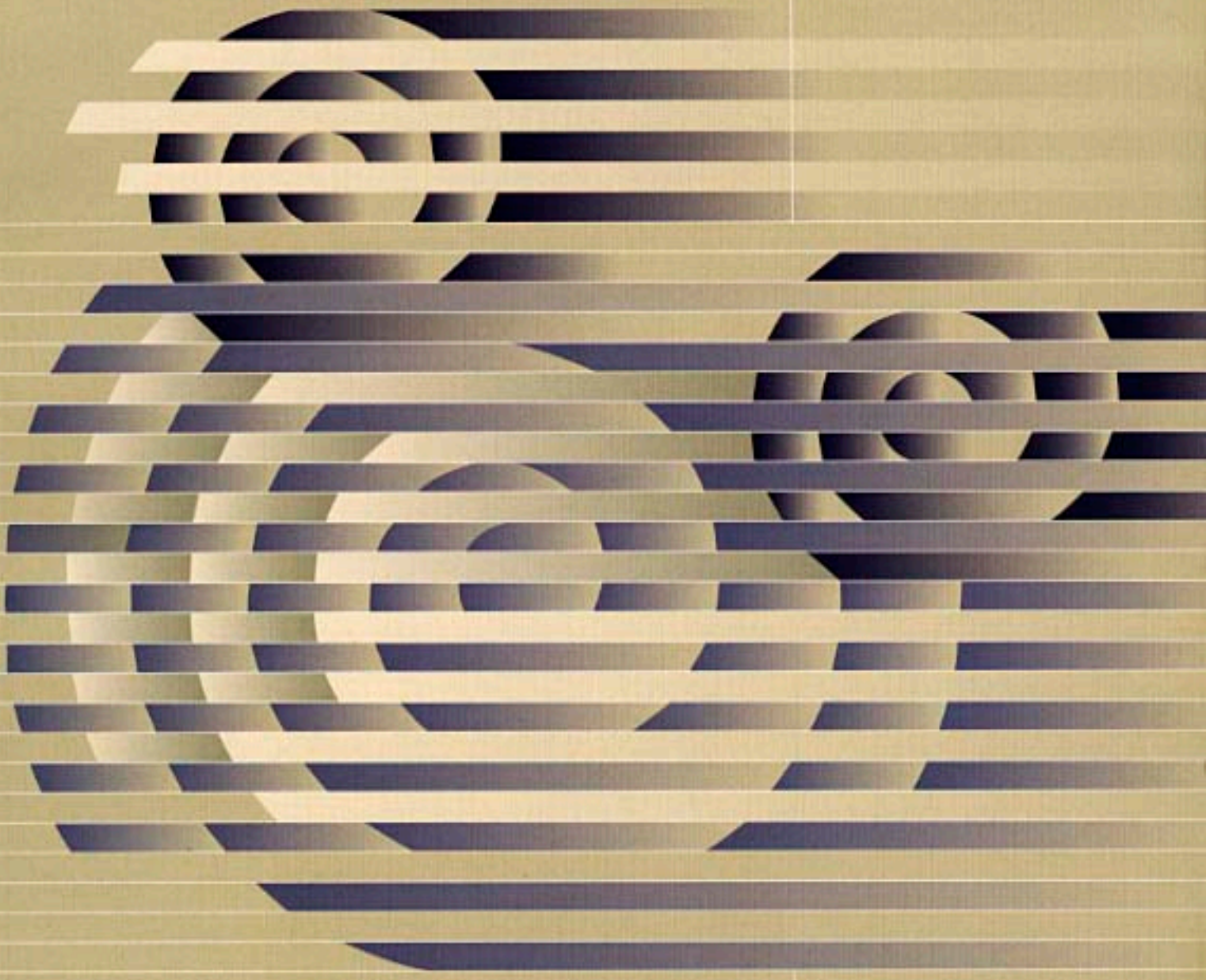
Telektronikk
3/4.2003

Network Planning



Telektronikk

Volume 99 No. 3/4 – 2003
ISSN 0085-7130

Editors:
Ola Espvik
Tel: (+47) 913 14 507
ola.espvik@telenor.com

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

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

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

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 Optimal as, Oslo

Circulation:
3,100

telenor

Contents

Network Planning

Guest Editorial

TERJE JENSEN

Terje Jensen

terje.jensen1@telenor.com

Front cover:
Network Planning

The overall basis for implementing a network plan is a global techno-economical optimal network solution created through an optimisation process. Such a global process involves a number of local disciplines and problems.

The artist Odd Andersen visualises the iterative optimisation process as a spiral reaching inwards to a desired target. In the same manner the local processes have to find their optimal solutions given global constraints. The local processes are visualised as minor spirals interacting with – and adjusting the path of the global process. This myriad of adjustments goes on until an acceptable global optimum is achieved. To illustrate the new plans' dependency on logical and physical network structures already present, Odd Andersen locates the optimisation process on top of the existing band of possibilities.

Ola Espvik, Editor in Chief

Chance favours a prepared mind!
– reflect on it ...

Spending some effort on systemising one's situation allows for a revelation of one's strengths and weaknesses. Moreover, it leads to steps to take in order to gain from "chances" by efficiently incorporating means into one's portfolio. Spotting opportunities and smoothly turning these into profitable operations is a key issue for any actor. As the market matures and the safe/traditional services and systems are deployed, new products and systems are required to maintain more profitable operation. In particular as the number of competing actors grows it is important not to lose out on the chances that turn up. Not every chance is to be seized though; only the ones that are expected to strengthen one's situation.

So how can network planning assist? First, in order to carry out the planning the *actual situation must be characterised,* hence describing the overall situation to the operator. This also outlines strengths and weaknesses. Second, strategic planning commonly applies scenario work that captures the possible future an operator may experience. The *scenarios* are characterised by a number of factors attached with *uncertainties*. These factors should therefore be particularly monitored as time evolves. Third, relating potential target states with the current situation reveals a *roadmap* describing specific events that may happen. These events may well reflect points where *decisions are to be made*. Naturally, several factors have to be evaluated when preparing for the decision, and these are also part of the network planning. Fourth, simply discussing and elaborating the plans is likely to *reveal further opportunities*. That is, gathering a number of individuals from different departments helps to form several profitable ideas.

The ultimate outcome of a planning exercise is quantified numbers related to specific network solutions. However, a *mixture of qualitative and quantitative evaluations* is included in the full-blown planning. One example of a fruitful mixture is to apply qualitative approaches for the wide set of options that might appear, while a more limited set is selected for the calculation exercise. The cases will also be taken into account when conducting a number of what-if evaluations. This also leads to the fundamental role of network planning as a basis for management decision support.

Network planning has a range of scopes, as described later in this issue. Several types of input and aspects have to be considered. In separating the time frames it is important that strategic, tactical and operational means are harmonised, which implies a need for good coordination between the commonly different groups managing the different scopes. A consequence of this is that *the strategic plans have to be connected to current operations* in order to show effects. This is one example that actions resulting from planning would be organised along a time axis and be reflected in the organisation.

Considering the operator's set of interdependent systems, products, traffic flows and customers, it is essential to have a *methodical support* of the planning exercises. Covering all combinations of options, the tasks frequently become too tedious for mere manual analyses. This also assists in assessing consequences and selecting better actions for urgent questions, like responding to competitors' product rollouts, sudden traffic increases, offers from vendors, and so forth.

More than 125 years after Alexander Graham Bell's invention in 1876, few people oppose the notion that the electronic communication means is one of the major indicators of a community's welfare. This shows the essence of having adequately operational telecommunication networks and corresponding systems in a nation. The number of people who are aware of the opportunities and are able to put the practical solutions into operation grows in importance as more candidate solutions and arrangements are faced and competition levels fluctuate. Hence, the speedy technological, market and service changes make it more challenging for a network planner to keep ahead of efficient network solutions. The convergence of telecommunications, information systems, broadcasting, user devices – partially fuelled by transition from complexity in hardware to software – requires that the network planning methodology is flexible and able to consider the holistic view and take into account future optional migration tracks.

The drivers for network planning include technology, markets, business and customer service. It is the financial turmoil in the telecom industry that has produced the latest changes. There is a

renewed interest in cost reduction and maximising use of resources which increases the need for network planning. However, this also comes at a time of great technological change with IP-based architectures and wireless access producing a great potential for new applications, and therefore additional planning challenges to be solved – and there is no sign that the pace of change is slowing.

The network planning area is too wide to be covered by a single issue. However, a number of aspects are treated in the following articles. The first article provides a *survey of network planning*, relating network planning to management/financial topics and the technical questions to be treated. Going directly to the core of the planning task a survey of between one and two *decades of implementing planning tools* is described in the subsequent article.

The next area considered places more weight on the *strategic perspective*. This is described by a survey article as well as a more in-depth description of *risk considerations*. Gone are the days with "never-failure" as the sole objective with less focus on delivery time-scales and solution cost. Risk management is now a key as technical and commercial risks are balanced against the need to meet market opportunities on time and in accordance with budget.

The essentials of markets must be understood as it profoundly affects network planning objectives. This is also documented in *articles on forecasting*, describing methodologies and results for obtaining estimates of number of customers and amount of traffic.

The final article returns to one of the central questions for backbone networks; that is, how to efficiently utilise the *capacity of transport networks*.

As stated above, a full coverage is beyond the scope of any single issue. The main objective of this issue is to provide insight into systematics and methodology applied for network planning. This also points to the need for awareness regarding adequate planning competence, defining relevant planning tasks and continuing assessment of input data and surveying the critical events revealed through the planning work. The wider scope of planning is taken on describing the need for on-going activities and links between different planning tasks. In particular, linkage between long and short term planning, as well as relations between different systems must be obeyed.

In order to provide the material in this issue of *Telektronikk*, interesting discussions with a wide range of individuals have been appreciated. The interest for the questions addressed during projects in later years clearly shows that network planning activities engage most people involved in electronic communications.

Enjoy your reading!

A Tale of a Technical Officer's Day

TERJE JENSEN

Dr. Terje Jensen (41) is Research Manager 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 on network planning, performance modelling/analyses and dimensioning.

terje.jensen1@telenor.com

Imagine being a technical officer – or maybe you are? A long list of questions are raised about how to get the highest return by smart spending of the company's money. In a broader perspective everyone involved in network evolution is faced with similar questions; how to organise the systems and system management and how to allocate resources to achieve company goals. An intuitive statement, however, is that the choices made have more dramatic effects when made at a higher level in the company. On the other hand, every function must work in a coherent way to allow for flexible, rapid and efficent service provisioning.

The following imaginary story tells about a workday of a technical officer, called TO. So, let us follow this imaginary technical officer during a day to see how various aspects of network planning could be utilised to ensure effective company operation in the shorter and longer term.

Introduction

Bleep – bleep, a message arrives on TO's mobile. The market survey group has spotted a competitor launching a new product. Immediate question: *"How do we respond to that?"* Well, our technical officer, we will call him TO for short, receives the news in a relaxed manner. Having gone through a set of possible scenarios, the one emerging has already been examined. And it turned out that the launched product does not allow sufficient profit margins compared to the ones already offered. All calculation results, internal as well as jointly with others, showed that this product will disappear from the market again – or the competitor will experience increasing loss. So a quick messaging reply was issued: *"Don't worry – stay happy – and have a look at the strategy document on scenarios point 6.4." "Well, being prepared is rarely a disadvantage,"* TO thinks as he continues his breakfast. Soon he receives a reply, *"Thanks – corresponding media action is under preparation". "Well, a nice start to a sunny day,"* TO thinks, pouring another glass of milk.

On the way to his office, TO happened to observe the cars changing lanes without signalling, causing sudden breaking and changing speeds. Like most systems involving humans or uncoordinated instances – efficiency measured on the individual level differs from that measured on the overall system level. In this case a single driver might believe that he is reaching his destination very fast, while the total effect on all the other cars on the same road may be that they are all delayed. This could well be applied to telecom systems. In fact, the inherent stochastic nature of systems, user behaviour and technical capabilities, competitors, governmental bodies, etc. further adds to the complexity of finding which steps to take to achieve company goals. Commonly these days, such goals are given as maximising growth profit. In a telecom system, however, traffic can mostly be directed more freely than what is accepted on the road.

Approaching this in a systematic manner, the questions can be arranged along a time axis. Then the number of options will likely grow with time. One pictorial model for this is to think of the future options as constructing a cone – a wider set of outcomes and actions are available in the longer time horizon (see Figure 1). On the other hand, one may also have a clearer picture of the major trends dominating the picture. Hence, by shifting the scope in the longer term, the choices to be made may be clearer on a general level when considering the technical areas. Besides, several other areas have to be taken into account even though technical decisions have to be made. Such areas include financial means, customer requests, competitors, regulatory issues, etc. Factors from these areas may well inspire additional actions relating to technological choices.

Still, although there are a lot of uncertainties a number of decisions have to be made. In order to ensure a robust company portfolio, several strategies could be pursued, although this must be a clear company choice and not simply something that happens because of lack of coordination. *"Hm, this is almost on a philosophical level,"* TO reflects.

Before Morning Coffee

Parking his car around 07.30 our technical officer TO notices that the car parking places are about to be re-arranged into a common pool. So far, department-specific places have been allocated. To TO, with his planning background, this seems like a natural step to take to increase the efficiency of the parking area; yet an example of the scale effect – loosely described as "the bigger the better". Again, some similarities between controlling vehicular traffic and telecom traffic could be recognised.

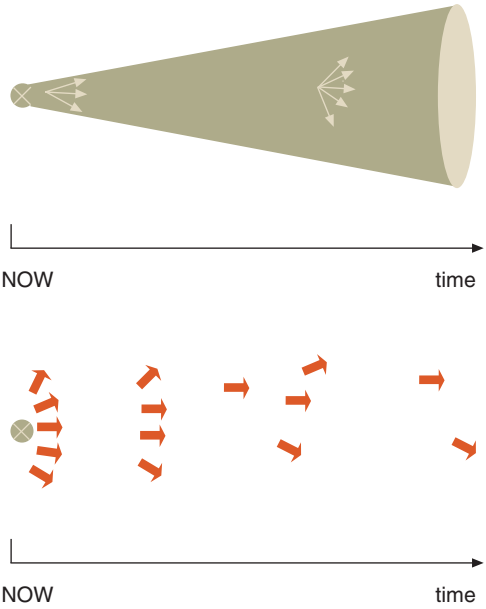


Figure 1 Depending on options selected available situations in the future may become wider (top). Commonly diverting recommendations and observations are seen in the shorter term, while longer term observations could well be more harmonised (bottom)

Leaving his car and approaching his office, TO reads today's lunch offer: "As many pancakes as you can eat for 5 €". As he is interested in inspiration for something to discuss during a lunch meeting with the secretary to the Ministry of Industry, this is something to reflect on, he thinks. In fact, tariffing models offered by the range of start-ups within the broadband area have reached the agenda of the National Assembly as part of the e-society discussions. Flat rates, their simplicity and options for cross subsidising between customer groups might be one of the subjects to bring up. Too many pancakes – gaining weight – too few carrots to provide a balanced diet, and less alert for the introduction of more services and meeting multinational competitors could be another subject to bring up.

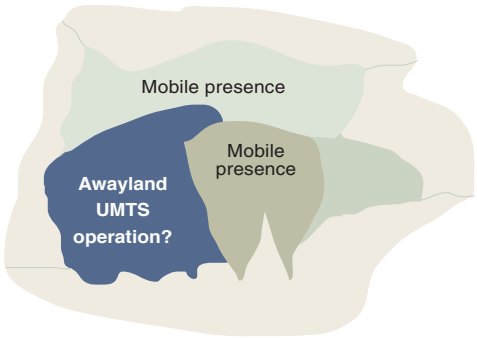


Figure 2 Awayland, indicating neighbouring countries where operations are in place or are being planned

Well, you never know where inspiration may come from ...

Going through his e-mails, a few news bulletins on mergers and network contracts attract his interest. Keeping an overview of the companies' dispositions is always a necessity in this fluctuating industry. Particularly interesting is the notice on a telecom manufacturer buying a video generator/storage company. Could this be another sign that content for broadband access will be provided in the near future? Or is it simply that the company wants to be involved in more links of the service delivery chain? So, all this goes into the memory bank for the on-going analysis of industry trends and possible use in discussions like the lunch meeting.

Before Lunch

Well, prior to earning his lunch rights some effort should be placed on the following question: Should one be engaged in UMTS operations in Awayland? The current status is that operations are running in two of Awayland's three neighbouring countries (see Figure 2). One reason for the question is that the interactions between the countries in that region seem to be steadily increasing.

So, the principal choices are: i) no engagements in Awayland, ii) application of UMTS licence for network operation, iii) establishment of a virtual operator simply offering the service to users without deploying a network infrastructure. These are the options available today. However, future options will become available depending on the choice made, see Figure 3. One of today's options is not to engage in any service offering. However, this is a decision that can be revised in the future.

When the UMTS network track is followed a second network deployment phase may be realised after two years and interactions with other wireless systems assumed to be intensified after three years. It is also an option to leave the UMTS network operation track after the second year by selling ownership of the company.

The virtual operator choice is also accompanied by future options; after two years of service offer

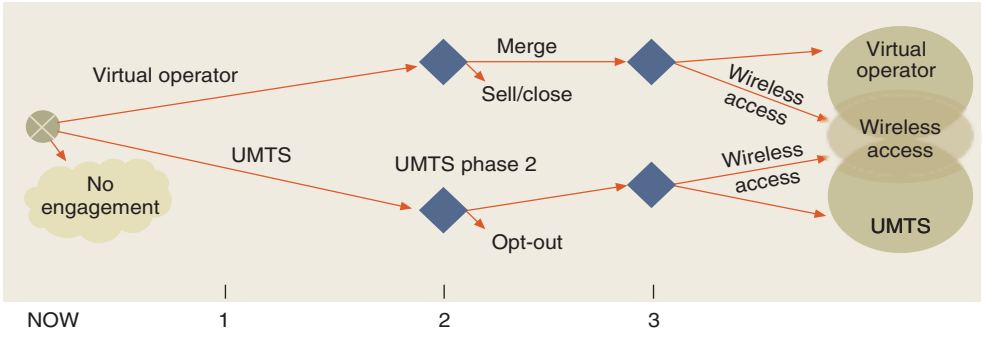


Figure 3 Options available in Awayland

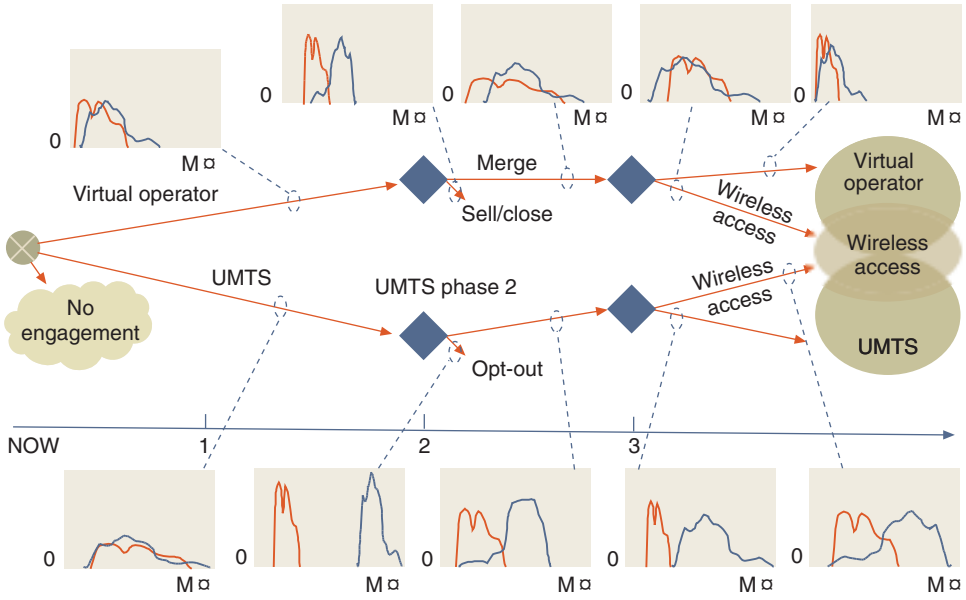


Figure 4 Illustrations of estimated cost and revenue distributions associated with each branch of the decision map
Note: distributions not in scale blue = revenue; red = cost

one would consider merging with others or sell the ownership. The same options are also seen after three years, although an additional option to enhance the operation with wireless access systems (similar to the UMTS track) is also present.

As we can see, four different combinations may happen: UMTS operation with or without wireless access, or virtual operator with or without wireless access. It is also believed that the track to be followed should be given a period of about two years before further steps are chosen. However, the actual criteria for making these decisions are rather based on market demand, competition level, price of equipment, and so forth. The current estimation shows that the chosen tactic can be followed in the next two years.

Preparing the basis for the current decision to be made, numbers should be estimated for each of the tracks – cost levels, income levels and probability estimates. Thanks to the harmonised way of preparing and describing the decision basis within the corporation, outlined in [Jens03a], TO does not have to spend any time explaining the questions to be answered in order to prepare this case to the corporate management. So basically, the task can be separated into problems to be addressed by a number of groups. However, even though the overall problem could look like a tedious assignment, efficient tools have been implemented to make most of the calculations. Moreover, the input data is readily obtained from survey activities in the company. Today's duty is to assess some more information for the UMTS track. The virtual operator track has already been covered. Meetings have also been held to estimate the likelihood levels. So most of the basis is complete, and an offer for UMTS equipment at significantly reduced prices has just been received, which has to be inserted into the calculations.

Facing this fairly simple decision map, all points and data can be inserted into the same calculation. The problem size also allows for using distributions for main variables such as demands, price levels, cost and timing of technical systems. Naturally, in several cases exact distributions are not known. However, running several combinations gives a broader foundation for making the decision.

Well, starting with the UMTS re-planning assisted by tools such as those described in [Lore03b], TO reflects on the times when no planning tools were available – every step required a lot of manual work, much longer time, and whether or not an efficient solution was identified depended a lot on the skills of the persons involved. Today it is more or less sufficient to insert the adjusted data on equipment costs. However, as always, some reflections of the results are needed to see whether a skilled eye could come up with improvements on the proposed solutions. After obtaining the new cost structure the results are inserted into the overall picture of cost/revenue for branches in the decision map, see Figure 4.

For each branch distributions of the cost and revenue are given – the amount of money is indicated along the *x*-axis and the probability level along the *y*-axis. These distributions are estimated by utilising different means such as studies on variations in demand, level of competition placing pressure on the tariffs, cost of hiring transmission capacity, and so forth. Other means include more subjective means as described in [Stor03a].

Completing these numbers and distributions, an evaluation is executed similar to the ones described for option theory in [Lore03a]. This returns that the UMTS track would give the highest return – almost three times the level

obtained if the virtual operator track were to be followed.

Although this could be run more or less mechanically, it is essential to have some additional thoughts on the case. TO's experience in this field motivates for further contemplation. Even though it is an extensive model, there are always some factors that are not fully captured. In this case there is a bill underway to the National Assembly of Awayland proposing to lift the requirement that data on the country's residents cannot be stored in a base abroad. In case this is lifted common bases can be utilised for several countries, further reducing the cost level. Another factor is that a general agreement with an equipment manufacturer may be finalised, also allowing for lower equipment and operation costs for the network operator. Both these factors are to be settled within the coming month. Settling these may eliminate some uncertainties, hence TO forwards a recommendation to the management board to postpone the decision.

On the other hand, the decision to go for a UMTS roll-out could be made based on the available results. However, there is no need to push through such a decision as there is enough lead-time in the plans to wait for about three months.

The complete documentation set together with a short note explaining the recommendation is finalised and submitted for the board's meeting next Monday. *"Well, better be prepared for what may be decided next month,"* TO thinks while swiftly going through the e-mails that have appeared during this morning's exercise.

Then, it's time for lunch …

Lunch Meeting

The small delegation from the Ministry appears in the lobby. As reservations have been made at the restaurant TO guides the group to the reserved seats. Well, perhaps one should avoid the pancakes for lunch although this morning's point could be applied. One could say that the lunch table is rigged in such a way that a screen is placed close by – "accidentally" running a slide show close to the topic to be brought up. Basically two issues are expected; firstly misconceptions regarding industry trends and globalisation, secondly the social effect of flat rates.

Without going into detail in the conversations, the pancake example comes in handy for the latter topic just as a few visual illustrations flash on the nearby screen. The documentation prepared providing examples of TO's view on these matters is also well received. Nice to have collected references from industry magazines and various statements – you never know when these can be applied.

After Lunch

One of the challenges to attack after lunch is "how to further deploy broadband services in the home market?" Broadband services have been offered for some time, but the overall corporate push in this area has not been completely harmonised. An example of the market adoption explanation comes to TO's mind, see Figure 5.

So, it is essential to have an understanding of where in the "life-cycle" the different product areas reside, as described in [Stor03a]. A major challenge, however, is estimating demands for products that have not yet been introduced or are in the initial phase. Prolonging the adoption rate illustration in Figure 5, phasing out products is an essential skill. As the presence of an operator continues, there is a tendency that more effort is placed on steadily developing new products and corresponding systems than cleansing the existing set of products and systems. TO reminds himself of the climate a few years ago – when slang expressions such as "cash in hand" and "burn rate" were uttered by several start-ups and proposed as measures of success. It seems that the companies staying on that bandwagon for too long ended up spending money on disperse activities, eventually facing quite an unsound portfolio of products and systems. TO almost laughs out loud when he remembers the situation in a neighbouring country where a company actually had six different products more or less addressing the same customer segment – and all these products were actually promoted by six

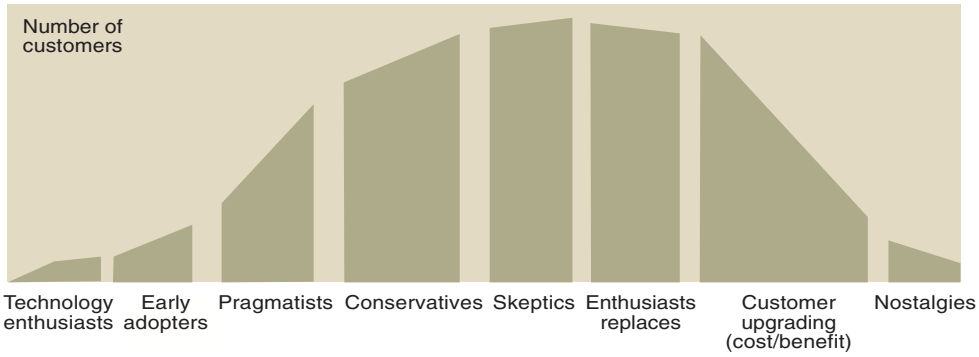


Figure 5 Illustration of adoption rate for a telecom service

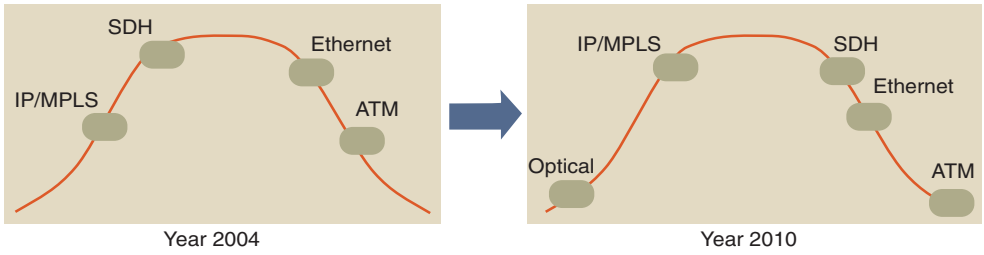


Figure 6 Imaginary example of "transport platform" life phase locations

different sales representatives holding meetings with the same customers. Well, some diversification is often good, but too much is surely not healthy in the long run.

So, returning to today's task – TO's company has been in the telecom market for some decades and a necessary product and system washout was carried out last year. The main results were plans to phase out products and systems. Besides the direct cost savings obtained when fewer systems are in operation, a dramatic increase in product rollout speed is achieved as the development and deployment procedures are harmonised. The effect on the sales side is also beginning to pay back as clear market communication strengthens reputation.

In order to make decisions on which product and system to remove from the list, a portfolio management perspective was introduced, see [Jens03a]. This allows for including several factors in the decision such as profit level and risks. Moreover, both shorter and longer terms were taken into account.

Still, while the overall plan is to reduce the number of systems and products, new ones are also considered. So today the analysis model is to be prepared for a new product group. The main question is whether or not to introduce this product group and in case the product is to be offered, how to do that.

In the overall mature market a new product would cannibalise existing products in some sense. Hence, the motivation is to avoid competitors capturing customers, improving the production efficiency (reduced costs allowing higher margins), appearing as an actor in the market forefront in accordance with trends, and so forth. Then, the evaluations must always consider the alternatives as before.

The product groups, called *Xrem*, can be supported by several candidate "transport platforms" such as SDH, IP/MPLS, Ethernet and ATM. On this level the functionality requirements are basically to be able to connect a number of points together allowing for a range of throughput, delay, jitter and dependability requirements. More advanced service features are also involved supporting custom-tailoring, multipoint configurations, self-management and so forth, but these requirements are topics for another day. During previous exercises on portfolio for the "transport platform" area, the candidates were organised in terms of which life cycle phase they were assumed to be in. Moreover, there was an overall plan describing which of these candidate systems to phase out and which belong to a core portfolio in the medium and longer terms. These were also backed by decisions of dominating vendors in this area. Assume the picture is like the one shown in Figure 6, but besides such illustrations the sizes of these systems must be considered. On the raising edge, a system's size grows. Two systems at the peak of their size, however, may be quite different in absolute size, e.g. measured by number of customers, traffic carried, number of nodes, etc. On the falling edge it must be decided whether systems should be reconfigured or simply kept without changes. This decision is strongly influenced by the total cost of the different options – that is both operational expenses and capital expenses have to be considered. Again, a risk evaluation must be present.

As *Xrem* is a new product we are faced with several challenges when making forecasts as described in [Stor03b]. Still, as elaborated in [Stor03a], it is simply a must to have numbers regarding future demands for different product groups. Naturally, different scenarios can be invented accompanied by different demand mixtures, see [Jens03b]. To some extent *Xrem* is expected to capture customers from SDH leased lines, ATM connections, Ethernet lines and IP-VPNs. A nice feature of *Xrem* is the very low cost (investments and operation/administration) that a customer has to cover. Therefore, the *Xrem* demand is expected to grow as the customers want to upgrade their current telecommunication configuration. Different sources are considered in order to estimate demand; interviews with main customers, discussions with vendors of user equipment, examination of traffic growth, financial situation in different industries, as well as background information from analysis agencies and trend descriptions.

Applying procedures as described in [Stor03b], a set of forecasts of the *Xrem* product as well as the other "competing" products is estimated, taking into account both the number of customers

and the traffic loads. So today's task is to prepare some reasoning around the forecasts. Having established a corporate approach for forecasts, TO prepares input for tomorrow's meeting on these issues. Then the forecasts are to be presented and consequences on other product groups assessed. This is a significant factor for estimating overall profit improvements of introducing the *Xrem* product. Other factors are those expressing the costs – both investments and operational expenses. Having all the various product groups within the corporation, decisions are to be made on the overall level, which is whether the overall profit will increase or decrease.

More challenges are faced because *Xrem* is a new product and limited information is available. So nearby products are looked at and the diffusion for the demand of *Xrem* may be estimated. Still, the price levels have to be taken into account arguing for assessing the product (cross) elasticities. The result is a product penetration ratio for *Xrem*. Then, factors providing the effective traffic loads must also be estimated, as described in [Stor03b]. For each area the number of customer sites, average and peak packet loads, etc. must be found in order to decide on the effective load and hence the needed capacity of network nodes and links.

Finding an optimal network design given the effective load, methods and tools as described in [Lore03b] can be applied. The result is typically locations, dimensions and handling of traffic. Several time periods could be considered, possibly running a multi-period optimisation algorithm. However, in several cases, much uncertainty is attached, and a simpler techno-economic approach could be applied (see e.g. [Stor03a], [Jens03a]) in order to provide a first assessment of the financial aspects. Treating the uncertainties more systematically, risk management procedures can be applied as described in [Lore03a] and [Jens03b]. It also needs verifying that realising the *Xrem* product still fits with the overall network strategy, complying with the set of decisions to be made. However, in case it turns out to be a strong case for realising *Xrem*, the overall network strategy should be adapted.

"Well, well," TO thinks – *"Having all this predefined and an organisation tuned into the same way of working greatly simplifies the evaluations and planning activities that need to be undertaken. Having demonstrated the swift work procedure, the top management have gained an interest in exporting this to other subsidiaries."*

So, there seems to be several options open in the future, allowing for more features and concerns to be incorporated in the work procedure. Returning his coffee cup to the dishwasher, TO thinks that better insight should be incorporated in phasing out products and networks/systems. In fact, clearer decision criteria for how and when systems should be switched off would greatly improve the evaluation procedure. As time goes, it is a lot easier to introduce new systems than to get rid of current systems. As TO takes on the overall portfolio perspective he thinks that having all the systems of various levels of success may in some cases be justified. But in most cases it is likely that too much attention from the operation and administration side is spent on less fruitful solutions. Again, it is a question of devoting more attention to the real challenges and allowing progress to be made in these matters without tedious managerial bother. For example, should one have a steady argument on when to start the dishwasher or which powder to use, the office would likely run out of clean cups. Moreover, if the dishwasher should be restarted every time someone returned with a dirty cup, the result of clean cups would also be delayed.

On his way out TO reflects on the advantages of having clear lines of duties and result handovers. Efficient network planning tasks – from strategy to operational matters – fit very well into this picture. Similar principles also apply to other areas of business – and lessons could be learned from those areas. *"Keeping one's eyes and mind open there is always more to learn,"* TO thinks; perhaps examples from the food industry could be applied in the telecom industry as well – something to bring up during tomorrow's "free style" discussions?

References

[Jens03a] Jensen, T. Network planning – introductory issues. *Telektronikk*, 99 (3/4), 9–46, 2003 (this issue).

[Jens03b] Jensen, T. Network strategy studies. *Telektronikk*, 99 (3/4), 68–98, 2003 (this issue).

[Lore03a] Lorentzen, R. Portfolio evaluation under uncertainty. *Telektronikk*, 99 (3/4), 99–109, 2003 (this issue).

[Stor03a] Stordahl, K. Forecasting – an important factor for network planning. *Telektronikk*, 99 (3/4), 110–121, 2003 (this issue).

[Stor03b] Stordahl, K. Traffic forecasting models for the incumbent based on new drivers in the market. *Telektronikk*, 99 (3/4), 122–127, 2003 (this issue).

Network Planning – Introductory Issues

TERJE JENSEN

Dr. Terje Jensen (41) is Research Manager 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 on network planning, performance modelling/analyses and dimensioning.

terje.jensen1@telenor.com

This article gives a brief introduction to a number of network planning issues. Besides several core technical areas, relations with telecom management and financial topics are also included. However, the list of issues that could be treated is quite long, and so is the vast amount of material that might be captured by the network planning title. A key message in this presentation is that network planning is multi-faceted and has relations with (almost) all other types of activities going on in a network operator. Hence, it is important to balance the system portfolio insight with skills in planning techniques as well as the financial means for comparing results.

Considering the required efficiency improvements most network operators are faced with, having an adequate network provides a crucial contribution. In fact, the systematic descriptions and treatment involved can allow for immediate improvements and reveal business opportunities.

1 Introduction

In its widest interpretation network planning has relations to every activity going on in a network operator. However, in a practical organisation, a number of clear interfaces are defined where units within the operator collaborate. This allows network planners to be concentrated on network dispositions – both in the short and long term. In order to ensure efficient networks, however, input and results have to be conveyed to other units.

As found in dictionaries a *plan* is an *arrangement for doing or using something, considered or worked out in advance*. Typically a plan is shown by a drawing/scheme. Then, to make a plan is to make preparations and hence consider something in detail and arrange it in advance. Referring to networks, a general interpretation could interact with most other tasks, although the planning task itself is focussed on how to evolve the network portfolio managed by an operator.

A fundamental gain from planning is certified resource utilisation. An effect of this is lower cost of deploying network elements, for example found by running a network optimisation program deciding where nodes should be located, nodes should be interconnected and traffic routed in order to minimise the overall cost. In the longer time scope a number of options will be looked at, including different trends of what the industry will look like, how competitors will behave, what customers will ask for, what vendors will offer, and so forth. Having a systematic description of these issues also allows for detecting new business opportunities for a network operator. That is, chances are revealed, and triggers/factors for when to go for these chances are described. Besides these upsides, other challenges can also be prepared for, such as presence of a disruptive system reducing the entrance threshold to the service-offering arena such that the power of newcomers may be overwhelming.

Relations between network planning and product development should be clear; the portfolio of networks is used as production means for delivering a set of products. This implies that features in the network portfolio give significant guidance on the products that are possible. It may also decide when a product group should be supported – or perhaps the product group should not be supported after all as the market is considered too small, other products will rapidly take over their positions, etc.

A network planner should take on the holistic perspective, considering how a portfolio of systems fulfils its overall purpose. This includes questions like:

- Which usage patterns will be seen?
- Which services should be offered?
- How will the competitors act?
- Will any regulatory changes happen?
- Which technologies should be introduced – and which should be phased out?
- What personnel competence is needed?
- Which tools should be applied to assist in the planning process?

The level of detail is strongly correlated with the time horizon; a shorter time horizon requires more concrete details and limited scope. This is illustrated in Figure 1. With a longer time horizon, less attention is given to individual tasks and more is assigned to having an integrating and holistic view.

