subjective testing.

In 1983 the FMK (Future Mobile Communication) committee was established in the framework of the

ISSN 0085-7130

Nordic Council to plan for a future digital land mobile radio system to follow after NMT in the Nordic countries. FMK also established a special subgroup (NR-FMK-T) to look into the speech coding issue. When SCEG was set up in 1985, the Nordic countries had already completed a coordinated subjective experiment to compare possible codecs for mobil radio.

A major result of the TF/ELAB collaboration [8,9] was a sub-band coder with sophisticated adaptive bit allocation as a candidate for GSM. Unfortunately, due to the very tight time schedule for the delivery of a hardware version at the CSELT laboratory session, some error sneaked into the algorithm. Considering that none of the sub- and coder could match we believe that this would not have changed the selection anyway.

References

1 Haug, T. How it all began. *Telektronikk*, 100 (3), 155–158, 2004. (This issue)

2 Maseng, T. *Telektronikk*, 100 (3), 161–164, 2004. (This issue)

3 Natvig, J E. Evaluation of six Medium Bit-Rate Coders for the Pan-European Digital Mobile Radio System. *IEEE Journal on Selected Areas in Communications*, 6 (2), 324–331, 1988.

4 Thierry, T, Tennehouse, D. *Estimating the computational requirements of a software GSM base station*. July 1, 2004. [online] – URL: http://citeseer.ist.psu.edu/336927.html

5 Natvig, J E, Hansen, S, de Brito, J. Speech Processing in the pan-European Digital Mobile Radio System (GSM) – System Overview. In: *IEEE GLOBECOM 1989*, November 1989.

6 Vary, P et al. Speech Codec for the European Mobile Radio System. *Proc. ICASSP 1988*, New York, 227–230.

7 Degener, J. *Digital Speech Compression – Putting the GSM 06.10 RPE-LTP algorithm to work*. July 8, 2004. [online] – URL: http://www.ddj.com/documents/s=1012/ddj9412b

8 Ramstad, T A. Sub-band coder with a simple adaptive bit allocation algorithm. A possible candidate for digital mobile telephony? *ICASSP Paris*, May 3–5, 1982.

9 Hergum, R, Perkis, A. *16 kbit/s speech coder for the GSM system*. Trondheim, ELAB, 1986. (ELAB report STF44 F86133)

10 ETSI. *GSM full rate speech transcoding*. Sophia Antipolis, 1992. (ETSI recommendation 06.10.)

Jon Emil Natvig (58) graduated from the Norwegian Institute of Technology (NTH) as siv.ing. in 1970 and Dr. Ing. in 1975. He has worked with Televerket/Telenor since 1972. In the period 1975–1990 his main work area was in speech QoS aspects, mostly related to digital voice transmission in satellite and land mobile systems. From 1985 to 1989, he chaired the Speech Coding Experts Group responsible for selecting and specifying the speech processing functions in the GSM system. Since 1990 his main work has been in the field of voice user interfaces. He is currently working with the usability and accessibility issues in the Products and Markets research program at Telenor R&D.

email: jon-emil.natvig@telenor.com

SMS, the strange duckling of GSM

FINN TROSBY

Finn Trosby is Senior Adviser in Telenor Nordic Mobile

Some introductory remarks

It might be appropriate to start with a warning: the reason for writing an article on 'the birth of SMS' is not to reveal a 15 year old story about huge achievements in terms of complex protocols and challenging combinations of radio, data and network design. The reader looking for that will inevitably be disappointed. The SMS or the 'Short Message Service' – as it has been labelled in every corner of GSM coverage – is definitely one of the simplest compounds of the GSM system.

The main reason for writing about the creation of SMS is because it is a story about innovation. SMS was indeed a true newcomer. All the other services of the GSM system – speech, fax and all the variants of circuit switched data – were well-known services, copied from the fixed network, in particular ISDN. SMS, as it was defined in terms of the stable versions of the relevant specifications, was an extremely simple messaging service tailor-made for GSM. It did not have its parallel or predecessor in any other system for offering mobile services to the public. The major part of the GSM community expected the circuit switched data and fax services to be the most important non-voice services, and regarded SMS to be more like an add-on that might increase the attraction of the GSM system without any commercial significance. The years to come proved it to be the other way round.

Books have for many years been published on the European mobile adventure during the last 20 years. A very good – and perhaps the most comprehensive – one is [1]. However, even the most complete volumes cannot cover every task of a huge endeavour like the GSM development. The fact that it covers more of the SMS design work well after the SMS specification was approved than before, gave me the final push to write some lines about just the period of time when I took part in the design work, i.e. from 1987 to 1990.

Background

Mobile communications of Europe and the US in the mid 80s were a true wilderness in terms of technologies and markets. In the area of speech services offered to the public, manual systems were replaced by the first automatic ones, giving 'mobile telephony' almost the same approach as the regular telephony that everybody had been used to. Paging services were steadily improving – both in terms of services and coverage.

In addition to the systems for public services offering, enterprises or organisations were making extensive use of a wide area of PMR systems. In the mid 80s most of those comprised

- A relatively simple radio network of one or a few base stations giving radio coverage to a limited area on a non-cellular basis;
- Medium complex protocol stacks, some offering analogue speech and some data services;
- A software application running on one or several host computers, being the primary fundament of the business itself (taxi companies, dispatch businesses, etc.);
- Interworking with the public networks – e.g. POTS – as a feature within a few of the systems.

At *Televerkets Forskningsinstitutt* – the Research & Development Department of Telenor at that time – I worked for a couple of years (1984–1986) as the project manager of a survey called Mobile Networks for Special Purposes (*Mobilt spesialnett*). The intention of the study was to explore the potential of mobile communications for other services than telephony. The survey comprised many activities – spanning from discussions with manufacturers and users of PMR systems in Europe and the US via an experimental system set-up by Televerket and Norwegian industry partners to specify a mobile messaging system. The system was to be connected to an X.25 network and provide both hosting of PMRs and extension of the X.400 service to mobile terminals. Together with extensive market analysis on mobile non-voice services in general and mobile messaging in particular the study reached a set of conclusions, from which one important was

- Offering mobile messaging within the framework of a public service portfolio may be a good idea since

 ISSN 0085-7130

- Mobile communications and messaging services make a very good match, since mobile users will be frequently out of coverage or turned off;

- Efficient store-and-forward mechanisms will be required to make the mobile terminal the prime target of crucial information to be delivered to the user as soon as possible. In this respect, it will to some extent outdate the fixed phone or the fixed data terminal handling the email.

• Offering mobile messaging jointly to both private and corporate segments may be a good idea since

- Unless attacking huge markets like the American, developing mobile transport services for just a set of business applications is aiming at bankruptcy. Either take this development to the US or make mobile communications for middle-sized markets that may attract both the private and professional segments;

- Enable business viability to new services like mobile messaging by seeking opportunities to bundle those in a flexible and non-complex way with a set of highly acknowledged services like telephony;

- Don't think too rigidly about new services like mobile messaging being for the corporate market and for professional use mainly. It may rather be the other way round; that take-up starts in the mass market and even in the long term supersedes the corporate market.

Start of work in CEPT and later ETSI

IDEG is established

In 1987 the work with the GSM specifications had taken some great leaps forward. The system architecture, the basic services, the characteristics of the radio interface, the signalling package – they all emerged with ever clearer contour. The GSM community had decided to establish three different working parties: WP1 – dealing with the services, WP2 – dealing with the radio aspects, and WP3 – dealing with the core network and the signalling aspects. The main group – or as it was to be called: GSM main body – was the group to a) survey the progress of the whole project, b) assign tasks to the working parties, and c) approve of the solutions produced.

There was but one outstanding domain that lagged behind: the detailed definition and specification of the data services. The responsibility had been allocated to WP3, but that group had its hands full with the huge challenges of establishing a complete package of all signalling functionality that might be required in order to fulfil the needs of the future GSM users. The GSM main body concluded that there was a need for another group to cater for the progress of data services definition. On May 20, 1987, the first meeting of IDEG – the Implementation of Data and Telematic Services Experts Group – was held in the city of Bonn. The group was chaired by Friedhelm Hillebrand from Detecon. IDEG had a somewhat blurred organisational status when created, but it soon became apparent that it was most convenient to give it the status of a working party, and eventually it was renamed WP4[1)].

IDEG soon defined four areas that the group had to concentrate on if it was to have a chance to catch up with the achievements that had been reached in the other three working parties

• Rate adaptation mechanisms;

• The radio link protocol (RLP), i.e. the protocol for carrying data at Layer 2 of the OSI model for the data services;

• The facsimile service within GSM;

• Message handling services that might be part of the GSM service portfolio.

There was allocated a so-called drafting group for each area[2)]. As a matter of coincidence, I was appointed chairman of the fourth one.

The tasks of the 'Draft Group on Message Handling'

GSM WP1 had left IDEG with two crucial specifications: GSM 02.02 [3], an overall description of the 'bearer services' of GSM; and GSM 02.03 [4], an overall description of the 'tele services' of GSM. [3] e.g. contained a variety of circuit switched data services, of which perhaps to some extent only the asynchronous non-transparent 9.6 kbit/s one ever came to practical use. [4] e.g. contained the framework of a set of messaging services: the fax message service, three services on short text message conveyance, and

1) *As the reader may know, naming of the GSM organisational entities within CEPT and later ETSI and 3GPP changed on several occasions. Thereby, e.g. WP4 became GSM4 and SMG4 in synchronism with the corresponding renaming of the other working parties.*

2) *A couple of additional drafting groups were also established, but on a very preliminary basis; one or two meetings only.*

a request that the GSM user should be able to access an MHS system.

The Draft Group on Message Handling – DGMH for short – which I was to chair, was to take responsibility for the short messaging and the MHS access.

The three services on short text messages were in [4] depicted as follows

1. Short Message Point-to-Point Mobile Terminated
2. Short Message Point-to-Point Mobile Originated
3. Short Message Cell Broadcast

The current version of GSM 02.03 listed these three as separate services with different level of importance: 1) – which was the service of carrying a text message through the network to the mobile terminal – was classified as one of the high priority services in GSM. 2) – which was the service of carrying a text message from the mobile terminal and through the network to an entity for further conveyance – should be optional for a GSM PLMN operator. 3) – which was the service of spreading a text message on a broadcast basis to all or a sub-set of the mobile terminals being within radio coverage of one or several base stations in the network – was for further study. It was further emphasized that all three services should exploit the capacity of the signalling channels of the radio path so that they should not face congestion due to ongoing circuit switched traffic – voice or data – of the mobile terminal.

In [4] of that time this was about all that was said about the short message services. Before the establishment of IDEG there had been sketches on architecture and how to accomplish Short Message Point-to-Point Mobile Terminated, but none of those documents were in the pile of the officially and approved guiding documents when the first meeting of IDEG was opened in Bonn.

The directives of [4] for defining MHS access were even scarcer. It merely said that specifications should be provided to allow the mobile user to exploit the services of 'MHS services'. It turned out quite quickly that integrating GSM and MHS did not require further GSM specifications. GSM users could very well access the User Agent of a X.400 MHS via GSM's own data services. A specification specifically on how to access MHS from a GSM terminal might perhaps represent a marginal improvement compared to relying on already established standards, but DGMH did not estimate this to be sufficient for suppliers to adopt this in their production plans. The MHS access activities within DMGH and IDEG were concluded in a technical report, probably the very first TR on data services of GSM. With that report, DGMH was allowed to leave the MHS access issue.

The objective of defining a Cell Broadcast service resulted in the required specifications, [6] and [9], at approximately the same time as the point-to-point services were approved. However, no core network transport mechanism was defined, and the GSM cell broadcast service was then left with an area that had to be based upon proprietary solutions. I think it is fair to say that both in IDEG and in DGMH, there was some hesitation among experts on how to design the broadcast service in a way that would be welcomed by operators. They were not troubled by possible technical problems, but rather by the feeling that it might be hard to find a viable business case. A sparkling contrast to this type of reluctance was demonstrated from Racal/Vodafone's side. It is impossible to touch upon the work with the cell broadcast service in those days without giving the very enthusiastic Alan Cox full credit for cell broadcast ever being defined. However, when the time came to implement the GSM network and its services, the scepticism of the GSM experts had contaminated the product development divisions of the mobile operators. Few operators ever implemented cell broadcast, and hardly anyone made it a commercial success. The destiny of this service is interesting and should give the supporters of e.g. future MBMS something to consider.

For the reasons indicated, I will leave cell broadcast with this and proceed with the mobile terminated and mobile originated service under the common acronym by which they gradually have been identified – SMS.

Items dealt with during the design of SMS

Service aspects – IDEG is given a considerable latitude in the SMS design

As stated above, the spring 1987 version of [4] did not reveal much of the basic perspectives of WP1 on the short message services. Extensive research by historians is outside the scope of this article, but probably there have been somewhat different opinions among the GSM delegations on the use of a text service. The Norwegian delegation e.g. filed a contribution in which it advocated the realisation of a service for telemetry applications, and I think other mobile operators had put forward similar proposals, but aiming at slightly different applications. The text in [4] is probably the result of their efforts to reach a consensus, leaving quite some freedom to the crew of designers.

Architecture

The Service Centre – a necessary entity

The point-to-point short message service – at least the mobile terminated part of it – would obviously be a store-and-forward service, since the mobile terminal might be turned off or out of coverage at the instance of delivery. Since it was explicitly stated that none of the regular network nodes of the GSM PLMN – such as the MSC or the BSC – should offer store-and-forward capabilities, there had to be an extra node with some genuine store-and-forward capabilities. So an additional node with the somewhat generic name Service Centre (SC)[3)] was added to the topology of GSM. The concept of the SC had lingered for some time also within WP1 before IDEG came to work, however without any specific characteristics. The procedure of short message transfer should then be SME ⇒ SC ⇒ MS and SME ⇐ SC ⇐ MS. The entity SME, the Short Message Entity, was whatever entity that might be connected to the SC in order to send or receive short messages, including a GSM MS. The last step was important, because it brought symmetry to the service aspects of the two components point-to-point mobile terminated and point-to-point mobile originated, thereby effectively integrating the two. It was therefore finally decided to comprise the two services in one service specification, namely [5].

The debate on why and how to distinguish between value-added services (VAS) and teleservices had been going on in Europe for quite some years, especially in the UK with its pioneering role in bringing market liberalism to the European telecommunications. The rigorous definition of those years implied that one should even say that an information stream was subject to value adding if it was in any way converted – even slightly re-formatted – or stored for some time. The issue immediately came up with the introduction of the SC. Where should it reside, within or without the PLMN? Due to the genuine VAS character of SMS, the UK delegation strongly opposed the first sketches of the architecture, where the SC was included in the PLMN, which they regarded as a platform for teleservices only. The operators of the other countries had at that time no strong opinions and neither had the manufacturers, so it was decided to logically locate it outside the PLMN. Since a more pragmatic view on VAS has gradually replaced the original one, one could question if the architectural design we chose at that time was the most feasible. In many countries, SMS was regarded both by operators and regulators as an add-on to the mobile telephony service and consequently being part of the same market. For the forthcoming standardization 1987–1990, it however resulted in a relaxed attitude towards making a mandatory specification of the interface SC – MSC, which may definitely be ranked as a shortcoming of the SMS standards from those years.

Long distance SMS

Another item of discussion was how the PLMN was to transfer the short message internally. For obvious reasons, the routing principles of a mobile terminated short message as well as a mobile originated short message became identical to the routing principles for a speech or data call set-up. Thus, there might be a long haul transfer MSC – MSC, similar to the connection of a telephone call within one PLMN or between two PLMNs. Should it be transferred by means of well defined mechanisms of user data transport like the X.25, or should one produce a certain operation within the signalling system of GSM – MAP – especially for short message transfer? The question was discussed both within DGMH and SPS/SIG, which was a body in ETSI that had responsibilities over a wide area of signalling tasks within both fixed and mobile networks. Several experts advocated the principal view that transfer of user data should not be mixed with signalling functionality, and recommended X.25 for this particular undertaking. The UK delegation in IDEG opposed that position. The UK operators were – unlike many of the other operators that took part in the GSM project – genuine mobile operators, with no fixed or data network operators within the same corporation. From their business point of view, SS No 7 was a free lunch, whereas X.25 was not. After some consideration, it was decided to base the short message transfer MSC → MSC on SS No 7 by adding an extra operation 'forward_short_message' to the repertoire of MAP operations. Retrospectively, users of the SMS should be grateful to the UK delegation for contributing to a correct decision, even if it might be for other reasons than the one mentioned above: the smooth and uncomplicated interconnect and international roaming on SMS stems from the choice to rely on the in-house capabilities of GSM.

Defining the length of a short message

The choice of MAP as the long haul carrier of the short message brought an end to another discussion that had been going on for some time: how long should the short messages be allowed to get? MAP was based upon TCAP, and thereby on the concept of bilateral operations, which were assumed to carry small weights in terms of user information. To

3) *Later changed to 'SMSC'.*

arrange for MAP operations to carry more load than the standard request – response sequence of TCAP allowed for would be both complex and cumbersome. But being applications based upon a signalling system of global coverage, operations within MAP or TCAP necessarily imply a substantial overhead. When analysing the 'forward_short_message' operation and removing the overhead, we found that there were somewhat more than 160 characters of the alphabet chosen (see below) left for user data. It was decided to round down the figure to the closest decade, and so the number 160 became the eventual size for regular SMS. The WP1 had earlier been leaving 128 characters as some very tentative request for the short message length, but it had no problems of increasing the limit to 160.

The short message over the radio path

But which GSM capabilities were required for sending or receiving the short messages over the radio interface? The answer was pretty much given by the requirement that short messages should flow freely to or from the mobile terminal whether the terminal was idle or busy with an ongoing call: it had to be on one of the signalling channels. I consulted my colleague Knut Erik Walter, who was at that time heavily involved in the work with the very essential [7], if he could take a look at what might be required in terms of specification work to cater for the SMS radio interface. Within a very short time he drafted [8], which was thereafter approved in WP3, and which I think stayed stable and without the need of any change for a very long time.

[8] allocates signalling channels SDCCH and SACCH according to Table 1.

Channel dependency	Channel used
TCH not allocated	SDCCH
TCH not allocated → TCH allocated	SDCCH → SACCH
TCH allocated	SACCH
TCH allocated → TCH not allocated	SACCH → SACCH opt. SDCCH?

Table 1 The impact that traffic allocation has upon choice of signalling resource to be used for the conveyance of the short message

[8] also allowed for the network to keep the signalling resources, e.g. in periods with frequent message traffic: "… the network side may choose to keep the channel and the acknowledged mode of operation to facilitate transfer of several short messages for or from the same Mobile Station. The queuing and scheduling function for this should reside in the MSC".

Reports, Messages_Waiting and some other features

As indicated above, most people outside DGMH seemed to regard SMS as a machine-to-person service mainly, e.g. as the main part of voice mail alerts. In that respect, there would be no need for any type of confirmation or acknowledgement of a short message arriving at its destiny. Fortunately, the DGMH crew appeared to have a perspective also for person-to-person messaging, and to have recognized the usefulness of being offered information concerning if and when the recipient actually received the message. A question arose at the stage of service definition in the case of a mobile recipient: should the confirmation be given at the event of manual actions taken by the user to display the message, or should it be given at the event of the terminal receiving the message? Picking the second alternative was an easy choice to make. The major challenge is to convey the message over the radio path at a time when the mobile is turned on. When this is achieved, the chance that it will somehow be destroyed before the user may read it is less than marginal.

From [2] I had learned that to make messaging effective for mobile communications, one has to provide for functionality to make the information transfer as swift and easy, meaning e.g. as far as possible to overcome annoyance of the inherent instability of the mobile terminal's contact with the network. I therefore proposed an additional interworking between the SC and the GSM network. When an attempt to transfer a short message to the mobile fails due to the mobile being turned off, the location registers take a note of the event together with the address of the SC that made the attempt. When the mobile user turns on his phone again, the location registers – provided that the operator is applying IMSI Attach / IMSI Detach – are notified and in their turn informs the relevant SC that it might be a good idea to repeat the transfer attempt. The feature was labelled 'Messages_Waiting', and aimed to be particularly useful for those who frequently would turn off their mobiles to reduce battery consumption, attend meetings or events where mobile phone calls were banned, etc

Other features that may be mentioned are

- Validity-Period, period that a short message stored in the SC due to absence of the receiving party should be kept before it might be deleted;
- Service-Centre-Time-Stamp, time when SC receives a short message to be delivered. Always

to be included in the short message delivered to the terminal;

- Protocol-Identifier, identifying which protocol to be performed at the application layer;

- More-Messages-to-Send, a Boolean included in the short message delivered to the terminal to tell if there are more messages in the SC still to be sent to the recipient.

The alphabet

Now, what should be the alphabet of the short message? The WP1 had in [4] made a reference to the ITU and ISO standards of International Alphabet no 5 (IA5), which were designed for what were anticipated to be the text services of the future, in particular MHS. In DGMH, we examined the IA5 standards, which were designed with the objective of providing different regions of the world suitable alphabets within the framework of adequate character lengths, in particular 8 bits. The exercise of finding a suitable alphabet for SMS occurred chronologically just after the corresponding work item in ERMES, who had approximately the same focus as DGMH had at that time: finding a sufficient set of characters for the western parts of Europe spending as few bits as possible. The ERMES alphabet was a result of picking the characters from the most used alphabets while still being able to wrap up the whole thing in a 7 bits notation. We therefore proposed to use the ERMES alphabet as default, but opened up in the protocol for the user to request other alphabets. Both IDEG and WP1 supported this proposal.

A review of the work leading up to approval

It may be worth while to try to summarize what was achieved, and mention the crew that made the results.

Merits and flaws of the SMS design

In my opinion, the merits of the SMS design were the following

- Simplicity, both in terms of functionality and in terms of architecture (e.g. only one SC in any MS → MS messaging);

- Merge of the two original point-to-point services into one service – SMS – with complete reciprocity 'mobile terminated' and 'mobile originated';

- An SMS based entirely upon in-house capabilities, e.g. SS no 7 instead of X.25;

- Reception confirmation for MS to MS messaging;

- Automatic delivery of waiting messages to a recipient just after he had switched on his mobile phone.

On the other hand, some major flaws are retrospectively not hard to pin-point:

- A protocol version number was not allocated at the transfer layer, requiring new versions to be backward compatible. Apparently, none of the DGMH members were well-experienced protocol experts!

- We were not bold enough in terms of exploiting future possibilities for MS to MS conversations, e.g. group chatting. Both address conversion (e.g. E.164 ↔ name@domain) and handling of distribution lists within the SC were discussed, but a number of people clearly expressed that we had gone far enough with our perspectives on SMS conversations!

- The same was the case with message templates, which was an idea inspired by transaction services within the X.400 domain and just very briefly and informally mentioned within the GSM and DGMH community. As with the above ideas, it did not have the necessary support to be pursued. However, it might have boosted SMS as a tool for *m*Commerce!

The 'SMS crew'

No individual expert or company should claim to be the 'father' or 'creator' of any service or major functionality produced during the GSM development. The GSM project was indeed a multi-national collaboration at its best. The cooperative working procedure was the case also for IDEG and DGMH. The latter consisted in my period as a chair of a group varying from 5 to 8 people, all dedicated and contributing to the ongoing work. I would in particular like to mention Alan Cox from Racal/Vodafone (later Vodafone), Kevin Holley from Cellnet and Eija Altonen from Nokia. I would also like to compliment Friedhelm Hillebrand for being an extremely good chairman of IDEG. Most of the IDEG participants in 1987 were not familiar with international collaboration like GSM, but in a very gentle and constructive way Fred encouraged them to immediately join in and do their best. Fred left the chair of IDEG for other GSM appointments in 1989, and was replaced by Graham Crisp from Plessey Networks and office Systems. Graham had chaired the draft group on rate adaptation mechanisms (TAIW, Terminal Adaptation and Interworking), and thus became the first person from industry who took a chair in CEPT. Graham, who had participated in IDEG from its first meeting, had exactly the same exquisite skills in chairing the group as Fred had exposed.

I would also like to appreciate colleagues of my own company – in particular Jan Audestad and Knut Erik Walter for swift responses to our requests on MAP upgrades ([10]) and the establishment of radio interface functionality ([8]), and a series of good advice along the line.

The tricky part: what can we learn from the SMS adventure – if anything at all?

Everyone knows stories about the strange random walk characteristics of business and technology development; the yellow stickers from 3M, the chat line of the Swedish phone company, and so on. Like those examples, many of them derived from internal mishaps and were just accidentally transferred to the production lines. Yet they became great successes.

The birth of SMS was definitely not due to a mishap or accident, even if the perception of SMS in 1987 was – as stated earlier – not very clear. Luckily enough, it was not excluded from the list. The story has a slight resemblance to those of the Norwegian fairy tale character Askeladden, who picks up all kinds of items that he encounters given the presumption that it may come to use some day. In the adventure they always do, resulting in a massive success. In real life, they sometimes pay off – as with the SMS. Trying to figure the same situation today, it is not hard to imagine the average modern executive immediately tearing the SMS concept of [4] into pieces: "When there is no extensive and convincing text of market analysis, there should be no further transfer to a lengthy and costly design and production process". The strange thing is that if one imagines the modern product development filtering on all other services than SMS, they might have passed the checkpoint procedures without difficulties. The speech service was a banker, no one doubted that there was a substantial potential of migrating telephony from the fixed to the mobile networks. The fax service also had a high standing: fax had been a popular service in the fixed networks for years! The circuit switched data service also had its fixed network parallels that made perspectives of a high usage probable. Thus, for all three services it would have been fairly easy to produce convincing arguments in the context of today's product development forums why they should all be profitable. In this way, we can very well envisage a situation where the methods of today would have accepted fax and circuit switched data – the failures – and discarded SMS – the success!

This should not be taken as polemic statements intended to give the impression that the participants of DGMH had some sort of ingenious formula or supernatural gifts that enabled them to see what nobody else saw: the full potential of SMS. Certainly, experience from earlier work and objectives had provided the group a hunch that messaging between mobile users might be a very good idea and worthwhile pursuing. However, no one within DGMH, IDEG or GSM was even close to comprehending the wilderness of applications that is provided by today's SMS. *m*Commerce, the flora of CPA based services, customizing the mobile handset by download of the required parameters, short message as the initiator of push services; none of those applications were thought of even vaguely. They just popped up because SMS was at hand, virtually from the start within and between any GSM network and easy to apply.

Finally trying to conclude, I would say that the success of SMS – unexpected among even the core GSM experts – might be associated with the following key words:

- Abundance and simplicity combined. The willingness to include *some* abundant dark horse that cannot be justified through clear-cut market analysis, but keeping in mind that also for an item of this category the rule applies that market appeal is proportional to simplicity;

- Hunch. The willingness to adopt, trust and support *some* ideas – not all, not even many – that give some elusive perception of great potential that cannot be justified through plain and clear-cut market analysis;

- Risk. The willingness to take *some* calculated risk when deciding upon the design;

- Seeing business in a broad and long-term perspective. The willingness to accept and endorse categories of work that open up vast new business areas, even if they will be available also for one's competitors and even if it may take several years before it pays back.

The development of GSM almost coincided with a huge paradigm shift in the business of telecommunications: leaving the age of monopolies and entering the age of the liberalised markets. The benefits of this transition – e.g. in terms of price reductions, more effective sales and distribution channels, and flexible and customer oriented production lines – have been emphasized ad nauseam, and will be contradicted neither by me nor by anybody else. But no change is entirely good or bad. With the shift mentioned above there was also something lost. The corporate environment that fostered the characteristics listed above for the ability to take substantial leaps forward – e.g. the

cardinal 'hunch' – was far more apparent in the dinosaur-like telcos of the past than it is the streamlined and ever cost reducing operating companies of today.

'Hunch' is what you get when – in between the tightly scheduled tasks of today's demands – you are allowed to stray into areas of terra incognita without almost any other purpose but to explore. The '*Mobilt spesialnett*' endeavour was one such exploration of mine, and it meant a lot to my qualifications for carrying out the objective that we were confronted with. I am sure that the other people involved with SMS – in WP1, IDEG and DGMH – had their corresponding strays, and that those were equally beneficial to them. The previous telco's could afford that luxury. The present ones cannot, and the soil is inevitably less fertile. Thus, today's SMS chatting crowd can be happy that the GSM system definition phase occurred well within the era of the previous regime. I'm not quite sure that the SMS sketches of 1987 would have passed the WP1 examination if its members had possessed the mindset of the operator community of 2004.

References

1 Hillebrand, F (ed.). *GSM and UMTS. The creation of Global Mobile Communication*. John Wiley, 2002.

2 *Mobile Networks for Special Purposes (Mobilt spesialnett)*. Study on messaging for mobile communications carried out at the R&D division of Televerket/Telenor in mid 80s.

3 *GSM 02.02 – Bearer Services (BS) Supported by a GSM Public Land Mobile Network (PLMN).*

4 *GSM 02.03 – Teleservices Supported by a GSM Public Land Mobile Network (PLMN).*

5 *GSM 03.40 – Technical Realization of the Short Message Service Point-to-Point.*

6 *GSM 03.41 – Technical Realization of the Short Message Service Cell Broadcast.*

7 *GSM 04.08 – Mobile radio interface layer 3 specification.*

8 *GSM 04.11 – Point-to-Point (PP) Short Message Service (SMS) Support on Mobile Radio Interface.*

9 *GSM 04.12 – Short Message Service Cell Broadcast (SMSCB) Support on the Mobile Radio Interface.*

10 *GSM 09.02 – Mobile Application Part (MAP) Specification.*

Finn Trosby graduated from the Norwegian Institute of Technology (NTH) as Chartered Engineer in 1970. He entered the R&D department of Televerket/Telenor in 1972, and since 1980 his main work area was in mobile communications, mostly related to system aspects relevant for new technologies. From 1987 to 1990, he chaired the Draft Group on Message Handling in the working party responsible for designing the data services in the GSM system. From 1990 to 1996 his main work was design of tools for the GSM operator. In 1996, he entered Telenor's mobile operator in Norway, Telenor Mobil, where he has worked since then with company strategy.

email: finn.trosby@telenor.com

The Norwegian GSM industrialisation – an idea that never took off

RUNE HARALD RÆKKEN

Rune Harald Rækken is Senior Research Scientist in Telenor R&D

While Sweden has Ericsson and Finland has Nokia, what did Norway get out of the intense participation, breakthrough of ideas and the acquired knowledge from the GSM specification work? From a scientific point of view, there is no doubt that the Norwegian effort put into the GSM work was a huge triumph. The choice of a narrowband solution for GSM gave a cost saving in the range of billions of kroner (NOK) during GSM deployment in the sparsely populated Norway, compared to wideband solutions proposed and strongly supported by other GSM players and even at political level. In the 1980s Televerkets Forskningsinstitutt (Telenor R&D) had a role as a national telecommunications industrial locomotive, and put a lot of effort into paving the way for the Norwegian GSM industry. The utilisation of the Norwegian GSM position from the national industry players is however nothing less than the largest disappointment that has been seen in the history of Telenor R&D. It is now the time to reveal some of the events that happened in the late 1980s when an effort was made to start growing a national GSM industry based on the research and standardisation effort done by Televerket and the Electronic Laboratory in Trondheim – ELAB.