On a technical level network planning includes designing, optimising and operating telecommunication networks. This will also assist decision-making on migration of network portfolio to maximise benefit for an operator – using a com-

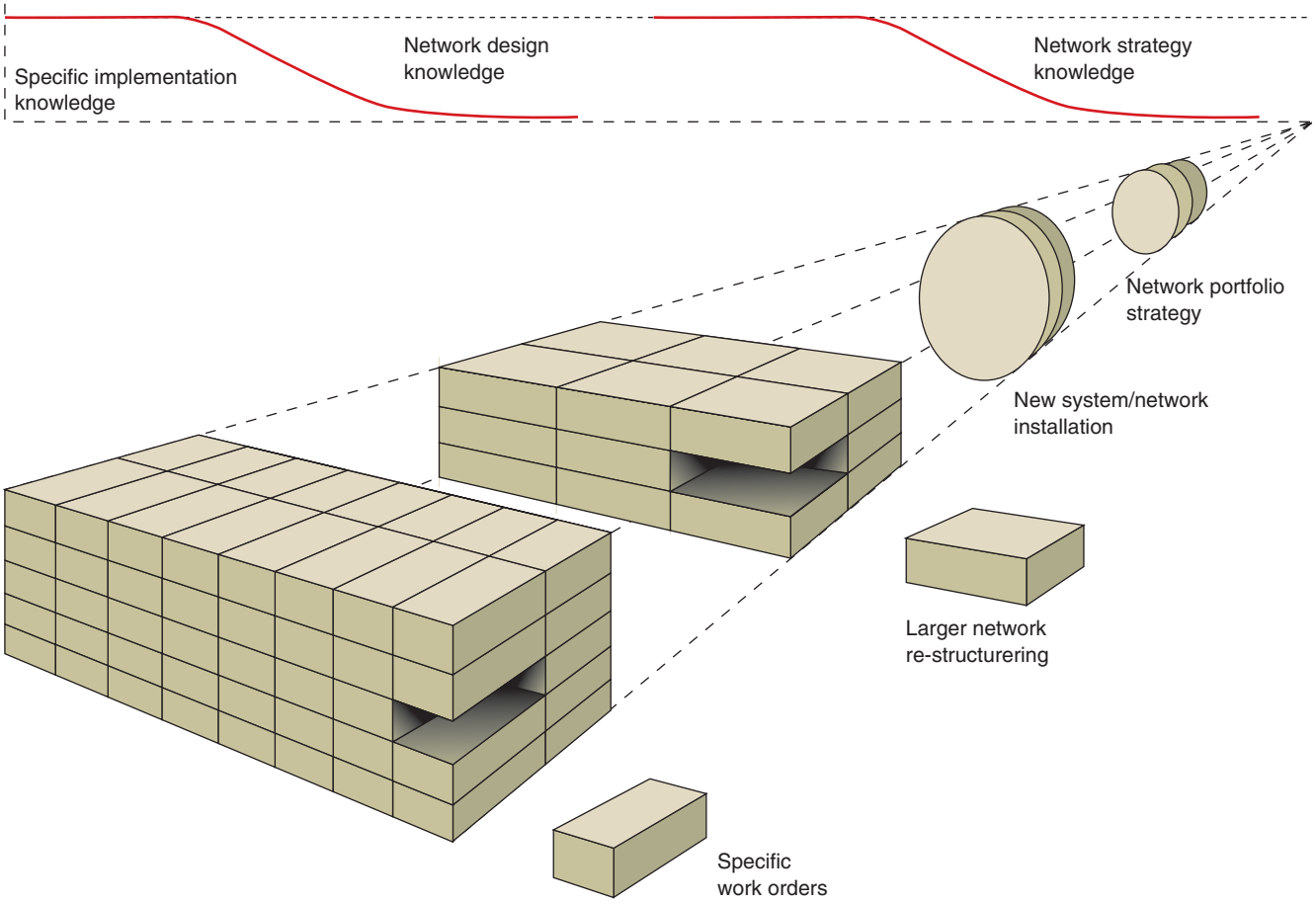


Figure 1 Different time horizons covered by network planning, relating to different level of system detail insights, design and strategies

bination of quantitative and qualitative evaluation. Commonly, quantitative evaluations are associated with the core of network planning, in particular for the short and medium term horizons. Key methods included in network planning are:

- Market and demand forecasting
- Teletraffic-related methods
- Economics engineering
- Operational research and optimisation
- Architecture and technology know-how

All these methods are combined in a complete planning exercise – incorporating shorter to longer terms (ref. Figure 1). Considering the planning scopes, in principle the complete system portfolio should be taken into account. However, some separation into segments – be it geographical, functional layer or system-dependent – is mostly carried out in order to limit the task. Still, it is important to bear in mind that the solutions recommended must fit into the overall portfolio managed by the operator. Besides finding efficient architectures, interconnection schemes, performance levels, SLA conditions, operational expenses and processes for support systems should be considered.

Having a full-fledged set of tools is essential to be able to fulfil all the planning objectives. The main resource, however, is the individual himself, drawing on his experience to choose the proper set of tools and the proper candidate descriptions. Having a capable network planner allows for i) strong arguments for future tasks, ii) contribution to development of suitable tools for future tasks, and iii) a presenter of recommended and optional evolution plans to the management for decision.

Relations between network planning and telecom management is discussed in the next section, with particular emphasis on financial issues. This is to acknowledge that most decisions are based on financial measures. Various planning scopes are handled in Section 3. Then, Section 4 gives an example related to Internet traffic engineering. A number of technical methods are treated in Section 5 – Section 15. Finally, a network planning tool from ITU-D is outlined.

2 Network Planning as Part of Telecom Management

Considering different time perspectives on network planning, there will also likely be interactions between the planning activities and management levels. A possible model is depicted in Figure 2.

A network operator, as an instance of an enterprise, would have a business plan. The business plan gives the financial indicators enabling the

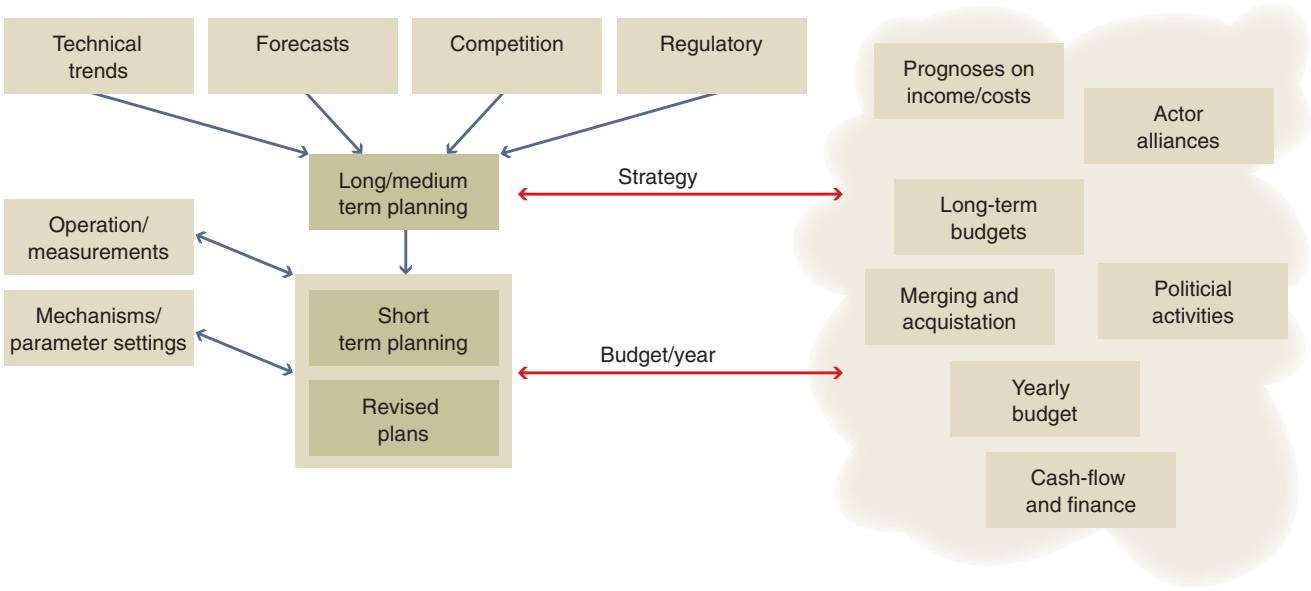


Figure 2 Illustration of relations between network planning and telecom management

manager to evaluate the financial performance of the enterprise in order to make further decisions. In brief, a business plan summarizes the results of the planning process for projects involved, including:

- Objectives to reach (future revenue, number of customers, etc.)
- Effort requested (human resources, equipment investments, etc.)
- Influence on performance indicators, in particular the financial indicators.

In a similar manner as network planning horizons, business plans can be divided into I) strategic business plan (decisions with a longer horizon, affecting all enterprise), II) tactical business plan (decisions for particular units/projects, affecting market segments, system choices, etc.), III) short-term business plans for management control (assisting monitoring of performance, follow-up of budgets, etc.).

The relations between network planning, planning for other areas and business plans, can be illustrated as an iterative process, see Figure 3.

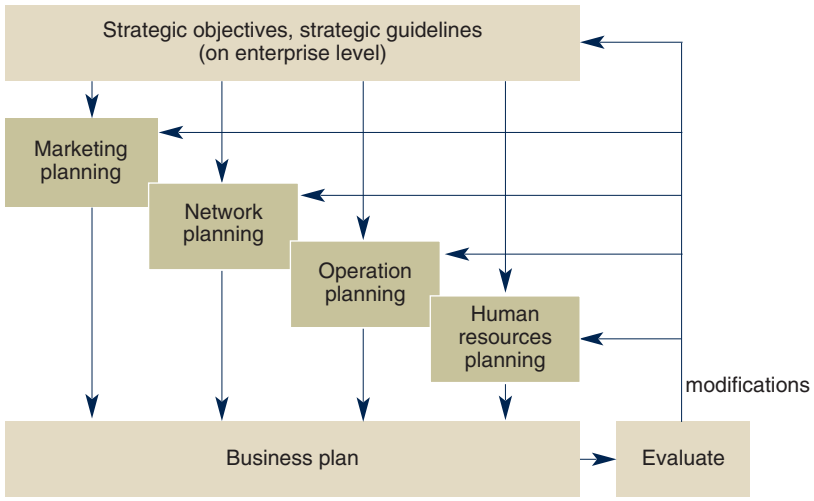


Figure 3 Relations between business plans, enterprise strategic objectives and area planning – improvements through iterations

Statement	Relations	Purposes (examples)
Income	Net income = revenues – expenses	• to analyze potential profit • is the profit high enough?
Balance sheet	Enterprise capital = assets – liabilities	• to analyze the financial structure • how to finance the migration • enough/too much equity? • enough/too much debt?
Cash flow	Cash balance = inflows – outflows	• to make payments at every due date • to have the right cash at the right time

Table 1 Elements of fundamental financial indicators

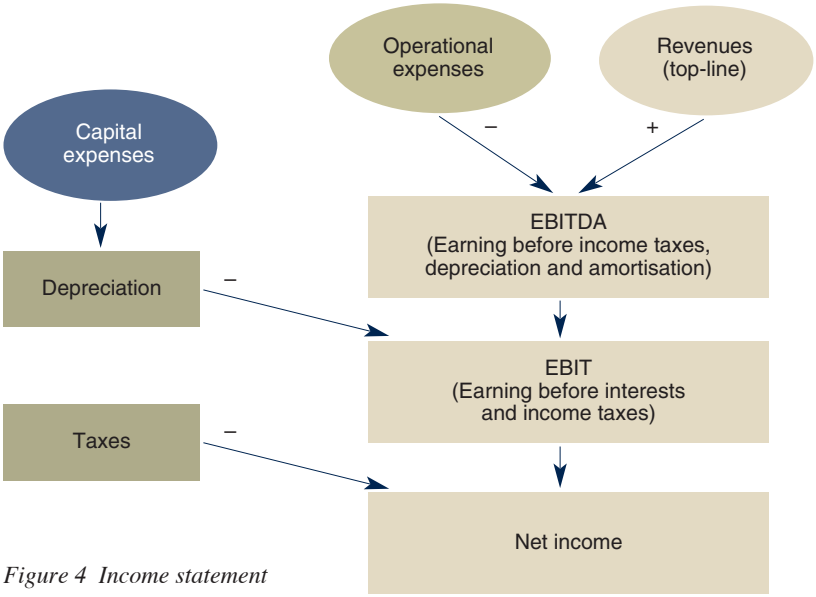


Figure 4 Income statement

Besides the outer iteration, there may also be interactions between the different planning groups. However, a fundamental message is that these have to be coherent in order to efficiently support the enterprise's objectives.

The elements in the business plan contain financial indicators, models for costs and revenues as well as sensitivity studies. These may be updated during planning exercises, both parameter ranges and modelling considerations; through evaluation of effects from activities, change of technical/organization/etc. solutions, portfolio adjustments, strength-weakness-opportunities-threats analyses, and so forth. As shown in the figure, iterations allows for steady improvements.

It is hard to describe business plans and business modelling without entering an arena of financial indicators. All fundamental financial indicators are carried out with elements as given in Table 1.

Figure 4 illustrates relations between key measures related to income. As seen the basic network-related costs can be divided into investment-related (capital cost) and operational

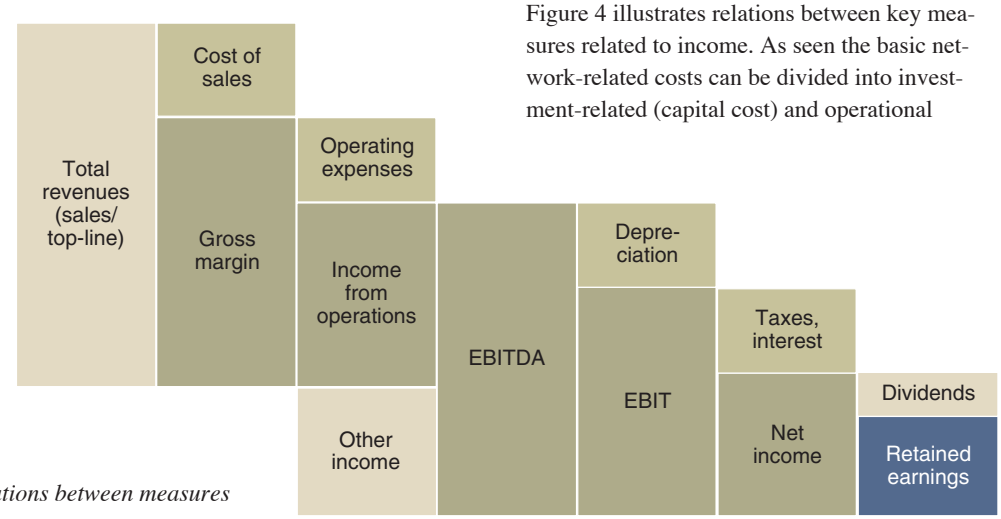


Figure 5 Income statement – relations between measures

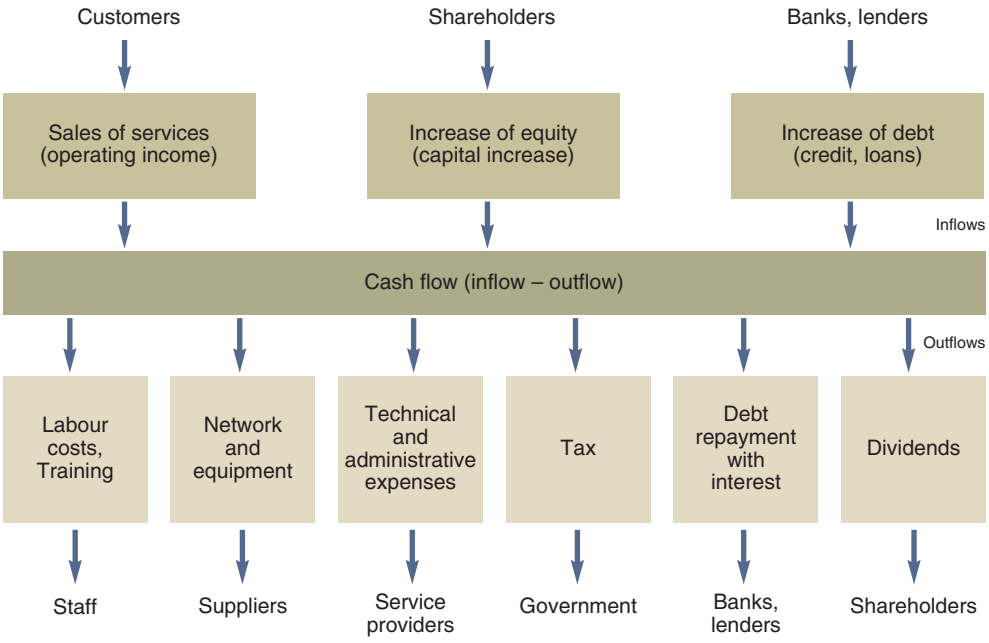


Figure 6 Cash flow statement (inflows and outflows)

related (rights of way, renting space, power, etc.). Similar illustrations for relations between central measures are given in Figures 5, 6 and 7. The indicators mostly used for economics/financial issues are:

- Internal rate of return (IRR)
- Net present value (NPV)
- Discounted payback period (DPP)
- Net cash flow (NCF)
- Discounted cash flow (DCF)
- Operating income
- Revenue per product group

Using the cash flow calculations for decisions is illustrated in Figure 8. From that example it is seen that introducing more services on the common infrastructure improves the cash flow as seen for the upgrade project. However, the total picture for the operation (including effects on other networks) must also be considered when an investment activity is decided. Similar cash flow calculations can also be made for other competition scenarios, e.g. with severe competition putting pressure on the price level that can be chosen for the different services.

Actions to consider to improve the cash flow for a project are to

- Increase the number of new customers, sales of connections and product bundles
- Limit the churn rate of customers by keeping the loyalty of the present customers
- Introduce new services and service bundles in the portfolio
- Minimize expenses (capital and operational)

3 Planning Scope and Life-cycle

Avoiding initiating a too tedious task, the planning scope should be properly defined. Tractable models state that the broader the scope the less detail should be considered. The scope can be limited in several directions, such as the geographic area, the network layers, the time period, etc. A number of directions are illustrated in Figure 9.

An example on network layering is shown in Figures 10 and 11. A finer or coarser separation into layers might be needed depending on the objective of the planning exercise. In the example shown in Figure 11 the lowest layer refers to buildings (including power, cooling, etc.) and cables. The next level refers to transport networks, such as SDH – possibly including wavelength division systems. On top of this are placed the networks of switches and routers; this also includes radio base stations. The top layer in this example is the service logic nodes, for example service control nodes in the intelligent network architecture, home location registers, and so forth.

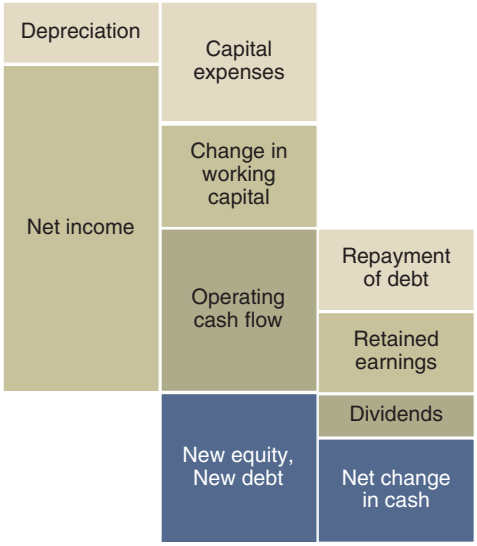


Figure 7 Cash flow calculations

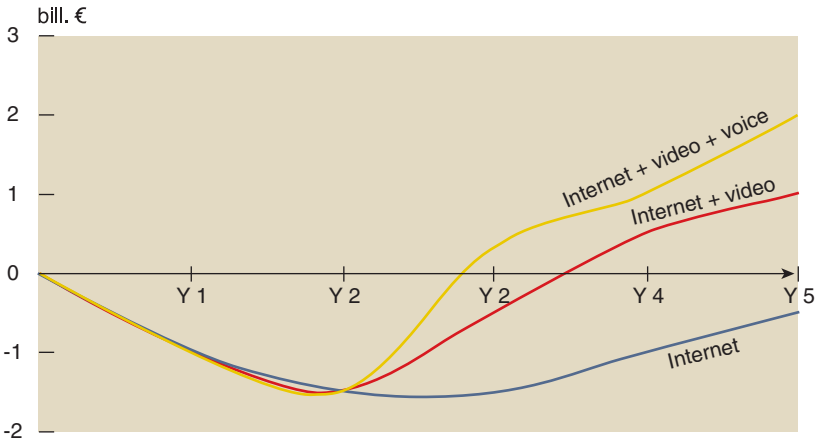


Figure 8 Cash flows for three network upgrade options

In ITU-T three planes have been described referring to various aspects of the network capabilities. Briefly, the main purpose of each of the planes are:

- *User plane* – to convey the user information (information from higher layers)
- *Control plane* – to control traffic flows and resource configurations
- *Management plane* – to manage network resources, including fault management, configuration management accounting management, performance management, security management.

The actual distinction between user data, control and management might be rather blurred in some cases. Any larger network would typically

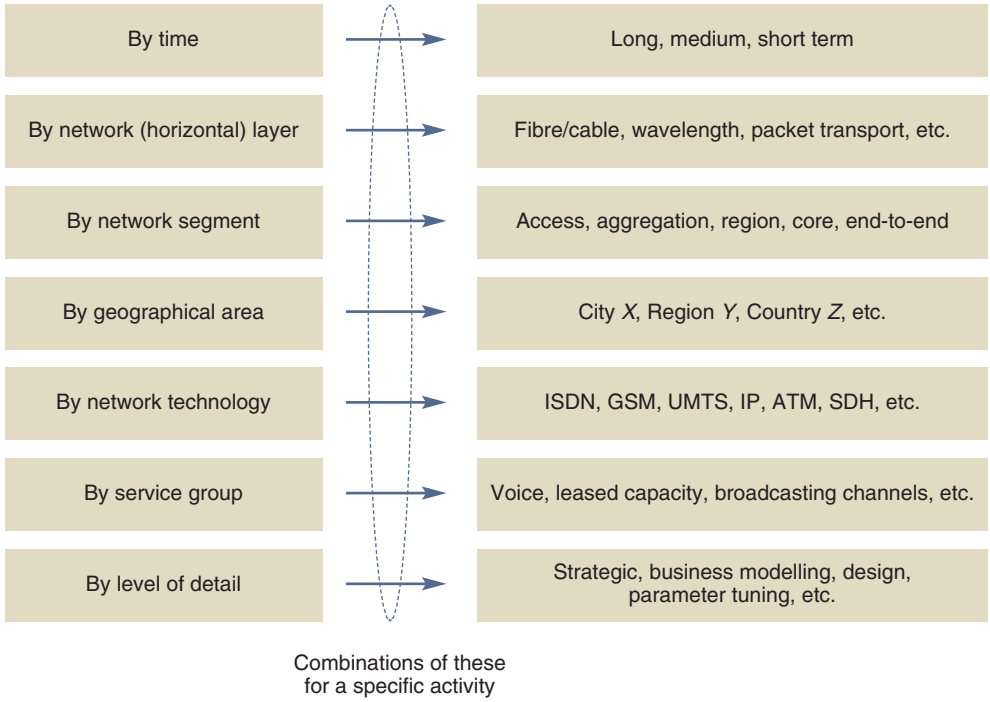


Figure 9 Directions for limiting the network planning task

have functions belonging to all these planes, although the way functions are implemented varies. An objective is to find efficient complete solutions and combinations of functions that incorporate all needed functionality.

In order to locate where to direct the information, *addresses* and *routing* functionality have to be present. So, addresses are used to identify a unit/interface, while routing is used to find how to direct the information towards the address (and the unit/interface it represents). Note that addresses and routing functions must be implemented for all three planes.

In DSL forum a "three-layer logical/functional architecture" has been elaborated as shown in Figure 10. This architecture is intended as a reference for delivery of differentiated and ensured services. Figure 10 shows an architecture for support of media-on-demand and conversational services. Three layers are defined: i) network layer, ii) common enabling services layer, and iii) application layer.

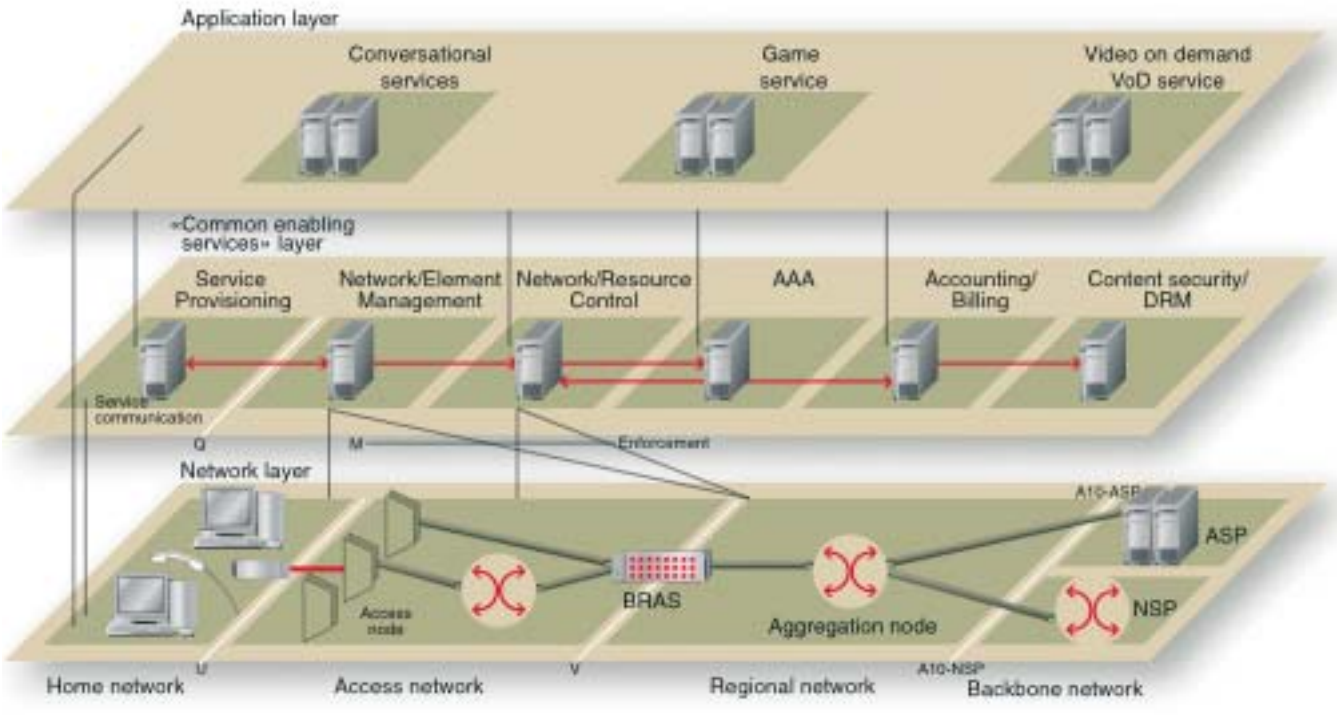


Figure 10 The three-layer logical/functional architecture, (from [DSL-058])

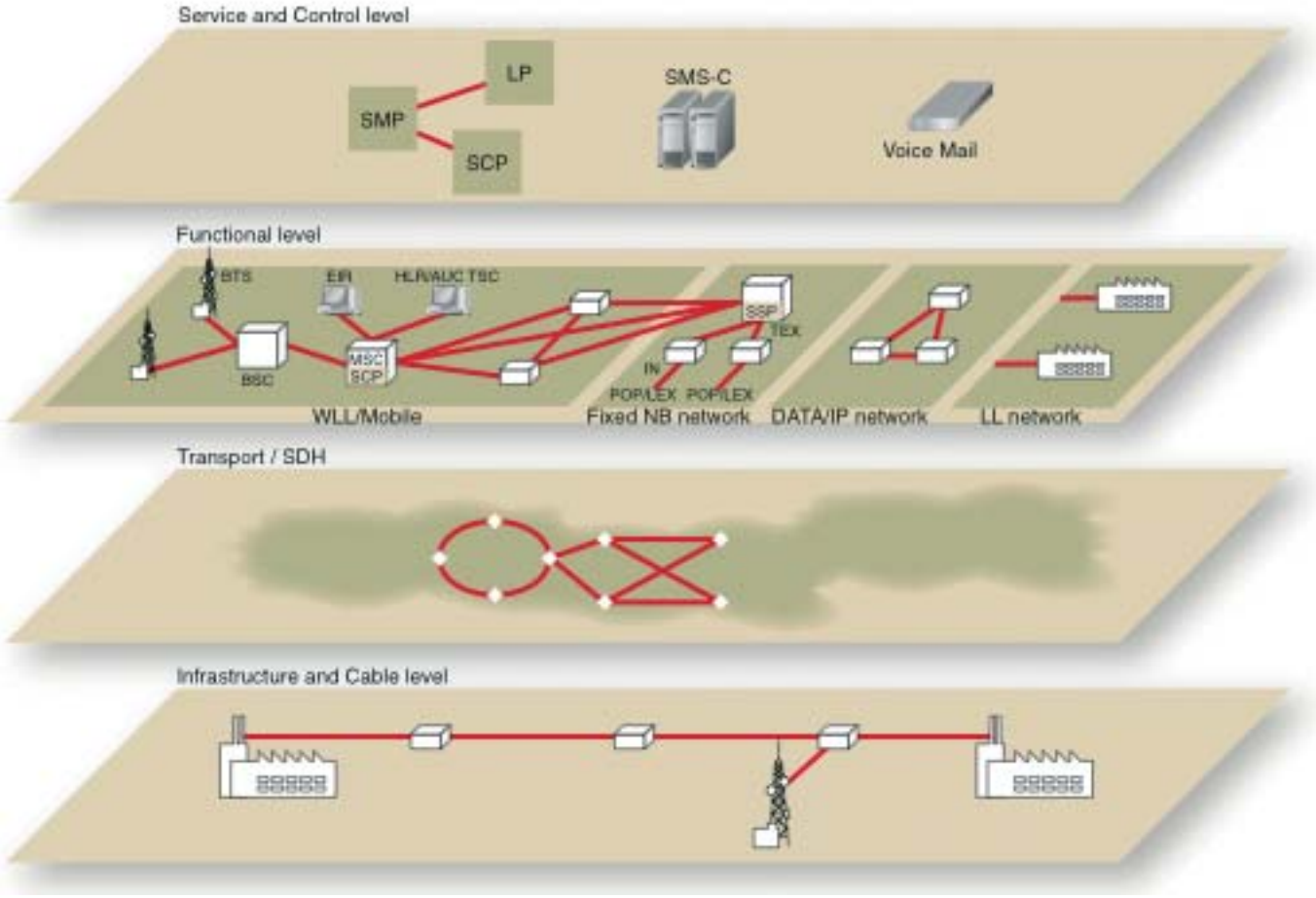


Figure 11 One way of defining network (horizontal) layers (from [Soto02])

A possible set of interactions between the layers as well as questions to be raised are given in Figure 12. However, the idea behind separation into layers is that the layers are examined in series; one-by-one. This would naturally not deal with the dependencies at the same time as solutions are obtained for each of the layers and might therefore imply that sub-optimal solutions are obtained.

As an ultimate goal, all the questions should be addressed in a single "go" for network planning. However, this is not a realistic case beyond the very simple networks. However, relations between current state, financial aspects, product forecasts and technical solutions must be considered, as exemplified in Figure 13. Note the set of procedures and corresponding tools to assist in conducting the various activities. A variation of this is also depicted in Figure 14.

A number of information elements are required as input and flowing between the different steps and scopes involved in network planning. Some samples are:

- Description of network architectures, node locations, product mapping, etc. resulting from longer term planning is related to solutions on shorter terms (commonly carried out iteratively between the activities).
- Planning results and information are transferred between management and operational support systems and network elements/architectures, e.g. based on measurements or number of nodes.
- Description of basic infrastructure information considered (location of customer sites, housing, power, cooling).

Figure 12 Illustration of questions and relations between network (horizontal) layers (from [Soto02])

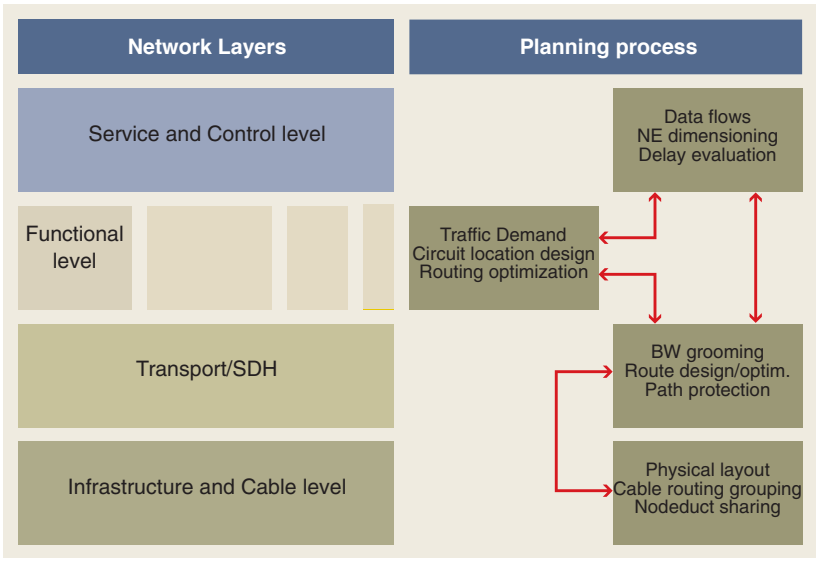
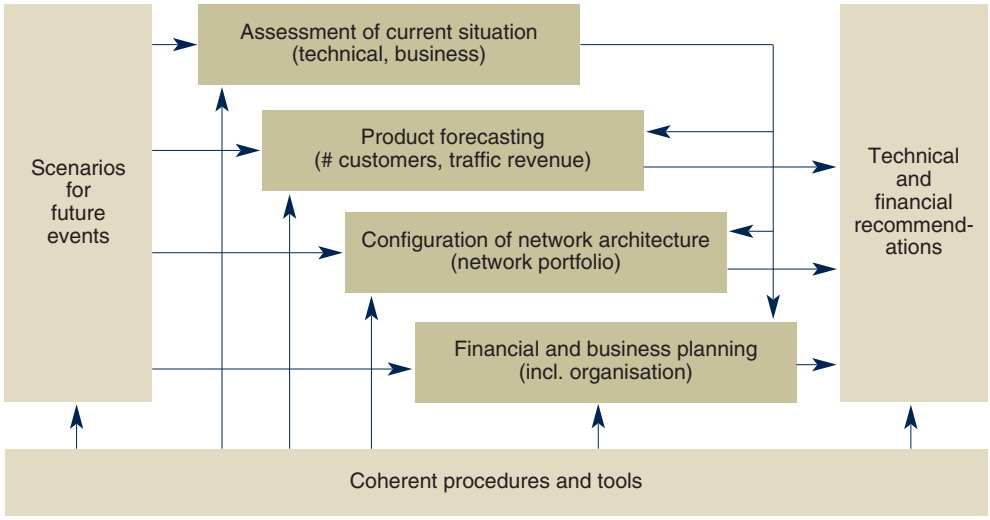


Figure 13 Integrating planning activities



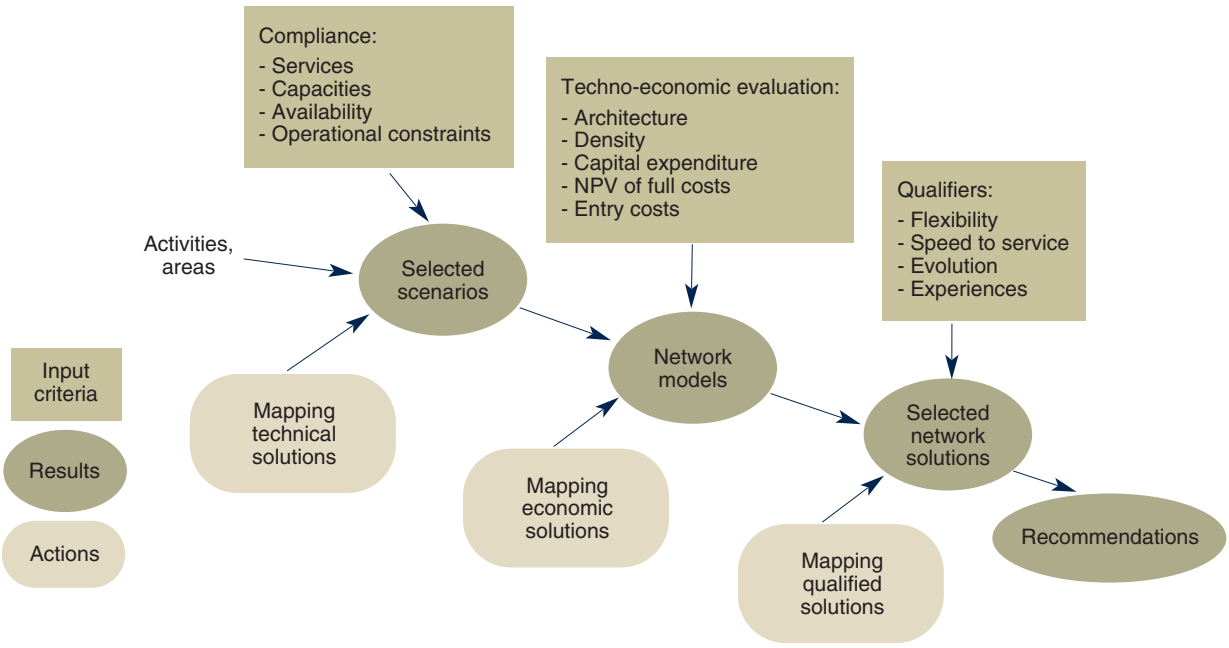


Figure 14 Illustration in steps towards recommending network architectures (from [Soto02a])

- Description of future scenarios – regarding product forecasting, technical issues, regulatory issues, competition, etc.
- Geographical and demographical information on households and other customer groups.
- Customer segmentation; characteristics, product bundles, usage patterns.
- Product portfolio, covering access-related and other service types, as well as their requirements/SLA-conditions.
- Network resource loads from service usage, covering all relevant resource types (nodes, links).
- Description of technical components available for network deployment; price, capacity/performance levels, functionality.
- Regulatory restrictions on service offerings and interfaces/unbundling, interconnection points.
- Description of existing network (nodes, links, cables) on locations, capacities, traffic, re-usability.
- Financial calculation data including interest rates, amortization periods, etc.
- Available labor force of relevant types.

A number of factors influence the choice of the better network configuration, ref. Figure 15. A general objective is to describe "the better network solution" considering the often opposing sets of requirements. It is important to note that the "better solution" for one period may not be the preferred option for another period. Nor may a "better solution" at all be wanted when both

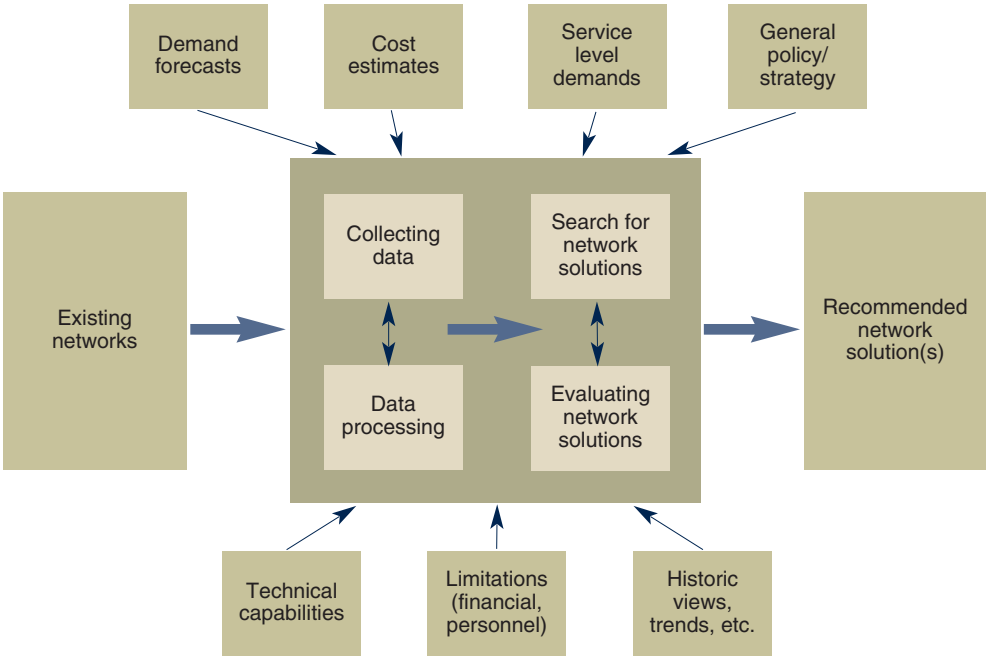


Figure 15 Sample of factors and activities influencing the better network solution(s)

the (historic/) current configuration and a longer term recommended solution are taken into account.