Introduction

Telenor (then Televerket) had been very active in the standardisation of GSM. Televerket had the chairmanship in SCEG (Speech Coding Expert Group), we participated in WP1 dealing with services, in WP2 dealing with radio subsystems, and Televerket also had the chairman and editors for specifications in WP3 (protocols). Also, Televerket financed the development of the radio modem designed by the Electronic Laboratory in Trondheim (ELAB), which paved the way for GSM as we know it today (see also Torleiv Maseng's article in this issue).

A lot of the ideas generated by Televerket and ELAB can be found in the GSM specifications. Hence the GSM system must be said to be a huge triumph, seen from the Norwegian researchers being involved in the process of specifying GSM. So, we would expect that Norwegian industry should be well positioned for entering the foreseen and important GSM market, growing a national GSM industry on the basis of the work done by Televerket and ELAB.

Trying to build a Norwegian GSM industry

Indeed, the idea of building a Norwegian GSM industry was elaborated, and a consortium was founded to join the Norwegian forces towards an industrialisation of GSM and to give Norwegian industry players a technological lead in the implementation of GSM. The consortium was founded by ELAB, Elektrisk Bureau (EB) Telecom, and Simonsen Elektro. Televerket's contribution to the consortium was to make available for the consortium the know-how of the technical foundation of GSM, and to be the customer of a trial system. At that time it was estimated that Televerket had put a research effort of 50 mill Norwegian kroner (NOK) into research and standardisation activities on GSM.

EB was one of the Norwegian industry giants in the area of electronics and telecommunications (manufacturing, amongst others, fixed line telephones and telephone switches). Simonsen was well known for designing and manufacturing high-end waterproof NMT telephones. The idea was that the consortium of ELAB, EB and Simonsen would design a GSM trial system that Televerket would buy and operate. This would be the first step to prepare the Norwegian telecom industry for entering the GSM market and take advantage of the Norwegian effort put into GSM. Televerket would contribute with know-how on the GSM standards, ELAB would contribute with know-how on implementation of GSM, EB would manufacture base stations and Simonsen would provide the prototype GSM mobile station.

The project started in 1988 with planning and specifications of a prototype system and whatsoever needed to industrialise GSM in Norway and take advantage of the relatively large effort that the Norwegian players had put into GSM standardisation for years.

The collapse

However, the outcome of the project on the Norwegian GSM industrialisation was not much to announce. The consortium more or less collapsed in 1989 when Ericsson, who also had their own activities on GSM, bought EB Telecom. At the same time, Simonsen had got into financial trouble – it was not easy being a small company having only one leg to

 ISSN 0085-7130

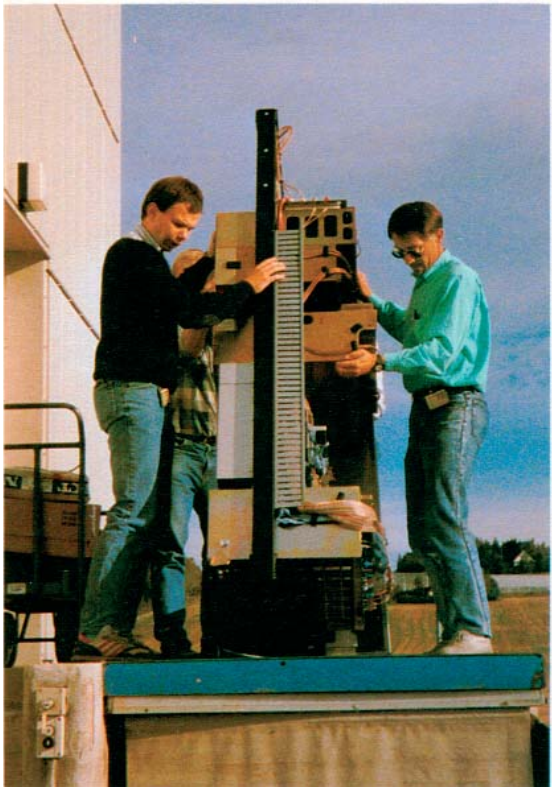
stand on – selling high-end mobile phones in a market where the big international players had dumped the price level of mobiles far below the production price of Simonsen phones.

Simonsen however delivered the promised prototype of a GSM telephone. But the company never managed to take GSM phones into production, and a few years later the name and company of Simonsen was history.

So the Norwegian Ericsson or Nokia never even came to birth. The large Norwegian companies withdrew from GSM development in respect for their foreign owners. The smaller companies never had the resources to overcome such a big task. And that was the brutal end of the Norwegian GSM industry fairytale.

The Nordic GSM trial system

Televerket promoted the Norwegian GSM consortium. But it was realised that the Norwegian telecom industry would not be able to deliver vital GSM system components in time for a GSM trial system that Televerket could use to gain practical experience on GSM before contracts on a commercial system had to be placed. Hence the strategy was changed. Together with the other Nordic telecommunications administrations Televerket started to plan a common Nordic GSM trial system.



Europe's first GSM trial system arrives at Televerket's Research Department (Telenor R&D) at Kjeller in 1989. The author (left) and Senior Engineer Geir Trøan from Televerket's Services Department are taking parts of the equipment into the building

The goal of the Nordic trial system was among other things to verify the radio subsystem of GSM and evaluate GSM performance in different environments (terrain types). The system was to be established in the Oslo area, and the plan was to have a branch to the Copenhagen area later on to evaluate GSM performance in flat terrain.

The next planned phase of the project was for the Nordic telecommunications administrations to perform national tests using the system.

It should be noted that the Swedish Televerket had their sole GSM field trial together with their national industry Ericsson. They however also joined the Nordic GSM trial project.

Several vendors were invited to give a tender and implementation plans for a Nordic trial system. In August 1988 the Danish company Storno was chosen as vendor of the GSM trial system. At that time Storno had been bought by Motorola, so in practice the trial system was a Motorola system. However, the Storno technicians were very amused talking about their sub-contractor Motorola.

The lead-time was very short, as the plan was for the trial system to be in full operation at the beginning of 1989. Motorola had the delivery of a corresponding GSM trial system in the UK.

The system was located at Televerkets Forsknings-institutt (Telenor R&D) at Kjeller outside Oslo, and consisted of one switch, four base stations, eight interference generators and three mobile stations (mounted in 19" racks).

Experimental technology

It soon became quite evident both that GSM at that time was an experimental technology, and that GSM is far more complex than NMT[1)]. We soon got into discussions with the Storno/Motorola technicians who wanted to stay at the test site day and night to make the system operate properly. We did not meet this request however. We regarded 16 working hours a day to be sufficient.

1) *As a curiosity, I can mention that in the second part of the 1970s, there were vendors claiming that NMT was so complex that it could never be implemented. I think that some people could get the same idea about GSM in those prototype days.*

The technicians often received new and updated code from Motorola in USA when they came to the test site in the morning, and there was a struggle to make the system behave in a stable way to give reliable and reproducible test results.

After a long period of alternate successful GSM calls and disconnected calls, and ever lasting system optimisation and improvement, acceptance tests finally started in the autumn of 1989. We did a lot of measurements to verify compliance with the specifications set for the test systems. It turned out, however, that the system was not 100 % stable, and sometimes it was impossible to make reproducible tests.

Hence, after a while it became evident that the trial system did not meet the expectations and the purpose we had when we entered the project. It turned out that we could not utilise the test systems to do a detailed mapping of the performance over to a commercially operated GSM system, to get the flying start as GSM operators that we wished for.

However, during the work with the trial system the representatives of the Nordic telecommunications administrations got a lot of hands-on experience working on GSM and what the implications of GSM might be. So did Motorola.

Bibliography

Annual report 1987. Kjeller, Telenor R&D (Teledirektoratets forskningsavdeling), 1988.

Annual report 1988. Kjeller, Telenor R&D (Teledirektoratets forskningsavdeling), 1989.

Annual report 1989. Kjeller, Telenor R&D (Teledirektoratets forskningsavdeling), 1990.

Collett, J P, Lossius B O H. *Visjon, forskning, virkelighet. Televerkets forskningsinstitutt 25 år*. Kjeller, Norwegian Telecom Research, 1993. ISBN 82-423-0268-5

For a presentation of the author, turn to page 169.

MOU of the GSM-MoU: Memorizing Old Undertakings of the GSM-Memorandum of Understanding

PETTER BLIKSRUD

Petter Bliksrud is Senior Adviser in Telenor Nordic Mobile

When the GSM specifications were mainly finished in the late 1980s, European operators were reluctant to invest in a new and immature system without knowing whether other countries would do the same. In order to move forward, the Western European community needed a declaration, and in 1987 the GSM Memorandum of Understanding (MoU) was signed by 13 countries. The article goes through the ideas and contents of the MoU document as well as the cooperation under it. The operators took on a heavy commitment and worked together for the success of GSM. The opportunity called, and it is not certain that the regulative framework, market conditions and business willingness will combine to allow similar undertakings again.

Background

The CEPT environment, in the early 1980s when the Groupe Spécial Mobile was born and fostered, was the cooperative framework of the governmental telecom monopolies. The responsibility of these monopolies was often twofold: In addition to operating the main national telecommunication systems they also had the regulative task of licensing possible private use of radio equipment and setting the functional requirements for such equipment. The latter engagement did not always include a commitment to actually imply these requirements in the national market. In fact, it happened more than once that a member tabled strict opinions in the discussions on a new radio standard and afterwards declined to introduce the system at home. In other words, the elaboration and adoption of a CEPT standard did not automatically oblige the member countries to the corresponding system deployment and operation. Manufacturing equipment to a CEPT recipe could then be a somewhat risky business for the vendors; and if the standard was comprehensive and on the borders of the state of the art, as the GSM standard was, investing in research, development and production could mean high risk and little business.

And in addition, the mobile telephony environment in Europe at that time was such that a number of countries had introduced analogue systems to build up their national markets. These systems normally differed from country to country, and cross-border roaming was no issue, except in the Nordic countries where the national operators/administrations had worked out their common NMT system. So why, one could ask and many did, should a country (i.e. a national operator, since monopoly was the general *modus vivendi*) invest in a new, immature system if there was no certainty of the other European countries following suit? Something new and alluring was necessary to convince the end users of the change: Roaming across European borders. And a further prerequisite for successful introduction of a new system was acceptable equipment prices, i.e. the economy of scale in mass production.

A new system was desired by politicians, as this would be an ample opportunity for introducing competition also within telecommunications and, of course, form a basis for a new deal in European industrial engagement.

So, the Western European community needed a declaration to move forward: The GSM Memorandum of Understanding was signed in a CEPT Plenary meeting on 7 September 1987 by 13 countries:

- The Quadripartite Group:
  France, Germany, Italy and the UK
- The Nordic Group:
  Denmark, Finland, Norway and Sweden
- The rest of CEPT Group:
  Belgium, Ireland, the Netherlands, Portugal and Spain (actually signing 3 days later)

Signing later, but not much later, were Austria, Greece, Luxembourg, Switzerland and Turkey.

The MoU document

The mediating hand in writing and bringing forward the MoU was that of the UK delegation.

The UK was one of the quadripartites, mentioned above, who were allied through an agreement of cooperation, originally based on the French-German GSM system research program. Members of this group had made considerable investments in the development of the GSM system, and it was vital for them in particular not to lose the momentum in the work towards network implementation. These four

ISSN 0085-7130 © Telenor ASA 2004

countries had internal meetings to coordinate their activities (and their marketing of them), in the CEPT work and towards the MoU initiative. From time to time, these meetings included Sweden, representing the Nordic countries. This extended participation was probably a recognition of the fact that the most advanced area in mobile communications was high up north and that suspicious jealousy, caused by lack of information, could impair the concerted effort towards the common goal.

The UK telecommunications manufacturing industry was indeed minor to the industry on the continent, but the UK was clearly leading in, and bringing experience from, a field of emerging importance: competition. The UK participation under the MoU was divided between two operators, BT/Cellnet and Racal/Vodafone, and one administration, the Department of Trade and Industry, to reign them and rein them both in case of inclinations to diverge opinions. This dual operator mode in the meetings was in the beginning clearly a strange one to the rest of the congregation, but a seed soon to grow in European telecom cooperation.

The original MoU document was made in two languages, French and English, but none to precede the other. This fact was signalling modernity with the French language no longer peremptory. The clause was not cumbersome, really, since detailed exegesis was never expected to become an issue. The other document clauses just stated the aims of the signatories, in broad terms, and set some rules on how to cooperate to fulfil them.

The Cooperation under the MoU

Signatories

The MoU document stated that a signatory could be any telecom operator, within CEPT, licensed in his country to provide "public digital cellular mobile telecommunications services" and, in addition, any CEPT telecom administration. In principle, the MoU cooperation has never been a pure Operator's Club, and the UK Department of Trade and Industry (DTI), the original regulator member, has contributed substantially to the work over the years. But few other Administrations/Regulators have seized the opportunity to become members and thus have information straight from the horse's mouth. The Norwegian Post and Telecommunication Authority (NPT) once tested participation in a plenary some time after the first initial phase of the MoU cooperation. They did not apply for membership afterwards, so maybe the information was not as juicy as we operators would like to believe.

The organisation of the work in the MoU group

The Memorandum prescribed meetings to be held and that a Chairman be elected for a period of six months and then each operator was in turn obliged to nominate a Chairman This clause was amended before every signatory was called upon in this respect. It is fair to say that some operators saw this requirement for nomination as an opportunity, not a burden, while others would rather procure equipment than chairmen. The first MoU meeting was held in London in October 1987, during a GSM Plenary, the so-called *candlelight meeting*, owing to a gale massively devastating power supplies in the city and forcing the meeting to work by candlelight. The MoU meeting was called by personal approach by the convenor, and the proposed agenda handwritten. The Danish representative missed this first meeting; he never investigated whether this was owing to his visiting the restroom just when the call went out (i.e. the convenor went around), or whether he was simply overlooked in the candlelight gloom.

The work of the MoU group was divided into a number of sub-groups. The actual tasks allocated to these groups were a mix between anticipated urgency, workloads and the prestige of having a sub-group chairman. The number was adapted to the most important participating operators (the Quadripartites plus two extras). In one instance, when the nomination of a chairman to a group was challenged, the matter was immediately solved by splitting the group. Not an unknown organisational technique.

Procurement

The document clearly reflected the industrial policy involved in the launch of the new digital system: the operators' procurement policies were to "encourage a strong competitive European industrial manufacturing base", of course properly within the constraints of international trade agreements and operators' own cost efficiency. Such a clause is perhaps not as easily digested among operators today, even if the real commitment to industrial bias is formally not very high. When the four Nordic signatories commonly procured a GSM test system that was not European, this clause did not create uneasiness.

It was an MoU goal that signatories should open the GSM service in 1991, i.e. about four years after the establishing of the MoU. It was a good choice, whether the result of clever reasoning or sheer luck. Some operators needed the new system as soon as possible as an alternative to the exhausted analogue network, to them four years was a long time. But the specifications were really immature in 1987, and a

shorter time would not be realistic, if the system should be expected to work properly.

The procurement rules to be followed were quite extensively described. In fact it is the only really detailed instruction in the MoU and it filled an annex of its own. The procurement process to be followed, by all signatories, included system validation phases where enough equipment should be procured to verify the system; first the air-interface and the interfaces between the base stations and the switch and then to procure more pieces to validate interfaces to other networks and interworking with other networks. Test systems appeared around Europe and tests were performed to the benefit of the operators, and actually of no less benefit to the infrastructure manufacturers.

The GSM system did work, but the first networks were actually out of work for quite a long time: there were no mobiles to be bought. Part of this delay was the type approval, or lack of it. Several test houses were established, preparing for type approval competition. However, it turned out that the necessary type approval equipment required investments and when there seemed to be more test houses than mobile types, investment willingness faded and resulted in some of the more solid operators buying the simulator under the auspices of the MoU cooperation.

The GSM standard

The MoU obligations of course referred distinctly to the CEPT recommendations for the digital system in the 900 MHz band. The document also instituted the MoU group as an instrument to help CEPT get out of possible deadlock situations caused by the requirement for unanimity. Voting was introduced in MoU with weighted voting rights for the national signatories, the weighting taken from a table of somewhat obscure origin. If a difficult situation should occur in CEPT, an MoU meeting could be called and a voting held on the subject, not unlike the voting in the European Song Contest and the signatories would all abide with the majority view in CEPT. Voting was never taken, the scheme served, in itself, as a deterrent.

Network deployment

The MoU document carried obligations on the signatories to cover particular areas in their countries in an early roll-out phase of the network, such as capital cities, main roads and main airports. This gave MoU members a possibility to criticize other members' coverage plans (and corresponding roaming possibility for neighbouring countries), and some could not resist the opportunity of such criticism. In one particular instant in the early days of the MoU, the discussion between two representatives became very heated before it all calmed down to a level of more normal roaming negotiations.

Promotion

The document also focussed on the promotional aspect. The signatories were required to make efforts to extend the GSM system to all territories of CEPT, i.e.Western Europe, as well as to provide information to the public in their own territories of the glory of the new digital, cellular system. It was even ventured, as proposed by the Nordic Administrations, to try to promote the standard for take up in the territories outside CEPT; that is to make the standard really international. A dedicated sub-group was given the task of working out generic material for such national promotion. This group was also tasked to come up with a better name for the system, marketwise. Groupe Spécial Mobile was not felt *piquant* enough for the broad public, even if it was in French. The exercise focussed on snappy names containing popular syllables like: *net, com, cell, mob* and *tel* either in combination or spiced with glorifying adjectives (*super, universal*). The surprise was not so much the number of permutations proposed, but the fact that any one of them would be in use somewhere in the world and thus would have to be bought from some hitherto unknown company. Since the acronym GSM was already ubiquitous, far more ubiquitous than GSM pieces of equipment and actually constituting a brand name, it was in the end decided to keep the letters but change their interpretation. "Global System for Mobile communications", however, leaves the last word, and maybe the most important one, somewhat in limbo.

Intellectual property rights

The signatories should align their policies on the Intellectual Property Rights (IPR) to ensure that the main interfaces in the standard were open. The governmentally funded research projects in some countries forbade the exploitation of patents and the participating operators and industry could be handicapped if others were free to apply patents. The MoU group therefore drew up a patent declaration to be issued with the national calls for tender, conditioning licence free conditions for the purchase. This turned out to be putting operator hands into a manufacturer beehive where patents are honey for trade. In particular Motorola objected and they indicated that they had applied for essential GSM licences, the contents of which could not be revealed according to the patent rules. The company and the MoU group met trying to agree in substance, and when that failed, on a text flexible enough to delay possible confrontations. The meetings were many in number and long in duration, but the avail was essentially that the problem of the antagonism between standard making

and patents were transferred to a more general level in the new European standard organisation (ETSI). The operators bought their GSM equipment without too much patent ado. The possible patents concealed in the deals were apparently exchanged on fair and reasonable terms.

Expansion of the MoU document

The work in the MoU Group turned out to be quite successful, so that clauses in the document, once goals to strive at, turned to constraints and the focus on Western Europe became too narrow. In 1991 the first non-European country, Australia, applied for membership.

Therefore, in 1991 an addendum was made to the original paper. Possible members from outside CEPT were released from the bonds to support European industry and the initial requirement to open the service in 1991 just had to be adjusted if one would avoid becoming a very closed interest group as time went by.

At that time the UK had once again been the vanguard in competitive regulation and had introduced three licences for GSM at 1800 MHz, also called PCN (Personal Communication Networks). These newcomers formed their own, competing operator organisation, to become, however, quite short lived before taken up in the real MoU. Their system (GSM in 1800 MHz) was not allowed to be called GSM 1800 until later and they had to abide for a while with the pseudonym, DCS (Digital Communication System).

New housekeeping rules were introduced. Chairmen were to be elected from volunteers, term of office prolonged to one year and it was necessary to put up a permanent secretariat. After a brief visit to the Hague, the MoU came to pick the fair city of Dublin as its home town.

The frontier era was over in the mid-nineties. The Memorandum of Understanding left the arena and the cooperation re-entered a new outfit, namely as an Association. It just had to be so, to cater for a substantial increase in member operators and new challenges and new undertakings. The Magna Carta of the organisation has now changed a bit. The old document had goals defined in detail and only few working rules. The new document has goals that are more generic and more loosely defined since they shall cover a plethora of members and should last longer. Details (whether or not the Devil is in them) have jumped from goals to rules.

Retrospective

If the telecommunications world is a stage, the play "GSM" has been a box office hit. The operators have both been playwrites and acting in main parts. Some would say that the manufacturers were the protagonists, as always, since vast equipment developments have conditioned the success. A touch of complacency by the operators should be condoned, though. We took on heavy commitments and worked together as it was required when opportunity called. It is not certain that the regulative framework, market conditions and business willingness will properly combine to allow similar undertakings again.

Petter Bliksrud (62) graduated from the Norwegian Institute of Technology in Trondheim, Norway, in 1968 and was subsequently employed by Norwegian Telecom. He has been involved in all areas of mobile communications, both terrestrial and satellite. In particular, he has participated in the ITU (CCIR) development of maritime mobile services and in international search and rescue satellite projects. In the 1980s he was engaged in the CEPT/ETSI specification work on the GSM system and in establishing the GSM Memorandum of Understanding and was in the succeeding decade actively participating in several international bodies on mobile communications. His present area of work within Nordic Mobile in Telenor concerns primarily questions related to national and international telecom regulations.

email: petter.bliksrud@telenor.com

ISSN 0085-7130

Introduction: From the archives

PER H. LEHNE

Per H. Lehne is Research Scientist at Telenor R&D and Editor in Chief of Telektronikk

Taking over the responsibility for an institution like *Telektronikk* soon made it clear that I had to educate myself in the history of the journal. Even though I had been involved with the journal since 1997, as editor for the Status section, I had not spent much time looking back into its past.

I am sure most of my predecessors have felt the same need. For me it became more intense when the 100th anniversary was coming up. Spending hours and days browsing the archives and reading previous issues and volumes made me also want to bring some of the knowledge out to our readers.

This section, 'From the archives', appears in this special Anniversary Issue and contains two articles from the early years. The idea is to illustrate how *Telektronikk* has imparted news and knowledge about important events and technologies in previous times.

The following articles describe two pioneering events in early telecom: The first transatlantic telegraph cable in 1858, and the first Northern European wireless telegraph established in Lofoten in North Norway in 1906. They are both based on early articles in *Telektronikk*'s predecessor, *Technical Information* in 1906 and 1907 respectively.

The transatlantic telegraph cable from 1858 broke the ground (or should we say: 'sea') for connecting the continents telegraphically. A lot of the physics of such a connection was unknown and had to be explored, and the problems had to be solved along the way. It involved contemporary engineers and scientists like Charles Wheatstone and William Thomson (Lord Kelvin). Even if it operated for only 27 days, it paved the way for permanent installations.

Wireless communications has become the most important business of today's telecom operators, Telenor included. Right from the beginning, Norway has been an early adopter of wireless technology to cover telecommunications needs. Just a few years after Guglielmo Marconi had demonstrated that radio waves could reach beyond the horizon, Telegrafverket installed the first North European permanent wireless telegraph. It was actually the second in the world.

Next year, in 2005, Telenor can celebrate its 150th anniversary. Snippets from the archives may well appear under the 'Kaleidoscope' heading together with other contributions of an historic nature.

For a presentation of the author, turn to page 9.

ISSN 0085-7130 © Telenor ASA 2004

The transatlantic telegraph cable of 1858 and other aspects of early telegraphy

PER H. LEHNE

Per H Lehne is Research Scientist at Telenor R&D and Editor in Chief of Telektronikk

From the very beginning, *Telektronikk* has made room for historical articles on inventions and major events in telecommunications. The first transatlantic telegraph cable was laid in 1858 and a comprehensive article on 'Cable telegraphy' from 1907 describes the project. It also describes other sub sea cable projects from the time of the early telegraph in the 1850s until 1907. Cables and cable technology was of major concern for the Norwegian Telegraph Administration for many years, because most of the investment cost in establishing the telegraph and telephone networks lay here.

Introduction

Both manufacturing and deploying a sub sea cable of approximately 3,000 km length was a risky and expensive project in the 1850s. There were a lot of unknowns and the first cable, which cost a total of £ 465,000, only worked for less than a month. This article contains a summary of the original article from *Technical Information* (*Telektronikk*, see [1]) in 1907. Additional information has also been gathered. A book from 2003, *'The Cable – The Wire that changed the World'* by Gillian Cookson [2] tells the full story of this daring project. However, we will start with some introductory notes on Samuel F.B. Morse and his 'electric' telegraph as well as the introduction of the telegraph in Norway.

The Morse telegraph

Samuel Finley Breeze Morse (1791 – 1872) is recognized as the inventor of the electric telegraph. Morse was an American artist, educated in England, but before that he had studied at Yale College, and was excited by lectures of the then newly-developing subject of electricity.

In 1832 Morse was on his way home from Europe on the ship *Sully*. Conversations with Charles T. Jackson, Professor in Physics at the University of Boston, gave Morse an idea. Jackson had been to Paris and heard the latest news on research in electricity. During the trip across the Atlantic Morse had developed and sketched the main concept for a *telegraph apparatus* based on electrical pulses.

It is believed that he had his first telegraph model working in 1835 in the New York University building where he taught art. In his model he used crude materials: an old artist's canvas stretcher to hold it, a home-made battery and an old clockwork to move the paper on which dots and dashes were to be recorded. With the aid of new partners, Morse applied for a patent for his new telegraph in 1837. Morse was discouraged from his art career and was devoting nearly all his time to the telegraph.

His telegraph was exhibited in New York in 1838, and there Morse transmitted ten words per minute. Instead of using the number-word dictionary (see frame about the Morse alphabet), he used dot-dash code directly for letters. The Morse code that was to become standard throughout the world had essentially been introduced [3].

The Morse code was dominant until the 1930s, but already in 1902, *Charles L. Krum*, a mechanical engineer and vice president of the Western Cold Storage Company, had started to develop the teletypewriter machine. This differed in many ways from the current interest of the telegraph engineers, which was to further develop the Morse telegraph. Krum equipped his *teletypewriter* with a keyboard similar to an ordinary typewriter. Speed telegraphs, like the Wheatstone – Creed Morse telegraph relied on special trained people and had problems with the synchronisation between the transmitter and receiver. The synchronisation problem was avoided by introducing start and stop signals for each combination of impulses (char-

In 1844, Morse demonstrated to the US Congress the practicality of the telegraph by transmitting the famous message "What hath God wrought" over a wire from Washington to Baltimore. The picture shows Morse's telegraph receiver, which he used at the demonstration in 1844 (Smithsonian National Museum of American History)

ISSN 0085-7130

The Morse alphabet (1832 – 1999)

The original Morse code from 1832 was a translation system consisting of two essential parts:

- A two-way code book or dictionary in which each English word was assigned a number (and in order to spell out proper names, unusual words, initials, etc., when necessary, each letter of the alphabet was also assigned a number), and
- A code symbol for each digit from 0 – 9 to represent that number.

The sender would convert each word to a number, send that number and the receiver would then convert it back again to the English word with a reverse dictionary.

Later, the dot-dash code was invented and was eventually named the *International Morse Alphabet* and adopted by the ITU.

The International Morse Alphabet has later gone through smaller changes and amendments. As an example, the international distress signal 'SOS' (as one sign) was made in 1906, when radiotelegraphy appeared. As late as in the 1960s special signs were introduced for e.g. first and last parenthesis.

Morse's system was practically supreme in transferring text from the start in the 1840s until 1910–20 when the telex arrived.

Also in Norway, when the State Telegraph was established in 1855, Morse's system was used. In the mid 1950s the use of Morse coding was completely left in the Norwegian network. On the wireless, the last fixed communication links were migrated to telex lines during the 1960s.

31 January 1999 was the last day that the International Telecommunications Convention demanded knowledge competence in Morse telegraphy aboard ships over a certain size.

acters). A new code was introduced, using five impulses per character. It was originally invented by *Émile Baudot* around 1874 and adopted by the CCIT as the 'International Alphabet No 1' (IA1). Later, it was improved by Donald Murray by adding extra characters and 'shift' codes. The code is still known as the 'Baudot code' also known as the 'International Alphabet No 2' (IA2). The code was used as long as the teleprinter service was operational. The term "baud" (a measure of symbols transmitted per second) is named after Baudot.

In 1934, CCIT's Plenary Assembly in Prague discussed the matter, and a test system for the new telex-service was demonstrated. The name 'telex' is short for 'teleprinter exchange' and became the international name for the teleprinter-subscriber telegraphy. The same year, a telex service was opened in Denmark, Germany and Switzerland using dedicated telegraph lines. Great Britain and the Netherlands used a different approach to combine it with the telephone service, but this was later abandoned. The teletypewriter exchange (TWX) service in the USA was introduced in 1930, but it did not apply with the recommendations from CCIT, thus making it difficult to interconnect the European and American telex networks.

The electric telegraph comes to Norway

The Norwegian State Telegraph (Statstelegrafen)[1)] was established January 1, 1855 with the first Morse telegraph line between Christiania (Oslo) and Drammen. Already in the first year of *Technical Information*, in 1904, the 50th Anniversary of the public telegraph in Norway is treated.

But before Statstelegrafen, in 1853, a telegraph line owned by the first railway company in Norway was opened between Christiania[2)] and Strømmen/Lillestrøm (approximately 20 km east of Oslo). *Telektronikk*, issue 2, 1961 contains the lectures from the Norwegian Telephone Engineers Conference (Telefoningeniørmøtet – NTIM) in 1960. In this issue, Head engineer *L. Saxegaard* from the Norwegian State Railways (NSB) introduces a lecture about the State Railways' telecommunications with this fact.

On September 19, 1853, the year before the first Norwegian railway ('Hovedbanen' between Christiania and Eidsvoll) was opened, the first railway telegram was dispatched from a workmen's barrack somewhere between Strømmen and Lillestrøm, east of Oslo. The sender was *Mr. Greener*, the telegraph constructor for the railway, and it was received by Greener's Norwegian assistant, *Christian Wiger*. The message was simple: "The telegraph in order between here and Christiania." Wiger later became NSB's telegraph inspector, a position he had for over 50 years.

This was not a Morse telegraph but used a double-needle system with a different alphabet. In March 1854, the line was opened also for private traffic, and it was frequently used for ordering transport. The price of a transport-telegram was 12 'skilling' (40 'øre')[3)] for 15 words. Other types of telegrams cost twice as much. Up to 70 telegrams per day were dispatched.

---

1) *Norw.: 'Statstelegrafen'. The name of the national Norwegian Telegraph and Telephone has changed throughout the years. We will use the Norwegian terms 'Statstelegrafen' for the early years and 'Telegrafverket', which was the name until 1969. Then, the name 'Televerket' and 'Teledirektoratet' came into use, which in English was common to name 'Norwegian Telecom (Administration)'. From January 1, 1995, the name of the company is 'Telenor', which makes no difficulties in English.*

2) *The Norwegian capital, Oslo, had the name 'Christiania' until 1925.*

3) *The Norwegian currency became 'krone' = 100 'øre' in 1875. Before this, the system was based on 'Speciedaler' = 120 'skilling'. 1 Norwegian 'krone' was defined to be 1/4 'Speciedaler', thus 30 'skilling'.*

After Statstelegrafen/Telegrafverket was opened in 1855, the development went fast. The following year, a line to Sweden was opened, and from there, Denmark and parts of the European continent could be reached. In a few years, lines were built from Oslo to Bergen and Trondheim, and during the next decade lines to North Norway were ready. Lofoten were in a special position when it came to expanding the telegraph in the north, because of the important fisheries in the area. That is the reason why Lofoten already in 1861 had a telegraph line, going from Sørvågen to Brettesnes (see also the article on the wireless telegraph in this issue [4]). This line was connected to the main national network in 1868. In 1870, the line had reached Norway's eastern outpost, the town of Vardø. In 1867, the first sub sea cable between Norway and Denmark was opened (see later in this article).