Lead times when ordering new equipment, limited work force for installation and testing and so forth can also influence the speed of introducing new network solutions or upgrade network elements.

As seen from Figure 15 data handling is needed at different stages such as:

- Collecting data from various sources
- Adaptation of data to the purpose at hand, e.g. for input to a calculation program
- Use of data when analysing the models as a simplification of the real world case
- Evaluation of results/output data from the calculations.

During all these stages, the data quality must be assessed, including the uncertainty attached and which values are the most likely ones. In order to do this, insight must be gained into the origin of the data as well as any processing on the way to model input.

Arriving at an understanding of the applications and services during the course of planning is important. However, a number of sensitivity analyses is typically run for different demand patterns to see how robust the network solution is with respect to these patterns. Some examples on applications related to TV and PC are shown in Figure 16.

Related to applications and customer demands are also the management of services and the formulation of adequate SLAs. What is needed for efficient service management? It is essential that the service management system works well with the network and element management systems, supporting the following tasks:

- Dynamic network representation – to handle new and out-of-service managed resources.
- Proactive data and event monitoring – to integrate data from a set of diverse vendor and protocol sources and identify potential problems.
- Automated topological and model-based reasoning – to understand the impacts of events based on connectivity and configuration.
- Cause – effect inference – to map between the effects of anomalies and their service impacts.
- Root cause analysis – to differentiate between symptoms of a problem and the true problem.
- Operator guidance – to assist to prioritising operator actions.
- Automated testing – to verify diagnoses while minimizing network impacts.
- Automated fault correction – to implement network reconfiguration to optimise service delivery.

These capabilities are essential for truly effective service level management. That is, not meeting the requirements, most of the event management

TV Focused Services	Typical bandwidth (upstream)	Typical bandwidth	Delay bound	Packet loss	On demand
Broadcast TV – e.g. MPEG2		2 – 6 Mbit/s[1)]	~1 s	10^{-5}	Yes
High definition TV – HDTV		12 – 19 Mbit/s[1)]	~1 s	10^{-5}	Yes
Pay Per View and NVOD – e.g. MPEG2		2 – 6 Mbit/s[1)]	~1 s	10^{-5}	Yes
VOD – e.g. MPEG2		2 – 6 Mbit/s[1)]	~1 s	10^{-5}	Yes
Navigator and EPG (can be locally launched and updated in non real time)		< 0.5 Mbit/s	N/a	N/a	No
Picture in Picture – two MPEG2 channels		Up to 12 Mbit/s[1,2)]	~1 s	~1 %	Yes
Picture in Browser – one MPEG2		Up to 9 Mbit/s[1,2)]	~1 s	10^{-5}	Yes
Personal Video Recorder PVR – replay MPEG2 file off hard disk		2 – 6 Mbit/s local[1)]	N/a	N/a	Yes
ITV – TV telephone features	< 64 kbit/s	< 64 kbit/s	< 400 ms (RTT)	~1 %	Yes
– TV browser		Up to 3 Mbit/s	N/a	N/a	Yes/No
– TV e-mail	128 – 640 kbit/s	Up to 3 Mbit/s	N/a	N/a	No
– TV Instant Messaging	128 – 640 kbit/s	Up to 3 Mbit/s	N/a	N/a	No
– TV Chat	128 – 640 kbit/s	Up to 3 Mbit/s	N/a	N/a	No
– TV on-screen notification		< 64 kbit/s	N/a	N/a	No
– TV interactive games	128 – 640 kbit/s	Up to 3 Mbit/s	~10 ms	10^{-5}	Yes
– TV Audio Juke Box		< 128 kbit/s	~1 s	< 1 %	Yes

Notes:
1) video compression advancements will enable more efficient encoding (1.5 to 3 Mbit/s)
2) more efficient solutions could be available

PC Focused Services	Typical bandwidth (upstream)	Typical bandwidth (downstream)	Delay bound	Packet loss	On demand
High Speed Internet Access (browsing, IM, Chat, FTP, VPN, access, etc.) – Residential (typically asymmetric) – SME/SOHO (typically symmetric)	 128 – 640 kbit/s Up to 6 Mbit/s	 Up to 3 Mbit/s Up to 6 Mbit/s	 N/a N/a	 N/a N/a	 Yes/No Yes/No
Server based e-mail	as above	as above	N/a	N/a	No
Live TV on PC		300 – 750 kbit/s	~1 s	~1 %	Yes
Video on Demand		300 – 750 kbit/s	~1 s	~1 %	Yes/No
Video Conferencing	300 – 750 kbit/s	300 – 750 kbit/s	< 400 ms (RTT)	~1 %	Yes/No
Voice/Video telephony	64 – 750 kbit/s	64 – 750 kbit/s	< 400 ms (RTT)	~1 %	Yes
Interactive Games	10 – 750 kbit/s	10 – 750 kbit/s	~10 ms	10^{-5}	Yes
Remote Education		300 – 750 kbit/s	~1 s	~1 %	Yes/No

Figure 16 Samples of applications (from [DSL-058])

responsibilities add to the already loaded human operators.

Mapping network events onto services and related SLAs allows for real-time management of the business. Key performance indicator can be polled periodically, and trend analysis of the historical values can be used for proactive SLA management, predicting limit violations. As a consequence of a violation of an SLA condition, any actions – not just reports – can be performed, allowing corrective measures to be automated. The service management system should include integrated object models of the network devices and their connection, the services, SLAs and customers. It should be able to perform what-if analysis for possible customers affected by a potential failure and be able to generate a list of all devices used by a customer or a group.

A network lifecycle could possibly also be thought of as several phases interacting. This is illustrated in Figure 17. Note, however, that the different time scales must also be considered in coherence.

4 Internet Traffic Engineering – Exemplifying Network planning

The text in this section is based on [RFC3272], giving an example of how network planning – on the shorter/medium term – can be presented. As stated there, the main goals of traffic engineering (TE) are to improve performance for IP-based traffic while still utilising the network resources efficiently. Internet traffic engineering is defined as

> *that aspect of Internet network engineering dealing with the issue of performance evaluations and performance optimisation of operational IP networks. Hence, measurement, characterisation, modelling and control of traffic are included.*

Capacity management and traffic management can be used for the optimisation done as part of TE. Capacity arrangement includes capacity planning, routing control and resource management. The resources include *link bandwidth*, *buffer space* and *computational power*. Traffic management includes nodal traffic control (e.g. traffic conditioning, queue management, scheduling) and other functions that regulate traffic through the network or control access to network resources. These activities can be carried out continuously and in an iterative manner. The activities are commonly divided into proactive and reactive. In the former preventive and perfective actions can be found, while corrective actions would be part of the latter.

The TE actions would operate on various time scales; coarse (days, years – like for capacity management), intermediate (ms, days – like for routing control), and fine (ps, ms – like for packet level processing).

The TE subsystems include capacity augmentation, routing control, traffic control and resource control. Input to the TE system would be network state variables, policy variables and decision variables etc. The challenge is to introduce automated capabilities that adapt fast and efficiently to changes in the network state, while stability is maintained. Performance evaluation is then a critical part of this, to assess the effectiveness of a TE method, to monitor a network state and to verify compliance with performance levels.

4.1 Settings

A number of settings for exercising TE activities are identified in [RFC3272]. To some extent these can also be considered as steps, see Figure 18. However, considering that TE activities are carried out continuously the different steps may be active at the same time, although possibly looking at different instances of time to implement the solutions into the network.

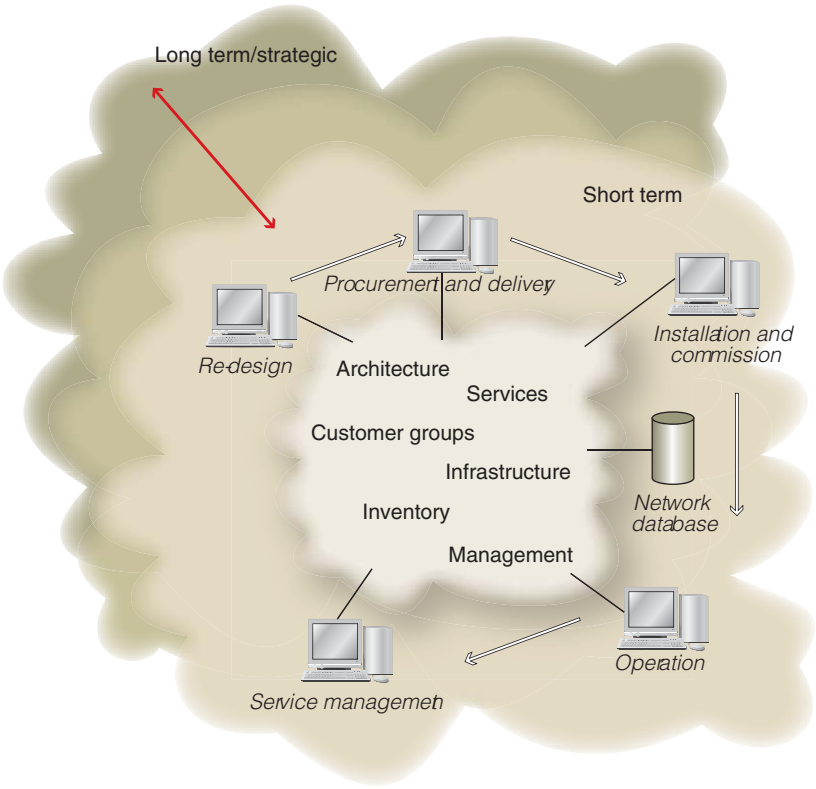
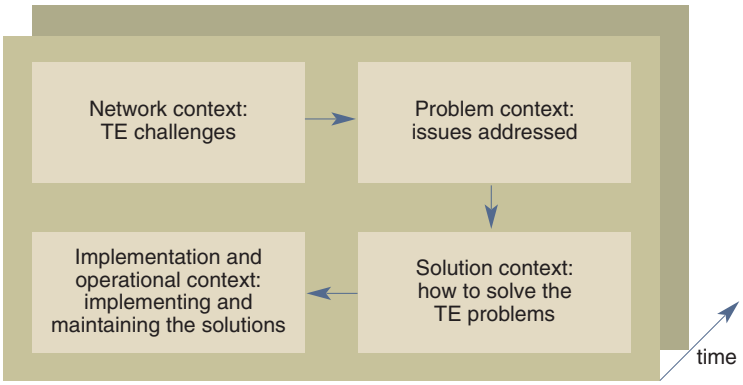


Figure 17 Example of network life-cycle

The settings described are:

- *Network* context; describing the situations where traffic engineering challenges are found. Such situations include network structure, network policy, network characteristics, network constraints, network quality attributes, network optimisation criteria, etc. A network can be represented as a system consisting of, i) a set of interconnected resources, ii) a demand representing the offered load, and iii) a response consisting of network processes, protocols and mechanisms that carry the offered load through the network. Several types of demand classes may be present, similar to traffic classes although also different customer types should be taken into account. This results in a request for differentiated service. The network resource and the traffic handling-related mechanisms do also have their characteristics.

Figure 18 Settings for exercising TE activities



- *Problem* context; defining the issues that TE addresses, like identification, abstraction, representation, formulation, requirement specification, solution space specification, etc. One class of problems is how to formulate the questions that traffic engineering should solve; how to describe requirements on the solution space, how to describe desirable features of good solutions, how to solve the problems and how to characterise and measure the effectiveness of the solutions. Another problem is how to measure and assess the network state parameters, including the network topology. A third class of problems is how to characterise and evaluate network states under a variety of scenarios. This can be addressed both on system level (macro states – "macro TE") and resource level (micro state – "micro TE"). Solving congestion is an essential part of performance improvement. Handling congestion can be divided into demand side policies (restrictive) and supply side policies (expansive).

- *Solution* context; elaborating how to solve the TE problems. This typically means evaluation of alternatives, which requires estimating traffic load, characterising network state, elaborating solutions on TE problems and setting up a set of control actions. The instruments relevant include i) a set of policies, objectives and requirements for network performance evaluation and optimisation, ii) a set of tools and mechanisms for measurement, characterisation, modelling and controlling traffic and allocation to network resources, iii) a set of constraints on the operating environment, network protocols and TE system, iv) a set of quantitative and qualitative techniques and methods for abstracting, formulating and solving TE problems, v) a set of administrative control parameters that may be managed by a configuration management system, vi) a set of guidelines for network performance evaluation, optimisation/improvement. Traffic estimates can be derived from customer subscription information, traffic projections, traffic models, and from empirical measurements. Polices for handling the congestion problem can be categorised according to the criteria i) response time scale (long – weeks to months, e.g. capacity planning; medium – minutes to days, e.g. setting routing parameters, adjusting Label Switched Path (LSP) design; short – ps to minutes, e.g. packet processing of marking, queue management), ii) reactive versus preventive, and iii) supply side (increase available capacity, redistribute traffic flows) versus demand side (control the offered traffic).

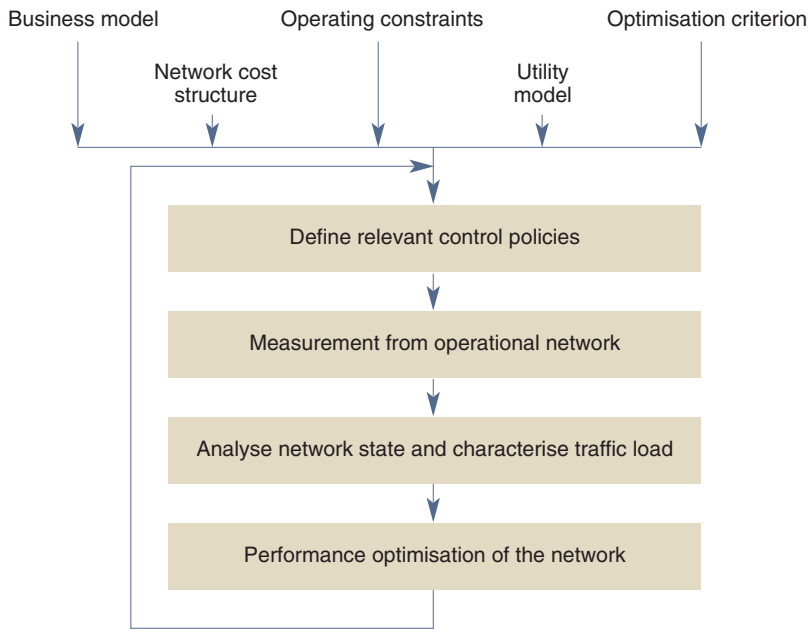
- *Implementation and operational* context; implementing the actual solutions, involving planning (e.g. a priori determined actions based on triggers), organisation (e.g. assigning responsibilities to different units and co-ordinating activities), and execution (e.g. measurement and application of corrective and perfective actions).

These context descriptions may also be looked upon as gradually getting more precise and closer to the implementation.

4.2 TE Process Model

A TE process model is presented in [RFC3272] and depicted in Figure 19 as an iterative procedure consisting of four main steps.

Figure 19 TE process model



The first phase refers to definition of *control policies*. These would typically depend on a set of inputs, like business model, network cost structure, operating constraints, utility model and optimisation criterion.

The second phase involves *measurements* in order to assess the conditions in the network; network state and traffic load.

The third phase consists of *analysing the network state and characterising the traffic load*. A number of potential models and analysing techniques may be relevant, for instance also looking at the timely and spatial distribution of the traffic load.

In the fourth phase, *performance optimisation* is done. This includes a decision process selecting and implementing a set of actions. Actions may work on the load demand, distribution of load and network resource configuration and capac-

ity. This may also initiate a network planning in order to improve network design, capacity, technology, and element configuration.

4.3 TE Key Components

The key components of the TE process model are:

- *Measurement* subsystem: Carrying out measurement is essential to provide feedback on the system state and performance. It is also critical in order to assess the service level provided (and QoS) and effect of TE actions. A basic distinction between monitoring and evaluation is to be observed; monitoring refers to provision of raw data, while evaluation refers to use of the raw data for inferring on the system state and performance. Measurements can be carried out at different levels of aggregation, e.g. packet level, flow level, user level, traffic aggregate level, component level, network-wide level, and so forth. In order to perform measurements systematically, several questions have to be answered, like [RFC3272]: Which parameters are to be measured? How should the measurements be accomplished? Where should the measurement be performed? When should the measurement be performed? How frequently should the monitored variables be measured? What level of measurement accuracy and reliability is desirable and realistic? To what extent can the measurement system permissibly interfere with the operational network conditions? What is the acceptable cost of measurements?

- *Modelling and analysis* subsystem: A central part of the modelling is to elaborate a representation of the relevant traffic characteristics and network behaviour. In case a structural model is used, the organisation of the network and its components are the main emphasis. When behavioural models are used, the dynamics of the network and traffic are the key issues. The latter model is particularly relevant when doing performance studies. Then adequate models of the traffic sources are also needed.

- *Optimisation* subsystem: Optimisation can be categorised as real-time and non-real-time. The former operates on short to medium time scales (e.g. ms to hours) and works to adjust parameters in mechanisms in order to relieve congestion and improve performance. Examples of means are tuning of routing parameters, tuning of buffer management mechanisms and changing Label Switched Paths (LSPs). Non-real-time is also seen as network planning, typically working on a longer scale. For both of these, stability and robustness are essential concerns.

Routing is a central component in efficient handling of traffic flows in an IP-based network. When introducing a number of service classes, some additional constraints can also be considered when deciding upon the possible routing. Examples of such constraints are available bandwidth, hop count, and delay. This implies that possible paths as seen from a router must have the corresponding attributes attached.

4.4 Requirements on TE Systems

[RFC3272] describes a number of requirements that a TE system should meet. Here a requirement is understood as a capability needed to solve a TE problem or to achieve a TE objective. The requirements are either non-functional or functional. A non-functional requirement relates to the quality attributes of state characteristics of a TE system. A functional requirement gives the function a TE system should perform in order to reach an objective or address a problem.

4.4.1 Non-functional Requirements

The generic non-functional requirements given in [RFC3272] are:

- Usability. This is a human factor aspect referring to the ease of deployment and operation of a TE system.

- Automation. Usually, as many functions as possible should be automated, reducing the human effort to control and analyse the information and network state. This is even stronger for a larger network.

- Scalability. The TE system should scale well when the number of routers, links, traffic flows, subscribers, etc. grows. This may imply that a scalable TE architecture is applied.

- Stability. This is an essential requirement for an operational system avoiding adverse results for certain combinations of input and state information.

- Flexibility. A TE system should be flexible both in terms of the optimisation policy and the scope. An example of scope is that additional classes should be considered in case these are introduced into the network. Another aspect of flexibility is that some subsystems of the TE system could be enabled/disabled.

- Visibility. Mechanisms to collect information from the network elements and analyse the data have to be present in a TE system. These would then allow for presenting the operational conditions of the network.

- Simplicity. A TE system should be as simple as possible; that is, considered from the out-

side, not necessarily using simple algorithms. Simplicity is particularly important for the human interface.

- Efficiency. As little demanding as possible, in terms of processing and memory resources, is requested. However, this also refers to the fact that a result from the TE system should be obtained in a timely manner.

- Reliability. A TE system should be available in the operational state when needed.

- Survivability. Recovering from a failure and maintaining the operation is requested, in particular for the more critical functions of a TE system. Commonly, this requires that some redundancy is introduced.

- Correctness. A correct response (according to the algorithms implemented) has to be obtained from a TE system.

- Maintainability. It should be simple to maintain a TE system.

- Extensibility. It should be easy to extend a TE system, e.g. when introducing new functions and when the underlying network is extended.

- Interoperability. Open standards should be used for the interfaces in order to simplify interoperation with other systems.

- Security. Means supporting integrity, information concealment, etc. have to be implemented.

As mentioned, some of these requirements may be mandatory while others are optional for a particular TE system.

4.4.2 Functional Requirements

Some functional requirements are also described in [RFC3272], such as those related to:

- Routing. An efficient routing system should take both traffic characteristics and network constraints into account when deriving the better routing schemes. A load splitting ratio among alternative paths should be configurable allowing for more flexibility in the traffic distribution. Some routes of subsets of traffic should also be controllable without affecting routes of other traffic flows. This is particularly relevant when several classes are present in the network. In order to convey information on topology, link characteristics and traffic load, several of the relevant routing protocols have to be enhanced. An example is constraint-based routing, which is gaining more interest. This addresses the selection of paths for packets and may work well with path-oriented solutions, that is LSPs.

- Traffic mapping. This refers to assigning traffic flows onto paths to meet certain requirements considering the set of relevant policies. A central issue arises when several paths are present between the same pair of originating and destination router. Appropriate measures should be taken to distribute the traffic according to some defined ratios, while keeping the ordering of packets belonging to the same application (or micro-flow).

- Measurement. Mechanisms for monitoring, collecting and analysing statistical data have to be in place. Such data may relate to performance and traffic. In particular, being able to construct traffic matrices per service class is an essential part of a TE system.

- Network survivability. Survivability refers to the capability to maintain service continuity in the presence of faults. This can be realised by rapid recovering or by redundancy. Coordinating of protection and restoration capabilities across multiple layers is a challenging task. At the different layers protection and restoration would typically occur at different temporal and bandwidth granularity. At the bottom layer, an optical network layer may be present, e.g. utilising WDM. Then, SDH and/or ATM could be present below the IP layer. Restoration at the IP layer is commonly done by the routing protocols, which may require some minutes to complete. Some means being proposed relate to MPLS allowing for faster recovery. A common suite of control plane protocols has been proposed for the MPLS and optical transport networks. This may support more sophisticated restoration capabilities. When multiple service classes are present, their requirements on restoration may differ introducing further challenges on the mechanisms to be used. Resilience attributes can be attached to an LSP telling how traffic on that LSP can be restored in case of failure. A basic attribute may indicate if all traffic trunks in the LSP are transferred on a backup LSP or some of the traffic is to be routed outside, e.g. following the routing protocols. Extended attributes may be introduced giving indications such as backup LSP is to be pre-established, constraints for routing the backup LSP, priorities when routing backup LSP, and so forth.

- Servers and content distribution. Location and allocation of content on servers have significant impact on the traffic distribution, in particular as long as much of the traffic is similar to client–server interactions. Hence, load bal-

ancing directing traffic on the different servers may improve the overall performance. This is sometimes called traffic directing, operating on the application layer.

- DiffServ issues. As DiffServ is more widely deployed, adequate TE systems become more critical to ensure that conditions in Service Level Agreements (SLAs) are met. Service classes can be offered by defining Per-Hop Behaviours (PHBs) along the path, exercising DiffServ in the nodes, in particular by configuring mechanisms like traffic classification, marking, policing and shaping (mainly in edge routers). A PHB is a forwarding treatment including buffer management and scheduling. In addition the amount of service capacity, e.g. bandwidth, allocated to the different service classes has to be decided upon. The following issues, from [ID_tepri], give some requirements on TE in a DiffServ/MPLS environment:

  - An LSP should provide configurable maximum reservable bandwidth and/or buffer for each supported service class.

  - An LSR may provide configurable minimum available bandwidth and/or buffer for each class on each of its links.

  - In order to perform constraint-based routing on a per-class basis for LSPs, the routing protocols should support extensions to propagate per-class resource information. When delay bounds is an issue, path selection algorithms for traffic trunks with bounded delay requirement should take delay constraints into account.

  - When an LSR dynamically adjusts resource allocation based on per-class LSP resource requests, adjusting weights for the scheduling algorithms should not adversely impact delay and jitter characteristics.

  - An LSR should provide configurable maximum allocation multiplier on a per-class basis.

  - Measurement-based admission control may be used to improve resource usage, especially for classes not having strict loss or delay/jitter requirements.

- Controlling the network. In order to see the effect of having a TE system, the relevant decisions must be introduced into the network. Control mechanisms may be manual or automatic, the latter being a goal for most. Network control functions must be secure, reliable and stable, in particular during failure situations.

4.5 TE Taxonomy

A taxonomy of TE systems is given in [RFC3272] according to the following criteria:

- *Time-dependent* vs. *state-dependent* vs. *event-dependent*. A static TE system implies that no TE methods are applied on the time scale considered. Therefore, it is commonly assumed that all TE schemes are dynamic (on the time scale looked at). A time-dependent scheme is based on timely variations in traffic patterns and used to pre-program changes in the traffic handling. A state-dependent scheme adapts the traffic handling based on state of the network, allowing for taking actual variations in the traffic patterns into account. The state of a network may be based on resource utilisation, delay measures, etc. Accurate information available is crucial for adaptive TE schemes. This information has to be gathered and distributed to the relevant routers. A challenge is to limit the amount of information that must be exchanged between routers, still allowing for sufficiently updated data in each of the routers to take the traffic handling decisions. Event-dependent schemes may lead to less information exchanges compared to state-dependent schemes. Then, certain events are used as input when updating the traffic handling, like traffic load crossing a threshold, unsuccessful establishment of an LSP, etc.

- *Offline* vs. *online*. In case changes in traffic handling do not need to be done in real time, the computations can be done offline, e.g. allowing for more thorough searches over the feasible solutions finding the best one to apply. On the other hand, when traffic handling is to adapt to changing traffic patterns, it is to be done online. For online calculations, relatively simple algorithms are applied leading to short response times until the updated traffic handling can be activated. Still the algorithm should present a solution that is close to the optimal one.

- *Centralised* vs. *distributed*. In a centralised scheme a central function decides upon the traffic handling in each of the routers. Then, the central function has to collect and return the information. In order to limit the overhead, infrequent information exchanges are sought; however, more frequent exchanges are asked for to keep an accurate picture of the network state in the central function. This results in a classic trade-off problem, finding the time interval for collecting and returning the information. A similar trade-off is also seen for the distributed scheme, although then the decisions are made by each router. A drawback of a centralised scheme is often that a single point of failure is introduced, implying that

the central function is available and has sufficient processing capacity for the scheme to work efficiently.

- *Local* vs. *global information*. Local information refers to a portion of the region/domain considered by the TE system. An example is delay for a particular LSP. Global information refers to the whole region/domain considered.

- *Prescriptive* vs. *descriptive*. When a prescriptive approach is used, a set of actions would be suggested by the TE system. Such an approach can be either corrective (an action to solve an existing or predicted anomaly) or perfective (an action suggested without identifying any particular anomaly). A descriptive approach characterises the network state and assesses the impact from exercising various policies without suggesting any specific action.

- *Open loop* vs. *closed loop*. In an open loop approach, the control actions do not use feedback information from the network. Such feedback information is used when a closed loop approach is followed.

- *Tactical* vs. *strategic*. A tactical approach considers a specific problem, without taking into account the overall solutions, tending to be ad hoc in nature. A strategic approach considers the TE problem from a more organised and systematic perspective, including immediate and longer-term consequences.

- *Intradomain* vs. *interdomain*. Interdomain traffic engineering is primarily concerned with performance of traffic and networks when the traffic flows crosses a domain, e.g. between two operators. Both technical and administrative/business concerns make such a TE activity more complicated. One example is based on the fact that Border Gateway Protocol version 4 being the (default) standard routing protocol does not carry full information like an interior gateway protocol (e.g. no topology and link state information). In a business sense it would not be likely that two parties, being potential competitors, would reveal all that data of its network. Another aspect is the presence of relevant SLAs that govern the interconnection, including description of traffic patterns, QoS, measurements and reactions. An SLA may explicitly or implicitly specify a Traffic Conditioning Agreement, which defines classifier rules as well as metering, marking, discarding and shaping rules (from TE framework).

A specific TE system can then be categorised by applying the criteria listed above.

5 Key Technical Activities – Overview

A number of basic methods are seen for structured handling of network planning tasks. These could be described as:

- Mathematical, analytical methods:
  - Optimisation: linear programming, integer programming, dynamic programming. Applied to find a better solution (“optimum”) given a number of criteria/constraints, objective functions, parameters to be found.
  - Complexity analysis, e.g. for scalability studies
  - Queueing theory, including Markovian processes. Applied for dimensioning, dependability, etc.
  - Network theory
- Simulations: Commonly used for complex, inter-dependent subsystems as they may become too involved for analytical treatment (or not solvable at all).
- Measurements/testing: Applied for actual implementations of systems in operation or configurations mimicking the system to be placed into operation.

A characteristic of modelling teletraffic phenomena is that fairly simple models can be treated by analytical means. Approximations are commonly introduced implying that the results achieved strictly are valid within certain regions of the system loads. Common simplifications are Poisonian distributed arrival rates, exponentially serving times, independence between arrivals, service periods, servers. These are seldom completely met, but usually allow observations with sufficient accuracy. On the other hand, it is important to be aware of the limitations of the applied methods, avoiding conclusions to be invalidly made.

Analytic algorithms may be used to identify principle behaviour or theoretic phenomena. Some of these observations may be fed into simulation models – also incorporating more dependencies and time-varying behaviour. Measurements, either by a laboratory or in the field, are however from the outset closer to the real system. The measurement system must then mimic the actual behaviour.

Then, commonly, observations from measurements can be abstracted to add to the knowledge as well as potentially be fed back into both analytic and simulation models. The three cate-

gories are mutually supportive and indispensable during the network planning and following the network operation.

Computer systems are regularly needed for network planning tasks of realistic cases because of number of candidates, number of calculations, etc. Another aspect is relations with databases containing information on the networks and systems to be planned. Relevant data must be fetched, sorted and stored for these bases.

Teletraffic theory can be defined as (ref. [ITU-D_TE]): *the application of probability theory to the solution of problems concerning planning, performance evaluation, operation and maintenance of telecommunication systems*. In a more general setting, teletraffic theory can be considered as a discipline of planning where the tools (stochastic processes, queueing theory and numerical simulation) are taken from the disciplines of operations research.

The *objective of teletraffic theory* is formulated as: *to make the traffic measurable in well-defined units through mathematical models and to derive the relationship between grade-of-service and system capacity in such a way that the theory becomes a tool by which investments can be planned.*

Hence, the task is to assist when designing systems given pre-defined grade-of-service levels, traffic demand and capacity of system components. The overall procedure for applying teletraffic theory as an inductive discipline is depicted in Figure 20. The models are established after observing the real systems. A number of parameters are derived, which may be compared to corresponding observations (or predictions) from the real system. In case of (stochastic confidence) correspondence, the model is said to be validated. If not, the model should be elaborated further. In general, this is called the research spiral.

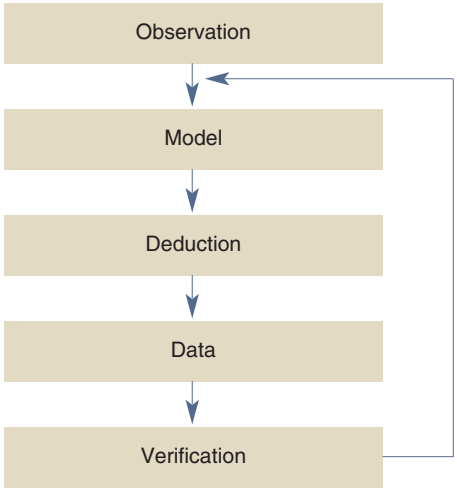


Figure 20 Teletraffic theory as an inductive discipline (from [ITU-D_TE])

The tasks involved in traffic engineering, as seen from ITU-D, are illustrated in Figure 21.

In the following sections a number of key areas and mechanisms are described. For an operational network, all of these areas should be considered.

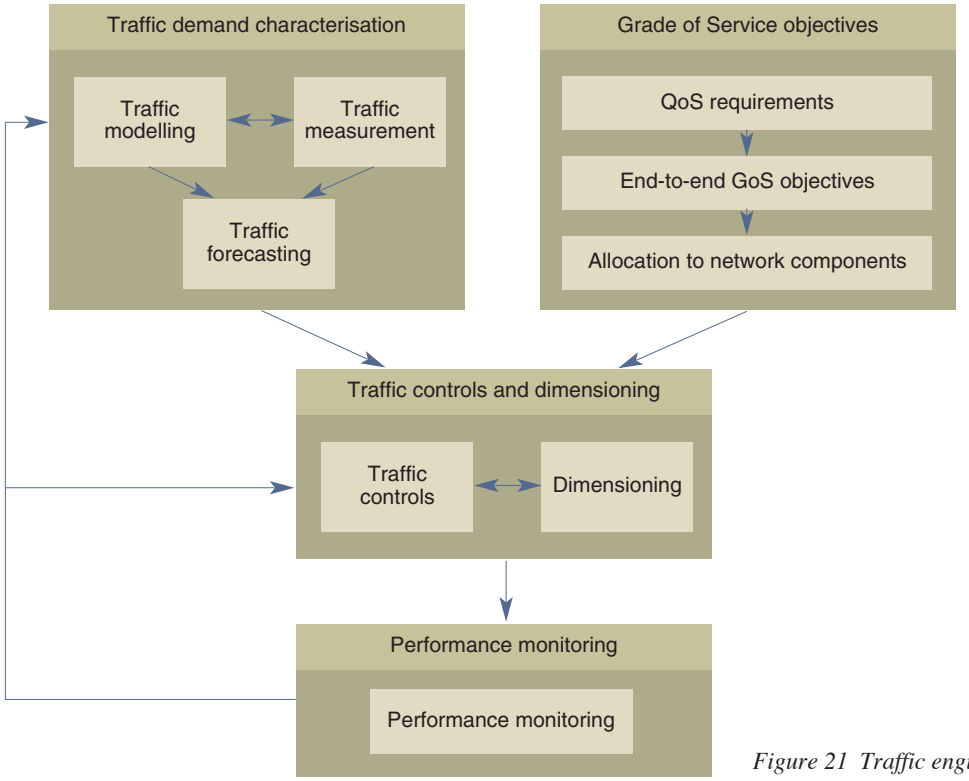


Figure 21 Traffic engineering tasks (from [ITU-D_TE])

6 Basic Traffic Terms

One commonly meets the term traffic when discussion network planning. Traffic is used to denote traffic intensity; that is, traffic per time unit. The definition is: *Traffic intensity – The instantaneous traffic intensity on a pool of resources is the number of busy resources at a given instant of time*. The pool of resources may be a group of servers. Note that one may get association of a discrete number of servers although the term is generally applicable also for a "continuum" of resources (e.g. for link bandwidth).

Statistical moment of the traffic intensity can be calculated for a time interval, say from 0 to T, e.g. for the mean value:

$$A_c(T) = \frac{1}{T}\int_0^T n(t)dt$$

where $n(t)$ expresses the (number of) occupied servers at time t.

Often the unit used for traffic intensity is erlang (referring to the Danish mathematician E.K. Erlang, 1878 – 1929). This unit is actually dimensionless.

Carried traffic, A_c – The amount of traffic carried by the group of servers during a time interval. This can be measured by observing the amount of servers being busy in that interval.

The total traffic carried during a time interval is named the traffic volume, e.g.

$$\int_0^T n(t)dt,$$

being equal to the sum of all occupation times inside the time interval. The carried traffic can never exceed the capacity of the server pool (i.e. a single server unit can never carry more than one erlang).

Offered traffic, A – The traffic that would be carried if no requests were to be rejected due to lack of server capacity. Two types of parameters are often seen to describe traffic:

- The request intensity, λ, which is the mean number of requests offered per time unit.
- The mean service time, s, for a request (note that this must be given with the same time unit as the request intensity).

The offered traffic is then equal to $A = \lambda \cdot s$. Again, it is recognized that traffic is dimensionless.

Rejected traffic, A_r – The difference between offered traffic and carried traffic.

That is, $A_r = A - A_c = A \cdot (1 - b)$, where b is the probability of a request being blocked/rejected.

In the general case, carried traffic is the only measurable dimension (as one does not actually know service times for requests being rejected).

Assuming that a system (pool of servers) has capacity C, the normalised load is given by

$$\rho = \frac{\lambda \cdot s}{C}.$$

The observed utilisation (i.e. referring to carried traffic) is in the interval [0,1].

In a multiservice system different request types may occupy different amounts of server unit capacity. This must be considered when the traffic is calculated. Assuming that traffic type i occupies a server capacity d_i, the offered traffic for K request types is found as:

$$A = \sum_{i=1}^{K} \lambda_i \cdot s_i \cdot d_i$$

Traffic characterisation is often done by establishing models that approximate the stochastic behaviour of the traffic flows. As few traffic parameters as possible is strived for (e.g. mean value, variance, etc.) while still being able to capture the pertinent characteristics of the traffic flow. In fact, the key trick in traffic modelling is to identify the simplifying assumptions to be made and the set of parameters to use.

When possible, measurements are conducted to validate the traffic models. On the other hand, measurements could only be done in order to assess the parameter values to be used, for example when the traffic models as such are fixed or applied due to previous experience.

For an expected future situation forecasting techniques are required in order to estimate the traffic demand. A list of ITU-T recommendations in this area is given in Box A.

Examples of characteristics of the traffic are persistence (on-demand or permanent), information transfer mode (circuit switched or packet switched), configuration (point-to-point, multipoint-to-multipoint, etc.), transfer bit rate, QoS parameter requirements, symmetry (unidirectional, bi-directional asymmetric, bi-directional symmetric), and so forth. These are all related to the session and connections being parts of the session.

A session can be initiated according to a process – such as the request initiation process – for example given by the arrival rate or duration.

Going further into the time scale, phenomena relating to individual information units (e.g. packets) are considered. These would be packet interarrival times or packet lengths.

A further dimension is needed for mobile systems where the users' whereabouts also need to be considered. That is, the spatial effect has to be taken into account.

The offered traffic varies as a function of time – in accordance with the activity of the environment of the system considered. For PSTN/ISDN the term Busy Hour has been defined. This refers to the highest traffic during a day. The Time Consistent Busy Hour (TCBH) is defined as those 60 minutes (determined with an accuracy of 15 minutes), which has the highest average traffic over a long period. With this definition it may well happen that the traffic offered in the busiest hour on some days is higher than the traffic in the TCBH.

The traffic variations can also be divided into different time scales such as:

- The 24 hours daily variations
- The weekly variations
- The yearly variations; some days or seasons are expected to have higher traffic such as New Year's Eve
- Traffic trends (increases or decreases)

Depending on how the system at hand operates, we may distinguish loss-systems from waiting-time systems. For the former, requests are rejected if all servers are occupied; this is typically the case for calls offered to trunk groups. For waiting-time systems the request is placed into a queue if it finds the servers occupied upon arrival.

A loss-system has a number of indicators to express the level of inadequacy:

- Request congestion: The fraction of requests that observes all servers busy upon arrival.
- Time congestion: The fraction of time all servers are busy.
- Traffic congestion: The fraction of the offered traffic that is not carried.

Typically, these indicators are used to establish dimensioning standards for server groups.

An inconvenience of waiting-time systems is the resulting waiting time. Not only the mean waiting time is of interest, but also the distribution of the waiting time. For example some short waiting time would barely be noticeable. The relation between the waiting time and the level of inconvenience for a user is not obvious.