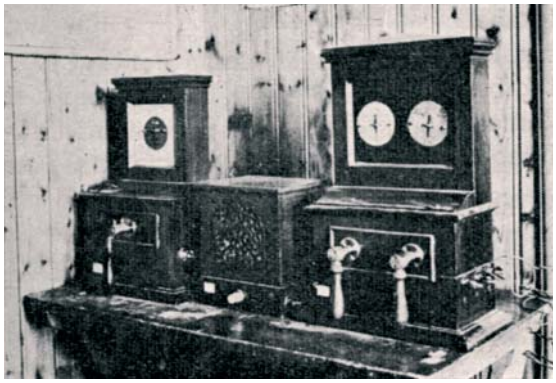
Telegrafverket's history has been covered on several occasions in *Technical Information* and *Telektronikk*. It is especially covered in 1955, on the occasion of the 100 years anniversary of Telegrafverket. Issue 1-3 in 1955 contains a 24-page article by the editor, *Julius Ringstad* (Editor from 1942 to 1957) and his successor *Nils Taranger* (Editor from 1957 to 1978).

Taranger later writes a retrospective article about telegraphy in *Telektronikk*, issue 3-4, 1963, in which he draws up the development of the telegraph until then. The spirit of the article is optimistic on behalf of the telegraph, and also competitive in the sense that he promotes telegraphy to the expense of the telephone, which has had a rapid growth in the period until 1962. In a sense, he foresees the need for data communications in the future, something in the endnotes he calls 'a kind of telegraphy'.

Morse telegraphy was gradually superseded by the telex system also in Norway, and the first manual telex exchange was opened in Oslo in February 1946. Initially it had 20 numbers, which was increased to 35 a month later. In 1946, there were 10 subscribers in Oslo, 4 in Bergen and 1 in Trondheim. The growth was slow, and in 1953, there were only about 290 subscribers in total. By the end of 1961, the number had grown to 1290.

In July 1957, the first automatic telex exchange was opened in Oslo with a capacity of 500 numbers. Bergen followed in April 1961, and the Trondheim exchange was under installation in 1962.

In Norway, the telex network was closed on June 30, 2000, however telex messages to and from the international telex network can still be sent and received via electronic mail. At the time of closing, the network had about 800 subscribers.



The first telegraph apparatus owned by the railway company operating 'Hovedbanen' between Christiania and Eidsvoll was used on the section between Christiania and Strømmen/Lillestrøm. It was of the double-needle type. On the right there are two 'needles', which moved right or left, according to the position of the two transmission handles below. When the handles were resting in the vertical position, the apparatus was in receiving mode (Illustration from Telektronikk, issue 2, 1961, p 87)

The first transatlantic telegraph cable

The following story is based on an article on cable telegraphy, which appeared in *Technical Information* in 1907. The author is unknown, as the rule was not to ascribe articles in the early volumes.

Already in 1842, Samuel Morse had tested a cable in New York harbour across part of East River through which he sent signals, and the following year he had suggested to the US government to establish an electric connection between Europe and America.

In 1855, the American telegraph network had reached Newfoundland, and on the European side, several places on the Irish west coast also had telegraph service. Thus, the possibility of inter connection was discussed on both sides of the Atlantic. Severe difficulties were reported, the shortest distance was 3,700 km and the sea depth reached down to 5,500 m. One was only in doubt whether it was possible to transmit telegraph signals on such a long distance with a feasible speed.

The Atlantic Telegraph Co. was incorporated in October 1856. The original capital was 300 shares at £ 1,000 each. Of this, Cyrus Field and John W. Brett took 25 shares each. The capital was increased by £ 50,000 a few days later. The total of £ 350,000 was secured by December 1856. At the end of 1858, all the company's assets had been spent, and with more liabilities than it could meet [2].

ISSN 0085-7130

Professor *Charles Wheatstone*[4] had, in his laboratory, measured the 'speed of electricity' to be 300,000 km/s. In telegraph wires, the speed had been determined to 30,000 km/s and the speed on the subterranean and sub sea connection between London and Dublin was found to be much lower. In 1854, Charles Bright had estimated the speed on gutta-percha insulated cables to be less than 2,000 km/s.

However, the doubts were resolved by *Edward O.W. Whitehouse*[5], who tested this by interconnecting several subterranean cables to form a length of approx. 3,700 km and achieved a speed of 210 – 270 signals per minute[6]. Morse himself participated in these tests during autumn 1856. After having slept on it, he estimated that eight to ten words per minute, or twenty messages an hour, could be sent between Ireland and Newfoundland [2].

In 1856, *John Watkins Brett*, *Charles Bright* and *Cyrus W. Field*[7] founded the *Atlantic Telegraph Company* (see frame). The necessary initial capital corresponding to 6 mill Norwegian kroner (NOK)[8] was raised in London, Liverpool and Glasgow. The following year it was raised to 8.5 mill and with a guarantee from the English State corresponding to 250,000 NOK per year and available ships for deployment.

Fig 3.

The cable used consisted of 7 strands of 0.7 mm copper wire covered with three layers of gutta percha. The outer diameter was just 10 mm. The core was wrapped in jute yarn soaked with a composition consisting of 5/12 Stockholm tar, 5/12 pitch, 1/12 boiled linseed oil and 1/12 common bees wax. Armouring consisted of 18 strands, each strand composed of 7 of the best charcoal iron wires, each 0.7 mm. The finished cable was then dipped in a heated mix of tar, pitch and linseed oil (Illustration from 'Technical Information', issue 1, 1907)

The production of the cable started in 1857 with a time schedule of four months. The distance between the landing points was determined to be 3,000 km, thus a cable length of 4,600 km was set. The European landing point was the island of Valentia[9], off the Irish southwest coast, and Trinity Bay in Newfoundland was chosen on the American side.

A lot of problems and improvisations were needed. For example, no housing was established on the beach for the manufactured cable, and it had to lie in the open exposed to the sun, which deteriorated the outer layers of the cable. This was probably an additional factor which influenced the rapid failure of the cable (see later).

Two ships were used for the deployment, the British Battleship, H.M.S. *'Agamemnon'* and the American steam frigate *'Niagara'*. With half of the cable on board each ship, the idea was to start in the middle of the Atlantic Ocean and lay the cable towards each side. For some reason, Whitehouse disturbed this plan, and the deployment started in Ireland on August 6, 1857. After several problems, the cable broke on 4,200 m depth after 620 km had been laid. The ships had to return to England where 1,300 km of new cable had to be manufactured.

On June 16, 1858, the deployment was started again, now from the mid Atlantic as originally planned. In this attempt, the cable broke three times, and they had to start afresh again. More than 1000 km of cable was lost and more cable had to be manufactured. On the

---

4) *Sir Charles Wheatstone (1802–1875) is for electrical engineers best known for the 'Wheatstone's Bridge', a method of doing precision measurements of electrical resistance. However, he was heavily engaged in the telegraph from 1835, and together with partner William Cooke, laid the first experimental line between Euston terminus and Camden Town station of London and North Western Railway on July 25, 1837. He had already in 1840 devised a method to connect Dover with Calais across the English Channel.*

5) *Edward O. Wildman Whitehouse was a retired medical doctor who had taken an interest in electricity and telegraph signalling. He had a practical approach to the problem and was later blamed for the failure of the cable by driving too much current through it. He was in conflict with another of the 17 directors of the project, Prof. William Thomson of Glasgow University, in 1892 to be knighted as Lord Kelvin for his achievements in science, among them the Atlantic telegraph cable project in 1866.*

6) *The article does not say what is meant by the term "signals per minute", but the same numbers are given in [2].*

7) *Cyrus West Field was an American entrepreneur from Massachusetts. By age 20, he was a partner in a paper manufacturing company and at 33 (1852) he had retired from business as a wealthy man.*

8) *At the time, the exchange rate between Norwegian kroner (NOK) and £ (GBP) was close to GBP = 20 NOK.*

9) *Today, Valentia is best known as a resort island. The cable station from 1865 was closed in 1965, and is now partly a factory and partly residential. In 2000 it was recognized as an IEEE Electrical Engineering Milestone, and a plaque is placed at the station. See: http://indigo.ie/~cguiney/valentia.html.*

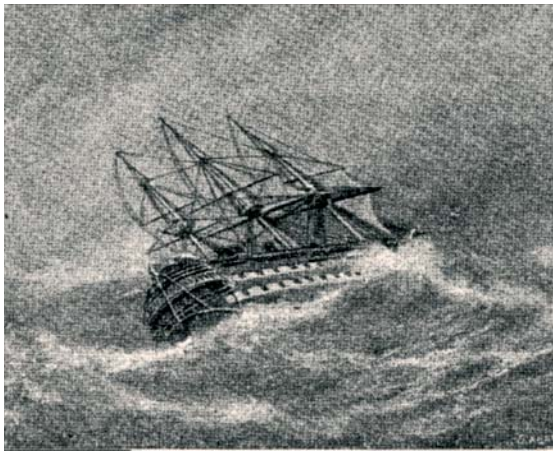
third attempt, which started on July 29, things went well. The *'Niagara'* landed in Newfoundland on August 5, and the *'Agamemnon'* reached Ireland at the same time after only one accident with the cable. Immediately after the cable ends had been brought ashore, the first signal was received. "The joy was great in England and America, and great festivities were held in both countries." Of course, congratulatory telegrams were sent between Queen Victoria and U.S. President James Buchanan.

A major error was however soon discovered 550 km from Valentia, and on the 1 September, the connection was broken. In total, 732 telegrams had been sent through the cable. One of these is apparently said to have saved the English government a sum corresponding to 900,000 NOK[10].

The first transatlantic telegraph cable which worked reliably for a longer time was laid and opened on July 27, 1866, this also between Valentia and Newfoundland; however the short-lived cable laid in 1858 shows the feasibility of such a project. The article from 1907 just briefly describes the 1866 cable.

Norway's first sub sea cable

In the early years after the telegraph had come to Norway, international traffic had to go through Sweden and Denmark, however the Danish-German War in 1864 demonstrated the necessity for multiple choices and redundancy. The first sub sea cable from Norway was then opened on July 1, 1867. In 1866, a licence was given to an English company for the deployment of a cable between Norway and Denmark. Obviously, a conflict has existed towards another powerful telegraph company, which wanted to take over this licence as well as a licence for a Danish-English cable. Because of this conflict, the cable could not be manufactured in England. A later licensee was the cable manufacturing company *Newall & Johnson*, and they were forced to use parts of an old cable, which had been used in Greek waters. The parts were sent to Arendal, on the south coast of Norway, where Telegraph Director *Carsten Tank Nielsen* and Telegraph Commissary *Jonas P.S. Collett*[11] evaluated the state of it. They could not approve it; however, it was sent to Copenhagen and



The British Battleship H.M.S. 'Agamemnon' experienced rough weather on its trip from the mid Atlantic until it reached the Irish coast on August 5, 1858 (Illustration from 'Technical Information', issue 1, 1907, p 4)

examined by Professor Thomsen[12], who found it to be all right. The cable was then laid and opened on July 1, 1866.

Sub sea cables both for telegraph and telephone are the subject of articles in *Technical Information* and *Telektronikk* on several occasions into the 1960s. Overview articles of different cables from Norway to the continent as well as the British Isles are published, as well as technical tutorials for specific problems.

A curiosity can be mentioned. In 1929, volume 9-10, a notice is printed under the signature 'He' about plans for a transatlantic telephone cable. It is a transcript of the *Journal of A.I.E.E*[13] and reports that the Bell Telephone Laboratories in 1928 had completed a deep-sea cable, which could be feasible for a transatlantic telephone connection. It is also reported that one is undertaking work to develop a cable system 'of this type' to connect London and New York City, and that the connection can possibly be made as early as 1932. Obviously, it took longer, and whether that stems from technical, economic or political problems shall remain unsaid. The first transatlantic telephone cable was completed in 1956, the Trans Atlantic Cable No. 1 (TAT-1), between Oban on the coast of Scotland and Clarenville, Newfoundland.

---

10) According to Cookson [2] one significant message was about the news that the Sepoy Mutiny in India had been put down, which halted the mobilisation of two British regiments from Canada. Nine words had saved the British Government £ 60,000.

11) Carsten Tank Nielsen was the first Norwegian Telegraph Director from 1855 – 1892. Jonas P.S. Collett was temporary Director for three and a half months in 1892 until Jonas Severin Rasmussen was appointed.

12) This could be Professor Julius Thomsen from the University of Copenhagen. He was a professor of chemistry from 1866, and he has a merit in being the first to suggest the modern form of layout of the Periodic Table.

13) A.I.E.E. – American Institute of Electrical Engineers – is the predecessor of IEEE until this was formed in 1963.

ISSN 0085-7130

References

1 Lehne, P H. Enlightening 100 years of telecom history – From 'Technical Information' in 1904 to 'Telektronikk' in 2004. *Telektronikk*, 100 (3), 3–9, 2004. (This issue)

2 Cookson, G. *The Cable – The Wire that changed the World*. Gloucestershire, UK, Tempus Publishing Limited, 2003. ISBN 0 7524 2366 5.

3 *Locust Grove – The Samuel F.B. Morse Historic Site*. September 22, 2004 [online] – URL: http://www.morsehistoricsite.org/history/morse.html

4 Lehne, P H. The second wireless in the world. *Telektronikk*, 100 (3), 209–213, 2004. (This issue)

Sources from Telektronikk's archives

Brief overview of the Norwegian Telegraph's technical development during its 50 years of existence (Kortfattet oversigt over Telegrafvæsenets tekniske udvikling under dets 50-aarige bestaaen). *Technical Information*, 0 (8), 57–74. 1904.

Cable telegraphy (Kabeltelegrafi). *Technical Information*, 3 (1), 1–17, 1907.

Trans Atlantic telephone cable (Transatlantisk telefonkabel). *Technical Information*, 25 (9-10), 70, 1929.

Taranger, N, Ringstad, J. The telegraph (Telegrafen). *Technical Information*, 51 (1-3), 30–54, 1955.

The Norwegian Telegraph through 100 years (Telegrafverket gjennom 100 år), Anniversary issue. *Technical Information*, 51 (1-3), 1955.

Saxegaard, L. The Norwegian State Railway's telecommunications (Statsbanenes telekommunikasjoner). *Telektronikk*, 57 (2), 87–98, 1961.

Taranger, N. Characteristics of the development towards modern telegraphy (Karakteristiske trekk ved utviklingen frem til moderne telegrafi). *Telektronikk*, 59 (3-4), 160–173, 1963.

Other sources

History of the Atlantic Cable & Submarine Telegraphy. September 11, 2004 [online] – URL: http://www.atlantic-cable.com

Lindley, D. *Degrees Kelvin. The Genius and Tragedy of William Thomson*. London, UK, Aurum Press Ltd., 2004. ISBN 1 84513 000 6

Norwegian Telecom Museum (Norsk Telemuseum). September 22, 2004 [online] – URL: http://www.norsktele.museum.no/index_ieoff.html

Rafto, T. *The history of the Norwegian State Telegraph 1855–1955 (Telegrafverkets historie 1855–1955)*. Telegrafverket, 1955.

For a presentation of the author, please turn to page 9.