Having finite financial means, it is often not cost-effective to build a system that meets any

Box A – Selected ITU-T Recommendations in the Area of Traffic Demand Characterisation

Traffic Modelling

- E.711 – User demand modelling
- E.712 – User plane traffic modelling
- E.713 – Control plane traffic modelling
- E.716 – User demand modelling in Broadband-ISDN
- E.760 – Terminal mobility traffic modelling
- E.523 – Standard traffic profiles for international traffic streams

Traffic Measurements

- E.490 – Traffic measurement and evaluation – general survey
- E.491 – Traffic measurement by destination
- E.500 – Traffic intensity measurement principles
- E.501 – Estimation of traffic offered in the network
- E.502 – Traffic measurement requirements for digital telecommunication exchanges
- E.503 – Traffic measurement data analysis
- E.504 – Traffic measurement administration
- E.505 – Measurements of the performance of common channel signalling network
- E.743 – Traffic measurements for SS no 7 dimensioning and planning
- E.745 – Cell level measurement requirements for the B-ISDN

Traffic Forecasting

- E.506 – Forecasting international traffic
- E.507 – Models for forecasting international traffic
- E.508 – Forecasting new telecommunication services

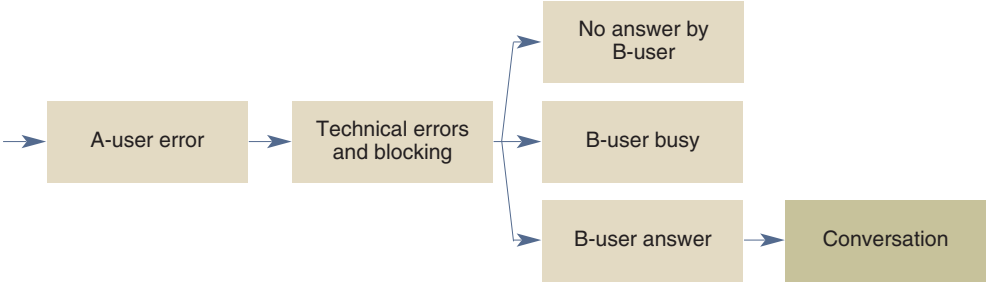


Figure 22 Sequence of potential events during a PSTN set-up

imaginable situation of request rates. One of the objectives of teletraffic theory is to find relations between the offered traffic, the system capacity and the resulting service level.

For a loss-system a block-diagram (in several ways similar to the ones used for dependability analyses) can be set up. An example is shown in Figure 22. Such diagrams may also be found for other services, e.g. for accessing a web page/site when browsing, although this service is rather implemented by waiting-time servers.

Naturally, the schema depicted in Figure 22 can also be used to study repeated call attempts. That is, the users' response to unsuccessful requests impacts the resulting load on the systems and hence an essential part of dimensioning the system.

One may make a distinction between characterisation and modelling, ref. [Macf02]. Characterisation refers to a phenomenological description of traffic, i.e. the analyses of measurement data, and the production of high-level statistical descriptions. Modelling covers the detailed low-level production of probabilistic models of traffic that can actually be used to make predictions about network and system behaviour. These two perspectives come together by the requirement of compatibility – the high-level profile derived from the low-level model must fit the experimentally observed characterisation.

On the shorter time scales, traffic descriptors are used as inputs for performance estimation and system sizing. A number of usefulness criteria should be fulfilled for such descriptors [Macf02]. Of most basic nature, the descriptors should allow for mathematically tractable handling and resulting in useful results. Moreover, they should be:

- Stable – a given stationary traffic flow should not have wildly fluctuating parameters from one day to another;
- Parsimonious – only a small number of parameters should be necessary;
- Comprehensible – the significance of the parameters should be easily understood;
- Aggregatable – the parameters of the superposition of two flows should bear a simple relation to those of its components;
- Scalable – natural growth in traffic should not result in complex changes in parameters.

7 Controlling Network Load

Generic requirements on the behaviour of a network overload control include:

- Convergence to a state which maximises the throughput of an overloaded resource subject to keeping its response times short enough to reduce the occurrence of user abandons. Achieving such convergence automatically over a realistic range of overload scenarios (including a range of overloaded resource capacitates and number of active traffic sources).
- Minimising ineffective traffic flows by selectively throttling flows experiencing high dropping ratios.
- Fairness between traffic flows in a certain sense.

Overload situations are typically stimulated by a few events:

- Media-stimulated, e.g. invitations to initiate sessions/accessing pages, casting votes, competitions, marketing;
- Emergency situations;
- Network equipment failures, including software/configuration errors;
- Auto-scheduled session initiations, including Denial of Service attempts.

In the absence of effective control mechanisms, such overload-triggering events would threaten the stability of network systems and cause a severe reduction in service levels. Ultimately, systems might fail and the service would not be available to users.

End-users are commonly intolerant with respect to long response times to their service requests. Typically, when a user has to wait more than a few seconds, re-attempts are made or the service request is abandoned altogether. Similar phenomena are found for different types of services, such as telephone calls, web page accesses and so forth.

User impatience combined with an excessively long response time can cause the effective throughput of a network resource to collapse to a low level when the original request load is high enough. This may come from the resource spending more of its capacity on partially processing the requests, not finalising any. Hence, a huge backlog of requests can occur, while most of the users' requests are not being completed before a certain threshold is met (the user releases, re-tries, protocol time-out, etc.).

The user model of this is often assumed to be described by (see Figure 23), i) the persistence

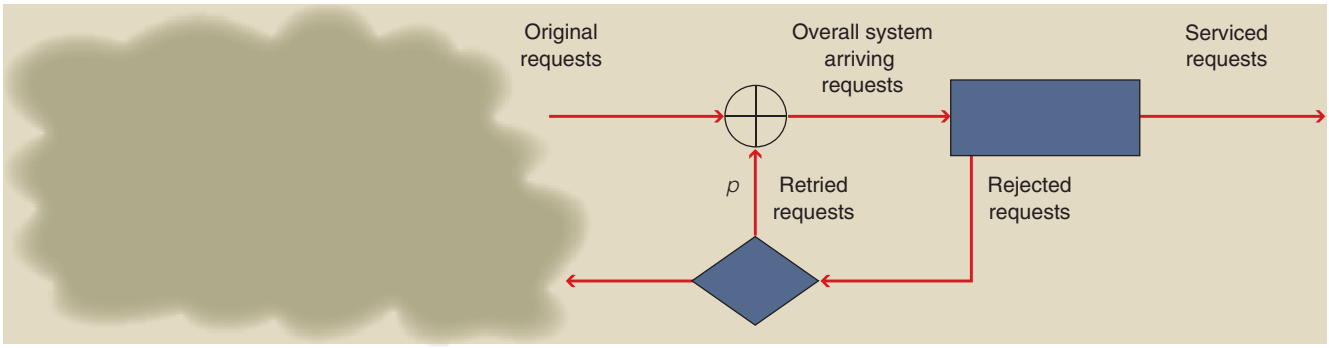


Figure 23 Schematic illustration of re-attempts contributing to the overall load on the system

probability, say p_i, that a new attempt will be made if the i^{th} attempt is not serviced, and, ii) the distribution of the time interval between the i^{th} and $(i+1)^{th}$ attempt.

If all attempts fail, then the mean number of attempts per initial intent is given by the series $1 + p_1 + p_1(p_2 + p_1(p_2(p_3 + \ldots$. If all $p_k = p$, this adds to $(1-p)^{-1}$. As an example, assume that p = 80 % (which is observed for some media-stimulated events. Then, on average five attempts per original intent is seen (if all attempts fail). Although being a simple example, it is seen that such phenomena easily result in a drastic increase in the number or requests on a resource. Hence, user persistence may lead to significant higher load, which even may result in overload in other parts of the network due to that those resources are not freed in order to serve requests that are successfully completed.

Expanding on the notion of an unsuccessful request:

P – the persistence
B – the probability of not completed service (e.g. no reply by B-side)

Say that a request requires s time units service when not completed and t time units when completed. Then the following can be seen:

Occupancy time when successful first time:
s; probability: $(1-B)$

Occupancy time when successful second time:
$(t+s)$; probability: $B \cdot p \cdot (1-B)$

Occupancy time when successful third time:
$(2 \cdot t + s)$; probability: $(B \cdot p)^2 \cdot (1-B)$

...

Occupancy time when successful i^{th} time:
$(i \cdot t + s)$; probability: $(B \cdot p)^{i-1} \cdot (1-B)$

The average occupancy time can then be found to be:

$$\frac{(1-B)\cdot(t+s\cdot(1-B\cdot p)}{(1-B\cdot p)^2}$$

Keeping the notion of A_c for carried traffic and A for offered traffic, $A = \lambda \cdot s$, gives:

$$A_c = A\cdot(1-B)\cdot\left(\frac{t/s+1-B\cdot p}{(1-B\cdot p)^2}\right)$$

That is, the "carried traffic" can very well become higher than the originally offered traffic, which may impact the dimensioning. Several means could be introduced to combat this situation, e.g. to reduce the persistence probability (e.g. providing a prompt to the user that re-requests should be postponed) or to increase the completion probability (e.g. attempted by providing a voice-mail service).

During varying request mixes/demand patterns, different types of resources in the network may become the bottleneck. In particular, within a network node, the resources may have different tasks; some taking care of peripheral interfaces/devices, some protocol processing, others routing updates, etc. Preserving a controlled service processing, it is essential to implement an adequate overload control mechanism. This mechanism is to reject requests in order to keep response times low and ensure service completion at satisfactory levels.

When media-stimulated events trigger a mass-call (voting, quiz, etc.) the call arrival curve usually has a fairly steep edge to a peak and a slower decrease, as illustrated in Figure 24. An essential feature of the overload mechanism is to avoid that the peak results in a too severe degradation in the service on this call type as well as the other call types utilising the same resources. Hence, the overload control kicks in at a certain load level – which is again related to a certain level of the call arrival rate. Then, a difference between the offered calls and the admitted calls can be observed at high load intervals.

The relation between offered and admitted calls is shown in Figure 24. Naturally, the exact shape of these curves depends on the resource types

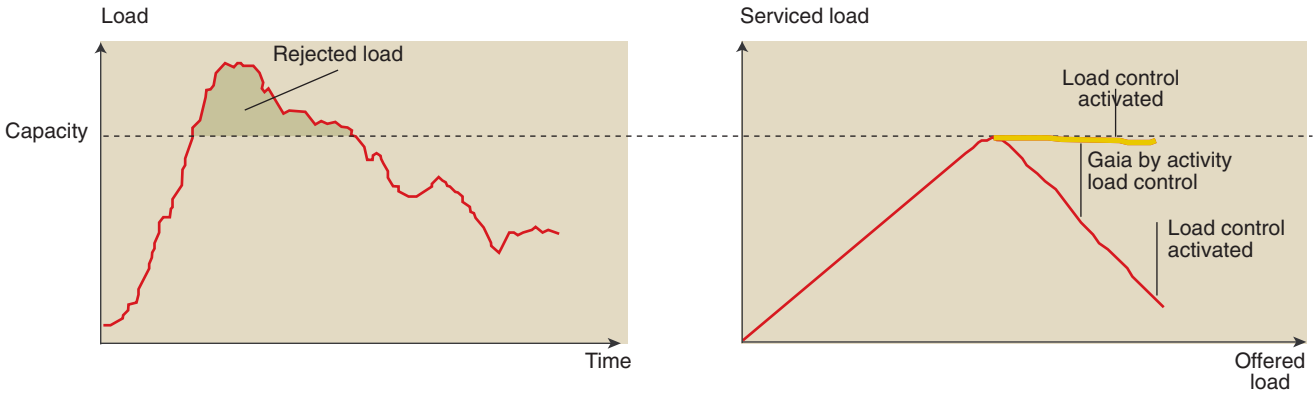


Figure 24 Illustration of request arrival for mass-calls and gains from load control mechanisms

and call types in question. In general, however, as the offered request rate increases, a point will be reached where the overload control is invoked and some part of the offered load is rejected in order to preserve the service capacity. As the load increases even further, the relative share of admitted requests decreases, although the trick is to keep the rate of admitted requests at almost the same level. Again, this depends on the implementation. For example, in case the same processor is used to run the overload algorithm, some initial processing would likely be needed for all arriving requests. This is to see what to do with them as well as to complete processing/serving of requests. Due to the initial processing per request the single processor will eventually be overloaded. That is, the rate of completed services will decrease, although this should happen at a higher load than would occur when no overload control was activated. In order to prevent the rate of completed services from decreasing, the load control – as well as any initial processing of every request – should be placed externally to the bottleneck resource. In order for the overload control to be effective, it needs to be automatic and embedded in the network resource. This is to react sufficiently rapidly to changes in demand or resource capacity. The mechanism must be able to handle variations in the request mixtures and the way requests are distributed over the relevant resources. For network planning objectives, all the parameters must be coordinately assigned in correspondence with targets for the different request types (reflected into service classes). These should then be according to the SLA conditions.

A number of requirements for overload controls are (adopted from [Whit02]):

- Maximise effective throughput of the bottleneck resource, during all ranges of offered request rates.

- Bounding response times related to the bottleneck resource, in particular during overload intervals. This implies that the overload controls have to react fast enough to prevent that the total load too greatly exceeds the processed load during the load increase phase.

As described earlier, using the reject rate as the controlled variable leads to convergence in the admitted rate. This is frequently combined with a leaky bucket algorithm due to its fairly simple implementation and effective response.

Overload controls are also expected in several coming applications, such as service control, e.g. by SIP or HTTP requests. On the other hand, it is important to distinguish between messages. For example, messages releasing resources (e.g. circuit disconnect message) should not be throttled as that could further worsen the load situation.

8 Traffic Controls and Dimensioning

An objective of traffic engineering is to provide efficient design and operation of the network while assuring that the demand is served within the requirements stated. Network dimensioning assures that the network has adequately configured resources, including the physical network elements and the logical network elements (e.g. virtual paths).

Several items are included in the traffic controls category ([ITU-D_TE]):

- Traffic routing: giving the set of routing choices and rules for selecting routes for different traffic types (including different source-destination pairs). Routing principles may be flat, hierarchical, fixed, dynamic, state-based, event-based, etc.

- Network traffic management controls: ensuring that the traffic throughput is maintained under overload and failure conditions. These may be restrictive or expansive. Restrictive controls, such as blocking or call gapping, throttle the traffic sources. The expansive controls typically reroute the traffic towards the network portions that are less loaded. Often,

real-time monitoring of the performance/load levels triggers these actions related to predefined levels. For example, when the number of calls on a link is equal to a given threshold, additional calls are rejected.

- Service protection methods: controlling the grade of service of certain traffic types by utilising discriminatory restriction of access to resource groups.

One example is to balance GoS between traffic types requesting bandwidth. Another example is to provide stability in networks without hierarchical routing schemes by restricting overflow traffic to an alternative route that is shared with firsts-choice traffic.

When several traffic types compete for the network resources, the requirements of the individual types have to be checked to ensure that their service levels should be met. Several systems support prioritisation of traffic, including separate criteria for admission control, scheduling and dropping rules.

Designing systems that continue to operate at maximum capacity even during overloads, various overload strategies can be introduced. For processing and signalling tasks leaky bucket, gapping and level priorities are seen. For information transfer modes, bandwidth reservation and circuit channels protection can be used, besides various routing separations.

Bandwidth reservations can for example be used in hierarchical networks with alternative routing where the primary traffic could be protected against load from overflow traffic (overflow traffic refers to flows that are rerouted onto secondary choices, e.g. when the primary route is congested). A common rule is that overflow traffic is only accepted if more than a given bandwidth is available.

Virtual channel protection includes two schemes; i) each traffic type is allocated a minimum bandwidth ensuring a minimum service level, and, ii) a maximum allocation per traffic type avoiding that a single traffic type dominates.

Packet-level traffic controls assure the packet-level GoS objectives under any network condition and that a cost-efficient GoS differentiation is made between services with different packet-level QoS requirements. A selection of ITU-T recommendations for traffic controls and dimensioning is listed in Box B.

In some respects, an incorrectly dimensioned network will have a degraded performance level in a similar manner as a network experiencing failures. Hence, dependability analyses and traffic analyses should be considered together. For dependability analyses the phenomenon that a fault may appear might be modelled as a process with given distributions, as well as the repair time of the failure. A fairly simple model of a system is expressed by a block diagram, where all essential components are reflected. Then, series and parallel branches are resolved to estimate the overall system dependability.

A growing part of telecom systems is made of software modules in different nodes interacting. Considering replication of data and program routines, a certain level of dependency may be present although from a (hardware) point of view the resources are completed. It is vital to consider the commonalities and dependencies in a dependability model to avoid that incorrect conclusions are drawn.

In one way, technical risk analyses are related to the dependability. Then the risk can be defined as the product of the probability and the consequence of an event. The overall risk becomes the sum over all the events. In several cases, quantitative estimates of the events are challenging to estimate, resulting in risk diagrams having consequence and probability on the axes – cate-

Box B – Selected ITU-T Recommendations for Traffic Controls and Dimensioning

Circuit-switched Networks

- E.510 – Determination of the number of circuits in manual operation
- E.520 – Number of circuits to be provided in automatic and/or semi-automatic operation, without overflow facilities
- E.521 – Calculations of the number of circuits in a group carrying overflow traffic
- E.522 – Number of circuits in a high-usage group
- E.524 – Overflow approximations for non-random inputs
- E.525 – Designing network to control grade of service
- E.526 – Dimensioning a circuit group with multi-slot bearer services and non overflow inputs
- E.527 – Dimensioning a circuit group with multi-slot bearer services and overflow traffic
- E.528 – Dimensioning of digital circuit multiplication equipment (DCME) systems
- E.529 – Network dimensioning using end-to-end GoS objectives
- E.731 – Methods for dimensioning resources operating in circuit switched mode

Broadband Integrated Services Digital Network

- E.735 – Framework for traffic control and dimensioning in B-ISDN
- E.736 – Methods for cell level traffic control in B-ISDN
- E.737 – Dimensioning methods for B-ISDN

Signalling and Service Control

- E.733 – Methods for dimensioning resources in Signalling System no. 7 networks
- E.734 – Methods for allocating and dimensioning Intelligent Network (IN) resources

gorised by small, medium and high. The risk profile of a company then decides which measures to apply trying to eliminate the events that are most unwanted.

A part of the network planning area is to formulate contingency plans. These may be expressed as; what to do when XX happens. Then, responsible units and their actions have to be clear to the parties involved. Moreover, all XX events should be captured. Examples of events are fire in a station, major vehicle accidents involving police, ambulance, etc.

9 Forecasting

Considering the steadily changing environment of a telecom actor, it is essential to have ideas on what the future situation will be like. In particular this is relevant for the demand patterns. Various forecasting techniques are applied to get an objective view of the future situation – this is commonly done by looking at the forces working in the past, present and future periods.

Two basically different approaches for making forecasts are:

- Quantitative – including trend and factor methods, analogy methods and potentials. A trend is a statistical description of a historic and relevant development that is prolonged into the future. A basic precondition for doing this is that factors that influence the development have been examined and their potentially changing magnitudes in the future captured in the predictions. An analogy method can be applied when the case at hand is believed to follow a similar evolution as another known case (e.g. another region/nation, another product, another customer segment, etc.). This method may also be applied when relating products with other industries. A potential for a product or customer segment (also referred to as saturation level) could possible be transferred from other industries or other product groups. For example, the number of internet subscriptions will not likely exceed the number of PCs.

- Qualitative – including perspective and scenario analyses, and interviews and questionnaires (such as Delphi and expert panels). The perspective method aims at describing likely development tracks commonly based on discussions on technological, market, social and economic evolution in the period. Hence, this will take today's standpoint and derive the potential future based on that view. In a scenario method a more or less remote future is taken on. Potential alternative futures – scenarios – are then elaborated, possibly without describing what happened between today and that point in the future. Hence, for a scenario one may say that a future standpoint is taken looking back towards today to reveal the causes of that scenario to appear. A characteristic of Delphi and expert panels is that several rounds are carried out, where the participants may get feedback on the statements of others in between. This is to try to achieve a convergence on opinions.

Traffic forecasting is necessary both for strategic studies as well as when designing systems on a medium term. Broadly, two basic approaches might be identified for forecasting:

- Descriptive approach, where the basic assumption is that the underlying process generating the demand in the past continues into the future. Statistical model is for example applicable in this case, whereby the time series (or demand history) is analysed for components such as the base or constant component, trends, etc. Then, this description is extrapolated into the future. Naturally, knowledge of any impending events should be included in the extrapolation.

- Explanatory approach, where casual relationships are built into the model. Several models can be found in the literature, such as wide econometric models to models predicting the purchasing patterns for a group of customers.

For all the forecasting a random error component (sometimes referred to as noise) is present. This composes the unexplained deviation of the data from the basic demand generating process.

When returning to the descriptive models, the following patterns are commonly considered when relevant, see Figure 25:

- Base – representing a central tendency of the time series at any given time

- Trend – exhibiting a consistent long-run shift of the average

- Seasonal – showing a repetitive variation, e.g. a one-year periodicity. This is commonly superposed upon a background process, e.g. showing a trend.

As time passes, new information is gained and the models can be updated. When doing this either the time origin can be fixed or the time origin can be shifted to the end of the current period. In the former, the parameters have to be recomputed every time in order to incorporate the latest observations. In the latter, also called adaptive smoothing, the parameters are updated by combining the old estimates and the latest forecast error recorded.

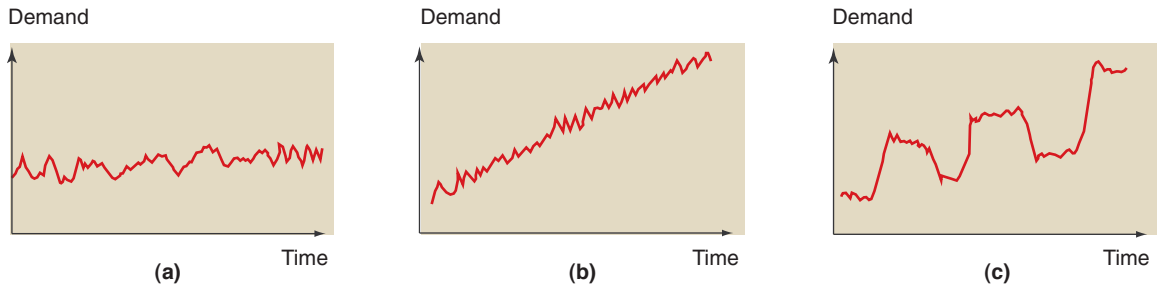


Figure 25 Demand patterns: (a) base, (b) trend, (c) trend and seasonal variations

A few forecasting techniques are:

- Regression methods – fit some hypothesized model, linear in its coefficients, to the time series. Least square estimators are commonly used for the coefficients. Multiple regression can also be used as an explanatory method of recasting when a causal relationship between the dependent variables and the independent variables is present.

- Moving average – computes an average of the constant number of past observations in order to eliminate the random variations or noise. As a new observation becomes available, the oldest observation is dropped (in order to keep their number constant) and a new average is computed.

- Exponential smoothing – assigning different weights to the observations according to an exponential decaying factor as the time distance increases. Hence, longer (historic) values get lower impact on the forecasted value.

- Bayesian methods – useful when faced with a lack of historical data at the beginning of the forecasting process. The approach is then to start with some initial subjective estimate of the parameters in the model, and use of Bayes' theorem to modify them in light of the actual data observed.

- Box-Jenkins models – set of autoregressive models in which successive observations are highly dependent. Three classes of models are considered; i) the autoregressive process, ii) the moving average (or the errors) process, and iii) the mixed autoregressive-moving average process. The choice of the correct model is made by examining the autocorrelation coefficients.

Both quantitative and qualitative forecasting techniques can be exercised. The first group includes models based on smoothing models, time series models (Box Jenkins approach), Kalman filter models, Regression models, Logit models, diffusion models and other econometric models. Examples of qualitative techniques include market surveys, scenario methods, analogy methods, expert methods and Delphi methods.

In case there are no historical data available, qualitative forecasting models are most relevant. Time series models and Kalman filter models are not relevant as they need a substantial number of (pre-)observations.

It is essential to know whether the peak periods for different services for different customer groups coincide as this affects the capacity needed in the network. This is illustrated in Figure 26, demonstrating a daily profile for residential and business customers.

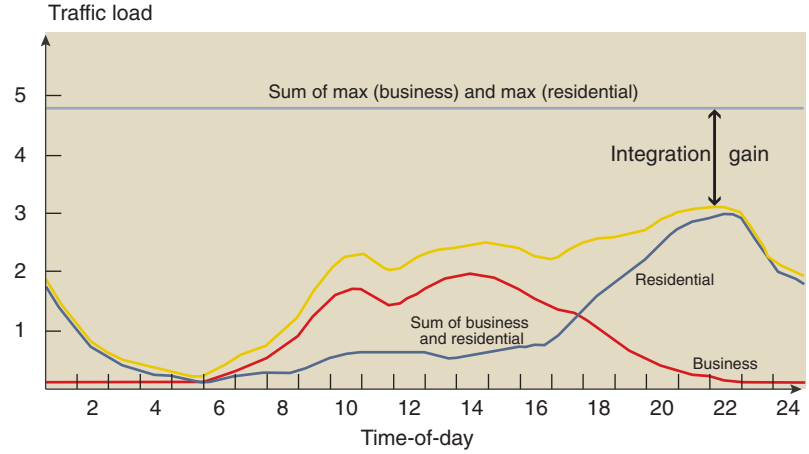
The averaging effect when measuring for a given interval must be handled. Some effects are illustrated in Figure 27.

ITU-T recommendations in the forecasting area are:

- E.506 – giving directions on forecasting including base data, social and demographic data. In case some samples are missing, guidance for obtaining these elements is found.

- E.507 – presents an overview of mathematical forecasting techniques: curve-fitting models, autoregressive models, autoregressive integrated moving average (ARIMA) models, state space models with Kalman filtering, regression models and econometric models. This recommendation also contains methods for evaluating the forecasting models and how to choose an appropriate one depending on available data, forecast period, and so forth.

Figure 26 Daily traffic profiles for different customer segments



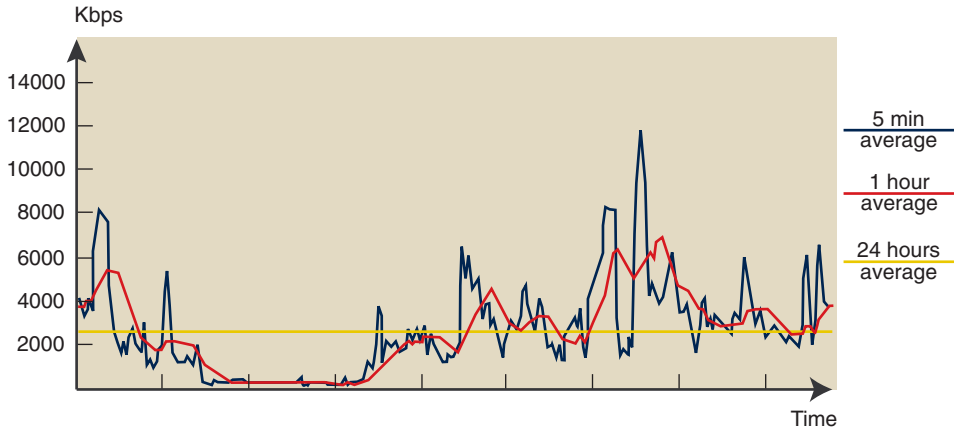


Figure 27 Effects from averaging over intervals (from [Soto02])

- E.508 – describes how to forecast new services where there are no historical data. Techniques described include market research, expert opinion and sectorial econometrics. The recommendation also discusses how to combine results from different forecasting techniques and how to adjust the results when first observations are obtained.

A central part of deriving demand characteristics is to elaborate a set of traffic matrices. These matrices give the amount of traffic flowing between sources and destinations. For N nodes N^2 traffic elements may complete the matrix. Basically, these traffic matrices do not give any information of the network topology or the traffic routing. Two approaches for obtaining a traffic matrix are i) direct measurements of source destination pairs, ii) partial information to infer the matrix elements.

In several networks, measurements of aggregates are carried out, such as the total traffic originating or terminating in a node. The individual elements in the traffic matrix may not be known, however; that is, the traffic between specific node pairs. Full details from measurements are not frequently seen as they involve a significant amount of measurement infrastructure, data storing and processing, hence likely to become a costly approach. Therefore, other procedures have been proposed only based on certain known data. These procedures are then started from these informed priors (referring to given information on the traffic matrix).

In a network with n nodes there are generally $n \cdot (n-1)$ source–destination pairs involving links in the network. If $\boldsymbol{L}$ is the traffic on links between nodes, $\boldsymbol{A}$ the traffic matrix between node pairs and $\boldsymbol{R}$ the matrix indicating whether a link is used on the way between a source and a destination, this can be written as: $\boldsymbol{L} = \boldsymbol{R} \cdot \boldsymbol{A}$. The problem, however, is that usually there are many more source–destination pairs than links in a network, leading to the equation being undetermined when the traffic matrix $\boldsymbol{A}$ is to be found (hence, theoretically there is an infinite number of feasible solutions for the traffic matrix).

Assuming that the total traffic originating and terminating in nodes are known, this is the same as saying that the column and row sums in the traffic matrix are known. Then, applying Kruithof's method, individual matrix elements can be estimated. The overall procedure is to iterately adjust matrix elements trying to make the column sums and the row sums to fit the predescribed values, respectively. Applying Kruithof's double factor method, a matrix element between two iterations is adjusted by the formula:

$$A_{ij}^{k+1} = A_{ij}^{k} \cdot \frac{S_t}{S_k}$$

where

A^{k+1}_{ij} – the adjusted element value in the traffic matrix (iteration $k + 1$)

A^{k}_{ij} – element value in the previous iteration

S_k – actual sum of elements in iteration k, for either column or row

S_t – target (column or row) sum to be achieved

Other approaches have also been proposed to estimate the individual traffic elements, but Kruithof's double factor method fulfils objectives such as (based on [ITU-D_TE]):

- Uniqueness – only one solution exists for a given forecasts.
- Reversibility – the resulting matrix can be reversed to the initial matrix with the same procedure.

- Transitivity – the resulting matrix is the same independent of whether it is obtained in one step or via a series of intermediate transformations.

- Invariance – regarding the sequence of nodes; their sequence may be changed without influencing the results.

- Fractioning – the individual nodes can be split into sub-nodes or be aggregated into larger nodes without influencing the result. Note that this is not completely met for Kruithof's double factor method, but deviations are small.

10 Performance Objectives and Service Constraints

Grade of Service (GoS) is defined as a number of traffic engineering parameters to provide a measure of adequacy of plant under specified conditions. GoS parameters are expressed as probability of blocking delay distribution, etc. Network Performance (NP) is defined as the ability of a network or network portion to provide the functions related to communications between users. GoS is the traffic-related part of NP, although NP also covers non-traffic related aspects (including dependability, charging, etc.).

NP and GoS objectives are derived from Quality of Service requirements. Basically, QoS should be user-oriented and, in principle, independent of the network. NP parameters though, are network-oriented; i.e. they can be used when specifying performance requirements.

A common approach followed so far by ITU-T is to define a so-called reference connection in order to derive GoS conditions for a network portion. Then, an end-to-end connection is partitioned into portions and the QoS requirements are divided into contributions from each of the portions. In several cases, an interface between two portions refers to an interface between two operators/providers. A number of recommendations address reference connections, including:

- E.701 – Reference connections for traffic engineering

- E.751 – Reference connections for traffic engineering of land mobile networks

- E.752 – Reference connections for traffic engineering of maritime and aeronautical systems

- E.755 – Reference connections for UPT traffic performance and GoS

- E.651 – Reference connections for traffic engineering of IP access networks

As described above, defining the reference connection is followed by elaborating GoS objectives. A number of recommendations are found for this purpose too:

- E.540 – Overall grade of service of the international pert of an international connection

- E.541 – Overall grade of service for international connections (subscriber-to-subscriber)

- E.543 – Grades of service in digital international telephone exchanges

- E.550 – Grade of service and new performance criteria under failure conditions in international telephone exchanges

- E.720 – ISDN grade of service concept

- E.721 – Network grade of service parameters and target values for circuit-switched services in the evolving ISDN

- E.723 – Grade-of-service parameters for Signalling System No. 7 networks

- E.724 – GoS parameters and target GoS objectives for IN Services

- E.726 – Network grade of service parameters and target values for B-ISDN

- E.728 – Grade of service parameters for B-ISDN signalling

- E.770 – Land mobile and fixed network interconnection traffic grade of service concept

- E.771 – Network grade of service parameters and target values for circuit-switched land mobile services

- E.773 – Maritime and aeronautical mobile grade of service concept

- E.774 – Network grade of service parameters and target values for maritime and aeronautical mobile services

- E.775 – UPT Grade of service concept

- E.776 – Network grade of service parameters for UPT

- E.671 – Post selection delay for PSTN/ISDNs using Internet telephony for a portion of the connection

- E.651 – Reference connections for traffic engineering of IP access networks

- Y.1540 – Internet protocol data communication service – IP packet transfer and availability performance parameters

- Y.1541 – Network performance objectives for IP-based services

Looking into these one finds a separation between the user data performance and signalling performance. Examples of parameters of the former are information transfer delay, information delay variation, information loss ratio, etc. Examples of the latter are set-up delay, disconnection delay, connection set-up blocking, etc. In most cases the GoS requirements refer to a fully operational network, not taking into account network failures. One exception is E.550 which combines the effect of availability and traffic congestion and defines parameters and target values that consider their joint effect.

11 Performance Testing

Performance testing belongs to the non-functional verification of a system operation. That is, not the *does the system deliver a proper result* (being a functional examination), but rather *how does the system deliver a proper result*. There may be several goals for carrying out performance testing, including:

- Identify bottlenecks and determine configuration for maximal performance

- Determine capacity of a system and effects under overload

- Examine effects of introducing additional modules, new releases, etc.

One advantage by performance testing is that it could work on the final system, avoiding testing on models which might – or might not – fully reflect the behaviour of the system. However, this is also a drawback as a complete testing then must be delayed until the system is fully configured – coming rather late in the development cycle. Moreover, the testing may become more expensive compared to testing on a model. This may also come in addition to real traffic being offered to the system, potentially impacting the customer. Due to the steady pressure on releasing new systems/products, vendors often choose not to conduct full testing, but rather restrict the examinations to so-called representative tests. However, these tests may not fully cover the operational situation the system might experience.

Performance modelling is the abstraction of a real system into a simplified representation to enable the prediction of performance. Note, however, that this term commonly means different things to different people. When new systems are to be introduced the components' capacity should be known. This can be obtained by measuring, under controlled conditions, or by modelling the components before the system is deployed. This allows for a more accurate dimensioning and provides a basis for following the performance during operation.

Performance and traffic analyses should be carried out for several phases of a system life, including:

- Design – the design principles may be studied to find which ones to choose

- Delivery – the specifications are to be verified by the receiver of a system

- Operation – performance under varying conditions, also when the system is enhanced or modified in other ways.

When an actual system is to be modelled, the amount of technical specification and data to be captured may be overwhelming, although more intriguing to concisely formulate in a model. In principle, a balance between the completeness of the model and its tractability can be faced. This makes the art of system modelling even more pleasing.

One goal of including performance engineering at all stages of the product life-cycle is to prevent problems. This also applies to systems in operation. Measuring during operation also gives additional insight as successful services tend to grow whereas unsuccessful ones usually shrink. Monitoring, comparing results to forecasts and the capacity can give early warnings of any system enhancements to be planned. Depending on the lead-time for enhancements this can be crucial during operation.

Detecting faults is also a result of performance engineering. Fault-tolerant systems, distributed systems and data inconsistence etc. may lead to degradation only, which could be detected with performance testing – possibly assisted by tracing capabilities.

In one way, when a system's data and processing become more distributed, the performance risk does as well. Rapid development and autogeneration of program code may also add to the performance risk. This underlines the need for performance engineering, in particular the performance prediction using adequate tools and models. The following benefits of performance modelling are listed in [Sing02]:

- Relatively inexpensive prediction of future performance
- Design support allowing objective choices to be made
- Decision support for further development of existing systems
- A clearer understanding of a system's performance characteristics
- A mechanism for risk management and reduction.

Relations with different phases of the system development are illustrated in Figure 28.

For analytic treatment, simulation and measurements several common aspects are seen, including ([Sing02]):

I Model requirements – what questions are to be answered by the model

II System understanding – gathering all the performance characteristics of the system or process

III Model design and development – translating the system design into a model design, determining what to include and to exclude

IV Data collection – usually takes the most (elapsed) time and should be started as soon as possible, e.g. low-level system data not readily available, such as the number of CPU seconds taken by a process to perform a particular task

V Model verification – ensuring the model design reflects the system design by testing the model at various levels and model walkthroughs in conjunction with the system designers, developers and users

VI Model validation – if possible running the model to reflect the real system using the same input and checking the similarity of the output "what if" scenarios, predictions and answers to the questions – exercising the model with specific inputs designed to answer the questions defined in the model r equirements

Iterations are often done for several of the stages in the process, e.g. as requirements are changed, a greater understanding gained and improved data quality of data is gathered. Simpler models can be produced early to give an initial prediction. As time evolves more design information and input data are available, models could be refined and accuracy gained.

12 Measurements and Monitoring

Basically, measurements are carried out in order to obtain quantitative estimates about the system behaviour, traffic characteristics, performance levels and so forth. A number of objectives of measuring can be identified, including estimating performance, characterising traffic flows, charging inputs, documenting SLA conditions, and so forth. Hence, different measurement schemes have been defined. Although a network may be correctly dimensioned, overload and failure situations are likely to occur where network traffic management actions have to be initiated. Moreover forecast errors or approximations, statistical variations in the demand, also lead to degraded service levels. GoS monitoring is needed to detect these problems and provide feedback for network design and traffic characterisation. Depending on the problems faced, network re-configuration, changes in routing patters, adjustment in traffic control parameters or network extensions may be initiated.