The second wireless in the world

PER H. LEHNE

Per H Lehne is Research Scientist at Telenor R&D and Editor in Chief of Telektronikk

Norway was early in adopting the wireless telegraph. Already in 1899 the Telegraph Administration had sent an engineer to study Marconi's system. The motivation of the Telegraph Administration was the communication needs of the large cod-fishing activity, which created a trade of substantial national value. The fishing ports and islands were hard to reach without laying expensive cables, and wireless communications were coming up as an attractive solution. The project of planning and deploying what became the second wireless telegraph line in the world was told by the first editor of *Telektronikk* the same year it happened, in 1906.

Introduction

The first Norwegian permanent wireless telegraph was established in Lofoten in Northern Norway between the fishing port of Sørvågen and the port on the island of Røst. This happened in 1906, and the same year, *Hermod Petersen*[1)], the first editor of *Technical Information* (*Telektronikk*, see [1]), writes an article about the project. Petersen was in charge of the project, thus in this case we can safely say that the story is told 'in real time' by the history makers themselves. Articles from this time are not filled with personal details, but are very sober and based on facts. However, it is possible to sense the time spirit of combined entrepreneurship and the conception of taking part in a nation-building project.

This article reviews the story of establishing Norway's first wireless telegraph in 1906, but before that we will shortly review Guglielmo Marconi's experiments, which culminated in the demonstration of the transatlantic wireless telegraph in 1901.

Marconi's experiments

From a telecommunications point of view, the wireless revolution started with the Italian scientist and engineer, *Guglielmo Marconi*'s experiment in 1901. On 12 December, Marconi received the first long-distance radio transmission at Signal Hill, St. John's, Newfoundland, 3,592 kilometres from the transmitter. This was situated at Poldhu in Cornwall on the Atlantic coast of England. Here, electrical engineer *John Ambrose Fleming*[2)] transmitted the Morse code signal for 'S' from across the Atlantic Ocean in England and Marconi heard it – three short clicks – through a radio speaker. Strong winds at Signal Hill made it difficult for Marconi and his assistants to launch the aerial on its kite, but at 12:30, 1:10 and 2:20, he was able to pick up the three dots that were being transmitted from Poldhu.

It was believed at that time that wireless transmission was limited to a range of approximately 200 miles (320 km), due to the curving of the earth's surface. When Marconi's experiment proved otherwise, it was a sensation. It proved conclusively that radiowave transmission was not bounded by the horizon. Shortly after this, contemporary scientists *Arthur Kennelly* and *Oliver Heaviside* suggested the existence of a layer of ionised air in the upper atmosphere (the Kennelly-Heaviside layer, now called the E region of the ionosphere) [2].

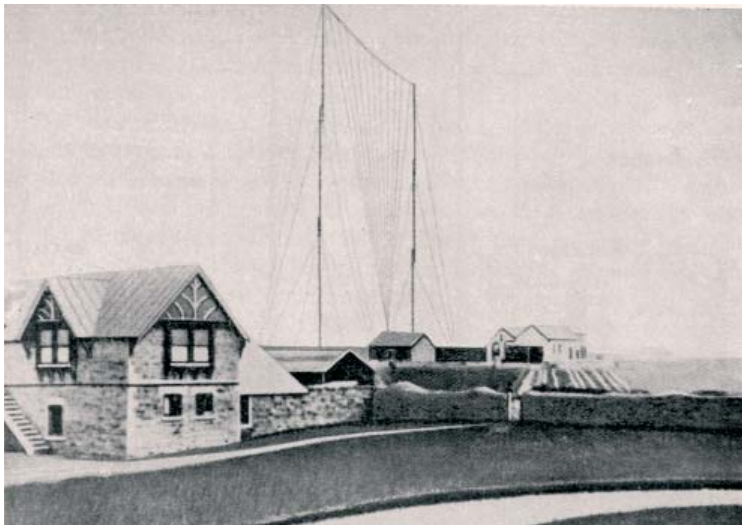
The size of the Poldhu station was tremendous. The antenna system consisted of about 400 wires sus-

Guglielmo Marconi (1874 – 1937) in the station at Signal Hill, Newfoundland in 1901

1) *The article uses the Norwegian term "vedkommende ingeniør", probably meaning the person who writes the story. Rafto [3] confirms that this was Hermod Petersen.*

2) *Sir John Ambrose Fleming was himself an inventor. In 1904, he demonstrates the vacuum tube diode, the 'pre-amble' to the radio valve, which appeared in 1906.*

ISSN 0085-7130



The first aerial erected for the transatlantic experiment at Poldhu station was a circular structure; however this system was destroyed in a storm before the experiment could take place. This temporary aerial was then erected and used for Marconi's experiment on 12 December 1901 (Photo from T. Rafto [4])

pended in an inverted cone from a 60 metre circle of twenty 60 metre high masts. This system was destroyed in a storm in 1901 and a temporary aerial was constructed using two of the original masts from directly opposite sides of the circle. This was the aerial configuration in place during the transatlantic experiments in December 1901 [3].

Sørvågen – an early Norwegian hub in wireless communication [6]

1861	Norway's first permanent 'fishing' telegraph. The 'Lofoten-line' was 170 km and was the country's first telegraph line outside the main national network. It connected nine fishing ports telegraphically.
1906	The first North European permanent wireless telegraph was established between Sørvågen and Røst – the second wireless in the world.
1908	Norway's first permanent ships telegraph. In 1906 and 1907, the staff at Sørvågen tried intensely to serve Emperor Wilhelm II's ship, the 'Hohenzolleren' in the best possible manner, however in vain. Contact was not established. This was regarded a shame for the young nation of Norway. On July 1, 1908, contact was established and we can probably thank the German Emperor for Sørvågen telegraph station being opened for communication with ships.
1928	Norway's first permanent wireless telephony station in Sørvågen communicated with another station at Lofotodden.
1946	Norway's first permanent, non-military, 'decimetre' radio link connection was established between Sørvågen – Værøy and Sørvågen – Røst. It used Telefunken equipment left behind by the Germans after WW II.

In 1894 Marconi had begun to study the works of *Heinrich Hertz*[3)]. He started experiments on the application of Hertzian waves to the transmission and reception of messages over a distance, without wires, late the same year at the Villa Griffone at Pontecchio, Bologna, Italy; the family home.

He progressively increased the distance for transmission and reception of signals: across a room, down the length of a corridor, from the house and then into fields. In the early summer of 1895 Marconi achieved signal transmission and reception over a distance of about 2 km, even when the line of sight was blocked by a hill. As long as visual contact was maintained, the waving of a handkerchief indicated success. When that was not possible anymore, a gun had to be fired.

In 1897 he founded his own company, the "Wireless Telegraph and Signal Company"[4)], from which he risked £ 50,000[5)] for the 1901 experiment [3].

In 1909 Marconi, together with Karl Ferdinand Braun was awarded the Nobel Prize in Physics "in recognition of their contributions to the development of wireless telegraphy" [5]. Many people were surprised and even disappointed that Marconi had to share the prize with his main competitor. Braun was the inventor of the cathode ray tube in 1897, and he was also the founder of Telefunken.

Lofoten 1906

In 1861, nine fishing villages had been connected to each other by telegraph lines during the Lofoten fishing season, the outermost being Sørvågen. In 1868, the Lofoten Line was linked up to the main Norwegian network, and from 1873, Sørvågen Telegraph Station was open all year round [6]. The Lofoten island of Røst was a centre point for the cod-fisheries and a cable, if it were to be laid across Moskenesstraumen, was estimated to cost NOK 100,000 [7].

Already in 1899 the Telegraph Administration had sent an engineer[6)] to study Marconi's system, motivated by the communication needs of the large cod-fishing activity, which created a trade of substantial national value. The fishing ports and islands were hard to reach without laying expensive cables, thus the new wireless came up as a very attractive solution. At this time, however, the evaluation was that

3) *In 1887, Heinrich Hertz (1857–1894) did the first experiments to prove Maxwell's theories right with the discoveries of the radio waves.*

4) *In 1900, the company became Marconi's Wireless Telegraph Company. The company later grew into a global communications and information technology company which was renamed Marconi plc in 1999.*

5) *£ 50,000 in 1901 corresponds roughly to £ 2.75 million today.*

6) *Principal engineer (avdelingsingeniør) Rødland [4].*

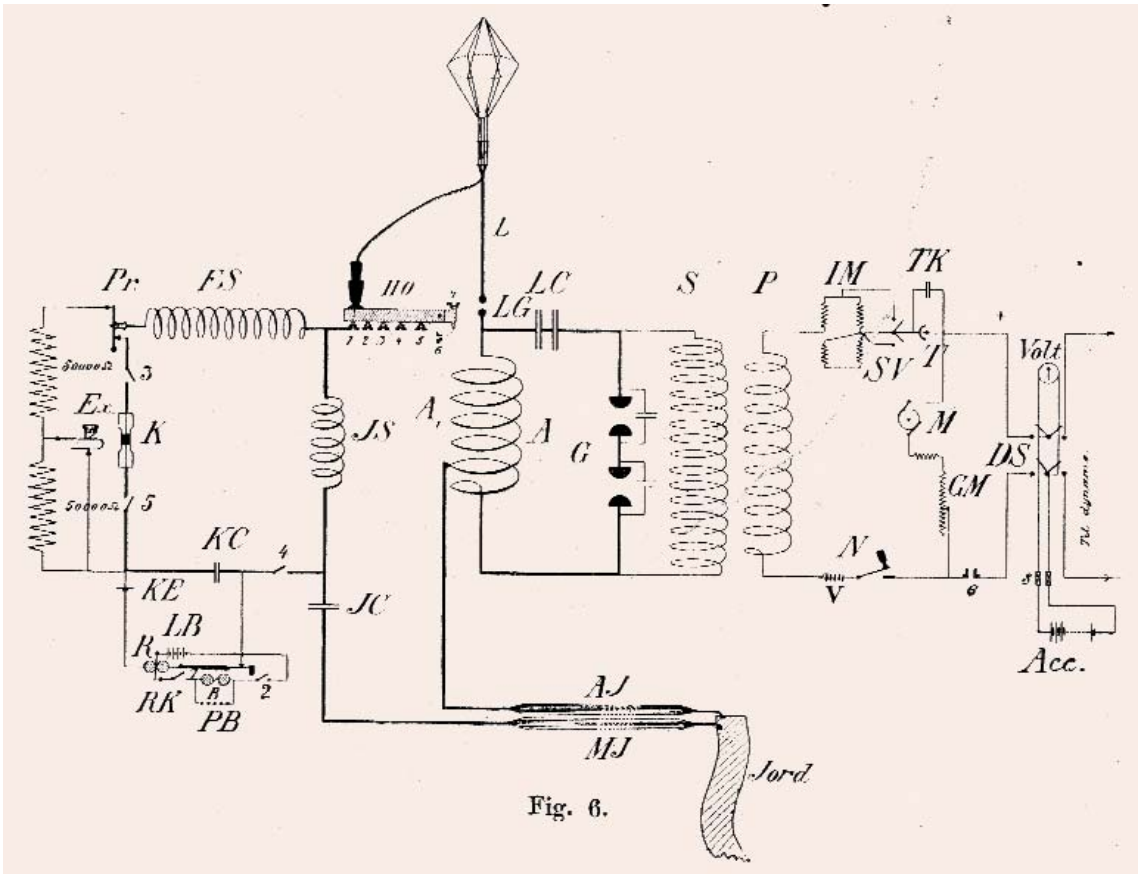
The picture shows Sørvågen Radio in 1906. Originally, both Sørvågen and Røst had 50 m radio-towers made of spliced spruce. It was secured by 28 wires in seven groups. The masts were approximately 40 m from the telegraph building. The upper part of the mast in Sørvågen was destroyed by storms several times, and the height was reduced to 40 m. (Photo from T. Rafto [4])

the available equipment was not satisfactory, especially when it came to reliability.

The first Norwegian experiments with wireless telegraph were carried out during the summer of 1901 by the Norwegian Navy, and the results were particularly rewarding. The equipment had improved significantly after Marconi started to face competition. In 1905, the Norwegian Navy established coastal radio stations at Tjøme and Flekkerøy. These were later taken over by the Telegraph Administration in 1910.

The same year as the Navy did their experiments, 1901, the Telegraph Administration sent Hermod Petersen up north to make inspections concerning the possibility of establishing wireless telegraphy connections to the fishing ports of Røst, Værøy, Træna and Grip. His examinations gave positive results for all cases. On March 21, 1902, the National Assembly granted the sum of NOK 15,000 to the Telegraph Administration to perform tests between the Lofoten fishing ports of Røst, Værøy and Sørvågen. Several companies made their equipment available for the tests, however not Marconi's company. The tests were performed during the autumn 1903. The tests confirmed the feasibility and that the German equipment from "Gesellschaft für drahtlose Telegraphie, system *Telefunken*", Marconi's main competitor (see above), satisfied the demands.

ISSN 0085-7130



The wireless telegraph equipment used in the Sørvågen – Røst connection established in 1906 is comprehensively described in 'Technical Information' the same year. This is a simplified diagram of the complete equipment, both transmitter and receiver (Illustration from 'Technical Information', issue 2, 1906, p 30)

The Sørvågen – Røst system operated at a wavelength of 520 m (approx. 580 kHz). The experiences so far had in general concluded that the 'short wave' was more advantageous; however Petersen concluded that on the connection between Sørvågen and Røst, the 'long wave' was better, supposedly due to the influence of the Lofoten Mountains.

At this time, the Telegraph Administration decided that the time had come to establish the connection between Sørvågen and Røst. Værøy was to be included after some experience had been collected. However, the local authorities at Værøy and Røst did not want to give a guarantee to the Telegraph Administration for "free services", and the construction was for the time given up. In 1905 the National Assembly granted the sum of NOK 22,000 for the establishing, provided the local authorities paid their share. Soon after, the county administration in Nordland took over the guarantee, and the case was closed.

The equipment was ordered from Germany and they came to start the installations on December 8, 1905.

The installation was finished on February 26, 1906 and opened for test the following day. On May 1, 1906, the connection between Sørvågen and Røst was opened, a distance of 59 km. The first North European permanent wireless telegraph also became the *second* wireless telegraph connection in the world that was permanently connected to the national telegraph network.

Hermod Petersen actually remarks that it is the second in the world ("As far as it is known") and that the world's *first* wireless telegraph line was opened in the spring 1905 between Saint Cataldo near Bari in Italy and Antivari[7)] in Montenegro, a distance slightly more than 200 km. It is interesting to note that this is also printed in a short notice already in volume 5, 1904, an indication that the sheets were both edited and printed with some delay. Since the Sørvågen-Røst article appears already in issue 2, 1906, this could not have been edited nor printed until the end of the year.

7) *Antivari is now the town of Bar, south of Dubrovnik.*

A short biography of Petersen is given in an article by Henrik Jørgensen in *Telektronikk*, issue 4, 1995. Petersen was a radio expert of international repute and in addition to his pioneering work at the Røst – Sørvågen connection, he was also the head of installation when the world's first Arctic radio station opened in Spitsbergen, Svalbard, in 1911. In 1917 and 1918, Petersen writes a series of 12 articles as a tutorial on the wireless telegraph where theories and principles are described comprehensively. This fills the volumes of *Technical Information* for a year.

Wireless telegraphy based on Morse was operated way up into the 1960s when these were migrated into telex systems, but telephone had to enter the wireless area sooner or later. In Norway, the first permanent radio telephone connection was opened in 1920 between Grip and Kristiansund.

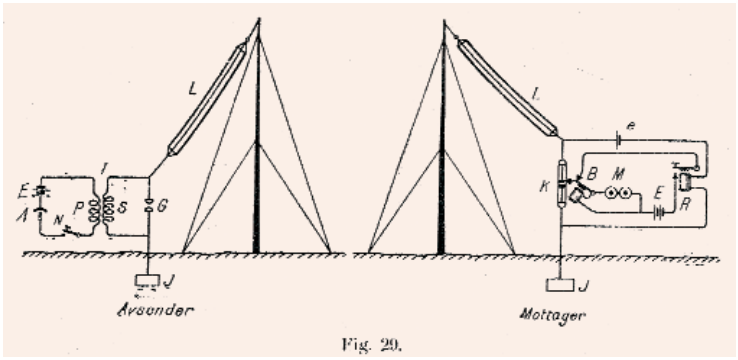


Fig. 29.