In ITU-T a number of recommendations are found relating to measurements and monitoring; a short list follows:

- E.490 – Introduction to the series on traffic and performance measurements, including a survey and application of results for short term (input to network management actions), medium term (inputs to maintenance and reconfiguration), and long term (foundations for network extensions).
- E.491 – Application of traffic measurements by destination for network planning, describing the approaches of call detail recording and direct measurements.
- E.492 – Traffic reference period
- E.493 – Grade of Service (GoS) monitoring
- E.504 – Describing operation procedures for measurements (by human operators and by underlying system).
- E.503 – Potential applications of the measurements and operational procedures for obtaining the analysis results.
- E.500 – Describing the principles for traffic intensity measurements, including a generalisation of the busy interval approach. Terms like estimates for daily peak traffic intensity, normal load and high load for month and year are introduced.
- E.501 – Describing methods for estimating offered traffic to a resource group and the ori-

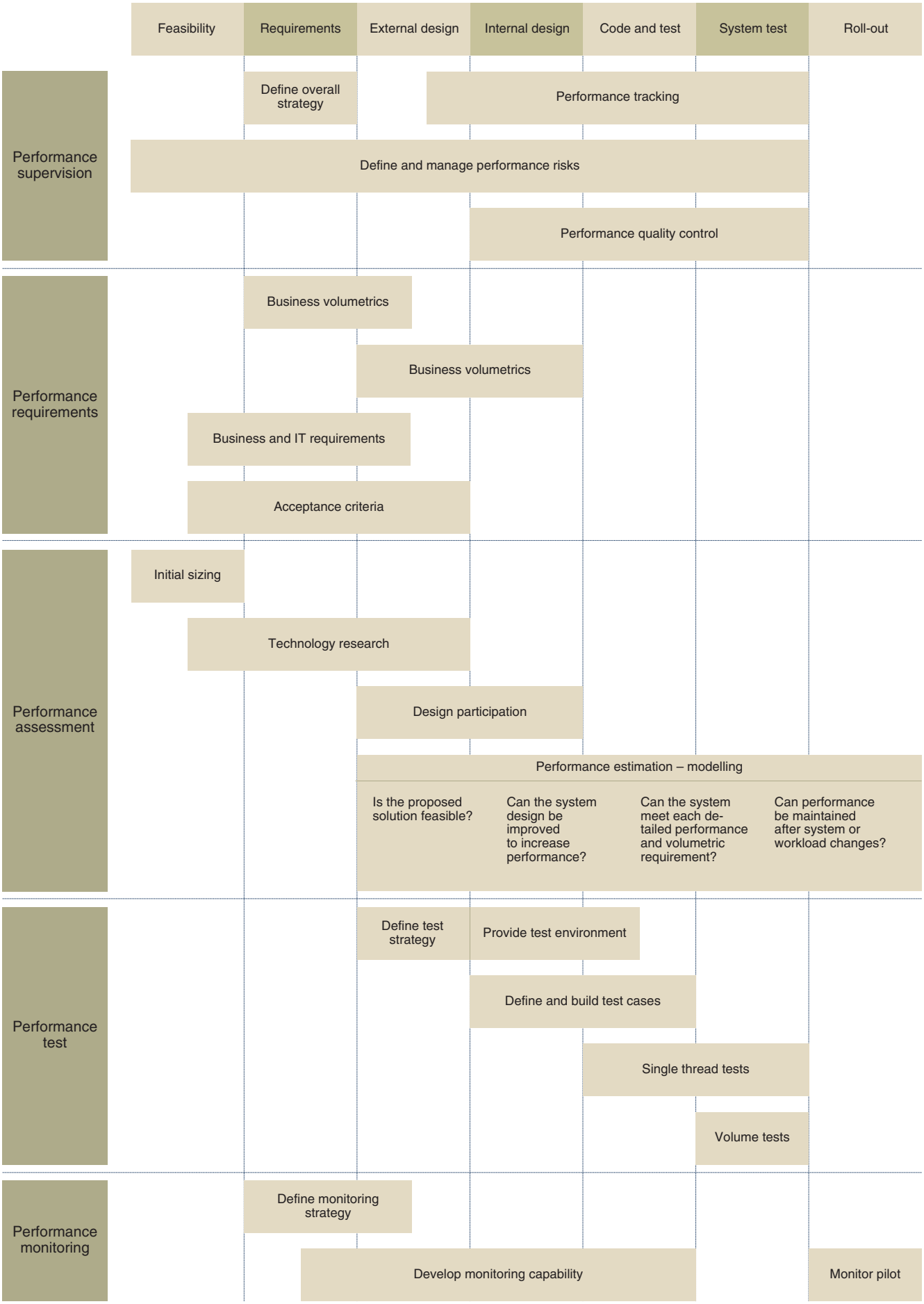


Figure 28 Performance activities related to system development phases (based on [Sing02])

gin-destination traffic demand based on measurements on a circuit group.

- E.502, E.505, E.745 – Specifies traffic and performance measurement requirements for PSND/SIDN, common signalling system no. 7 networks, B-ISDN exchanges, respectively

- E.743 – Identifies the subset of measurements (of E.505) to be used for dimensioning and planning of common signalling system no. 7 networks.

The type of measurements and parameters to be estimated must be in accordance with the objective of the measurements. Defining the measurement procedure implies balancing the technical and administrative efforts against information gained. According to the stochastic nature of traffic, a time limited measurement gives observation of certain realisation of the traffic process. Thus, this can be seen as a sample of the underlying traffic process. Repeating the measuring one often sees different values. Therefore, one can commonly only state statistical measures of the traffic process, e.g. a mean value, variation. Having full information of the traffic process implies that the distribution function is known.

A traffic process that is discrete in state and continuous in time can, in principle, be measured by combining:

- Measurements of number of events, e.g. number of requests, number of errors.

- Measurements of time intervals, e.g. processing times, session times, waiting times.

Two broad categories of measurement methods are; i) continuous measurements, and, ii) discrete measurements. Another categorisation is, i) passive measurements, and ii) active measurements. In the latter case test traffic is typically injected into the system, possibly influencing the resulting system behaviour and performance and facing the challenge of defining patterns of the test traffic for it to mimic the process that is to be estimated.

A general challenge for sampling, including both measurements and simulations, is to derive an estimate of the level of confidence that should be placed on the values estimated. Let us assume that we have N independent and identically distributed observations $\{X_1, X_2, \ldots, X_N\}$ of a random variable with unknown finite mean value m and finite variance σ^2.

The mean and variance of the sample are defined as:

$$\hat{X} = \frac{1}{N} \cdot \sum_{i=1}^{N} X_i, s^2 = \frac{1}{N-1} \cdot \left[\sum_{i=1}^{N} X_i^2 - N \cdot \hat{X}^2 \right].$$

Both $\hat{X}$ and s^2 are functions of a random variable and hence, random variables themselves, defined by a distribution called the sample distribution. The first is the main estimator for the population mean value;

$$E[\hat{X}] = m$$

s^2/N is an estimator of the variance of the sample mean, $\hat{X}$, that is:

$$\hat{\sigma}^2 \left[\hat{X} \right] = \frac{s^2}{N}$$

The accuracy of an estimate of a sample parameter is described by the confidence interval. The confidence interval specifies how the estimate is placed relatively to the unknown theoretical value with a given probability. Given the assumptions above, this confidence interval of the mean value becomes:

$$\hat{X} \pm t_{N-1,(1-\alpha/2)} \cdot \sqrt{\frac{s^2}{N}},$$

where $t_{N\text{-}1,\,(1\text{-}\alpha/2)}$ is the upper $(1 - \alpha/2)$ percentile of the t-distribution with $N - 1$ degrees of freedom. $(1 - \alpha)$ is called the level of confidence and expresses the probability that the confidence interval includes the unknown theoretical mean value. When N becomes larger the t-distribution converges to a Normal distribution.

Note the assumption of independence of the sampling events. This assumption might be broken if the observations are made too close as the values then can be correlated. Similar for simulations, different starting seeds should be applied to strive for independence of the simulation traces.

13 Traffic Handling Mechanisms

Three main groups of traffic handling mechanisms are:

- Classification: At the ingress, the traffic is classified on a packet-by packet basis into one of the defined service classes. Typically an application may be looked for by means of static field information, e.g. in the packet headers.

- Traffic conditioning (policing and marking): Commonly some form of policing to some or all of the defined service classes, to ensure that usage remains within prescribed levels. This is particularly vital for the more appreciated classes, which should often receive better treatment, and thereby result in higher cost to the network operator. Policing refers to decid-

ing on a packet-by-packet basis whether the packet stream for a particular flow lies within a specified contracted profile. The profile is commonly defined by mean rate and peak rate/peakedness. If the contracted terms are broken, appropriate actions are taken. Actions may be to drop packets, to delay packets and to mark packets.

- Differential treatment according to service class: At every point along a path, mechanisms are deployed to ensure that each class gets the appropriate treatment. This is commonly achieved by means of two complementary mechanisms; i) establishing separate queues for each class and using scheduling algorithm, and ii) providing a means to intelligently discard within a particular queue. For the former, an algorithm has to determine which queue to serve next. Examples of such algorithms are priority queueing combined with class-based weighted fair queueing and a modified deficit round-robin. For these a priority queue is used for some classes and this queue is served whenever there are any packets in it. For the second class of mechanisms it is commonly differentiated between in-contract and out-of-contract packets.

14 Network Dimensioning

A well-known principle was published by Moe in 1924: *the optimal resource allocation is obtained by a simultaneous balancing of marginal incomes and marginal costs over all sectors*. This principle is applicable to all kinds of productions, also when referring to network resources and allocation of these onto various traffic types.

Two portions of the problem can be recognized: i) given a limited amount of resources, how to allocate these among the traffic types? ii) how many resources should be allocated in total?

An example (from [ITU-D_TE]) allocates traffic onto links to nodes. Given a node having connections in K directions. Assume that the cost of a connection of a node i is a linear function of the number of circuits n_i:

$$C_i = c_{oi} + c_i \cdot n_i, \qquad i = 1, 2, ..., K$$

The total cost for that node is:

$$C(n_1, n_2, ..., n_K) = C_0 + \sum_{i=1}^{K} c_i \cdot n_i$$

where C_o is a constant.

The total amount of carried traffic is a function of the number of circuits: $A = f(n_1, n_2, ..., n_K)$.

Adding more resources increases the carried traffic;

$$\frac{\partial f}{\partial n_1} = E_i \cdot f > 0.$$

In a pure loss system $E_i \cdot f$ expresses the improvement function that is positive for a finite number of circuits due to the convexity of Erlang's B formula.

One wants to minimise the cost C for a given carried traffic A:

Min$\{C\}$ given $A = f(n_1, n_2, \ldots n_K)$.

Introducing Lagrange multiplier ν and $G = C - \nu \cdot f$, that is:

$$\text{Min}\{G(n_1, n_2, \ldots, n_K)\} = \min\{C(n_1, n_2, \ldots, n_K) - \nu \cdot [f(n_1, n_2, \ldots, n_K) - A]\}$$

A necessary condition for the minimum solution is:

$$\frac{\partial G}{\partial n_1} = c_i - \nu \cdot \frac{\partial f}{\partial n_1} = c_i - \nu \cdot E_i f = 0$$

for all i. That is,

$$\frac{1}{\nu} = \frac{E_1 \cdot f}{c_1} = \frac{E_2 \cdot f}{c_2} = ... = \frac{E_K \cdot f}{c_K}.$$

Hence, a necessary condition for the optimal solution is that the marginal increase of the carried traffic when increasing the number of circuits (improvement function) divided by the cost of a circuit must be equal for all groups.

In case the different traffic directions have different income levels, these have to be included by replacing the unit cost c_i by the fraction c_i / g_i where g_i is the income level.

Denoting the revenue with $R(A)$ and the cost with $C(A)$, the profit is given by:

$P(A) = R(A) - C(A)$. A necessary condition for optimal profit is then given by

$$\frac{\partial P(A)}{\partial A} = 0,$$

implying

$$\frac{\partial R}{\partial A} = \frac{\partial C}{\partial A}.$$

Hence, the marginal income should be equal to the marginal cost to achieve optimal profit. This can be expressed as:

$P(n_1, n_2, ..., n_K) =$

$$R(f(n_1, n_2, ..., n_K)) - \left[C_0 + \sum_{i=1}^{K} c_i \cdot n_i \right],$$

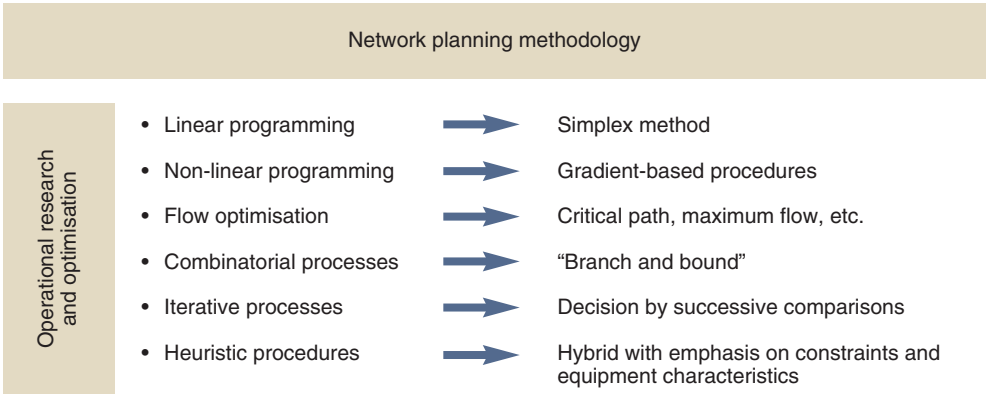


Figure 29 Some optimisation methods (based on [Soto02a])

and hence the optimal solution is achieved when

$$\frac{\partial P}{\partial n_i} = \frac{\partial R}{\partial A} \cdot E_i f - c_i = 0.$$

Using the results above, this can be written as:

$$\frac{\partial R}{\partial A} = \nu.$$

The factor ν is the ratio between the cost of one circuit for the traffic which can be carried additionally if the group were extended by one circuit. That is, circuits should be added to the group until the marginal income equals the marginal cost ν.

15 Optimisation Techniques

As described earlier, optimisation is a pivotal element in network planning. Here a defined objective is to be optimised, commonly under a set of specified constraints. An overview of some optimisation methods is shown in Figure 29.

One of the founding approaches is the simplex method, which was invented in the late 1940s. By this the use of mathematical programming techniques, and in particular linear programming methods quickly gained acceptance and popularity. On the other hand, uncertainties are commonly associated with the inputs and the relations between subjects of the systems to be analysed. Furthermore, in many applications of mathematical programming, the performance of the optimal solution could well be drastically compromised if the actual state of the nature turns out to be different than the specifications of the model and input.

15.1 Locations and Links

Deploying a new network, both locating nodes and interconnecting nodes are to be found. A simple example is shown in Figure 31.

Assuming that the intermediate nodes, index i, are to be located as well as the connections between the two neighbour node sets, a simple optimisation formulation can be given as:

$$min \sum_i f_i \cdot w_i + \sum_k \sum_i d_{ki} \cdot x_{ki} + \sum_i \sum_j c_{ij} \cdot y_{ij}$$

where

f_i – cost of installing node at location i

w_i – variable indicating whether node i should be established ($w_i = 1$) or not ($w_i = 0$)

d_{ki} – cost of link between location k and location i

x_{ki} – variable indicating whether link k_i should be established ($x_{ki} = 1$) or not ($x_{ki} = 0$)

c_{ij} – cost of link between location i and location j

y_{ij} – variable indicating whether link ij should be established ($y_{ij} = 1$) or not ($y_{ij} = 0$)

Straightforwardly, this could be attacked by the simplex method. However, as integer constraints are attached to all the variables, certain adaptations must be incorporated. One approach is the branch and bound technique, where an integer-constrained variable is assigned the value 0 and 1, respectively, and the rest of the undetermined

Box C – Some Network Terms

Networks are commonly modelled by directed graphs (digraphs) whose vertices represent the network switching nodes and whose directed edges represent the transmission links. A digraph is said to be 2-edge connected in case it takes the removal of at least two edges to break the graphs into more than one disconnected component.

A graph is planar if it can be drawn on the plane in such a manner that no two edges intersect (have a common point other than a vertex). The resultant embedded graph is called a plane graph. Any connected planer graph embedded in a plane with n vertices ($n \geq 3$) and m edges has $f = 2 + m - n$ faces. A face is a region defined by the plane graph. f denotes Euler's number. The number of faces f includes ($f - 1$) inner faces and one outer face (the unbounded region). The protection cycles are then a set of facial cycles with special orientations.

Box D – Network Topologies

A number of topologies are relevant for telecommunication networks, see Figure 30. These are often applied on different portions of the network also in a hierarchical manner. For example, the mesh network may be used between transit telephone exchanges, the star network in the access network from a local exchange to its subtending concentrators, while a ring structure is commonly found in SDH networks for redundancy arguments. Bus networks may be seen for local area networks as well as subtending concentrators in a chain.

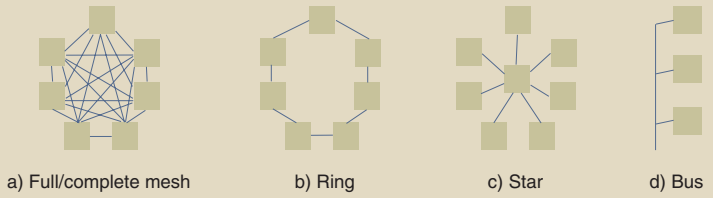


Figure 30 Examples of network topologies

Besides these fundamental topologies, hierarchical levels can be introduced where each of these levels in principle can have different topologies.

variables are not necessarily obeying the integer constraints. Introducing the integer constraints always limits the options and hence a "worse" objective function value can be achieved. Therefore, all branches that give a worse optimum estimate than a complete solution found so far can be omitted from the further search for the optimal configuration.

An extension of the problem formulation above is to introduce link and node capacities; modelling the relations between node and link size and overall costs. This implies that the amount of traffic flowing is to be explicitly considered. This also means that the cost parameters are not given as a single value, but have to be calculated depending on the overall traffic and capacity. Calculating needed link and node capacity, one is often faced with non-linear relations. For example, the scale effect in several cases states that the overall capacity needed when a number of traffic flows are aggregated is less then the sum of the needed capacity for these traffic flows individually.

Considered a somewhat more complicated example that the one shown in Figure 31, further variables have to be introduced when the traffic flows are to be routed between the nodes (note that Figure 31 only gives one possible route between a node i and a corresponding node k). For a more complete network – allowing for any-to-any terminal nodes – different routes may exist between the terminal nodes.

As the parameters may be involved in a non-linear manner for more advance problems, several principles have been proposed for solving such problems:

- Kuhn-Tucker
- Fibonacci search
- Golden search
- DSC-Powells method
- Steepest decent
- Simulated annealing
- Tabu search

15.2 Stochastic Programming

It can be shown that ordinary sensitivity or parametric analyses do not provide an adequate approach for supporting decision-making when uncertainties are involved. One way of approaching this is to define models that explicitly take the uncertainty into account. And, this is where stochastic programming models come into the picture. Commonly levels of certainty are associated with all variables.

A schematic formulation of a random (linear) programming problem is:

$$\min c \cdot x$$

such that

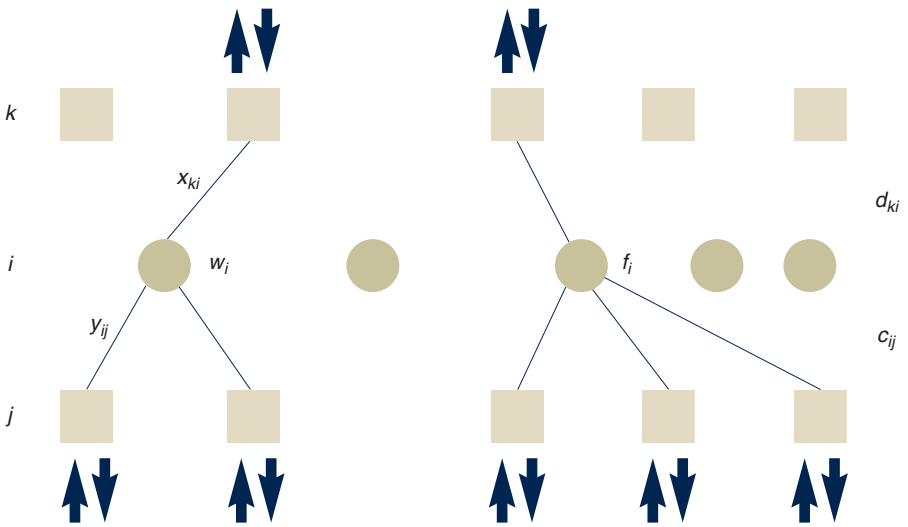


Figure 31 Illustration of locations and interconnections for intermediate nodes (index i)

$A \cdot x \geq b$,
$T(w) \cdot x \geq h(w)$,
$x \in X$

Here $c \in \mathcal{R}^{n1}$ and $b \in \mathcal{R}^{m1}$ are known vectors, A is a known $m1$ x $n1$ matrix, and $X \subseteq \mathcal{R}^{n1}$ is some subset that may contain integrality restrictions (that is the corresponding element of x must then take on integer values only). The uncertainty is reflected through the way the coefficient matrix T and the corresponding right-hand side h of the second set of constraints depend on the outcome of random event(s) w. This implies a measurable mapping between two spaces such as $T : W \rightarrow \mathcal{R}^{m2 \text{ x } n1}$ and $h : W \rightarrow \mathcal{R}^{m2}$, defined in a corresponding way.

An essential statement is that we seek related values of x that are only based on information known at the point of decision. That is, the decision cannot be based on the actual outcome of the random event w. The information concerning this future event is assumed to be known through its probability distribution.

This also implies that the notion of optimality is no longer absolutely obvious as the objective to be optimised as well as the formulation of the constraints depend on the decision-maker's attitude to risk.

15.2.1 Chance-constrained Programming

Consider the same optimisation formulation as above, except that the second set of constraints is to hold with a probability of at least α, where $\alpha \in [0,1]$. That is,

$\min c \cdot x$

such that

$A \cdot x \geq b$,
$P[T(w) \cdot x \geq h(w)] \geq \alpha$,
$x \in X$

This is formulated as a chance-constrained stochastic program where the second set of constraints is referred to as a joint chance-constraint or a joint probabilistic constraint. One way of generalising this is to introduce individual constraints:

$P[T_k(w) \cdot x > h_k(w)] \geq \alpha_k, k = 1, \ldots, m2$

Where T_k is the k^{th} row of T, h_k is the k^{th} component of h, and α_k is the probability associated with the k^{th} constraint.

15.2.2 Two-stage Stochastic Programs with Recourse

For two-stage stochastic recourse programs the decision can be divided into two groups; one group, called first-stage decision, that must be made without certain knowledge about some random parameters of the mode, and another group, called second-stage decision, that can be taken after uncertainties have been revealed.

Consider the example above; in the first stage the second set of constraints can be violated. These constraints, however, are considered at the second stage through some corrective or recourse actions $y(w)$ that are considered after the actual outcome of w is known. Assuming one wants to minimise the sum of direct cost $(c \cdot x)$ and expected resource costs, this can be formulated as:

$\min c \cdot x + E[q(w) \cdot y(w)]$,

such that

$A \cdot x \geq b$,
$T(w) \cdot x + W(w) \cdot y(w) \geq h(w)$,
$x \in X, \quad Y(w) \in Y$.

Here q and W are measurable mappings onto corresponding spaces, in a similar manner as h and T above.

Referring to the formulation above, c is the first-stage cost, q is the second-stage cost, T is the technology matrix, W is the recourse matrix, and h is the second-stage right-hand side. In general, the two-stage stochastic recourse model applies to a wider formulation.

The division of decisions into two sets may reflect an inherent nature of an overall problem – showing the timing of decision. That is, some decisions may be made "immediately" whereas others can be postponed until uncertainty has been disclosed, or at least until additional information on uncertain parameters is available. Hence, this represents a simple dynamic decision process. The first-stage decisions must be made without certain knowledge of random parameters and must be chosen as to minimise the sum of direct cost and the expected value of future cost. The future cost, on the other hand, is determined when an optimal second-stage decision is made after observation of the outcome of the random parameters:

Decision on x -> observation of $q(w)$, $h(w)$, $T(w)$ and $W(w)$ -> decision on y.

15.2.3 Multi-stage Stochastic Programs with Recourse

The decision process outlined in the previous section may be generalised to include a number of stages; that is, an alternating sequence of decisions and observations of random variables.

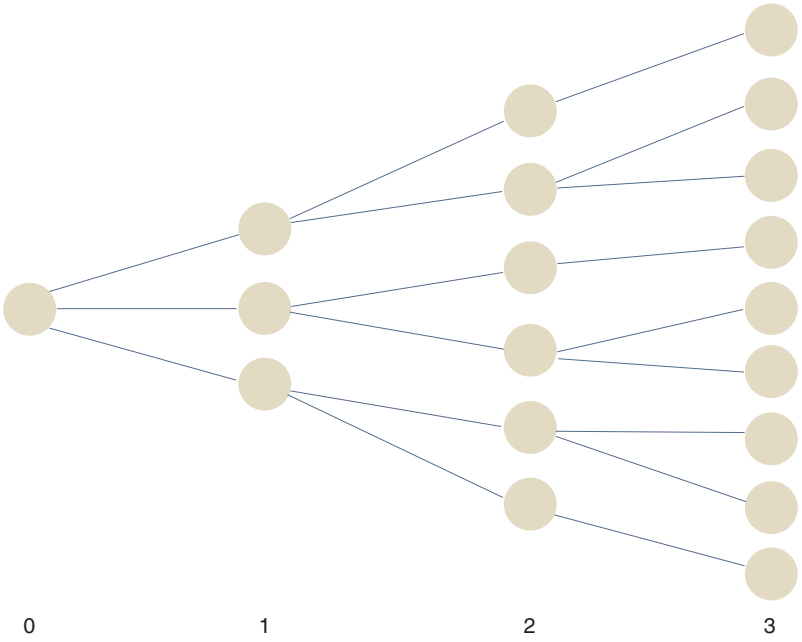


Figure 32 Scenario tree with 4 stages and 9 scenarios

X_t denotes the decisions to be made at stage t and y_t the random variables to be observed at the same stage. Similar as for two-stage programs this can be illustrated as:

observation of y_1 -> decision on x_1
observation of y_2 -> decision on x_2
…
observation of y_N -> decision on x_N

For some cases it is assumed that the constraints obey a Markovian criterion, i.e. the ones applicable at stage t are coupled to previous stages only through stage $t - 1$.

This formulation intuitively invites the scenario tree description, see Figure 32.

Even though this kind of formulation looks inspiring for thinking of future options in network planning, it turns out that the algorithms for solving the corresponding problems often must be tailored to the problem at hand. Hence, few attempts have been made to devise general-purpose algorithms for multistage stochastic integer programs.

15.3 Minimum Risk Problem

An objection to the recourse approach is that the expected cost is to be minimised. However, a risk-resistant decision-maker may prefer to minimise the probability of total cost exceeding some prescribed threshold value, say f (e.g. the bankruptcy level or simply a budget limit). This could be considered by introducing the corresponding measure in the objective function:

$$\min c \cdot x + E[q(w) \cdot y(w)] + k \cdot Q(x, f),$$

such that

$$A \cdot x > b,$$
$$x \in X,$$

where $Q(x, f)$ expresses the probability that the total cost exceeds the specified level f and k gives the weight placed on this term.

In this manner both the expected total cost and the probability of the "budget" being exceeded are considered.

16 Example of Network Planning Tool and Approach

PLANITU ([PLANITU]) is a computer tool for dimensioning and optimisation of networks, aiming at minimum cost designs for:

- Location and boundaries of exchanges
- Selection of switching and transmission equipment
- Circuit quantities, traffic routing, switching hierarchy
- Choice of transmission paths.

The main focus is telephone exchanges/nodes (concentrators, local exchanges, transit exchanges), transmission systems and traffic routing.

An overall picture of functionality and input/output is shown in Figure 33. The planning process is outlined in Figure 34.

The main groups of input data, although depending on the situation at hand, are:

- Existing network configuration: i) exchange locations and boundaries, ii) exchange and transmission equipment, iii) geographical layout of subscriber and inter-exchange network.

- Demand forecasts: i) subscribers – location and category, ii) traffic – quantities and dispersion (traffic matrices).

- Switching equipment: i) capacity – subscriber lines, trunk lines, call attempts, etc., ii) costs – subscriber – trunk links, exchange units, ii) traffic handling specification, iv) floor space requirements.

- Transmission equipment: i) capacity, ii) cost – system, terminal equipment, repeaters, interfaces to other systems, iii) attenuation and loop resistance.

- Buildings and ducts: existing situation, potential extensions.

- Quality considerations: i) Grade of Service, ii) transmission plan.

Due to the interdependencies between the cost contributions, such as number and location of exchanges and transmission equipment, it is hard to optimise the overall network in one go. Hence, an iterative procedure has been defined for PLANITU allowing for a modularised approach where sub-optimisation is used for each stage. This is illustrated in Figure 35.

17 Concluding Remarks

This article has provided a brief introduction to the topic of network planning. Although relations with enterprise management, financial issues and a number of more technical areas are listed, there is a vast amount of literature and lists of supplementary issues that could be included. For example, the optimisation methods themselves have been treated in an extensive list of books over the years.

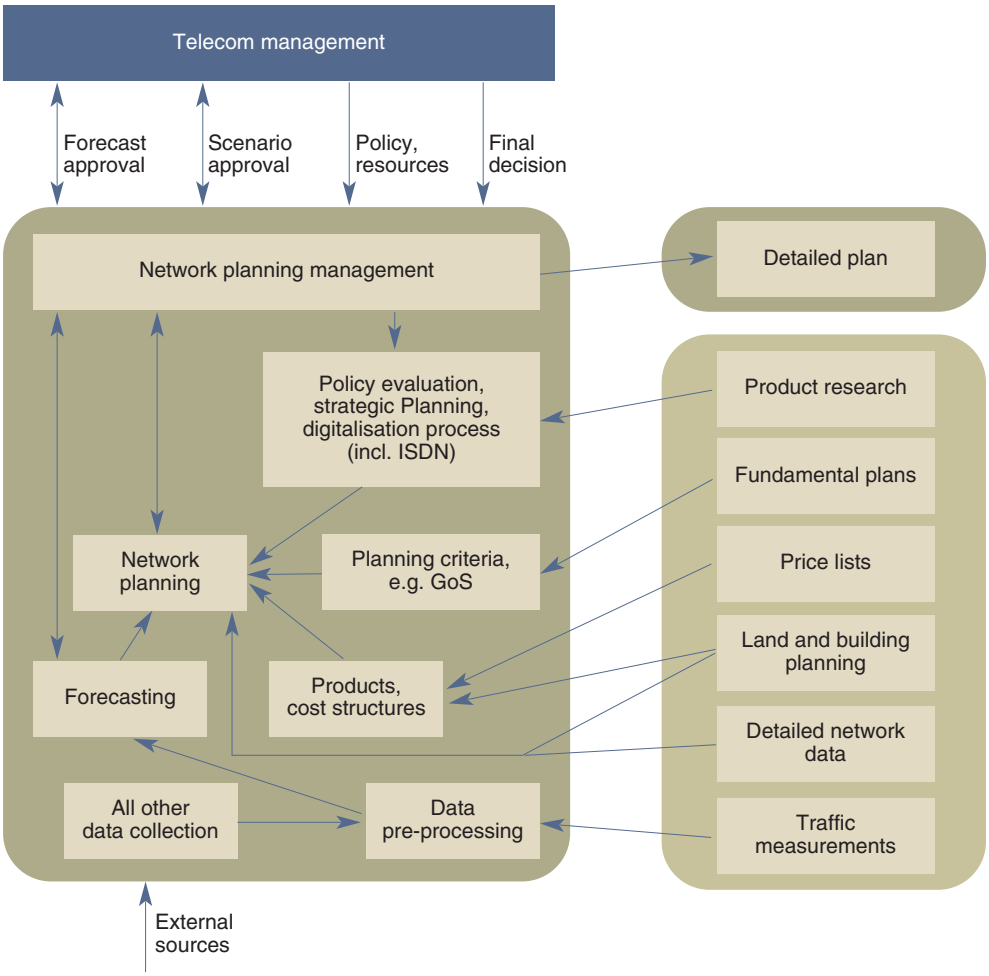


Figure 33 Task blocks and interfaces for network planning (adapted from [PLANITU])

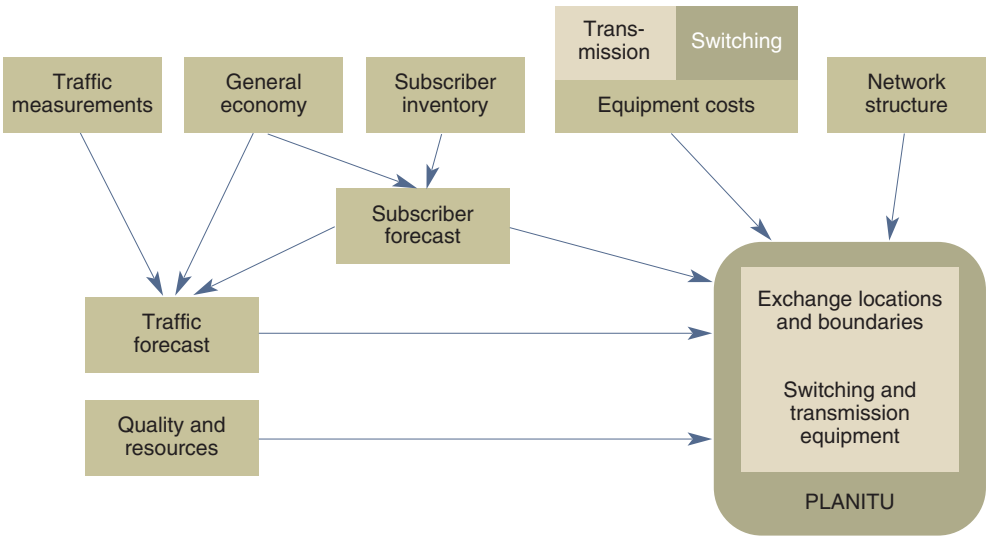


Figure 34 The network planning process (from [PLANITU])

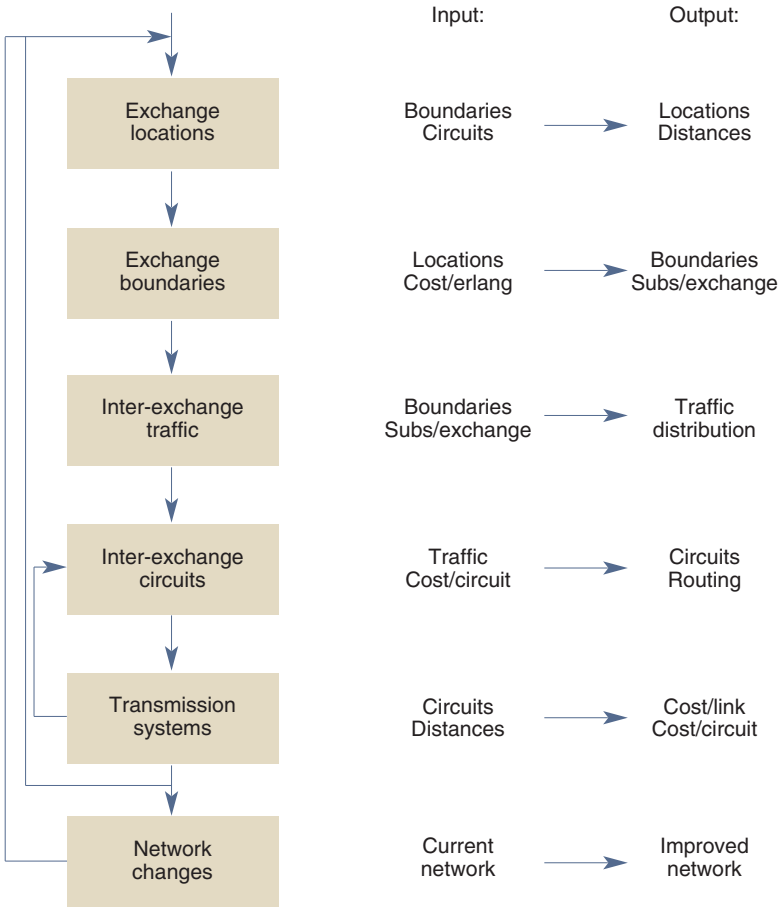
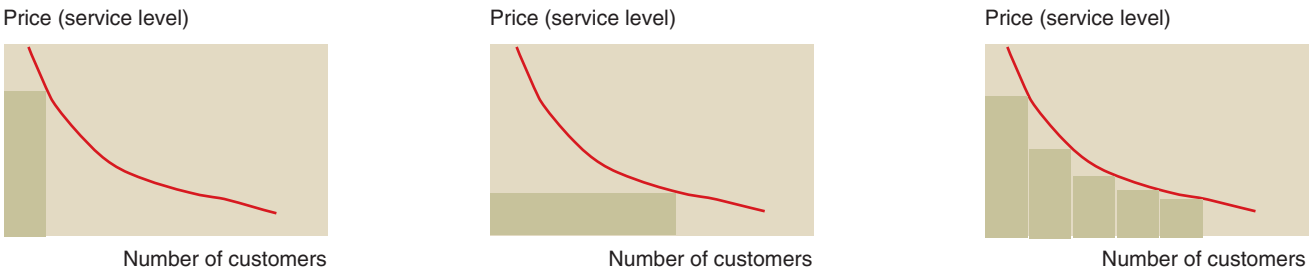


Figure 35 Activities for exchange network planning, based on [PLANITU]

A key message in this presentation is that network planning is multi-faceted and has relations with (almost) all other types of activities going on in a network operator. Hence, it is important to balance the system portfolio insight with skills in planning techniques as well as the financial means for comparing results. In addition comes the different time horizons one has to deal with – from strategic planning to tuning of parameters for traffic handling.

A key point for an operator is to provide the set of services asked for by the customers – and this should be done more cost-effectively and faster than the competitors. As shown in Figure 36 the customers' demands can often be arranged along a scale with the corresponding service levels. Whether this is done by a single network or not is not shown in the illustration.

Figure 36 Illustrating the awareness of diverse demands and offering adequate service levels and price levels may improve the operator's margins



The increasing awareness of network planning is particularly crucial when efficiency improvements are to be implemented in an organisation. In fact, the systematic views that network planning techniques provide also support the smoother interactions between units within the organisation as well as between the operators and other actors in the market.

References

[DSL-058] DSL forum. *Multi-Service Architecture & Framework Requirements*. September 2003. Technical Report TR-058. (www.dslforum.org)

[Macf02] MacFadyen, N W. Traffic characterisation and modelling. *BT Technology Journal*, 20 (3), 14–30, 2002.

[PLANITU] *ITU-D PLANITU Network Planning Software*. Februar 25, 2004 [online] – URL: www.itu.int/ITU-D/planitu/index.html 2003

[RFC372] Awduche, D et al. *Overview and Principles of Internet Traffic Engineering*. IETF RFC 3272, May 203.

[Sing02] Singleton, P. Performance modelling – what, why, when and how. *BT Technology Journal*, 20 (3), 133–143, 2002.

[Soto02] Soto, O. *Integrated Planning methodology. Scenario analysis and data gathering*. ITU-D workshop on network planning strategies for evolving network architectures, Nairobi, Kenya, 7–11 October 2002.

[Soto02a] Soto, O. *Requirements for decision making in network evolution, Strategic planning and new Technologies, Solution mapping for geo-scenarios*. ITU-D workshop on network planning strategies for evolving network architectures, Nairobi, Kenya, 7–11 October 2002.

Optimization-Based Network Planning Tools in Telenor During the Last 15 Years – A Survey

RALPH LORENTZEN

Ralph Lorentzen (67) is daily leader of the company Lorentzen LP Service which does consulting work within mathematical programming. He retired July 1, 2003 from his job as senior scientist in Telenor R&D. Ralph Lorentzen graduated in mathematics at the University of Oslo, where he worked for some time as assistant professor giving lectures in statistics and operations research. He worked as distribution planner in Norske Esso, as principal scientist at Shape Technical Centre, as systems engineer in IBM, and as chief consultant in Control Data Norway, before joining Telenor R&D in 1985.

ralph.lorentzen@tiscali.no

1 Introduction

The rapid technological development in the telecommunications field during the last two decades made it necessary for the operators to repeatedly reevaluate the structure, design and application of their networks. In order to establish cost-effective network design and utilization many telecommunication operators developed and used optimization-based network planning tools. This happened also in Telenor. The rationale was that the frequent technological shifts did not give the planners sufficient time to acquire an intuitive feeling of what constituted good network designs. So the requirement for comprehensive, cost-effective and robust designs excluded simple back of the envelope or spread sheet calculations. The idea was that mathematical models could be used to optimize network structures and thus contribute to reduce costs, ensure network reliability and improve operational efficiency.