Marconi system from 1917 (Illustration from 'Technical Information', issue 11, 1917, p77)

References

1 Lehne, P H. Enlightening 100 years of telecom history – From 'Technical Information' in 1904 to 'Telektronikk' in 2004. *Telektronikk*, 100 (3), 3–9, 2004 (This issue.)

2 Bedi, J E. Guglielmo Marconi. In: *The Froehlich/ Kent Encyclopedia of Telecommunications*. NY. USA, Marcel Dekker, 10, 361–368, 1995.

3 *Marconi Calling – Interactive web museum*. September 22, 2004 [online] – URL: http://www.marconicalling.com

4 Rafto, T. *The history of the Norwegian State Telegraph 1855–1955 (Telegrafverkets historie 1855–1955)*. Oslo, Telegrafverket, 1955.

5 *Nobelprize.org – The Nobel Prize in Physics 1909*. September 22, 2004 [online] – URL: http://nobelprize.org/physics/laureates/1909/index.html

6 *Norwegian Telecom Museum, Sørvågen (Norsk Telemuseum, Sørvågen)*. September 27, 2004. [online] – URL: http://www.lofoten-info.no/telemuseum/

7 *Norwegian Telecom Museum (Norsk Telemuseum)*. September 22, 2004. [online] – URL: http://www.norsktele.museum.no/index_ieoff.html

Sources from Telektronikk's archives

Petersen, H. The wireless telegraph installation Røst – Sørvågen (Det trådløse telegrafanlæg Røst – Sørvågen). *Technical Information*, 2 (2), 23–39, 1906.

Petersen, H. Radio telegraphy (Radiotelegrafi). *Technical Information*, 13 (8–12), 57–90, 1917 and 14 (1–8), 1–61, 1918.

Birkeland, A M. The radio telephony station at Grip (Grip Radio-Telefonstasjon). *Technical Information*, 15 (9–11), 69–96. 1919.

Other sources

Simons, R W. Guglielmo Marconi and Early Systems of Wireless Communications. *GEC review*, 11 (1), 37–55, 1996.

For a presentation of the author, turn to page 9.

Terms and acronyms

2G Second Generation (mobile network)	Refers to the family of digital cellular telephone systems standardised in the 1980s and introduced in the 1990s. They introduced digital technology and carry both voice and data conversation. CDMA, TDMA and GSM are examples of 2G mobile networks [1].
3G Third Generation (mobile network)	The generic term for the next generation of wireless mobile communications networks supporting enhanced services like multimedia and video. Most commonly, 3G networks are discussed as graceful enhancements of 2G cellular standards, like e.g. GSM. The enhancements include larger bandwidth, more sophisticated compression techniques, and the inclusion of in-building systems. 3G networks will carry data at 144 kb/s, or up to 2 Mb/s from fixed locations. 3G will standardize mutually incompatible standards: UMTS FDD and TDD, CDMA2000, TD-CDMA [1].
4G Fourth Generation (mobile network)	A term used for future wireless systems providing broadband mobile access, with high mobility and bit rates up to 100 Mb/s.
ARQ Automatic Repeat reQuest	A standard method of checking transmitted data used on high-speed data communications systems. The sender encodes an error-detection field based on the contents of the message. The receiver recalculates the check field and compares it with the received one. If they match an 'ACK' (acknowledgement) is transmitted to the sender. If they do not match, a 'NAK' (negative acknowledgement) is returned, and the sender retransmits the message [1].
ATM Asynchronous Transfer Mode	A high bandwidth, low-delay, connection-oriented, packet-like switching and multiplexing technique. ATM allocates bandwidth on demand, making it suitable for high-speed connections of voice, data and video services. Access speeds are up to 622 Mb/s and backbone networks currently operate at speeds as high as 2.5 Gb/s. Standardised by ITU-T [1].
Availability	Dependability with respect to readiness for usage; Measure of correct service delivery with respect to the alternation between correct and incorrect service [2].
AWGN Additive White Gaussian Noise	Usually used to denote thermal noise in a communications system. The noise has unlimited bandwidth with constant noise spectral density (white). The noise is uncorrelated with the useful signal, and can be added on a power basis to determine resulting noise + signal power. The amplitude of the noise is stochastic, Gaussian distributed.
B3G Beyond 3rd Generation	A term used for the future wireless systems providing broadband mobile access, with high mobility and bit rates up to 100 Mb/s. It differs from the 4G term in that B3G is mainly used to describe a future integration of a multitude of wireless standards.
BER Bit Error Rate	The ratio of error bits to the total number of bits transmitted.
CCIR Comité Consultatif International des Radio Communications	The International Radio Consultative Committee of ITU, called ITU-R as from 1992. http://www.itu.int
CDMA Code Division Multiple Access	A digital, spread spectrum, packet based access technique generally used in radio systems. CDMA is used in certain cellular systems, like e.g. IS-95, and is also in 3G systems like UMTS

CELTIC
Cooperation for a European sustained Leadership in Telecommunications

CELTIC is a five year EUREKA cluster project, which started work in November 2003. The initiative is supported by most of the major European players in communication technologies. The main goal of CELTIC is to maintain European competitiveness in telecommunications through collaborative R&D. CELTIC projects are characterised by a holistic approach to telecoms networks, applications, and services. CELTIC is open to any kind of project participants from all EUREKA countries.
http://www.celtic-initiative.org/

CEPT
Conférence Européenne des Administrations des Postes et des Télécommunications

The *European Conference of Postal and Telecommunications Administrations* – CEPT – was established in 1959 by 19 countries, which expanded to 26 during its first ten years. Original members were the incumbent monopoly-holding postal and telecommunications administrations. CEPT's activities included co-operation on commercial, operational, regulatory and technical standardisation issues.
In 1988 CEPT created ETSI, The European Telecommunications Standards Institute, into which all its telecommunication standardisation activities were transferred.
In 1992 the postal and telecommunications operators created their own organisations, Post Europe and ETNO respectively. In conjunction with the European policy of separating postal and telecommunications operations from policy-making and regulatory functions, CEPT became a body of policy-makers and regulators. At the same time, Central and Eastern European Countries became eligible for membership of CEPT. With its 45 members CEPT now covers almost the entire geographical area of Europe. http://www.cept.org/

Confidentiality

Absence of unauthorized disclosure of information [3].

CORBA
Common Object Request Broker Architecture

A standard defined by the Common Object Group. It is a framework that provides interoperability between objects built in different programming languages, running on different physical machines perhaps on different networks. CORBA specifies an Interface Definition Language, and API (Application Programming Interface) that allows client/server interaction with the ORB (Object Request Broker).

COTS
Commercial Off The Shelf

A product that is used "as-is". COTS products are designed to be easily installed and to interoperate with existing system components.

CRC
Cyclic Redundancy Check

A process used to check the integrity of a block of data. A CRC character is generated at the transmission ends. Its value depends on the hexadecimal value of the number of binary 'ones' in the data block.

CSI
Channel State Information

Near instantaneous information about a transmission channel's properties. Mostly used in radio communications, especially applied to smart antennas and MIMO systems. CSI may contain data about e.g. signal loss, frequency selective variations and multipath conditions.

CSNR
Channel Signal-to-Noise Ratio

The ratio between signal level and noise level on a transmission channel.

DAB
Digital Audio Broadcasting

Terrestrial digital radio system, also called T-DAB developed by the EUREKA 147 Project in the period 1986–1993. It offers near CD-quality sound, more stations and additional radio and data services. Standardized by ETSI and recognized by the ITU in 1994. Regular broadcast in several European countries. Operates in the frequency bands 47–68 MHz, 87.5–108 MHz, 174–240 MHz and 1452– 1479.5 MHz. First transmissions from 1995. Intended to replace conventional FM broadcast.
http://www.worlddab.org,
http://www.eurekadab.org

dB
Decibel

One tenth of a Bel, where 'Bel' refers to Alexander Graham Bell. A unit of measure of signal strength, usually the relation between a transmitted signal and a standard signal source. Decibel was invented by the Bell System to express gain or loss in telephone transmission systems. Decibel is a logarithmic unit, which defines the level of gain or loss in signal strength in a circuit or link, or device compared to a reference value.

$$[dB] = 10 \log_{10} p / p_{ref}$$

Where p denotes the actual power level, and p_{ref} denotes the reference power.

Dependability
The ability to deliver a service that can justifiably be trusted. The service delivered by a system is its behaviour, as it is perceived by its user(s); a user is another system (physical, human) that interacts with the former at the service interface. The function of a system is what the system is intended for, and is described by the system specification [2].

DFT
Discrete Fourier Transform
Discrete representation of the Fourier transform. The Fourier transform is a linear transformation which can be used to convert signals and signal representations between the time- and frequency domain by the use of harmonic functions. The function used in F is the complex exponential function, $e^{j2\pi f}$.

DVB
Digital Video Broadcasting
An international digital broadcast standard for TV, audio and data. DVB can be broadcast via satellite, cable or terrestrial systems. It has been initially used in Europe and the Far East. http://www.dvb.org/

DVB-T
Digital Video Broadcasting – Terrestrial
The terrestrial version of DVB.

ECN
Explicit Congestion Notification
A congestion avoidance scheme that uses marking packets instead of dropping them in the case of incipient congestion. The receivers of marked packets should return the information about marked packets to the senders, and the senders should decrease their transmit rate. To avoid heavy congestion, routers mark packets with probability depending on an average queue length. IETF RFC 3186. http://www.ietf.org

EGPRS
Enhanced General Packet Radio Service
GPRS is an enhancement to the GSM mobile communications system that supports data packets. GPRS enables continuous flows of IP data packets over the system for such applications as Web browsing and file transfer. *Enhanced* GPRS (General Packet Radio Service) uses the modulation technique 8PSK (8 Phase Shift Keying) to increase the achievable user data rate for operation on EDGE networks. http://www.etsi.org

ELAB
Elektronikklaboratoriet ved Norges Tekniske Høgskole
The Electronics Laboratory at The Norwegian Institute of Technology. Research Institute. Now part of SINTEF. http://www.sintef.no

End-to-end reestablishment
Reestablishment of information flows in a network after a failure by establishing new end-to-end virtual or physical paths, connections, or routing schemes for (at least) the affected flows. Also denoted “End-to-end restoration” and “Global repair”.

ERMES
European Radio Message System
ETSI standard for European wide area paging operating in the VHF band. Ready in 1993, but was not deployed in any substantial manner due to the success of GSM SMS.

Error
Part of a system state which is liable to lead to failure; manifestation of a fault in a system [2].

ESS 1
Electronic Switching System no. 1

ETSI
European Telecommunications Standards Institute
A non-profit membership organization founded in 1988. The aim is to produce telecommunications standards to be used throughout Europe. The efforts are co-ordinated with the ITU. Membership is open to any European organization proving an interest in promoting European standards. It was e.g. responsible for the making of the GSM standard. The headquarters are situated in Sophia Antipolis, France. http://www.etsi.org

Failure
Deviation of the delivered service from compliance with the specification; Transition from correct service delivery to incorrect service delivery [2].

Fault
Adjudged or hypothesized cause of an error; Error cause which is intended to be tolerated or avoided; Consequence for a system of the failure of another system which has interacted or is interacting with the considered system [2].

FEC
Forward Error Correction

A technique of error detection and correction in which a transmitting host computer includes a number of redundant bits in the payload (data field) of a block or frame of data. The receiving device uses the extra bits to detect, isolate and correct any errors created in transmission.

FFT
Fast Fourier Transform

FFTs are fast DFT Algorithms.

Flooding

Self-healing achieved by flooding the network with requests in a search for available capacity to replace that of a failed link.

Gb/s
Gigabits per second

Binary term denoting digital data rate of $1{,}024^3 = 1{,}073{,}741{,}824$ bits per second.

GMDSS
The Global Maritime Distress and Safety System

The GMDSS became fully effective on 1 February 1999 and is a worldwide network of automated emergency communications for ships at sea. All oceangoing passenger ships and cargo ships of 300 gross tonnage and upwards must be equipped with radio equipment that conforms to international standards as set out in the system. The basic concept is that search and rescue authorities ashore, as well as shipping in the immediate vicinity of the ship in distress, will be rapidly alerted through satellite and terrestrial communication techniques to a distress incident so that they can assist in a co-ordinated SAR (Search And Rescue) operation with the minimum of delay. http://www.imo.org

GMPLS
Generalized Multi-Protocol Label Switching

Generalized Multi Protocol Label Switching is applied to MPLS signalling that is used not only to support packet based paths but other technologies such as Optical MUX (Multiplexer), ATM and Frame Relay switches.

GPRS
General Packet Radio Service

An enhancement to the GSM mobile communications system that supports data packets. GPRS enables continuous flows of IP data packets over the system for such applications as Web browsing and file transfer. Supports up to 160 kb/s gross transfer rate. Practical rates are from 12–48 kb/s. http://www.etsi.org

GSM
Global System for Mobile Communications

A digital cellular phone technology that is the predominant system in Europe, but is also used around the world. Developed in the 1980s by CEPT and ETSI. GSM was first deployed in seven European countries in 1992. It is operating in the 900 MHz and 1.8 GHz bands in Europe and the 1.9 GHz PCS band in North America. GSM defines the entire cellular system, from the air interface to the network nodes and protocols. As of January 2004, there were more than 1 billion GSM users in more than 200 countries worldwide. http://www.gsmworld.com/, http://www.etsi.org

HDR
High Data Rate

HPA
High Power Amplifier

HW
Hardware

IA5
International Alphabet no 5

Now International Reference Alphabet No. 5 (IRA5) and previously International Alphabet No. 5 (ISO 646) is defined in ITU-T T.50 and is a 7-bit character code. There are multiple national versions. http://www.iso.org, http://www.itu.int

ICT
Information and Communication Technologies

IDFT
Inverse Discrete Fourier Transform

The inverse linear transformation of the Discrete Fourier Transform (DFT).

IFFT
Inverse Fast Fourier Transform

The inverse linear transformation of the Fast Fourier Transform (FFT).

IFIP
International Federation for Information Processing

A non-governmental, non-profit umbrella organization for national societies working in the field of information processing. It was established in 1960 under the auspices of UNESCO as an aftermath of the first World Computer Congress held in Paris in 1959. Today, IFIP has several types of Members and maintains friendly connections to specialized agencies of the UN system and non-governmental organizations. Technical work, which is the heart of IFIP's activity, is managed by a series of Technical Committees and Working Groups. http://www.ifip.or.at

IMO
The International Maritime Organization

Established in 1948 at an international conference in Geneva.The original name was the Inter-Governmental Maritime Consultative Organization, or IMCO, but the name was changed in 1982 to IMO. The IMO Convention entered into force in 1958 and the new Organization met for the first time the following year. The purposes of the Organization, as summarized by Article 1(a) of the Convention, are "to provide machinery for cooperation among Governments in the field of governmental regulation and practices relating to technical matters of all kinds affecting shipping engaged in international trade; to encourage and facilitate the general adoption of the highest practicable standards in matters concerning maritime safety, efficiency of navigation and prevention and control of marine pollution from ships". The Organization is also empowered to deal with administrative and legal matters related to these purposes. http://www.imo.org

IMSO
The International Mobile Satellite Organization

The organizational part of former INMARSAT.