The approach was said to be pragmatic but thorough; heuristic but nevertheless intelligent. The aim was not to give a black box answer, but a deeper understanding of the issues and of the consequences of proposed solutions. The optimization tools were not meant to give a final answer, but to constitute instruments to be integrated into the decision process of the planners.

Common to all the tools were that they were based on mathematical programming. The different network components were modeled as decision variables with associated costs. Technological constraints were modeled by equations and inequalities that the decision variables had to satisfy. Commercially available optimization software together with tailor-made program modules were used to find cost-effective solutions. The fact that the tools were to be used by others than the ones who developed them implied that they had to be incorporated in high quality user interfaces and linked up to relevant data sources in an appropriate way. In general the effort spent on the development of user interfaces exceeded the effort spent on implementation of optimization algorithms by a factor of three to four.

In 1986 Telenor R&D together with the consultant company Veritas established a group of three people who started building optimization models for key network design problems. Telenor R&D was mainly responsible for the modelling whilst Veritas was given the task of implementing the models. Later on the consultants at Veritas moved to the consultant company Computas, and the cooperation with Computas has continued to this day.

From the outset a good cooperation was established with network operators in the field who allocated time and energy to specification and testing.

The initiative was triggered by a situation where a local cable TV network planner compared two alternative network layouts which both seemed reasonable, but discovered afterwards that the cost of one of the layouts was twice the cost of the other. The group worked together on and off over a period of about 15 years and designed and implemented a series of network planning tools based on integer programming. Some of the tools will be described in the sequel, namely KABINETT, ABONETT, FABONETT, PETRA, MOBANETT, and MOBINETT.

2 KABINETT – A Cable TV Network Planning Tool

2.1 General

At the time when the cable TV network planning tool KABINETT was developed the networks should have a tree structure. Manual planning was time-consuming, and the planners had capacity for analyzing a few alternatives only. One important feature of the planning problem was to secure that the subscribers received sufficient voltage with a minimum number of amplifiers. Also, the cost of civil work constituted a large portion of the total cost. Of course, the planning problem became much simpler later when the decision was made to go for star networks in lieu of tree networks.

2.2 The Cable TV Network Design Problem

Here will be given a short description of the cable TV network design problem KABINETT attempted to solve.

We were given a signal source and a set of subscribers. The signal source was to be connected to the subscribers by copper cables via *amplifiers*, *splitters* and *taps*.

The amplifiers, splitters and taps had to be placed in *cabinets*. There were two types of cabi-

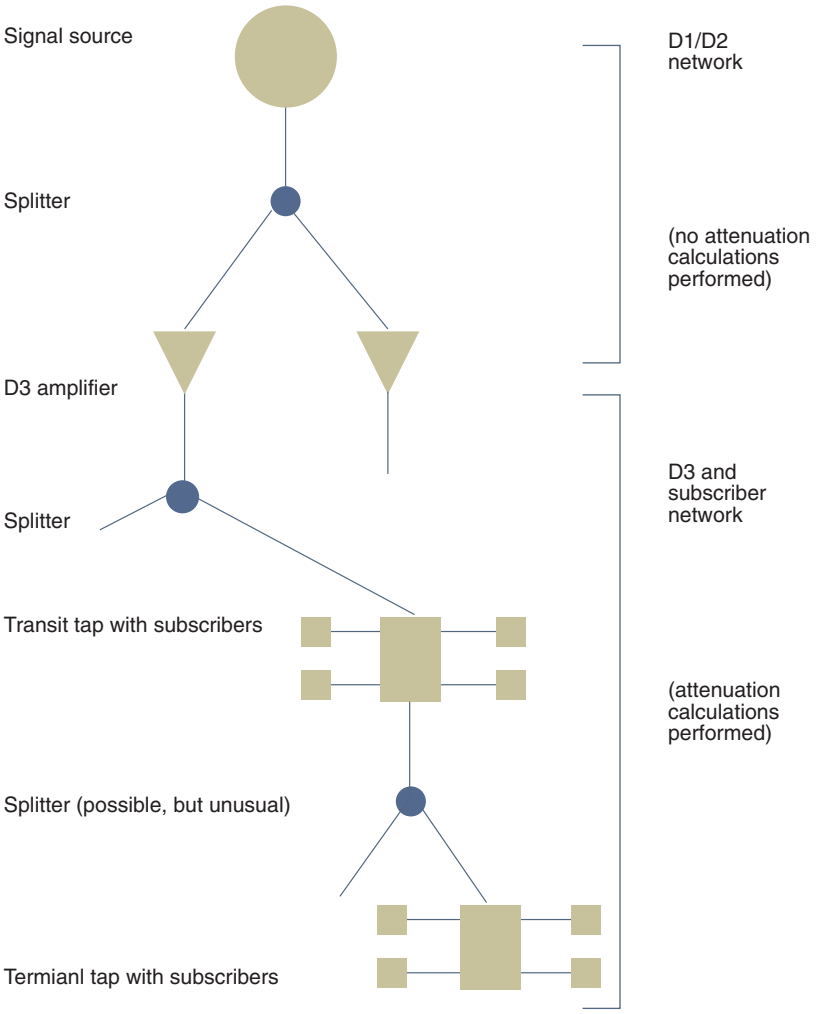


Figure 2.1 Possible network structure

nets, large cabinets and small cabinets. Amplifiers must be placed in large cabinets while splitters and taps could be placed in small cabinets. A cabinet could (in addition to the amplifier in large cabinets) contain an arbitrary number of splitters and taps.

The cables were placed in *trenches*. The trenches formed a network which was called the *trench network*. The trench network was assumed to have a tree structure.

In KABINETT the signal passed via exactly one amplifier on its way from the source to the subscriber. These amplifiers were called D3 amplifiers and were placed in the network by KABINETT such that the signal that the subscribers received had sufficient voltage. In reality it could be necessary to place additional amplifiers between the signal source and the D3 amplifiers. In KABINETT there was no voltage calculation on the stretch between the signal source and the D3 amplifiers. Consequently KABINETT did not make any analysis of the requirement for amplifiers in addition to the D3 amplifiers. However, once the D3 amplifiers had been placed, this was generally straightforward. The users were assumed to have access to a calculation program (which did not perform any optimization) for voltage and attenuation calculations between the signal source and the D3 amplifiers, which they could use as an aid for deciding where to place additional amplifiers.

From the signal source the signal might go via splitters on its way to the amplifiers. This part of the network will here be called the *D1/D2 network*.

From an amplifier the signal could go via splitters on its way to the taps. There were two types of taps, namely *terminal taps* and *transit taps*. Terminal taps had *subscriber ports* from which the signal could only proceed directly to subscribers. A transit tap had an additional port, the *D3 port*, from which the signal could proceed to a splitter or to another tap. Normally the D3 port would be connected to another tap and not to a splitter. A subnetwork connecting a D3 amplifier to the taps underneath it will here be called a *D3 network*. If the D3 amplifier had sufficient voltage to support its associated taps with their subscribers, the D3 network was said to be *feasible*. Otherwise it was said to be *infeasible*. All the D3 networks that KABINETT proposed had of course to be feasible.

Every subscriber was connected directly to a subscriber port on a tap. The subnetwork connecting the subscribers to the taps will here be called the *subscriber network*.

Based on the location of the signal source, costs and locations of candidate trenches, costs and candidate locations of D3 amplifiers, splitters and taps, costs and attenuations of candidate cable types, and subscriber voltage requirements, KABINETT tried to find the least costly network design. The user could specify parts of the design and leave it to KABINETT to complete the design at the lowest possible cost.

The algorithms used in KABINETT were heuristic, so a theoretically optimal solution to the network design problem was not guaranteed.

Figure 2.1 shows a possible network structure. The splitters could have more than two exit ports, and the exit ports could have different attenuations. There were also in general many types of taps where each tap type was characterized by its number of exit ports and the attenuation associated with each individual exit port. One could have several splitters and transit taps in series, but normally the signal would not proceed from a tap to a splitter.

Ideally the trench network, the D1/D2 network, the D3 networks and the subscriber network shold be optimized together. Instead a separate

cost minimization of the trench network was done first. Then an optimization of the subscriber network was done followed by the optimization of the D3 networks, and finally the D1/D2 network was determined. The separate optimization of the trench network was partly justified by the fact that usually well above 70 % of the total network cost was the cost of trenches.

2.3 Input

The input to KABINETT specified

- the location of the subscribers to be connected to the network
- the location of the signal source node
- the location of the nodes where D3 amplifiers, splitters and taps could be placed
- trench route alternatives
- cable types
- splitter types
- tap types
- the D3 amplifier type
- cabinet types

The trench routes consisted of trench sections, each of which was characterized by geographical location, length, and cost per metre.

The cable types were characterized by attenuation and cost per metre.

The splitter types were characterized by the number of ports and the attenuation at each individual port.

The tap types were characterized by the number of ports and the attenuation at each individual port and whether they were transit taps or terminal taps.

KABINETT operated with one type of D3 amplifier which was characterized by cost and the voltage at the exit port.

There were two cabinet types, namely large and small. Each type was characterized by cost.

2.4 Output

The output from KABINETT described the proposed solution to the cable TV network design problem. It showed

- the trenches to be dug
- what types of cables should be used where
- where D3 amplifiers should be placed
- what type of splitters and taps should be placed where
- which subscribers should be connected to which ports on which taps
- a table showing labour costs and bill of materials

The output could be fed into an interactive calculation program (not described here) for placement of amplifiers in the D1/D2 network. This program also gave a schematic description of the solution.

2.5 Determining the Trench Network

The trench network cost minimization problem is an example of the classical Steiner problem in undirected graphs. A network of candidate trenches was given, and the problem was to find the cheapest possible subnetwork (which must necessarily be a tree) which connected a subset of nodes. The nodes to be connected included the signal source node and the subscriber nodes. In addition the planner could include amongst the special nodes some of the *equipment nodes*, i.e. candidate nodes for placement of D3 amplifiers, splitters and/or taps.

There could be subscribers who were connected to the rest of the candidate trench network by two or more candidate trenches. However, one might want to avoid using these candidate trenches for transit cables to other subscribers. This was achieved by multiplying the cost of these trenches by a large number before the optimization.

In KABINETT an approximate solution to the Steiner problem was found using Rayward-Smith's algorithm [1].

2.6 Determining the Subscriber Network

The model used for the subscriber network was a classical capacitated plant location model.

The subscribers were to be connected to taps by cables. In this module all taps were considered to be terminal taps. One cable type only was considered for each subscriber. In practice the planner would assume a default cable type for all subscribers in an initial run. As a result of inspecting the initial solution he could change the cable type for some of the subscribers and run the optimization again.

Each candidate tap was given a cost which was the sum of the tap cost, the cost of a small cabinet, and the cost of the copper cable needed to connect the tap to the closest equipment node in the direction of the signal source node.

The cost of connecting a given subscriber to a tap in a given equipment node was set to the sum of the material and labour cost associated with connecting them with the appropriate cable.

Attenuation was not considered in the subscriber network optimization. Voltage and attenuation

considerations were postponed to the D3 network optimization.

The subscriber network optimization model is described in mathematical terms below.

The following notation is used:

Subscripts

n denotes equipment nodes,
s denotes subscribers,
t denotes terminal tap types.

Costs

c_{ns} denotes the cost of connecting subscriber s to equipment node n;
c_{nt} denotes the cost of placing a terminal tap of type t in equipment node n. This cost is the sum of the tap cost and the cost of a cable from the tap to the nearest equipment node on the route to the signal source node.

Capacities

p_t denotes the number of subscriber ports on a tap of type t.

Variables

$x_{ns} = 1$ if subscriber s should be connected to a tap in equipment node n, and = 0 otherwise.

$y_{nt} = 1$ if a tap of type t should be placed in equipment node n, and = 0 otherwise.

Using this notation the optimization problem can be formulated as follows:

$$\text{minimize } \Sigma_{ns} c_{ns} x_{ns} + \Sigma_{nt} c_{nt} y_{nt} \quad (2.1)$$

subject to

$$\Sigma_n x_{ns} = 1 \quad (2.2)$$

$$\Sigma_s x_{ns} \leq \Sigma_t p_t y_{nt} \quad (2.3)$$

$$x_{ns} \leq \Sigma_t y_{nt} \quad (2.4)$$

where x_{ns} and y_{nt} are 0 – 1 variables.

This optimization problem was solved using a conventional branch and bound code.

The planner could preset part of the solution by fixing some of the x_{ns} or y_{nt} variables to 1.

When the subscriber network was determined, the voltage requirement at the taps were calculated and recorded.

2.7 Determining the D3 Network

2.7.1 General

The method for determining the D3 networks consisted of three algorithms, namely the *construction algorithm*, the *partitioning algorithm* and the *move/exchange algorithm*.

Typically, KABINETT started with attempting to cover all the taps with one D3 amplifier. If this failed, two D3 amplifiers were tried, etc. If the minimum number of D3 amplifiers which KABINETT found was needed was K, and the parameter 'extra' was set to a positive integer, then KABINETT tried to find the least expensive solution with K, $K + 1$, …, or K + extra amplifiers.

The construction algorithm took as input a subset of the taps with their voltage requirements and placed a D3 amplifier in a node such that either a feasible D3 network was formed or an (infeasible) D3 network was formed where the voltage requirement in the D3 amplifier node was as low as possible.

The partitioning algorithm was a one pass algorithm for partitioning the taps into a specified number of subsets and constructing a (feasible or infeasible) D3 network for each of the subsets.

The move/exchange algorithm was an iterative algorithm which tried to improve on the solution found by the partitioning algorithm. The move/exchange algorithm tried first of all to make an infeasible solution feasible, and tried thereafter to find a solution with reduced cost.

Before we treat these algorithms we will describe the splap concept.

2.7.2 Splaps

In the D3 network KABINETT could place more than one splitter and/or tap in an equipment node. The result was a collection of splitters and taps which could be characterized by its total cost, a set of D3 ports with their individual attenuations, and a set of subscriber ports with their individual attenuations. Such a collection of splitters and taps in a node was in KABINETT denoted a *splap* (see Figure 2.2).

Before the D3 network optimization was started KABINETT compiled a list of candidate splaps based on the splitter and tap types in the input. As mentioned earlier the D3 port of a transit tap would normally be connected to another tap and not to a splitter, and the splap list was based on collections of splitters and taps where this was the case. The planner decided the size of this list by specifying the maximal number of subscriber and D3 ports of the splaps.

A splap *S* was said to be inferior to another splap *S'* if the numbers of subscriber and D3 ports of *S* were less or equal to the corresponding numbers for *S'*, and the attenuations of the ports of *S* were greater or equal to corresponding ports of *S'*. The splap list would not contain any splap which was inferior to another splap on the list.

2.7.3 The Construction Algorithm

First an outline of the algorithm will be given:

Initially the signal source node was used as what is called an attraction node.

The voltage requirements in the nodes in the trench tree were calculated starting from the taps and proceeding towards the attraction node. Whenever a junction node in the tree was reached a splap which minimized the voltage requirement was placed.

If a node was reached where the voltage requirement was greater than the output voltage of the D3 amplifier, this node was made the attraction node the first time this happened. All subsequent voltage calculations were then made proceeding towards the new attraction node.

Based on criteria to be detailed below a D3 amplifier was then placed in the vicinity of the attraction node such that a (feasible or not feasible) D3 network was formed.

The algorithm is described in more detail below.

Algorithm for constructing a candidate D3 network for a subset of taps:

1 Initialize the attraction node to be the signal source node, label all edges in the trench tree as untreated, label all the taps as treated and label the remaining nodes in the trench tree as untreated. The treated nodes are characterized by the fact that their voltage requirement is established.

2 If all nodes are treated, go to 4. Otherwise, find an untreated node *n* which is farthest away from the attraction node. Pick an untreated edge connecting *n* to a treated node, calculate and save the contribution to the voltage requirement of *n* caused by this node (i.e. the voltage requirement of the treated node plus the attenuation in a cable between *n* and the treated node), and label the edge as treated. Repeat until all the edges coming into *n* from treated nodes are treated. If then there is only one treated node adjacent to *n*, set the voltage requirement in *n* equal to the contribution to the voltage requirement of *n* caused by this node. Otherwise, place in *n* a feasible splap which minimizes the voltage requirement of *n*, and establish this as the voltage requirement of *n*. In both cases label node *n* as treated. Go to 3.

3 The first time a voltage requirement of a node is calculated which exceeds the voltage delivered by a D3 amplifier, change the attraction node from the signal source node to this node. Go to 2.

4 If the voltage requirement of the attraction node does not exceed the voltage delivered by a D3 amplifier (i.e. the attraction node has not been changed from the signal source node), place a D3 amplifier in the signal source node and terminate the algorithm. Otherwise, try iteratively to place a D3 amplifier in nodes obtained by fanning out from the attraction node until either a node is found whose voltage requirement does not exceed the voltage delivered by a D3 amplifier or a local minimum of the voltage requirement is found (which is larger than the voltage delivered by a D3 amplifier). In both cases a D3 amplifier is placed in the node found.

If the voltage requirement of the node where the D3 amplifier was placed exceeded the voltage delivered by a D3 amplifier, the resulting D3 network was infeasible. In that case the *infeasibility* was defined to be equal to the difference between the voltage requirement and the D3 amplifier voltage. If the D3 network was feasible, the infeasibility was defined to be 0.

2.7.4 The Partitioning Algorithm

In KABINETT several variants of the partitioning algorithm were implemented. Here, however, only one of these variants will be described.

The partitioning algorithm took as input a positive integer *K* which represented the number of D3 networks.

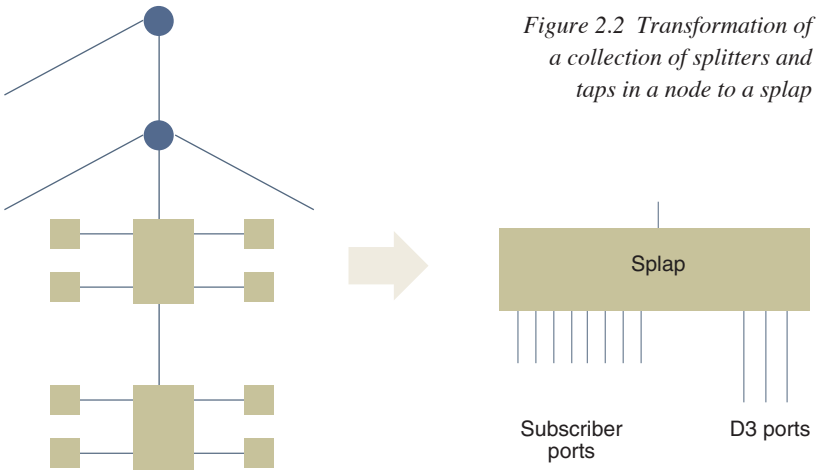


Figure 2.2 Transformation of a collection of splitters and taps in a node to a splap

A description of the partitioning algorithm follows.

Partitioning algorithm for creating K (feasible or infeasible) D3 networks:

1 Start with all the K networks empty. Select the K taps which are farthest apart, i.e. the K taps $t_1, \ldots, t_K$ which maximize $\min_{i<j} d(t_i, t_j)$. Include one of these taps in each of the K networks. Go to 2.

2 For each network find its *close node*, i.e. the closest node which has not been included in any of the networks. If none of the networks have a close node, terminate (every node belongs to a network). Otherwise, order the close nodes in a list according to how close they are to their associated network. Label all the close nodes as untreated. Go to 3.

3 If there are no untreated nodes, go to 5. Otherwise, go to 4.

4 Pick the first untreated close node on the list. If this close node does not contain a tap, include the node in its associated network, and go to 2. If the close node contains a tap and the node can be included in its associated network without making the network infeasible (this is checked by the construction algorithm), make this inclusion and go to 2. Otherwise, label the close node as treated and go to 3.

5 Include the first close node on the list and go to 2.

2.7.5 The Move/Exchange Algorithm

When a solution was found by the partitioning algorithm, it was natural to try to improve on it. Whether the solution is feasible or not, attempts were made to change it into a feasible solution with the same number of D3 amplifiers, and with the lowest cost possible. This was the function of the move/exchange algorithm.

In the description of the move/exchange algorithm the following concepts will be used:

Let N_1 and N_2 denote D3 networks and let t denote a tap.

If N_2 is infeasible with t in N_1, let $r_1(t, N_1, N_2)$ denote the reduction in the sum of infeasibilities obtained by moving t to N_2.

If N_1 is feasible with t in N_1, and if N_2 is feasible after t is moved to N_2, let $r_C(t, N_1, N_2)$ denote the reduction in cost obtained by moving t.

If N_1 is infeasible with t in N_1 or if N_2 is infeasible after t is moved to N_2, let $r_V(t, N_1, N_2)$ denote the reduction in the sum of voltage requirements obtained by moving t to N_2.

Now we can describe the move/exchange algorithm.

Move/exchange algorithm between a collection of D3 networks:

1 If all networks are feasible, go to 2. Otherwise, calculate $r = \max r_1(t, N_1, N_2)$ where the maximization is done over all N_1, N_2 and t in N_1. If $r > 0$, make a move which gives the infeasibility reduction r. Repeat until either all networks are feasible or $r \,''\, 0$. If all networks are feasible, go to 2. If $r \,''\, 0$ and 3 has been executed – stop, otherwise, go to 3.

2 Calculate $r = \max r_C(t, N_1, N_2)$ (as usual maximization over the empty set is defined to be $-\infty$), where the maximization is done over all t, N_1 and N_2 for which $r_C(t, N_1, N_2)$ is defined.
If $r > 0$, make a move which gives the cost reduction r. Repeat until $r < 0$. If 4 has been executed, terminate. Otherwise, go to 4.

3 Calculate $\max r_V(t, N_1, N_2)$ and $\max r_V(t, N_2, N_1)$, where the maximizations are made over all t, N_1 and N_2 for which the expressions are defined. Let t_1 and t_2 be taps which maximize the two expressions respectively. If moving t_1 from N_1 to N_2 and t_2 from N_2 to N_1 results in a reduction in the sum of infeasibilities, do these moves. (This is an 'exchange'.) Repeat until all networks are feasible or no reduction in infeasibility results. If all networks are feasible, go to 4. Otherwise, go to 1.

4 Calculate $\max r_C(t, N_1, N_2)$ and $\max r_C(t, N_2, N_1)$, where the maximizations are made over all t, N_1 and N_2 for which the expressions are defined. Let t_1 and t_2 be taps which maximize the two expressions respectively. If moving t_1 from N_1 to N_2 and t_2 from N_2 to N_1 results in a reduction in cost, do these moves. Repeat as long as cost reduction results. Then go to 2.

2.8 Determining the D1/D2 network

Once the D3 networks were determined the determination of the D1/D2 network was straightforward. KABINETT just cabled up the trench tree which connects the D3 amplifiers with the signal source node, and placed appropriate splitters wherever necessary.

As mentioned earlier no voltage and attenuation calculation was done in the D1/D2 network, so KABINETT did not place amplifiers in this network.

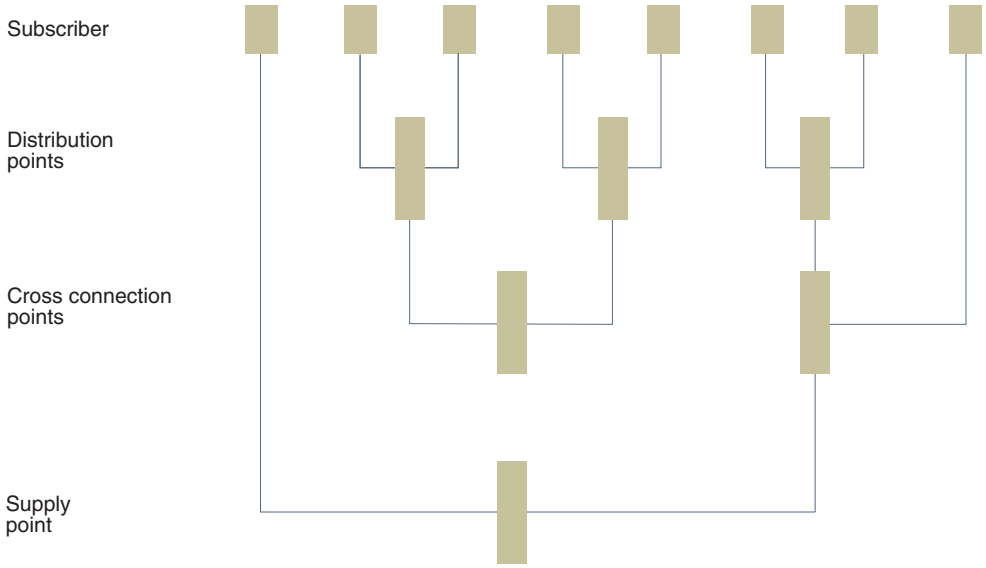


Figure 3.1 Outline of a subscriber network with one supply point at subscriber switch level

2.9 Possibilities for Manual Modification of the Solution

One of the weaknesses of KABINETT was that the placement of the splaps in the construction algorithm was not optimized. The solutions tended to have too many splaps, which implied too many cabinets, in the D3 networks.

Facilities were therefore incorporated into KABINETT's user interface which made it easy for the planner to change locations of splaps and amplifiers in the D3 networks and to check the feasibility of modified solutions. The user could also change the cable type to be used on a specified stretch. These facilities were described in KABINETT's user manual.

3 ABONETT– A Subscriber Network Planning Tool

3.1 General

After the modeling of KABINETT was completed it was found that a similar tool could solve the problem of extending a telephone subscriber network to connect new subscribers in a cost-effective way. The absence of attenuation problems made it possible to be somewhat more ambitious in the modeling and go for a simultaneous optimization of the trench and cable network. The resulting tool was called ABONETT. It was found that ABONETT resulted in solutions which on the average were 10 % less costly, and that the work involved in the design phase was reduced by at least 50 %.

3.2 The Subscriber Network Design Problem

Here will be given a short description of the subscriber network design problem that ABONETT attempted to solve.

The terminology used will first be described.

We were given a set of new *subscribers*, each of whom was to be connected by cables to one of a set of alternative *supply points*.

The network between a supply point and the subscribers connected to this supply point had a hierarchical structure.

The subscribers could be connected to *distributors* placed in *distribution points* or directly to cross connectors placed in *cross connection points*. The subscribers which were connected to distributors were called *ordinary subscribers* whilst the subscribers which were connected directly to cross connectors were called *special subscribers*. The distributors were connected to cross connectors, and the cross connectors were connected to subscriber switches or RSUs. All connections were made with cables selected from a set of cable types with different capacities.

A supply point could be a cable from a subscriber switch, an RSU, an existing cross connector or an existing distributor. Figure 3.1 illustrates a subscriber network where the supply point is a subscriber switch or an RSU.

The cables were placed in trenches and possibly in ducts. There could be a requirement for placing cables or empty ducts in trenches leading from supply points up to specified *cable or duct termination points* in order to facilitate future extensions of the network.

There was also a requirement for reserve capacity in cross connection points and distribution points, also in order to facilitate future extensions.

The ABONETT user specified the subscribers, candidate supply points, candidate trenches, candidate cross connection points and distribution points, and requirements for cables/ducts to specified cable/duct termination points.

ABONETT tried to find a network design which satisfied all the given requirements and which was as inexpensive as possible.

The ABONETT user had full control over the solution in the sense that he could specify as many characteristics of the design as he wanted. In the extreme he could specify the solution completely and use ABONETT just to check the validity of the solution and calculate its cost.

3.3 Input

The input to ABONETT described the subscribers to be connected to the network, the supply point(s), cross connection and distribution point alternatives, reserve capacity requirements, cable or duct requirements (if any), and trench route alternatives.

The trench routes consisted of trench sections, each of which was characterized by geographical location, length and the required trench type. In certain cases it was desired to reserve trench usage. If a subscriber could connect to the network via more than one trench alternative through his property, it was important to prevent ABONETT from considering the possibility of using these trench alternatives as "transit trenches" for reaching other subscribers. Technically this was treated in ABONETT by adding a high artificial cost to such trenches.

The trench types were described by cost per metre.

The subscribers were characterized by their geographical location, the capacity they required, whether they should be connected to a distribution point or directly to a cross connection point, and whether the adjacent trench sections should be reserved.

The supply points were characterized by their geographical location, whether they were switches/RSUs, cross connectors or distributors, capacity, reserve capacity requirement, and maximal distance to cross connection points, cable termination points, distribution points or subscribers they could offer connection to.

The cross connection points were characterized by their geographical location, the types of cross connectors which could be placed there, the required reserve capacity in percent, and maximal distance to distribution points or special subscribers they could offer connection to.

The distribution points were likewise characterized by their geographical location, the types of distributors which could be placed there, the required spare capacity in percent, and maximal distance to subscribers they could offer connection to.

The cross connector types, distributor types, cable types, and ducts were characterized by capacity and cost.

3.4 Output

The output from ABONETT described the proposed solution to the subscriber network design problem.

It listed the supply points, cross connection points and distribution points with the equipment selected by ABONETT, which trenches should be dug, and which cables and ducts should be used. A simplified bill of materials was also given.

3.5 The Trench and the Cable Networks

Two networks were introduced, namely an undirected trench network and a directed cable network.

The edges in the trench network were the trench sections, and the nodes were the end points of the trench sections.

The nodes in the cable network represented existing supply points, existing and candidate cross connection and distribution points, subscribers, and cable/duct termination points. In addition an extra node was introduced, namely the root node.

The arcs in the cable network connected the root node to the supply points, the supply points to the cross connection points and cable/duct termination points, the cross connection points to the distribution points and special subscribers, and the distribution points to the ordinary subscribers.

The cable network thus constituted a rooted directed acyclic graph.

3.6 The Trench, Cable and Combined Optimizations

In order to keep the computation time within reasonable limits three optimizations were made; namely the *cable optimization*, the *trench optimization*, and the *combined optimization*.

In the cable optimization ABONETT tried to find the cheapest cabling which satisfied all the requirements under the assumption that all the trenches are available free of charge.

In the trench optimization ABONETT first tried to find the cheapest trench network which connected together the subscribers, the supply points, the cross connection points, the distribution points, and the duct/cable termination points. Then ABONETT tried to find the cheapest cabling in this trench network which satisfied all the requirements.

In the combination optimization ABONETT considered both trench costs and cable costs. However, the only trenches which were taken into consideration were those selected either in the trench optimization or in the cable optimization (or both). Trenches which were selected both in the trench and cable optimization, and in the trench optimization were selected to contain either two or more cables or at least one cable terminating at a subscriber, are made available at no cost. All other trenches which were selected either in the cable or the trench optimization were made available at their real digging cost if they belonged to a circle in the resulting trench network, and at no cost if they did not belong to a circle.

The solution resulting from the combination optimization was ABONETT's suggested solution to the network design problem.

3.7 Finding the Cheapest Possible Trench Network

In the trench optimization the cheapest trench network connecting the subscribers, cross connection points, distribution points and the duct/cable termination points was sought. It is obvious that the cheapest network is a tree.

This was an example of the classical Steiner problem in undirected graphs. In ABONETT an approximate solution to the Steiner problem is found using Rayward-Smith's algorithm [1].

3.8 The Integer Programming Model for Finding the Cheapest Cabling in a Given Trench Network

The problem of finding the cheapest cabling in a given trench network was in ABONETT formulated as an integer linear programming problem.

The following notation was used:

$x_{ij}^t = 1$ if there is a cable of type t carrying signals from node i to node j, and 0 otherwise.

a^t is the capacity of cable type t.

n_{ij}^t is the capacity used in cable type t from node i to node j.

b_i is the capacity of supply point i.

p_j is the capacity required in the cable leading to subscriber or cable termination point j.

c_{ij}^t is the cost of a cable of type t carrying signals from node i to node j. This cost included the cost of any technical equipment (cross connector, distributor) this cable necessitated in node j, and it was calculated under the assumption that the cable followed the shortest path from i to j in the given trench network. For the trench optimization this shortest path is the only path leading from i to j. For the cable optimization the shortest path was calculated using Dijkstra's algorithm.

Ducts were looked upon as a "cable type" which served a special class of "subscribers", namely duct termination points.

The integer linear programming problem can be formulated as follows:

$$\text{minimize } z = \sum_{ijt} c_{ij}^t x_{ij}^t$$

subject to

$$\sum_{it} a^t x_{ij}^t \geq \sum_{kt} n_{jk}^t \tag{3.1}$$

if j is a cross connection or distribution point,

$$\sum_{it} n_{ij}^t \geq p_j \tag{3.2}$$

if j is a subscriber or a cable termination point,

$$\sum_{jt} n_{ij}^t \leq b_i \tag{3.3}$$

if i is a supply point,

$$n_{ij}^t \text{ " } a^t, \tag{3.4}$$

$$x_{ij}^t = 0 \text{ or } 1, \; n_{ij}^t \text{ integer.} \tag{3.5}$$

3.9 Approximate Solution of the Integer Programming Problem by Lagrange Relaxation

In order to reduce the integer linear programming problem in the previous chapter to a problem which can be handled by graph theoretical algorithms, Lagrange relaxation combined with subgradient optimization was used. We relaxed the constraints expressing capacity requirements while we retained the requirement that the network should connect the root node with all the subscribers and the cable/duct termination points.

Before relaxation we made an approximation: We replaced the constraints

$$\sum_{it} a^t x_{ij}^t \geq \sum_{kt} n_{jk}^t \quad \text{by} \tag{3.6}$$

$$\sum_{it} a^t x_{ij}^t \geq \sum_{kt} a^t x_{jk}^t, \tag{3.7}$$

and we disregarded the other constraints involving n_{ij}^t.

This had the effect of reducing the subproblem to a pure 'cabling problem'. The *n*-s, which constituted the corresponding 'usage solution', were uniquely determined by the cabling.

However, when we updated the dual variables corresponding to the modified constraints above in the subgradient method, we did this based on to what extent the following inequalities are satisfied:

$$\sum_{it} n_{ij}^t \geq \sum_{kt} n_{jk}^t. \tag{3.8}$$

In this way we were led to solve a sequence of cabling subproblems which were Steiner problems in directed acyclic graphs. We elected to use Wong's method [2] augmented by Pacheco and Maculan's solution improvement procedure [3] to solve these subproblems. For the sake of completeness we describe in 3.11 and 3.12 the specialization of these methods to acyclic graphs.

3.10 The Integer Programming Model for Simultaneous Cable and Trench Optimization

In the combined optimization cables and trenches were optimized simultaneously. The problem was again formulated as an integer linear program. This model is an extension of the model described in the previous chapter.

Several trench paths between nodes were made eligible as alternatives in the optimization. First of all the trench paths found in the cable and trench optimizations were eligible. New paths up to a prescribed maximum number were generated by shortest paths whilst successively blocking trench sections which were given a cost in the combined optimization. Here paths containing few of these trenches were given priority.

The following notation was used:

The eligible trench paths between node *i* and *j* are referred to by the superscript *r*.

x_{ij}^{rt} =1 if there is a cable of type *t* carrying signals from node *i* to node *j* following trench path *r*, and 0 otherwise.

c_{ij}^{rt} is the cost of a cable of type *t* following trench path *r* from *i* to *j*.

n_{ij}^{rt} is the capacity used in cable type *t* from node *i* to node *j* following trench path *r*.

b_i is the capacity of supply point *i*.

p_j is the capacity required in the cable leading to subscriber or cable termination point *j*.

y_g = 1 if trench section *g* is dug, and 0 otherwise.

d_g is the cost of digging trench section *g*.

δ_{ij}^{gr} = 1 if trench section *g* is on the trench path *r* from *i* to *j*, and 0 otherwise.

The variables y_g were only defined for the trench sections which were available at their true digging cost in the combination optimization.

The integer linear programming problem was formulated as follows:

$$\text{minimize } z = \sum_{ijrt} c_{ij}^{rt} x_{ij}^{rt} + \sum_{g} c_g y_g \tag{3.9}$$

subject to

$$\sum_{irt} a^t x_{ij}^{rt} \geq \sum_{krt} n_{jk}^{rt} \tag{3.10}$$

if *j* is a cross connection or distribution point,

$$\sum_{irt} n_{ij}^{rt} \geq p_j \tag{3.11}$$

if *j* is a subscriber or a cable termination point,

$$\sum_{krt} n_{jk}^{rt} \leq b_i \tag{3.12}$$

if *i* is a supply point,

$$n_{ij}^{rt} \text{ " } a^t, \tag{3.13}$$

$$\delta_{ij}^{gr} x_{ij}^{rt} \text{ " } y_g, \tag{3.14}$$

$$x_{ij}^{rt}, y_g = 0 \text{ or } 1 \text{ integer.} \tag{3.15}$$

Ducts were again looked upon as a "cable type" which served a special class of "subscribers", namely duct termination points.

This problem was again solved by Lagrange relaxation where in addition the relationship between cables and trenches were relaxed, and where the same type of approximations were made as when the trench network was given.

3.11 Wong's Dual Ascent Method for the Steiner Problem Specialized to Rooted Directed Acyclic Graphs

Here we describe Wong's dual ascent method specialized to rooted directed acyclic graphs in the form it is implemented in ABONETT.

Let $G(N, A)$ be rooted directed acyclic graph where arc (i, j) has length $c(i, j) \geq 0$.

Let r be the root node, S the set of special nodes to be connected to the root node and $v(s)$ an auxiliary function to be defined for the nodes s in S.

Initialization:

Set $v(s) = 0$ for all s in S.
Set $d(i, j) = c(i, j)$ for all lines (i, j).
Let G' be the subgraph with node set N and no lines.

1) For all nodes s in S, let $C(s)$ be the set of nodes from which there are directed paths to s in G'. Choose an arbitrary s in S such that r does not belong to $C(s)$. If there is no such s, go to 3). Otherwise, find nodes i and j such that i does not belong to $C(s)$, j belongs to $C(s)$, $d(i, j) = \min d(k, 1)$ for k not in $C(s)$ and 1 in $C(s)$.

2) Set

$v(s) = v(s) + d(i, j)$
$d(k, l) = d(k, l) - d(i, j)$ for k not in $C(s)$ and l in $C(s)$.

Include the line (i, j) in G' and go to 1).

3) Find the shortest directed spanning tree in G' from r to the nodes which can be reached from r in G' and prune this tree.

It can be remarked that $\sum v(s)$ becomes a lower bound on the minimum length of a Steiner tree.

3.12 Pacheco/Maculan's Improvement Algorithm for the Steiner Problem Specialized to Rooted Directed Acyclic Graphs

Pacheco and Maculan have designed an algorithm which in many cases improves significantly the solution of the Steiner problem which results from Wong's algorithm. We describe here the specialization of this algorithm to rooted directed acyclic graphs in the form it has been implemented in ABONETT.

Let $G(N, A)$ be a rooted directed acyclic graph where arc (i, j) has length $l(i, j) \geq 0$. Let r be the root node, and let S be the set of special nodes to be connected to r in a Steiner tree.

Definition:
The *Steiner length* of a directed spanning tree in $G(N, A)$ is equal to the sum of the length of the arcs in the Steiner tree it contains.

Definition:
The *Steiner function* $f(i, j)$ defined for the arcs in a directed spanning tree is equal to the number of special nodes which can be reached via the arc (i, j) in the Steiner tree it contains.

Definition:
A *fragment* in a directed spanning tree with Steiner function f is a maximal chain of arcs with the same value of f. A node belongs by definition to the fragment which the incoming arc belongs to.

The definition above secures that every node belongs to one fragment only.

Definition:
The *segment* of a node is the subchain of the fragment to which the node belongs that terminates at the node.

In a directed spanning tree where (u, v) is an arc, the predecessor u of v is denoted by $p(v)$.

The algorithmic iterates from one spanning tree to another such that the Steiner length either is reduced or stays constant while the prospect of obtaining Steiner length reductions in later iterations is increased.

We operate with type 1 and type 2 iterations which we describe below:

Type 1 iteration:

We consider a node i and an out-of-tree arc (i, j). If we obtain reduction in the Steiner length by replacing $(p(j), j)$ by (i, j), this is done, and the iteration is terminated.

If $f(p(j), j) = 0$, the Steiner length will remain unchanged by such a replacement. Nevertheless. the replacement is carried out and the iteration terminated if the length of j's segment is reduced by it.

Assume now that $f(p(j), j) > 0$, $f(p(i), i) = 0$, and that the change in the Steiner length which would result by replacing $(p(j), j)$ by (i, j) is $\delta \geq 0$. Then we look for nodes j' such that (i, j') are out-of-tree arcs and such that replacing $(p(j'), j')$ by (i, j') would contribute to a reduction of the Steiner length. If we can find a set of such nodes whose total contribution Δ to the reduction of the Steiner length is greater than δ, all the replacements are carried out, and the iteration is terminated.

If the largest total reduction Δ we can find is $\geq \delta$, we back up one node from i. If $\Delta - \delta \geq l(p(i), i)$, then the iteration is terminated without carrying out any replacement. Otherwise we look for out-of-tree nodes j' such that replacing $(p(j'), j')$ by $(p(i), j')$ would contribute to a reduction of the Steiner length. If we can find a set of such nodes whose total contribution to the reduction of the Steiner length is $> \delta - \Delta$, all the replacements are carried out, and the iteration is terminated. Otherwise we back up one node from $p(i)$ and continue in the same way. If this process does not terminate in accordance with the criteria given, we terminate it without replacements when we reach the beginning of i's segment.

Type 2 iteration:

Here we consider a node n not in S with $f(n, j) > 0$ for at least two j-s. We will try to find an alternative spanning tree with less Steiner length where $f(p(n), n) = 0$. This is done by searching for one suitable replacement arc for each fragment with $f > 0$ which succeeds n.

First we calculate the length l of the fragment which ends in node n. For each fragment starting at node n we then search for the candidate out-of-tree replacement arc leading to the fragment which would increase least (or reduce most) the Steiner length. If the sum of these increases is less than l, the replacements are carried out, and we obtain a reduction of the Steiner length.

The improvement algorithm consists of carrying out iterations of type 1 and 2 until no replacements can be made.

3.13 Post-Processing the Solution

Experience has shown that Lagrangian relaxation and subgradient optimization not necessarily yield acceptable primal solutions. Therefore a simple post-processing of a selection of solutions obtained in the final stages of the iteration process was done, and the cheapest solution obtained in this way was selected.

Typical elements in the postprocessing were:

- Increase the cable capacity to all nodes with insufficient cable supply;
- Reduce the cable capacity to all nodes where this is possible;
- Move a subscriber to another distributor if this is profitable;
- Replace a cross connector or distributor with one of another type if this is profitable;
- Try to eliminate nodes with low utilization.

3.14 Related Work

In the combination optimization, only a subset of the trench sections are considered at their real cost. A formulation was implemented where all cable and trench options were considered simultaneously.

4 FABONETT– Planning the Access Network

When it was decided to place Service Access Points (SAPs) and ring structures in the access network, Telenor R&D was requested to make a planning tool, which could assist in finding cost-effective access network designs.

A local switch (*LS*) and a set of main distribution points (*MD*s) together with a set of what we call special subscribers (*SS*s) were given. The MDs and the SSs could be connected directly to *LS* or via service access points (*SAP*s) where multiplexing was done. FABONETT operated with copper cables, fibre cables and, by abuse of language, radio cables. An SS should be connected to *LS* or to a SAP by either fibre or radio cables. An MD should be connected to *LS* or to a *SAP* by copper cables. A sequence of cables which connected an MD or an SS to a SAP or to *LS*, or which connected a SAP to another SAP or to *LS*, was called a *connection*. The SAPs may belong to SDH rings, which must go through LS. In an SDH ring there were connections between pairs of contiguous SAPs in the ring, and connections between *LS* and SAPs adjacent to *LS* in the ring. Each cable was placed in a sequence of contiguous *trace sections*. A trace section was characterized by its *cost*, *length*, *type* and one or more *section codes*. The trace section type determined inter alia which cable types that could be placed in the trace section. Typical trace section types were *conduits* and *ducts* (existing or new), *trenches* of different categories, *air cable sections* and *radio sections*. A section code was simply a positive integer. Two trace sections shared a common section code if events causing damage to the two trace sections were assumed to be positively correlated. A connection inherited the section codes from the trace sections used by the cables forming the connection. Two connections belonging to the same SDH ring could not share a section code. FABONETT operated with PDH SAPs and SDH SAPs. All SDH SAPs were assumed to contain *add/drop multiplexers* (*ADM*s). If an MD was directly connected to a SAP, the SAP had to contain *RSS*/*RSU*. An SDH SAP could be a *Transmission Point* (*TP*). A TP did not contain RSS/RSU. Consequently, SSs and other SAPs, but no MDs, could be directly connected to a TP.

All SDH rings passed through *LS* and were either *STM1* or *STM4 SNCP* rings. The SDH SAPs were ordered hierarchically: SDH SAPs

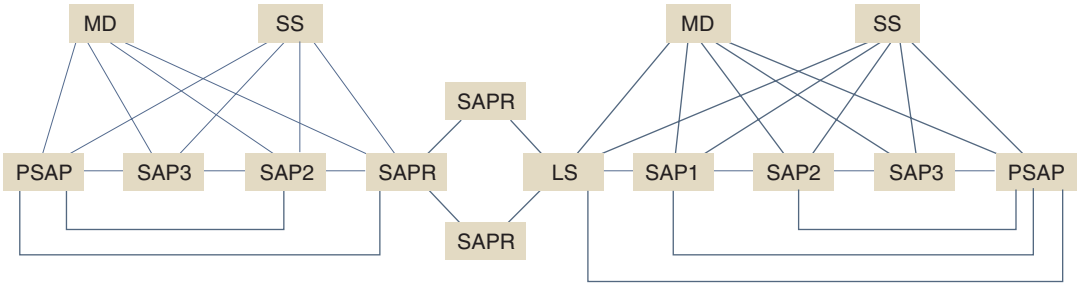


Figure 4.1 Schematic view of the access network structure

which could only be connected directly to *LS* or placed in SDH rings through *LS* were called *S1 SAP*s. SDH SAPs which could only be connected directly to *LS* or S1 SAPs were called *S2 SAP*s. SDH SAPs which could only be connected directly to *LS*, S1 SAPs or S2 SAPs were called *S3 SAP*s. S1 SAPs could belong to STM1 or STM4 SDH rings or not belong to rings at all. S1 SAPs which belonged to rings were called *ring SAP*s. By abuse of language *LS* was also denoted as an *S0 SAP*. If an S2 (S3) SAP *S* was connected to an S1 (S2) SAP *S'*, it was said that *S* was *subordinate to S*'. PDH SAPs could be directly connected to *LS* or to SDH SAPs. There were two categories of PDH SAPs, namely those which required only a single connection, and those which required double connection (i.e. two connections with no section code in common) back to *LS*. The PDH SAPs would normally be connected to *LS* in a way specified by the user. If this was not done, a PDH SAP which required double connection to *LS* would be allowed to be singly connected to a ring SAP. An MD could be directly connected through copper cables either to *LS* or to a SAP which was not a TP. A PDH SAP could only have connected to it MDs which either were connected to it in the existing network or which the planner explicitly connected to it. There could be subscribers connected to an MD who required the MD to be connected directly either to *LS*, to a ring SAP or to a PDH SAP with double connection to *LS*. A circuit connecting an SS to LS belonged to one of two types, namely *regular circuits* and *singular circuits*. The singular circuits had to be directly connected to *LS* through fibre. The regular circuits could be connected directly to *LS* or to a SAP through fibre or radio. Some SSs could require that their regular circuits were connected directly either to *LS*, to a ring SAP or to a PDH SAP with double connection to *LS*. A PDH SAP could only have connected to it SSs which were connected to it in the existing network or which the planner explicitly connected to it. By convention it was said that an SS was connected to a SAP (or *LS*) if the SS's regular circuits were connected to the SAP (or *LS*).

FABONETT did not propose new PDH SAPs. It could, however, propose that PDH SAPs, which in the existing network were directly connected to *LS*, should be connected to an SDH SAP instead. Furthermore, FABONETT/SDH did not invent possible locations for candidate SAPs and candidate trace sections. All candidate SAPs and trace sections must be provided by the user.

Based on the location of *LS*, locations of MDs, SSs, existing cables, existing and candidate trace sections, existing and candidate SAPs, FABONETT tried to find the least costly network design. The design problem was formulated as an integer program which was solved by a combination of linear programming with dynamic row and column generation, branch and bound, and heuristics.

Figure 4.1 gives a schematic view of the network structure.

Since FABONETT did not necessarily solve the design problem to a theoretical optimum, the planner had to inspect the solution and sometimes make model reruns with slightly altered input. The planner could for example question FABONETT's selection of a particular SAP candidate and wish to make a rerun with this SAP excluded. FABONETT's input format made this possible without erasing the SAP candidate from the input. Or the planner could question the correctness of connecting a particular SS directly to LS. A rerun could then be made where the planner specified which SAPs the SS should be allowed to connect to.

The planner had a problem of a dynamical nature. In establishing the best network structure the development of the demand structure over time had be taken into consideration. FABONETT was a static 'one shot' model. Some simple features for 'dynamical use' were, however, built into FABONETT. The planner could give selected SAP and trace section candidates the label 'preferred' and give them a bonus. Then two FABONETT runs were made. First a 'future run' was made where the circuit demands represent some future point in time. Then the main run was made where some or all the candidate SAPs and trace sections chosen in the future run were labelled 'preferred' and given a suitable bonus.

A detailed description of the FABONETT model and solution algorithms may be found in [4].

5 RUGSAM/RUGINETT/PETRA –Transport Network Planning

Up to the early 1980s the routing and grouping of circuits in what is now called the transport network was done manually. Telenor had experimented with the use of planning tools from other telecommunication administrations without much success, and it was decided to give Telenor R&D the task of developing an optimization tool. The development of the tool went through several phases. In the first phase routing and grouping of circuits in a pure PDH network was considered. The tool RUGSAM consisted of two optimization models where the first model optimized routing without considering grouping. Then the model was changed into a single model, RUGINETT, which optimized routing and grouping simultaneously. RUGINETT was generalized to consider a combined PDH and SDH network. In RUGINETT the cable and radio connections were considered as given, and the tool proposed cost-effective routing and grouping of demands in the given network. Eventually it was found that this scope was too narrow, and that the RUGINETT methodology could be generalized to also propose new trenches, ducts, cables, and radio connections, and thus become a combined network design and network utilization tool. Even if the basic ideas were the same, the generalization was so fundamental that the tool was renamed and was called PETRA. PETRA contained an optimization model based on integer programming which proposed the installation of new components like multiplexers, cables and radio connections and ring structures, and at the same time how the network should be used to service forecasted demands. PETRA did not pretend to solve the integer program to optimality. However, the user interface allowed the planner to make modifications to the proposed solutions and check feasibility and costs.

We shall briefly describe the basic concepts that were used in PETRA. The term *route objects* were used as a common name for connections and equipment. The route objects were partitioned into *connection objects* (short: *connections*) and *equipment objects* (short: *equipment*). Equipment objects were situated in *nodes*. Every route object belonged to a *route object type*.

Connections were generally routed on other connections in the connections hierarchy and possibly on nodes according to a set of inputted *routing rules*. In general, however, existing connections or connections specified by the planner could have partial routing only, or no routing at all. The routing rules were visualized by an acyclic *routing graph*. The nodes represented route objects. An arc from node *R* to node *R'* indicated that a route object of type *R* could be routed on a route object of type *R'*. One or more weight factors could be associated with arc (*R*, *R'*) in the graph indicating the fraction of the different capacities of type *R'* taken up by a route object of type *R* routed on *R'*.

Connections that service circuit demands directly were called *demand connections*. The planner could specify that demand connections associated with a certain demand could or could not be routed on by other connections. If a demand could be routed on, it was denoted *accessible*. Otherwise it was denoted *inaccessible*. Accessible demands were treated in the following manner in the optimization module:

- For every accessible demand, one or more connections (which might be routed on) without routing were established which the optimization regarded as existing. Normally no existing connections were routed on these connections, but the planner could specify any partial use of them.

- Each accessible demand was changed into an inaccessible demand where a possible requirement for diversified routing was maintained.

- After the optimization (which operated with inaccessible demands only), whatever was routed on the accessible connections without routing was shifted over to the connections which covered the created demands.

A connection belonged to exactly one *connection type*. Typical connection types were

- different variants of 2 Mb groups/circuits
- x Mb PDH groups/circuits/transmission systems
- x Mb SDH virtual containers/multiplex sections/groups/circuits
- x Mb SDH ring sections
- x Mb WDM groups
- cables and radio links
- XDSL connections, point-to-multipoint connections, conduits and ducts

A connection could be *one-way*, *two-way* or *undirected* and was characterized by capacities, costs and how it was governed by the routing rules. In particular, one-way connections could be routed on one-way, two-way, and undirected connections, whilst two-way connections could be routed on two-way and undirected connections only, and undirected connections could be routed on undirected connections only.

The number of possible connections is huge, and it would be impossible to introduce them all as decision variables in the optimization. They were therefore generated dynamically through

a column generation technique. Another means of keeping the number of variables down was to operate with *connection sets* in lieu of connections, where a connection set was a set of connections with identical routing.

The (mixed integer) optimization problem could in principle be solved to optimality by a branch-and-generate algorithm. However, the very size of the problem made this prohibitive, and the problem was instead solved by a combination of linear programming and heuristics.

Like the access network planner, the transport network planner had a problem of a dynamical nature. In establishing the best design, the development of the demands over time had to be taken into consideration. Like FABONETT, PETRA was a static 'one shot' model, and the same features for 'dynamical use' were built into PETRA. Two PETRA runs could be made, first a 'future run' where the demands represent some future point in time, and then a main run where some or all the route objects chosen in the future run could be labelled 'preferred' and given a bonus.

A detailed description of an early version of RUGINETT may be found in [5].

6 MOBANETT – GSM Access Network Planning

Like for the other networks we have discussed, planners of mobile networks in Telenor had found that manual planning of the GSM access network was time-consuming, and that they would have capacity for analyzing a few alternatives only. Therefore Telenor R&D was again given the task of developing a suitable PC-based planning tool. The result was the tool MOBANETT which attempted to find a GSM access network which minimized total cost. Like the other tools, MOBANETT did not pretend to solve the cost minimization problem to optimality. However, the accompanying user interface allowed the planner to make modifications to the solution found by MOBANETT and check feasibility and cost.

6.1 General

MOBANETT consisted of several modules which could be put into one of two categories:

- optimization modules which found solutions to the cost minimization problem by mathematical and graph teoretical methods;
- modules which enabled the user to interface with MOBANETT via (input and output) tables and which prepared the data formats suitable for the optimization modules.

6.2 The GSM Access Network Design Problem

The mobile subscribers can connect to a set of given *Base Transceiver Stations (BTSs)*.

Several BTSs can be connected together forming a rooted tree. The BTS sitting at the root of such a tree is called an *anchor BTS*.

Each anchor BTS must be connected to a *Base Stations Controller (BSC)*, possibly via a *Digital Access Cross Connect System (DACS)*. In MOBANETT each BTS is characterized by a number of 64 kbit/s radio channels. The connection must have sufficient capacity to be able to carry the total number of radio channels for all BTSs in the rooted tree associated with the anchor BTS.

Each BSC must in turn be connected to a *Mobile Services Switching Centre (MSC)*. The connection must have sufficient capacity to carry the traffic from the BSC subject to a given blocking probability.

Each MSC must be connected to one or two *Main Switches (FS2s)*. The connection must have sufficient capacity to carry the traffic from the MSC subject to a given blocking probability and a given so-called *redundancy factor*. To simplify the presentation we shall assume that each MSC should be connected to two FS2s.

An MSC and a BSC can be *collocated* in order to reduce cost.

Thus the GSM access network has a tree structure. This structure is shown in Figure 6.1.

6.3 Cost Minimization, General Description

In Figure 6.1 the network to be designed lies between the two horizontal lines.

The locations of the FS2s and the BTSs and the number of radio channels between the anchor BTSs and the BSC/DACS were input to MOBANETT. MOBANETT tried to determine

- where MSCs, BSCs and DACSs should be placed;
- which BTSs should be connected to which DACS/BSCs;
- which BSCs should be connected to which MSCs;

in order to obtain a feasible access network at the lowest possible cost.

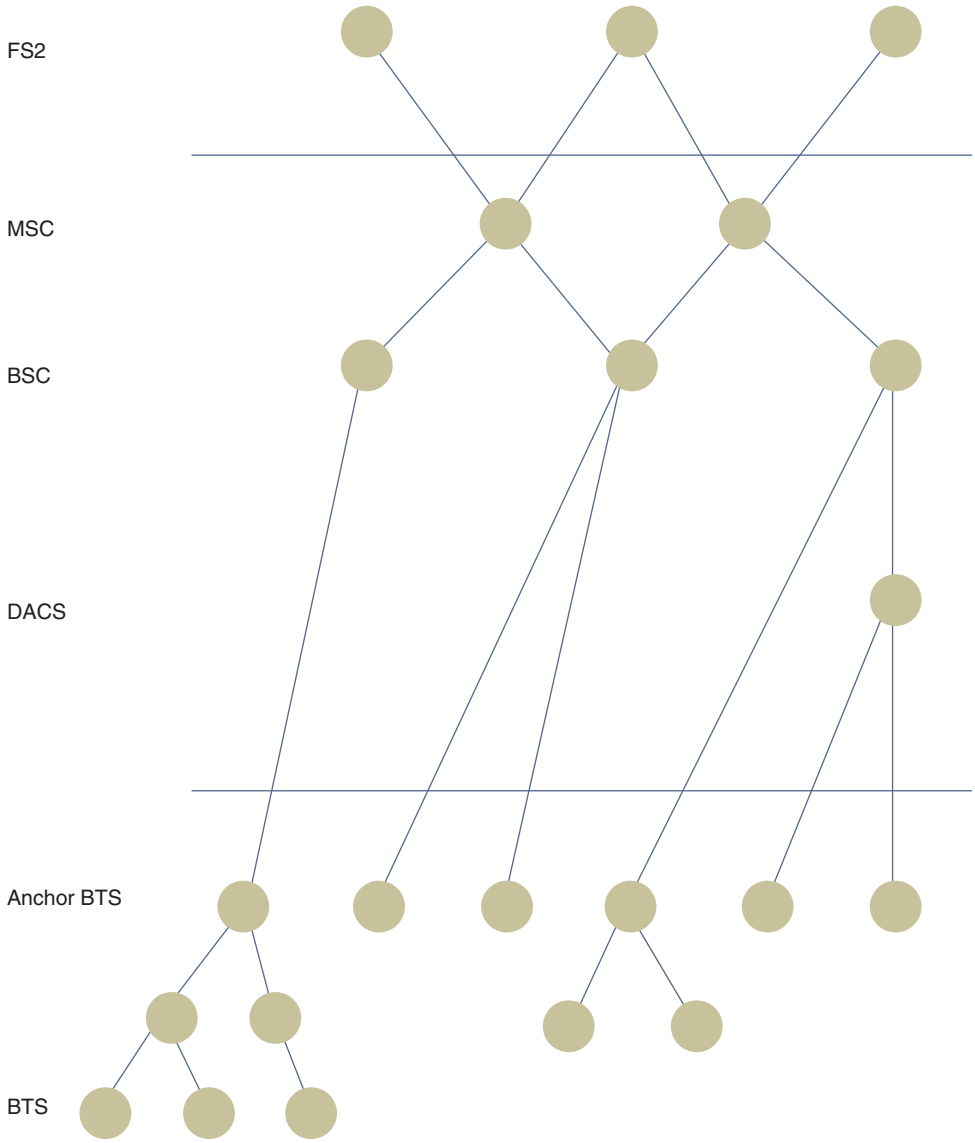


Figure 6.1 GSM access network structure

The cost minimization was done with algorithms which relate to a particular graph called an *options graph*. An example of an options graph is shown in Figure 6.2. Here the nodes between the two horizontal lines represent *candidate DACSs, BSCs* and *MSCs*, and the edges represent *candidate connections*. Each of the candidate connections can carry a number of 64 kbit/s circuits. Between BTSs and BSCs each circuit can carry one radio channel, while between BSCs and MSCs, and between MSCs and FS2s, each circuit may carry F (usually 1 or 4) channels where F is set by the user. Each of the edges in the options graph had a cost function associated with it which gives the connection cost as a function of the number of circuits connected. The algorithms tried to find the subtree of the options graph which at the lowest possible cost connects the BTSs via DACS/BSCs and MSCs to two FS2s. The two FS2s which a candidate MSC should be connected to (if it was selected) were input to the optimization although MOBANETT would propose which two FS2s to use. In the options graph a DACS may be connected to one BSC only. So if one wanted to model several optional connections for a candidate DACS, the DACS had to be duplicated (an example of which is shown to the right in Figure 6.2).

6.4 Mathematical Model

6.4.1 Notation

The following notation will be used:

Subscripts:

- t: BTS
- d: DACS
- m: MSC
- $b(d)$: the BSC to which DACS d is connected

Constants:

- k_t: number of radio channels to be connected from BTS t
- e_t: the traffic measured in erlang at BTS t
- f: 1 – blocking probability in BSC and MSC
- F: number of channels per BSC – MSC circuit, and per MSC – FS2 circuit

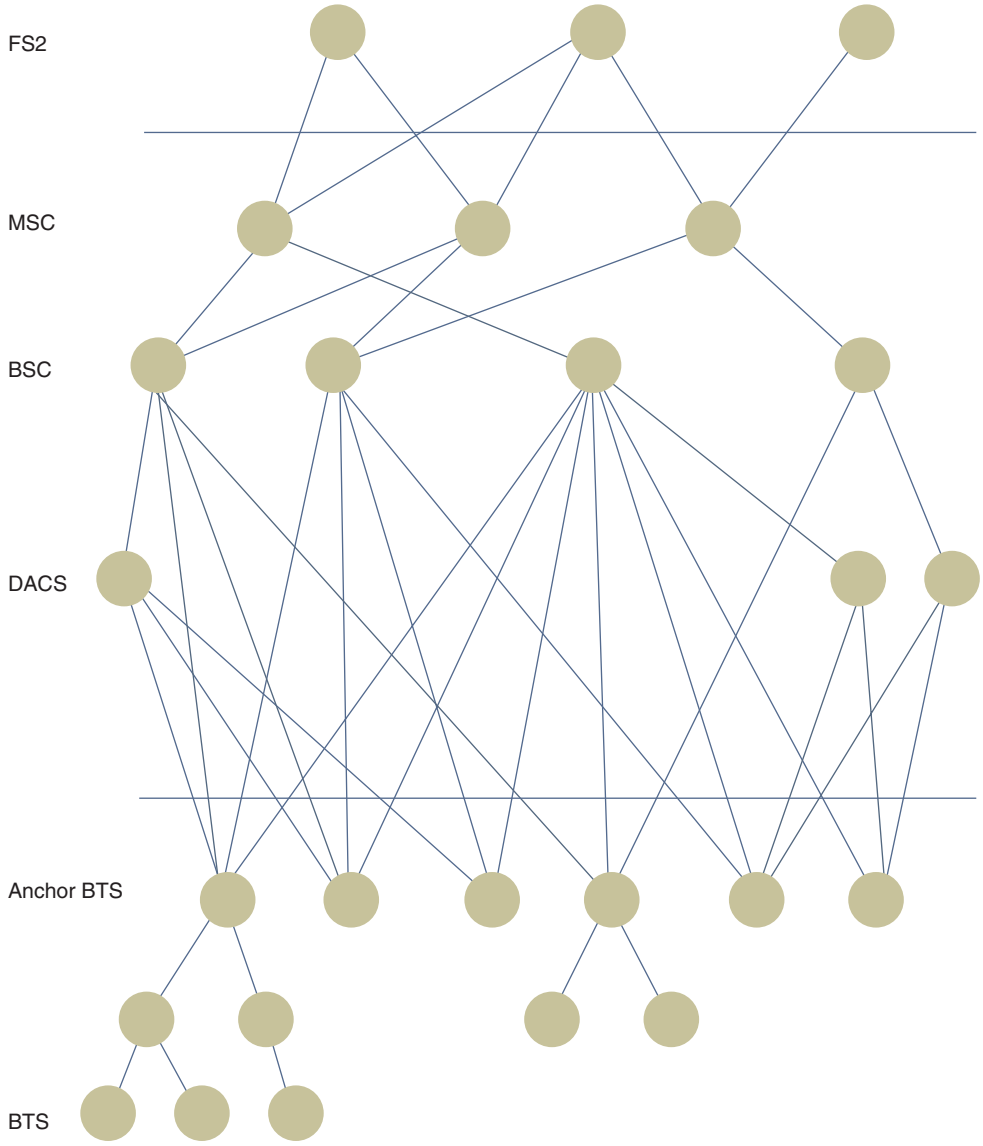


Figure 6.2 Options graph

R_m: redundancy factor between MSC m and FS2

c_{mk}: cost of connecting k channels between MSC m and FS2

Decision variables:

x_{td} = 1 if BTS t is connected to DACS d, and 0 otherwise

x_{tb} = 1 if BTS t is connected to BSC b, and 0 otherwise

x_{dk} = 1 if there are k channels between DACS d and BSC $b(d)$, and 0 otherwise

x_{bmk} = 1 if there are k channels from BSC b to MSC m, and 0 otherwise

x_{mk} = 1 if there are k channels from MSC m to FS2, and 0 otherwise

Cost coefficients associated with the variables:

c_{td}: cost of edge between BTS t and DACS d

c_{tb}: cost of edge between BTS t and BSC b

c_{dk}: cost of connecting k channels between DACS d and BSC $b(d)$

c_{bmk}: cost of connecting k channels from BSC b to MSC m

The cost coefficients will reflect connection costs as well as costs of DACS, BSC and MSC. We shall return to the cost structure later.

Auxiliary variables:

e_b: offered traffic at BSC b

e_m: offered traffic at MSC m

Linearized erlang functions:

T: traffic in erlang

k: number of channels

$E(T)$: number of channels needed as a function of T (for given f)

E_0 and E: constants in the linearization of $E(T)$

$E(T) = E_0 + ET$ for $T > 0$, $E(T) = 0$ for $T = 0$

with inverse

$E^{-1}(k) = -E_0 / E + k / E$ for $k > E_0$, $E^{-1}(k) = 0$ for $T = 0$

6.4.2 Optimization Model

The cost minimization problem was formulated as follows:

$$minimize\ \Sigma c_{td}x_{td} + \Sigma c_{tb}x_{tb} + \Sigma c_{dk}x_{dk} + \Sigma c_{bmk}x_{bmk} + \Sigma c_{mk}x_{mk}$$

subject to

$$\Sigma_t k_t x_{td} \text{ " } \Sigma_k k x_{dk} \quad \text{(balance in DACS } d\text{)} \quad (6.1)$$

$$E(e_b) \text{ " } F\Sigma_{mk} k x_{bmk} \quad \text{(balance in BSC } b\text{)} \quad (6.2)$$

$$R_m E(e_m) \text{ " } \Sigma_k k x_{mk} \quad \text{(balance in MSC } m\text{)} \quad (6.3)$$

where

$$e_b = \Sigma_{t,d\in b} e_t x_{td} + \Sigma_t e_t x_{tb}$$

and

$$e_m = \Sigma_b f E^{-1}\ (F\Sigma_k k x_{bmk})$$

Using the formulas for $E(T)$ and $E^{-1}(k)$ transforms (2) and (3) to (4) and (5):

$$E\ \Sigma_{t,d\in b} e_t x_{td} + E\Sigma_t e_t x_{tb} \text{ " } \Sigma_{mk}(Fk - E_0)x_{bmk} \quad (6.4)$$

$$R_m f\Sigma_{bk}(Fk - E_0)x_{bmk} \text{ " } \Sigma_k(k - R_m E_0)x_{mk} \quad (6.5)$$

Note that we must have $k \geq E_0 / F$ in BSCs and $k \geq R_m E_0$ in MSCs. This defines the minimal values k_{min} which k can take.

Thus the optimization model becomes:

$$minimize\ \Sigma c_{td}x_{td} + \Sigma c_{tb}x_{tb} + \Sigma c_{dk}x_{dk} + \Sigma c_{bmk}x_{bmk} + \Sigma c_{mk}x_{mk}$$

subject to

$$\Sigma_t k_t x_{td} \text{ " } \Sigma_k k x_{dk} \quad \text{(balance in DACS } d\text{)} \quad (6.6)$$

$$E\ \Sigma_{t,d\in b} e_t x_{td} + E\Sigma_t e_t x_{tb} \text{ " } \Sigma_{mk}(Fk - E_0)x_{bmk}$$
(balance in BSC b) (6.7)

$$R_m f\Sigma_{bk}(Fk - E_0)x_{bmk} \text{ " } \Sigma_k(k - R_m E_0)x_{mk}$$
(balance in MSC m) (6.8)

where

$k \geq E_0 / F$ in BSCs and $k \geq R_m E_0$ in MSCs.

6.5 Cost Structure

A detailed presentation of the cost elements which the planner gives via tables to MOBANETT is not given here. Prior to the optimization the costs of components (DACS, BSC, MSC) and connections as a function of the number of radio channels c were approximated by functions $C(c)$ which in general were sums of up to three terms. These terms were

a constant part = C,
a linear part Lc, and
a 'saw tooth' part $T(c) = T \cdot (1 - c(\text{mod } 30) / c)$

Thus $C(c) = C + Lc + T(c)$.

The saw tooth part reflected cost elements which depended on the number of 2 Mbit/s circuits.

We shall now describe the individual cost functions and how they were allocated to the edges in the options graph.

6.5.1 DACS Cost

The cost of a DACS consisted of a cost per 2 Mbit/s circuit connected to a BTS or a BSC. The cost decomposed into a linear part and a saw tooth part. The linear part translated into a constant part which was put on each of the edges connecting to BTSs. The saw tooth part was put on the edge connecting to the BSC.

6.5.2 BSC Cost

The cost of a BSC consists of a fixed cost plus a cost per radio channel coming from the BTSs which are connected to it (possibly via DACS). The cost decomposed into a constant part and a linear part. The constant part was put on the edges connecting to MSCs. The linear part translated into constant parts on the edges between BTSs and the BSC and on the edges between BTSs and the DACSs associated with the BSC.

6.5.3 MSC Cost

The cost of an MSC consisted of a fixed cost plus a cost per 2 Mbit/s circuit connecting to a BSC or FS2. The cost decomposed into a constant part on the edges to the FS2s, a linear part on the edges to the BSCs, a saw tooth part on each of the edges to BSCs and a saw tooth part for each of the edges to FS2s.

6.5.4 Cost of Connection Between BTS and DACS/BSC

The cost of the connection between a BTS and a DACS/BSC consisted of a cost per 64 kbit/s or per 2 Mbit/s. In both cases this translated into a constant cost on the edge between the BTS and DACS.

6.5.5 Cost of Connection Between DACS and BSC

The cost of the connection between a DACS and a BSC consisted of a cost per 2 Mbit/s. This decomposed into a linear part which translates into constant parts on the BTS – DACS edges and a saw tooth part on the DACS – BSC edge.

6.5.6 Cost of Connection Between BSC and MSC

The cost of the connection between a BSC and an MSC consisted of a cost per 2 Mbit/s. This

decomposed into a linear part which translated into linear parts and saw tooth parts on the BSC – MSC edges.

6.5.7 Cost of Connection Between MSC and FS2

The cost of the connection between an MSC and an FS2 consisted of a cost per 2 Mbit/s. This decomposed into a linear part which translated into a linear part and a saw tooth part on the MSC – FS2 edge.

6.5.8 Collocation of BSC and MSC

It could be cost effective to place a BSC and an MSC in the same node. This was treated simply by introducing into the options graph additional nodes for a candidate MSC and a candidate BSC with reduced costs and a zero cost connection between them.

6.5.9 Shifting Linear Edge Costs Towards BTS

It was believed that the optimization algorithms function better if the cost elements were shifted to edges in the options graph which are as close to the BTSs as possible. We shall now describe how this shifting is done in principle.

We have already seen how linear costs associated with connections between DACS and BSC were changed to constant costs on DACS – BTS edges. We similarly translated the linear costs on the FS2 – MSC edges to corresponding MSC – BSC edges. In order to do this it was necessary to find the number k of channels between FS2s and an MSC as a function of the number k' of channels between the MSC and MSCs. Inequality (6.5), which could just as well have been written as an equality, gives

$$R_m f \Sigma_b (Fk' - E_0) = k - R_m E_0 \qquad (6.9)$$

where the sum is over the BSCs connected to the MSC.

This gives

$$k = R_m E_0 + \Sigma_b (R_m f F k' - R_m f E_0) \qquad (6.10)$$

A linear term Lc on the edge MSC – FS2 thus translates into

- a constant term $-LR_m f E_0$ on each BSC – MSC edge
- a linear term $LR_m f F k'$ on each BSC – MSC edge
- a constant term $LR_m E_0$ on the FS2 – MSC edge

Linear costs on the edges between MSCs and BSCs cannot be moved towards the DACS/BTSs in the same way because a BSC can be connected to several alternative MSCs. The best we can do is for a BSC to move a part of each linear MSC – BSC cost equal to the minimal linear MSC – BSC cost for this BSC. This was done after all other costs had been allocated and possibly shifted. Let L_{min} be the minimal coefficient for the linear costs on the edges from MSCs and let L be the corresponding coefficient for the edge to an arbitrary MSC. The linear cost coefficient on the BSC – MSC edge was changed to $L - L_{min}$.

Considerations analogous to those above give the number k of channels between BSC and MSC equal to

$$k = E_0 / F + E\Sigma_{t,d \in b} e_t / F + E\Sigma_t e_t / F.$$

The cost shifting thus resulted in

- a constant term $L_{min} E e_t / F$ on each BTS – DACS edge
- a constant term $L_{min} E e_t / F$ on each BTS – BSC edge
- a constant term $L_{min} E_0 / F$ on each BTS – DACS edge
- a linear term $(L - L_{min})k$ on the BSC – MSC edge

6.6 Candidate BSC and MSC Nodes and Connections

In order to relieve the planner of the tedious task of inputting all locations for candidate BSC locations a simple algorithm was developed which proposed candidate BSC nodes.

The geographical locations of the BTSs and FS2s were given as input to MOBANETT. In addition it was possible to give as input the locations of NMT base stations and NMT switches (MTXs). All these geographical points had the possibility of becoming candidate BSC nodes.

The MTXs were automatically made into candidate BSC nodes, and the planner would add on more candidate BSC nodes of his choice. Thereafter he could apply the algorithm below.

Algorithm which proposes additional candidate BSC nodes:

1 Draw the largest circle possible around every candidate node (with the node as centre) so that the total traffic generated by the BTSs in the circle is below a given limit (the exclusion limit), and exclude all points in the circles from the possibility of becoming candidate nodes.

2 If every geographical point is either a candidate node or excluded, stop. Otherwise draw the largest possible circle around every geographical point (with the point as centre) which is not yet a candidate node and which is not yet excluded, so that the total traffic generated by the BTSs in the circle is below a given limit (the inclusion limit). Define the point with the smallest circle as a candidate BSC node. Go to 1.

After the candidate BSC nodes had been decided upon the planner selected a subset of these, possibly augmented by some other geographical points, as candidate MSC nodes.

The planner could also get some assistance in setting up candidate connections. MOBANETT would for every MSC candidate propose connections to the two closest FS2s and to the closest candidate BSC nodes up to a given number and within a certain distance. Furthermore MOBANETT would for each candidate BSC propose connections to the closest BTSs up to a given number and within a given distance. Connections via DACS were also proposed according to certain criteria which we shall not go into here.

6.7 Solving the Optimization Model

6.7.1 Use of Lagrange Relaxation

The optimization model as defined in 3.2 was solved by using Lagrange relaxation and subgradient optimization. The technicalities connected with the use of the subgradient method are standard and will not be described here. Inequalities (1), (4) and (5) were relaxed. However, the condition that the BTSs must be connected to FS2s in a tree structure was retained as a constraint. The relaxed problem thus became a classical Steiner problem in a rooted directed acyclic graph. The root node was an auxiliary node connected to the candidate MSC nodes where the cost of the connection reflected the cost of connecting the MSC to its two FS2s, and where the special nodes to be connected to the root represented the BTSs.

6.7.2 Solving the Steiner Subproblem

Since the Steiner subproblem had to be solved a substantial number of times a heuristic is used. The heuristic chosen was Wong's dual ascent algorithm [2] followed by Pacheco-Maculan's solution improvement algorithm [3], both specialized to rooted directed acyclic graphs.

7 GSM Frequency Planning by MOBINETT

The requirements for GSM services increased rapidly in the early 1990s, and it was realized that it was crucial to have access to a good PC-based tool for frequency planning. Telenor R&D had earlier experimented with algorithms for finding the least number of frequencies necessary to carry a given traffic in a network with a given set of BTSs with sufficiently low level of interference. Telenor R&D was then given the task of instead making an optimization tool which assigned a given number of frequencies to each BTS such that the level of interference was acceptable and minimized. The input was a set of admissible frequencies (which needed not be contiguous) and a symmetric compatibility matrix which for each pair of BTSs indicated whether they interfered on neighboring frequencies, on same frequencies only, or not at all. MOBINETT went through several stages.

The first variant considered only BTSs which supported neither baseband nor synthesizer hopping. This problem was formulated as a pure integer programming problem which was initially solved by a combination of linear programming and heuristics. For realistic cases the number of variables was reasonably small (about 10,000) whilst the number of constraints could be rather large (> 150,000). Since the linear programming software we used performed better with a high number of variables and a low number of constraints than vice versa, the transposed problem (where the dual variables had to be integers) was solved instead. In order to solve the problem to optimality the heuristic was replaced by a dual 'branch and cut and generate' software package that was designed and implemented at Telenor R&D.

As new BTSs which supported baseband and synthesizer hopping became available, MOBINETT had to be modified accordingly. The compatibility matrix was replaced by two asymmetric interference matrices, one describing interferences from interfering BTSs to victim BTSs on the same frequency, and one describing corresponding interferences on neighboring frequencies. Thresholds were set for acceptable interference. This necessitated a complete remodeling of the problem, and new solution algorithms had to be implemented. The new MOBINETT tried to allocate frequencies to cells such that the frequency requirements were satisfied and such that a weighted sum of interference contributions above given thresholds were minimized.