IMT-2000
International Mobile Telecommunications 2000

The global standard for third generation (3G) wireless communications, defined by a set of interdependent ITU Recommendations. IMT-2000 provides a framework for worldwide wireless access by linking the diverse systems of terrestrial and/or satellite based networks. It will exploit the potential synergy between digital mobile telecommunications technologies and systems for fixed and mobile wireless access systems. http://www.itu.int

INMARSAT
The International Maritime Satellite Organization

The international organization formed in 1979 responsible for the international maritime satellite communications system. Changed name to IMSO in 1994, and the operations were privatized to Inmarsat Ltd in 1999. http://www.inmarsat.com

Integrity

Absence of improper system state alterations [3].

INTELSAT
The International Satellite Organization

Established August 20, 1964 on the basis of agreements signed by governments and operating entities. It established the first commercial global satellite communications system and changed the way the world connects. http://www.intelsat.com

IP
Internet Protocol

A protocol for communication between computers, used as a standard for transmitting data over networks and as the basis for standard Internet protocols. http://www.ietf.org

IPO
Initial Public Offering

The first sale of stock by a company to the public.

ISDN
Integrated Services Digital Network

A digital telecommunications network that provides end-to-end digital connectivity to support a wide range of services, including voice and non-voice services, to which users have access by a limited set of standard multipurpose user-network interfaces. The user is offered one or more 64 kb/s channels. http://www.itu.int

ISI
Inter-Symbol Interference

The interference between adjacent pulses of a transmitted code.

ITU
The International Telecommunication Union

On May 17, 1865, the first International Telegraph Convention was signed in Paris by the 20 founding members, and the *International Telegraph Union* (ITU) was established to facilitate subsequent amendments to this initial agreement. It changed name to the International Telecommunications Union in 1934. From 1948 a UN body with approx. 200 member countries. It is the top forum for discussion and management of technical and administrative aspects of international telecommunications. http://www.itu.int

Jgroup/ARM
Java Group/ Autonomous Replication Management

A specific Java based group communication system with Autonomous Replication Management.

LAN
Local Area Network

A network shared by communicating devices, usually in a small geographical area.

Layering

System architecture approach where the total functionality of the system is divided into layers, each layer providing a well defined subset of the functionalities and where a layer uses the services from the layer(s) below. E.g. ISO-OSI model and transport networks. Typically, lower layers deal with issues of the physical environment, higher layers provide functions with increasing abstraction, and the system's services are provided by the top layer.

LDPC
Low Density Parity Check

Binary error correcting codes defined in terms of low density parity check matrices. Defined by R. Gallagher in 1962, and rediscovered in 1996 by D. MacKay and R. Neal. Current interest is in the use in wireless communications and inter-planetary communication.

LOS
Line of Sight

This term is often associated with radio transmission systems indicating there is a clear path between the transmitter and receiver.

MAC
Medium Access Control

The lower of the two sub layers of the Data Link Layer. In general terms, MAC handles access to a shared medium, and can be found within many different technologies. For example, MAC methodologies are employed within Ethernet, GPRS, and UMTS

MAD
Mobility, Adaptivity and Dependability

Maintainability

A system's ability to undergo repairs and modifications [3]. Measure of the time needed to rectify failures for a given O&M environment.

MAP
Mobile Application Part

A protocol that enables real time communication between nodes in a mobile cellular network. A typical usage of the MAP protocol would be for the transfer of location information from the VLR (Visitor Location Register) to the HLR (Home Location Register).

MARISAT
Maritime Communications Satellite

A pre-INMARSAT maritime satellite system, established by the USA in 1976. Now part of Boeing Satellite Systems Inc. http://www.boeing.com/satellite

Mb/s
Megabits per second

Binary term denoting digital data rate of $1{,}024^2 = 1{,}048{,}576$ bits per second.

MBMS
Multimedia Broadcast/ Multicast Service

A broadcast/multicast service defined for UMTS.

MIMO
Multiple Input Multiple Output

An arbitrary wireless communication system in which the transmitting end as well as the receiving end is equipped with multiple antenna elements.

MPLS
Multi-Protocol Label Switching

MPLS has been developed to speed up the transmission of IP based communications over ATM networks. The system works by adding a much smaller "label" to a stream of IP datagrams allowing "MPLS" enabled ATM switches to examine and switch the packet much faster.

MSC
Mobile Switching Centre

A Mobile Switching Centre is a telecommunication switch or exchange within a cellular network architecture which is capable of interworking with location databases.

NEC
Nippon Electrical Company

Japanese electronics company, manufacturer of i.a. satellite earth stations. http://www.nec.com/

NTA
Norwegian Telecommunications Administration

Norwegian 'Teledirektoratet'. The name and organization of Televerket's (Telenor's) administrative headquarters in Oslo 1969–1994. http://www.telenor.com

NTNF
The Royal Norwegian Council for Scientific and Industrial Research

The predecessor of the Norwegian Research Council (NFR) (1946–93). http://www.forskningsradet.no/

O&M
Operation and Maintenance

The processes and functions used in managing a network or element within a network.

OFDM
Orthogonal Frequency Division Multiplexing

A spread spectrum technique that distributes the data over a large number of carriers spaced apart at precise frequencies. This spacing provides the "orthogonality" in this technique, which prevents the demodulators from seeing frequencies other than their own. The benefits of OFDM are high spectral efficiency, resiliency to RF interference, and lower multi-path distortion. This is useful because in a typical terrestrial wireless scenario there are multipath-channels (i.e. the transmitted signal arrives at the receiver using various paths of different lengths). Since multiple versions of the signal interfere with each other (inter symbol interference (ISI)) it becomes very hard to extract the original information. OFDM is sometimes called multi-carrier or discrete multi-tone modulation. It is the modulation technique used for digital TV in Europe, Japan and Australia. It is used in DAB, ADSL and WLAN 802.11a and g and WMAN 802.16 standards.

OSI
Open Systems Inter-connection

Refers to the 7 layer reference model developed by the ISO. The reference model breaks communication functions down into one of seven layers, each layer providing clearly defined services to adjacent layers. They are often referred to as Layer 1 through to 7:

1 Physical layer
2 Data Link Layer
3 Network layer
4 Transport layer
5 Session layer
6 Presentation layer
7 Application layer

OxS
Optical x Switching

x ∈ {C, B, P}and C: circuit, B: burst and P: packet

PAN
Personal Area Network

A group of personal devices, communicating seamlessly together. This is also referred to as a WPAN (Wireless Personal Area Network), an example being the Bluetooth™ system.

PAPR
Peak-to-Average Power Ratio

Also known as the 'crest factor'. The ratio between the peak and average power level applied to a carrier frequency.

PLMN
Public Land Mobile Network

Common notation in the 80s of a land mobile network of any category that was used to offer public services.

PMR
Private Mobile Radio

Radio communication systems used by small to medium sized groups of users.

POTS
Plain Old Telephone Service

A very general term used to describe an ordinary voice telephone service. See also PSTN.

Protection

Techniques to make networks fault tolerant where dedicated spare path is established between the end nodes of the protected path or sub-path.

PSTN
Public Switched Telephone Network

Common notation for the conventional analogue telephone network.

PTT
Post and Telecommu-nication operator/ authority

Common notation for the postal and telecom regulatory authority in a country. Previously common term for the national postal and telecom operator.

QoS
Quality of Service

The *"degree of conformance of the service delivered to a user by a provider, with an agreement between them"*. The agreement is related to the provision/delivery of this service. Defined by EURESCOM project P806 in 1999 and adopted by ITU-T in recommendation E.860 [4]. http://www.itu.int, http://www.eurescom.de

Reconfiguration
Techniques to make networks fault tolerant based upon a centralized management, where a network management system supervises and controls all network resources, and reconfigures the paths in the network to maintain the traffic flows when a network element fails.

Reliability
Dependability with respect to continuity of service. Measure of continuous correct service. Measure of time to failure [2].

Rerouting
Self healing achieved by letting the traffic handling mechanisms of the network and the user equipment re-establish individual packet flows/connections after they have been disrupted by a network failure, e.g. Internet (re)routing.

RF
Radio Frequency
Any frequency within the electromagnetic spectrum normally associated with radio wave propagation.

RLP
Radio Link Protocol
Radio link protocol of the circuit switched data services in GSM.

RSC
Recursive, Systematic, Convolutional

Rx
Receiver
The terminator of any signal on a transmission medium.

SACCH
Slow Associated Control Channel
A GSM signalling channel that provides a relatively slow signalling connection. The SACCH is associated with either a traffic or dedicated channel. The SACCH can also be used to transfer SMS (Short Message Service) messages if associated with a traffic channel.

Safety
Dependability with respect to the non occurrence of catastrophic failures. Measure of continuous delivery of either correct service or incorrect service after a benign failure. Measure of the time to catastrophic failure [2].

SC
Service Centre
The node catering for the store-and-forward and service capabilities of SMS.

SDCCH
Stand-Alone Dedicated Control Channel
This channel is used in the GSM system to provide a reliable connection for signalling and SMS (Short Message Service) messages. The SACCH (Slow Associated Control Channel) is used to support this channel.

SDH
Synchronous Digital Hierarchy
A standard technology for synchronous data transmission on optical media. It is the international equivalent of the North American SONET (Synchronous Optical Network). It is a method of transmitting digital information where the data is packed in containers that are synchronized in time enabling relatively simple modulation and demodulation at the transmitting and receiving ends. The technique is used to carry high capacity information over long distances.

Self-healing
A common denominator for several techniques to make networks fault tolerant which have distributed control and require no dedicated pre-reserved transmission capacity.

SINR
Signal-to-Interference +Noise Ratio
The power ratio between the useful signal level (C) and the sum of interference signals (I) and thermal noise level (N). Often expressed in dB.

$SINR = C / (I + N)$

SISO
Single Input Single Output
Radio transmission system including a single transmitter antenna and a single receiver antenna.

SME
Short Message Entity
Any type of device that may act as an originator or a recipient of a short message in GSM.

SMS
Short Message Service
A means by which short messages can be sent to and from digital cellular phones, pagers and other handheld devices. Alphanumeric messages of up to 160 characters can be supported [1].

SNR
Signal-to-Noise Ratio
The power ratio between the useful signal level (C) and the thermal noise level (N). Often expressed in dB.

$SNR = C / N$

SOLAS
International Convention for the Safety of Life at Sea

The SOLAS Convention in its successive forms is generally regarded as the most important of all international treaties concerning the safety of merchant ships. The first version was adopted in 1914 in response to the Titanic disaster, the second in 1929, the third in 1948 and the fourth in 1960. The 1960 Convention was adopted on 17 June 1960 and entered into force on 26 May 1965. It was the first major task for IMO after the Organization's creation and it represented a considerable step forward in modernizing regulations and in keeping pace with technical developments in the shipping industry. The main objective of the SOLAS Convention is to specify minimum standards for the construction, equipment and operation of ships, compatible with their safety. http://www.imo.org

Span re-establishment

Reestablishment of affected information flows in a network after a failure by replacing the part of a virtual or physical path or connection affected by the failure, by new sub-paths or connections. Also denoted "Span restoration" and "Local repair".

SPC
Stored Program Control

Common in all modern telephone switches. Stored software (program) controls the computer or microprocessor, which in turn controls the operation of the switch [1].

SS no 7
Signalling System no 7

A CCS (Common Channel Signalling) system defined by the ITU-T. SS7 is used in many modern telecom networks and provides a suite of protocols that enables circuit and non-circuit related information to be routed about and between networks. The main protocols include MTP (Message Transfer Part), SCCP (Signalling Connection Control Part) and ISUP (ISDN User Part). http://www.itu.int

STBC
Space-Time Block Code

Special case of Space Time Codes (STC). STC schemes use a number of code symbols equal to the number of Tx antennas in MIMO. These are generated and transmitted simultaneously, one symbol from each antenna. The symbols are generated by a *space-time encoder* such that by using an appropriate signal processing and decoding procedure at the receiver, the diversity gain and/or the coding gain is maximized. STBC uses block codes and was described by S. Alamouti in 1998.

STSK
Skandinavisk Telesatelitt-komité

The Scandinavian Tele-satellite Committee, established in 1964.

SW
Software

Term denoting code and programs not hardwired into the equipment, but giving instructions and data for a computer's central processing unit.

TCAP
Transactions Capabilities Application Part

Transaction Capabilities Application Part enables the deployment of advanced intelligence in the network by supporting non circuit related information exchange between SP (Signalling Point) using the SCCP (Signalling Connection Control Part) connectionless service.

TCP
Transport Control Protocol

Transport layer protocol defined for the Internet by Vint Cerf and Bob Kahn in 1974. A reliable octet streaming protocol used by the majority of applications on the Internet. It provides a connection-oriented, full-duplex, point-to-point service between hosts. http://www.ietf.org

TINA
The Telecommunication Information Networking Architecture

A developing standard which is intended to resolve issues of integration between TMN (Telecommunications Management Network) and IN (Intellligent Network) standards and concepts. TINA focuses on the definition and validation of an open architecture for worldwide telecom services through a flexible software architecture for both end-users and network management services. http://www.tinac.com

Tx
Transmitter

The source or generator of any signal on a transmission medium.

UHF
Ultra-High Frequencies

Notation used to denote the frequency band from 300 to 3,000 MHz

UMTS
Universal Mobile Telecommunications System

The European member of the IMT-2000 family of 3G wireless standards. UMTS supports data rates of 144 kb/s for vehicular traffic, 384 kb/s for pedestrian traffic and up to 2 Mb/s in support of in-building services. UMTS was initiated by ETSI, but is now standardised by the Third Generation Partnership Project (3GPP). http://www.3gpp.org/

UNINETT	The academic network interconnecting the Norwegian universities with the global internet.
WAN Wide Area Network	A network that provides data communications to a large number of independent users spread over a larger geographic area than that of a LAN (Local Area Network). It may consist of a number of LAN connected together.
WARC World Administrative Radio Conference	Global conferences held every two to three years by the ITU Radiocommunication Sector (ITU-R) to revise the Radio Regulations, the international treaty governing the use of the radio-frequency spectrum and the geostationary-satellite and non-geostationary-satellite orbits. In 1995 it was renamed WRC – World Radiocommunication Conference. The next conference will be in 2007. http://www.itu.int/ITU-R/conferences/wrc/index.asp
WLAN Wireless Local Area Network	This is a generic term covering a multitude of technologies providing local area networking via a radio link. Examples of WLAN technologies include Wi-Fi (Wireless Fidelity), 802.11b and 802.11a, HiperLAN, Bluetooth and IrDA (Infrared Data Association).
WWRF Wireless World Research Forum	Created in 2001 as a result of the work of the Wireless Strategic Initiative (WSI), a partly EU-funded project. Started by Alcatel, Ericsson, Nokia and Siemens. The objective of the forum is to formulate visions on strategic future research directions in the wireless field, among industry and academia, and to generate, identify, and promote research areas and technical trends for mobile and wireless system technologies. http://www.wireless-world-research.org
WWW World Wide Web	An international, virtual network based information service composed of Internet host computers that provide on-line information. Created at CERN, Geneva in 1991. http://www.w3c.org/
X.25	The ITU specifications defining packet switched services. A widely available, low speed, packet switched data service. http://www.itu.int
X.400	The ITU specifications defining message-handling services (MHS). A universal protocol for email. X.400 defines the envelope for email messages so all messages conform to a standard format. http://www.itu.int

References

1 Newton, H. *Newton's Telecom Dictionary*. San Fransisco, USA, CMP Books, 2003.

2 Laprie, J-C (ed). *Dependability: Basic Concepts and Associated Terminology*. Dependable Computing and Fault Tolerant Systems, Vol-5. Springer Verlag, 1992.

3 Avizienis, A, Laprie, J-C, Randell, B. Fundamental concepts of dependability. In: *Position Papers for the Third Information Survivability Workshop – ISW-2000*. The Institution of Electrical and Electronics Engineers, Inc, October 24–26, 2000.

4 ITU-T. *Draft new ITU-T Recommendation E.860 (formerly E.SLA) – Framework of a service level agreement*. Geneva, Switzerland, ITU, 2003.