The planner could partition the cells into subsets and solve the allocation problem for one subset at a time. The basic approach was to formulate the subproblems as integer programs which was solved by classical branch and bound. In addition, postprocessors based on tabu search and simulated annealing were implemented. MOBINETT is at the time of writing still in use.

8 Conclusion

We see that Telenor R&D over the years has been involved in the establishment of network planning tools for most parts of the physical network. This had not been possible without the participation and support from dedicated network planners in Telenor. The main hurdles have been:

- The varying quality of the data in the network databases. A spin-off effect of the planning tool development has been a substantial increase in the accuracy of some of the data sources.

- The responsibility for the planning of the different networks was in the past decentralized. The local planners had a variety of responsibilities, and it was often difficult for them to allocate the time necessary to acquire and maintain the necessary familiarity with the tools. Also, the ICT equipment available to the planners at the local level was not always sufficient for making effective use of the tools.

There are reasons to believe that these hurdles will be easier to overcome in the future:

- The requirement for high accuracy in the network databases is becoming more and more pronounced, not only because of requirements from planning tools.

- There is a tendency to centralize the network planning activity. This implies that planners can dedicate more of their time to get the most out of sophisticated planning tools.

- The reliability, capacity and speed of ICT equipment has become more than sufficient to support the optimization algorithms that form the motor in modern network planning tools.

Finally it should be observed that the development, implementation, and use of network planning tools ought to be an ongoing process. Both the demands for new services and the technology that can be used to serve them evolve at increasing speed.

9 References

1 Rayward-Smith, V J, Clare, A. On Finding Steiner Vertices. *Networks*, 16, 283–294, 1986.

2 Wong, R T. A Dual Ascent Approach for Steiner Tree Problems on a Directed Graph. *Mathematical Programming*, 28, 271–287, 1984.

3 Pacheco, O I P, Maculan, N. Metodo Heuristic para o Problema de Steiner num Grafo Direcionado. *Proceedings of the III CLAIO*, Santiago, Chile, August 1986.

4 Lorentzen, R. Mathematical Model and Algorithms for FABONETT/SDH. *Telektronikk*, 94 (1), 135–145, 1998.

5 Lorentzen, R. Mathematical Methods and Algorithms in the Network Utilization Planning Tool RUGINETT. *Telektronikk*, 90 (4), 73–82, 1994.

Network Strategy Studies

TERJE JENSEN

Dr. Terje Jensen (41) is Research Manager 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 on network planning, performance modelling/analyses and dimensioning.

terje.jensen1@telenor.com

Every network operator needs to have a network strategy covering the complete network portfolio. The strategy must also be operational, meaning to be related to decisions and actions in near-time. One main goal for having a strategy is to be prepared for chances that can be revealed with time. That is, the strategy is likely to assist in detecting business opportunities.

A number of methods can be applied when elaborating the strategy, including scenarios, cost/benefit calculations and risk assessment – these aspects are briefly presented in this article.

1 Introduction

In today's world, change and uncertainty seem to be constants. On the one hand this is nice, as improvements are captured in the evolution process. On the other hand, the ongoing dynamics is hard to capture in network planning with a longer time horizon. In fact, a claim could be that there is no point at all in looking a few years ahead as for (almost) certain, the future will not be exactly as predicted. It is fair to say that this claim is missing the main objectives of carrying out strategic planning, for networks as well as for other areas. Besides being incorrect, the claim disregards the benefits following from a systematic analysis of a company's surroundings and future options.

A motivation for evaluating the coming available choices is to use the results to elaborate a robust action plan to cover triggers for making decisions or to initiate actions as well as optional actions. As described later a robust plan supports a flexible roadmap of system/network evolution, that is what to do for each of the systems given certain factors. Hence, a consequence of this is to try to postpone final decisions until the actions have to be started. This gives the option to not start the action or carry out another action.

In an economic sense, there are at least two motivations why flexibility should be included in an investment evaluation: Firstly, the estimate of the activity value is improved, better reflecting the actual characteristics of the challenge. Another motivation is the improved insight gained for uncertainties and hence a better understanding of the flexibility and the options expected. Commonly real options are applied to include these options in the evaluation. Bringing real options into the equation increases the level of activity values, although the increase differs for the different nature of activities.

The following sections give an overall description of strategy work, both the intention (Section 2) and the overall approach (Section 3). A number of criteria are needed to choose between the different options, as elaborated in Section 4. As presented in Section 5, a sound strategy must also relate to trends at various levels. Section 6 presents how scenario work can be applied in order to assist when deriving strategies, followed by examples in Section 7. A quantitative approach is then given in Section 8. Assessing risk is also an essential element in any strategy as described in Section 9. Some overall discussions are then given in the last two sections.

2 Relating Long and Short Term – the Scope and the Challenges

It is fundamental to understand that the strategic evaluations and planning are not carried out for their own purposes per se, but need to be related to the current situation in order to get practical implications. Naturally a number of steps in the strategy can be located some time ahead, and hence do not necessarily influence today's activities. This means that one result is to sort out the decisions to be made in a timely manner. However, one main result is to devise the actions needed in the actual situation an actor finds itself in. The overall process is depicted in Figure 1.

A number of general steps can be identified:

i. The "frame" of the task is described by the elements: a) Current network portfolio, b) Forecasts and trends, c) Set of (optional) target states. Naturally, these are interrelated as a potential target state is influenced by one's position in the existing situation and a scenario for further development. That is, a scenario-based approach will likely apply for this step. Included in the current portfolio goes also an assessment of short-term development, e.g. given through a "historic" description in the traffic/user development in a system.

ii. The "gap" between the current situation and the target states is explained with a set of possible paths. That is, a path will tell a story for how the current network portfolio will

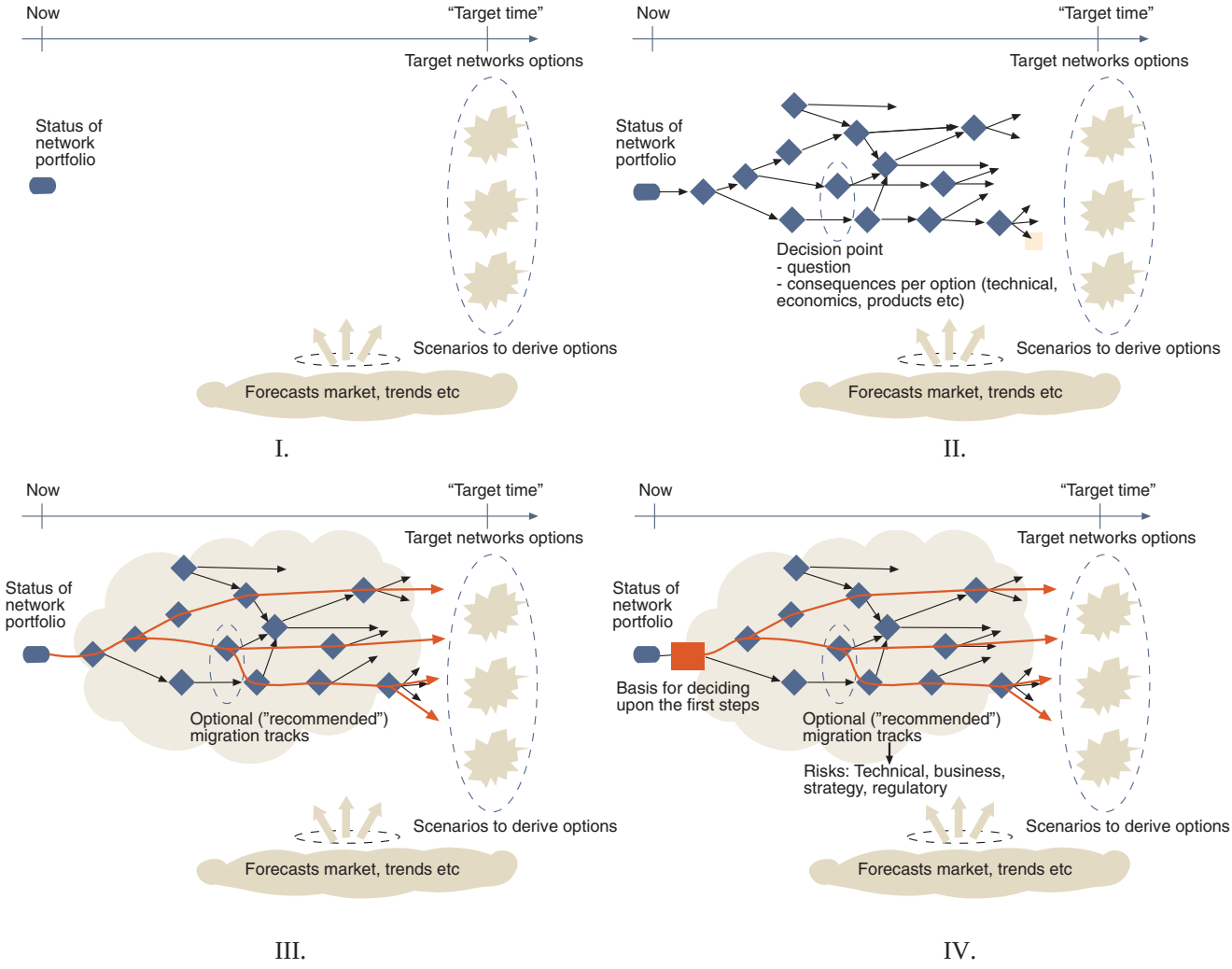


Figure 1 Illustration of step-wise approach using the target options for deriving recommended next steps for today's networks

migrate in order to reach a target state. During a migration, a number of decisions has to be made and solutions implemented. A decision is indicated by a diamond in Figure 1. For each decision, in principle, other choices could have been made, possibly resulting in branches to the path. A decision point should be estimated in time (i.e. when should the decision be made), in addition the relevant question has to be described together with consequences of each of the relevant outcomes of the choices made (consequences in terms of technical, economic, product, etc. issues).

iii. Considering the set of optional (migration) paths and set of decision points, a decision map can be drawn. This then contains the set of candidate paths than can be followed towards the targets. Commonly, quite a few of the theoretical paths are not likely for practical reasons, including financial measures. Hence, a number of real candidates can be described, of which fewer can possibly be recommended migration tracks. These candidates may further be compared according to agreed upon measures, such as net present value, internal rate of return, financial needs, etc.

iv. Looking at the candidate tracks, a number of risk factors can be attached to each track – describing technical, business-related, strategic, regulatory and other risk phenomena. As shown later, these risk factors can be described qualitatively or included in a quantitative way. So, carrying out this exercise, it is essential to relate the observations made to how to make the first step for the current network portfolio. That is, the results are used when making the choice of how to proceed with today's networks. It is also important that the information is revisited and revised as necessary on a regular basis; including the target options, the decision map and the risk evaluations.

Working with a horizon for a given number of years, it is natural to start by gaining insight into trends and key drivers for the network evolution. However, quite a few optional network solutions are expected. Being able to select an appropriate migration of an operator's network portfolio, a set of evaluation criteria has to be defined. To

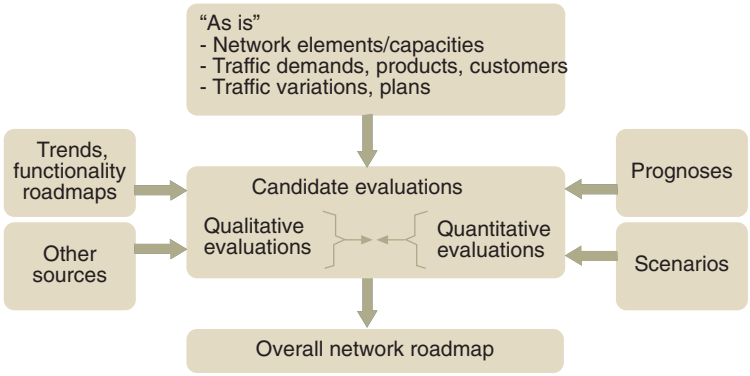


Figure 2 Input and results for evaluation (qualitative and quantitative)

some extent, these criteria can be quantified although some qualitative assessments will also be necessary.

An overall objective of an actor is to level the profit level, both short- and long term. Then, facing the brutal reality of long-term economics, the basic question of which networks and systems to base the operator's future operation on is catching up. As commonly seen elsewhere, managing to terminate a (still) profitable system seems like a harder decision to make than to start up a new system. Hence, the conglomerate of systems that emerge may grow. In addition, several of the systems seem to gradually increase overlap of their application area. This means that in the future more candidates will enter the arena, capable of supporting a spectre of products including those already carried today. One consequence is that the answer to which system to base the future on is steadily more difficult to find.

On the other hand, a few trends seem to dominate the industry, such as IP, optics, DSL and mobile/wireless. But, in view of the current network portfolio managed by an operator and the need to constrain investments and operations to profitable ones, there is still a question in where to invest in order to arrive at a future-proof portfolio. Moreover, the overall risk should be acceptable, considering the profit level sought.

Bearing in mind the overall objective, the goals of the strategic evaluations are to:

- Describe an operator's target network portfolio,
- Identify smart timing of new investments and migration of existing platforms,
- Establish a foundation for decisions on investment and operational aspects to keep a low cost level:
  - Obtain lower overall investment levels, balanced between different areas with corresponding timing (access/transport/service platform, support systems),
  - Ensure complete solutions; consider running costs as well as investments.
- Ensure that an operator maintains the leading position for introducing innovative/holistic network solutions adapted by the broadband market (and others), by enabling rapid service provision.

The priority placed on the different initiatives has to be future-proof, meaning that the steps taken have to fit into a longer-term "journey" approaching a target state of an operator's operation. This basically captures the motivation as well as the overall challenge currently faced by several operators.

3 General Approach

An overall schematic illustration of the strategic evaluation task is given in Figure 2. This is to describe evaluations needed to make the decisions shown in Figure 1. The following main groups of work items are:

a) Description of *status and current plans* ("as-is") for the networks/systems looked at. Configurations at the selected locations are described in terms of network elements (manufacturer, capacity, links, etc.), traffic flows and traffic variation.

b) Elaboration of *optional futures/scenarios* in terms of i) service demands/prognoses, ii) trends for network solutions (and corresponding functionality), and, iii) major uncertain issues (captured by the scenarios).

c) *Other sources*, e.g. from analysis companies, standards, vendors, and so forth.

d) Elaboration of *coherent plans for network roadmaps* considering the dependencies between different systems/networks utilising their strengths.

In order to elaborate plans for network roadmaps the candidates have to be evaluated. For this project, both qualitative and quantitative evaluations have to be conducted. Placing this evaluation in the centre, several inputs have been identified, as shown in Figure 2. The overall results, implying the recommended roadmaps, are then supported by economics as well as functionality arguments, and other arguments related to business and regulatory issues.

4 Criteria for Selecting Migration Branches

4.1 Overall – Financial

In detailing future network candidates, a number of options are revealed in terms of choosing which candidates to base a further network migration on. Hence, there is a need to define a set of criteria to use when selecting which options are the better ones. Some criteria are described in the following.

One basic criterion is the profit levels expected by taking certain steps. Requirements on the expected profit are given considering the accompanying risk level. Hence, higher risk would likely ask for a higher profit level than a lower risk action. Profit can be estimated in different ways. For example, assuming that network solution is irrelevant to income level, the cost of the solution is to be minimised. In general, however, various solutions may support different levels of service and product type. The income side must therefore also be considered in the equation.

A fundamental challenge on the income side is that the price level is influenced by many factors outside the operator considered, such as competitors and regulator. Theoretical models and analyses exist for similar configurations and will not be dealt with in this article.

Concerning financial aspects a number of topics should be looked at, such as:

- Financing needs, e.g. peak funding,
- Value and profitability evaluations, e.g. net present value, internal return rate, payback time,
- Sensitivity analyses of major factors.

As for risks, the technology risk addresses one area. However, a number of additional areas could also be treated:

- Political and market economic risk
- Market and commercial risk (including regulatory)
- Partner risk (also including vendors)
- Financial risk
- Organisational risk (to follow the activities and realise profits)

In order to realise the profit, appropriate steps have to be prepared for in a timely manner, like organisational efficiency, increased revenue, reduced cost, etc. Moreover, exit strategies have to be elaborated to cover alternative steps/paths to follow in case some of the planned steps turn out to be unwanted later on in the process. Which exit strategies that are possible should also be discussed. This corresponds with the map of decision points shown in Figure 1.

The interplay with vendors and customers must also be obeyed. That is, trends among the users have to be observed and possibly matched in order to increase the service demands. Likewise, choices and prioritisations made by the vendors have to be followed, as it could turn out very costly to install and maintain a system from a vendor that is leaving the market or decides not to support that system in the future.

Keeping in mind the broader set of criteria, a few criteria groups are treated in the following (Figure 3):

- Product-related: capabilities of supporting relevant products,

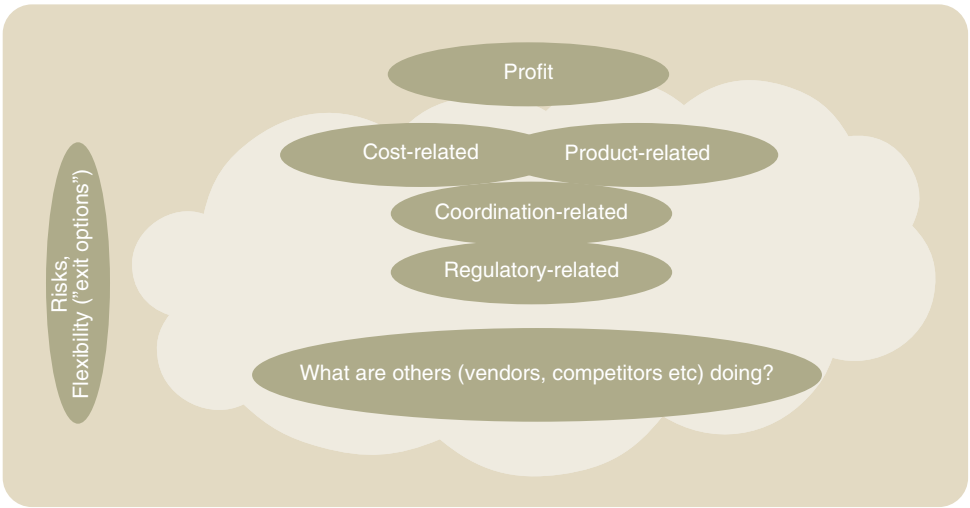


Figure 3 Selected criteria classes

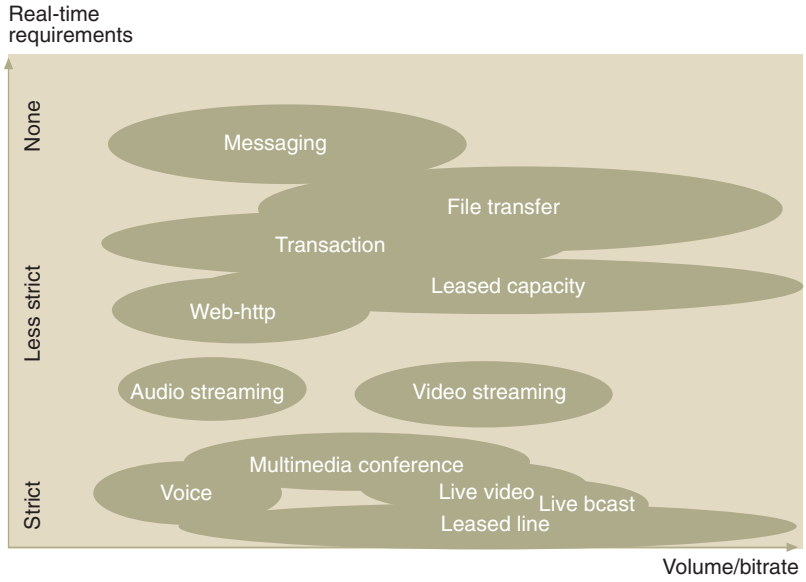


Figure 4 Organising individual end-user services

- Cost-related: the investment and operational costs associated with the network solution (and migration),
- Regulatory-related: would there be any regulatory restrictions or exposure involved,
- Co-ordination-related: are there any effects from the network solutions onto business matters (internal trading, etc.).

The actual criteria weights to use might also depend on the market situation/position, like whether an established operator's situation or a "green field" region is considered. Hence, a somewhat wider scope may be taken on where there is no "history"/legacy systems. To some extent this would also be valid when the overall service portfolio and operation of an actor are looked at – deciding upon which products to offer and which market segments to address. When a set of legacy systems is involved, more emphasis might be placed on managing the systems efficiently and relating the products and systems (production means). This also includes decisions regarding the operations support systems.

4.2 Product-related

All product types relevant for the evaluation have to be considered. This includes not only products seen by end-users, but also products to other operators or departments within an operator when relevant. Moreover, a product may also be bundled together with other products to compose a "new" product.

In the following are treated more product characteristics relating to transport. Basically, such a categorisation could be done in a number of ways. However, to simplify the task, the following factors could be used:

- Timing requirements: Covers real-time requirements, potentially with further variants on timing values (delay and delay variation).
- Volume/bit-rate: The volume is an indicator of the amount of data to be transferred. Considering constraints on the timing, this can also be expressed as requirements on the range of bit-rate needed.
- Interface type/technical solution: Covers types of protocol, etc., and will likely be mostly related to products between operator units.

Some of these may be used to separate products within categories. The timing requirements would likely be a significant factor in addition to the volume/bit-rate. Some concerns may also be related to the content/medium carried and how the two end-points relate to each other (e.g. topology/scope – covering relations between the end-points, like one-to-one, one-to-*N* and *M*-to-*N*).

Some examples using the first two factors are looked at in the following (Figure 4):

- Voice: Carrying speech conversations, which implies interactivity between a set of participants and strict real-time requirements on carrying the voice samples. Currently, most of the voice traffic is carried in PSTN/ISDN and GSM. Typical bit-rates are in the range 8 – 64 kbit/s.
- Live video (VoD): Representing transfer of video in real time between a source and a receiver. This implies fairly strict real-time requirements. Typical bit-rates vary in the area of 500 kbit/s – 6 Mbit/s.
- Multimedia conference: Representing a set of media types (voice, video, text, image, etc.) that are to be transferred between a set of participants in the conference. Typically there are interactivity requirements, although some media types will have less strict real-time requirements. An example of aggregated bit-rates from a single participant is in the range 64 kbit/s – 2 Mbit/s. Some services are offered in a dedicated videoconference network today. In addition, a bundle of ISDN channels may also work as transport. A NetMeeting-like application would also address this product category.
- Video-streaming: Representing the transfer of video content with a play-out buffer in the receiver. Hence, this buffer captures some slack in the transfer-rate variation. The average bit-rate may be the same as video, 500 kbit/s – 6 Mbit/s.

- Audio-streaming: Representing the transfer of audio with a play-out buffer in the receiver. Typical applications would be radio and music. Today radio stations can be listened to through the Internet. Typical bit-rates 16 – 128 kbit/s.

- Web – http: Browsing implies some level of interactivity with a user selecting objects to be transferred. This means some requirements on time for transferring the objects, although no real-time requirements as such. Typical effective (average) bit-rates in the range 15 – 100 kbit/s (although peak bit-rates are higher, say 64 – 500 kbit/s).

- File transfer: This product category covers regular file transfers and more business critical applications, e.g. related to outsourcing. For the latter a high-bandwidth connectivity of 100 Mbit/s or more could be demanded. File transfer represents transfer of (larger) files without real time requirements. Examples of file sizes are 0.5 Mbyte (document), 4 Mbyte (mp3 file), 1 Gbyte (video movie).

- Messaging: Exchange of information between users without real-time requirements. A store-and-forward principle is introduced, allowing for storage of messages in intermediate servers.

- Transactions: Transactions represent messages containing text, images, etc. Examples of volumes are found in the range a few hundred bytes to a few Mbyte. No strict real-time requirements are attached, although an acknowledgement is commonly conveyed to the source.

- Live TV broadcasting: Representing today's broadcasting and future digital systems. Example of capacity is 6 Mbyte per channel. Strict real-time requirements (direction towards the receiver).

- Leased line: Representing leased line services provided, e.g. by the SDH network. Today's bit-rates include 64 kbit/s – 155 Mbit/s and up to wavelengths. Strict real-time requirements are given.

- Leased capacity: Representing a "pipe" from ingress to a set of egress points. Bit-rates may be as for leased line; however, the real-time requirements are less strict.

More detail can be considered when both directions are quantified, see Figure 5. Considering the set of applications used simultaneously also gives indications of which access solutions (bit-rates upstream and downstream) can be chosen.

In several of these products, a number of traffic flows could be involved, each with its separate characteristics. In addition, other aspects could also be considered, including:

- To what extent mobility (and portability) is supported. The highest bit-rates would not be easily provided on wireless/mobile links (except for broadcast networks).

- The topology between the involved communication parties. For conference, multicast and broadcast services, there would be several parties involved. Hence, allowing for multi-party destinations may require corresponding functions in the network. Collection networks may also be considered, where a single receiver gathers information from several sources.

- The dynamics in establishment/release of communication sessions. Two variants are on-demand (controlled by user) and permanent (fully controlled by the operator/provider). Traditionally signalling has been used for the former, while management activities are invoked for the latter. However, a combination of signalling and management procedures may the applied.

- Degree of dependability. Two main dependability measures are availability and reliability. Requirements on dependability may differ for the different products and also vary for the different customer types. An example of a product with fairly high availability is an alarm service (considered to belong to the "transaction" product category).

- For products between operators/providers, these may be aggregation of the end-user products (like for a wholesale operation). The interface type may also be specified. A number of interface types could therefore be specified.

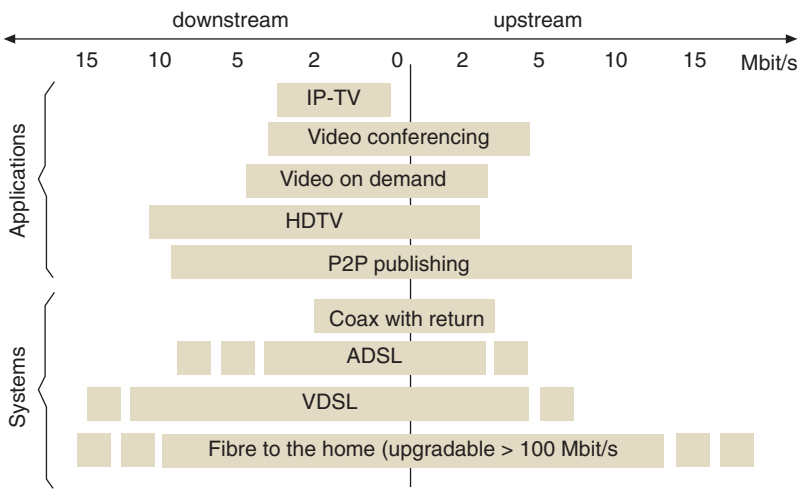


Figure 5 Matching applications to be supported and access solutions (examples)

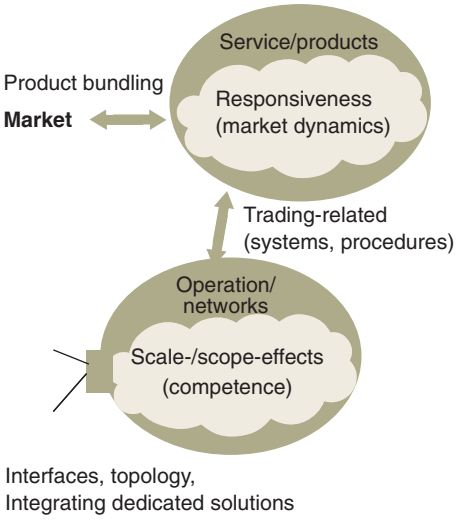


Figure 6 Factors related to co-ordination between an operator's business units

4.3 Costs

There are a number of reasons why a network operator will invest in telecom equipment. Firstly, new equipment could be installed providing the new services that cannot be provided by current systems. Secondly, an existing system could be replaced in order to produce the services more efficiently in a new system. Naturally, the first and the second will also be combined in several cases. A third reason is to expand an existing system in order to cover more traffic load or customers. This is also likely to provide the new versions (and hence a combination of the first and second as well).

The cost components come in various flavours, including:

- Equipment cost – both hardware and software as well as other vendor costs. In several cases recurrent license and maintenance costs are given, or a cost level depending on the amount of traffic/carried load.

- OSS costs, as a consequence of introducing new equipment – this can be either completely new OSS, or adaptations and integration related to existing systems.

- Installation costs

- Operational costs such as:
  - Cost of human resources
  - Other necessary infrastructure (e.g. transport layer support, fibre, copper line, etc.)
  - Necessary vendor support and licenses
  - Training
  - Equipment footprint, reflecting building rental fees
  - Power and cooling requirements
  - Recurring way of right costs

On the other hand, when replacing/modernising an existing system, there is also the potential re-use or sale of redundant equipment which could be taken into the equation. Then, the revenue/income side may also be included in the overall equipment cost calculations.

4.4 Co-ordination-related

For an operator having a wide set of systems, the set of customers to support is also likely to be diverse in terms of requirements. However, there is a steady trend to integrate common solutions, arriving at single systems within an operator, potentially supporting a range of customer segments and product types. This is likely to require co-ordination between business units within the operator in order to successfully deliver services to their respective customers. A simple example is the use of a common IP network supporting both residential and business customers, potentially having different dependability requirements.

A number of factors influence the co-ordination between units co-operating to deliver a product, see Figure 6. The main ones discussed in the following emphasise "trading" between units:

- *Centralisation vs. distribution:* Centralised solutions are better for relatively simple, common tasks with small requirements for individual changes and adaptation. For some other cases, distributing the solutions is the better choice.

  Centralised solutions imply a risk of becoming "least common denominator", i.e. not fully compliant with individual needs. Also, (too) late introduction of new functionality that only applies to some application areas is a potential outcome when focusing on centralised solutions only. The downside of distributed solutions, obviously, is the risk that functions or tasks that otherwise could have been combined, may be duplicated. On the other hand, distributed solutions allow for different time frames when introducing changes and upgrades, taking into account varying "local needs". That is, needs reflecting individual business units.

  As long as the size of the task/area is above a minimum, a distributed solution will be best suited to adapt to variation in requirements and changes in environment. The scale/scope effect, however, would influence the minimum level, including equipment cost/utilisation, human resources, etc.

- *Outsourcing:* Outsourcing a particular task and thus becoming one of several buyers provides some additional challenges:

- A well-defined and professional supplier/customer relation must be defined.

- The buyer's possibilities of having problems solved rapidly may be reduced.

- The buyer's possibilities of prioritising and controlling error corrections and upgrades are limited.

- *Internal trading:* A "demander" of a certain service within an operator has fundamentally three alternatives: i) buy from an external unit, ii) buy from another internal unit, iii) build it yourself. Whenever constraints are imposed on internal trading (i.e. internal trading has to be applied), the challenges in defining the relationship are particularly great. A healthy climate for internal trading requires freedom and openness for all parties, i.e. any agreement must be commercially viable for supplier and customer. This requires flexibility and incentives for the business units involved.

- *Scale and scope issues:* The following factors may be considered for the scale effect ("the bigger, the better"):

  - Co-location of equipment may be easier, including common power supply, ventilation, etc.

  - Operation/maintenance staff and gathering of competence

  - Planning expertise

  - Vendor contacts – prices on equipment and support

  - OSS-related – number of systems and need for interconnecting the systems

  - Common arrangements for interconnection with other operators/providers

  The scope effect, advocating collective functions, may allow for an easier introduction of new products and transfer of products between different networks. Hence, this would affect the service bundling challenge, in particular across different business units' responsibilities.

The co-ordination effects also address a generic challenge for a system design, sketched in Figure 7; the effort spent during the initial design and set-up of a system commonly eases the effort needed to introduce changes to a system. This also goes for the procedures and schemes to follow between units – both within an operator and between the operator and others; an efficiently designed procedure commonly pays back in the long run.

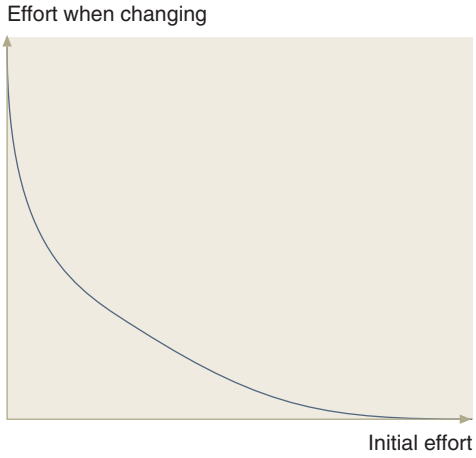


Figure 7 Trade-off effort during initial phase and modification/upgrading/operational phase

4.5 Regulation-Related (Including Competition)

Providing public telecommunications services and use of radio frequency spectrum are fairly strictly regulated. Operators have obtained licenses in Norway to provide public services. Moreover, a few operators are considered to have significant market power in several areas and thereby have to obey additional constraints, like open and announced conditions for interconnection.

Examples of areas subject to regulation are:

- Provision of public mobile services (NMT, GSM, UMTS)
- Provision of public telephony service
- Provision of leased line services
- Unbundling of local access lines (LLUB)

In addition, an operator may have to obey the Universal Service Obligation (USO) with respect to public telephony, leased line services up to 2 Mbit/s and access to digital telecommunication network (the latter referring to interconnection as opposed to leased line).

Following additional rules due to holding Significant Market Power (SMP) imply that more issues have to be taken care of:

- Transparency
- Non-discriminating
- Cost-based pricing
- Separate accounting for different services

SMP regulation is valid both for fixed network and mobile network services. For these services reporting to the regulating authority has to be taken care of.

5 Trends

5.1 General

Trends can be observed in several areas, including customer demands, market situation and technology/system. Besides these, others may be regulatory, financial, macro-economic, etc. A few samples of trend statements are given in the following. These are by no means exhaustive.

- *Customer demands:* What will customers ask for in the future? Besides being a "million-dollar" question on the business level, it is also a topic that concerns many of the groups engaged in network evolution. A few examples of trend statements are:

  - The overall market for fixed telephony has become rather flat (in terms of revenue) in developed countries and a decrease is expected.

  - The rapid growth of voice in GSM seen previously seems to be taking a break in Western Europe while the steady migration of voice and narrowband traffic from fixed to mobile networks has characterized the last decade.

  - New broadband accesses are primarily based on xDSL. Steadily increasing bandwidth demands due to increased penetration of broadband access and heavier use of the network.

  - Increasingly more cost-oriented business customers, implying a trend of moving towards on-demand services to partly replace leased-line services. This is particularly true for small and medium enterprises (SMEs). Other customers will also look for means to reduce cost, like changing to less expensive interface cards (e.g. Ethernet-interfaces).

- *Market situation:* In several regions, the providers are about to, or have recently gone through a consolidation phase. This implies that several smaller actors have given up or are being bought out by others. On the other hand, there are also several actors entering the telecom area such as the power supply companies and local communities installing high capacity network capabilities at the same time as other facilities are placed into the ground.

  In addition to these come the global players and companies operating in other regions. These may, assisted by their sheer size, be able to operate at lower cost bases, realise higher development capacity and exercise greater purchasing power. Hence, an on-going alliance trend, mergers and acquisitions are expected. At the same time there will frequently be newcomers starting up within certain niches of the market.

- *Technical issues:* Steadily miniaturized electronics allow higher processing power and memory/storage capacity. Hence, more intelligent terminals emerge, including terminal types with potential communication needs. This is seen by wireless communication being integrated in various device types and that machine-to-machine communication seems to grow.

  In the network equipment area, convergence of equipment capabilities is observed. That is, several systems (and vendors) have included most options within their roadmap. Two examples are provision of speech (telephony) and Ethernet-based services. These may be provided by several combinations of network equipment. Another development is manifestation of acknowledged interfaces between modules allowing for interconnecting units from different vendors. It also allows potentially different market segments to converge in the sense that more services can be offered and requested in several segments.

The struggle between services provided and the application of services should also be noted, see Figure 8. That is, services and applications may be seen in a continual pursuit; where the service provider tries to offer adequate services (bundles) while the applications try to utilise expect-

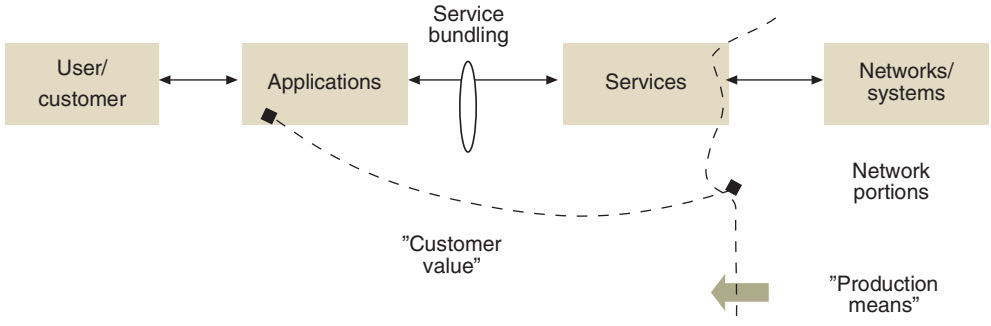


Figure 8 Relation between services (delivered by networks/systems) and applications

	2001		2000		1999	
	revenue (mill NOK)	per cent	revenue (mill NOK)	per cent	revenue (mill NOK)	per cent
Analogue/Digital (PSTN, ISDN, ADSL)	13,668	36 %	12,802	38 %	13,313	45 %
Mobile telephony	9,531	25 %	7,197	21 %	5,468	18 %
Leased line	1,065	3 %	902	3 %	810	3 %
Satellite and TV distribution	3,879	10 %	3,245	10 %	2,584	9 %
Other network-based	2,633	7 %	2,215	7 %	1,593	5 %
IT service and installation	5,009	13 %	4,738	14 %	3,501	12 %
Advertisements, etc.	1,266	3 %	1,555	5 %	1,588	5 %
Others	1,388	4 %	1,040	3 %	987	3 %
Sum	38,439	100 %	33,694	100 %	29,844	100 %

Table 1 Telenor operational-related revenue (excluding asset sales)

	2001	2000	1999	1998	1997
GSM – subscriptions in Norway, end of year					
Fixed subscriptions	1,210,000	1,145,000	1,003,000	944,000	803,000
Prepaid	1,027,000	911,000	732,000	316,000	68,000
Churn rate (related to fixed)	12.5 %	12.7 %	14.2 %	13.1 %	13.9 %
Mobile – originated traffic, Norway, mill. minutes					
GSM	2,969	2,298	1,801	1,279	711
NMT	64	108	174	271	331
ARPU GSM per month, NOK					
Total	340	3,381	341	366	401
Fixed subscriptions	494	473	440	400	401
Prepaid	154	1,652	157	169	0
Telephony lines (fixed) in Norway, end of year					
Analogue (PSTN)	1,527,000	1,680,000	1,908,000	2,167,000	2,324,000
Digital (ISDN)	1,735,000	1,590,000	1,228,000	755,000	410,000
Telephony traffic (fixed) in Norway, mill. minutes					
Domestic, excl. Internet	10,567	11,612	12,371	12,911	11,923
Internet dial-up	4,974	5,667	4,255	2,059	1,079
International	383	387	415	386	379
To mobiles	1,412	1,295	1,246	967	727
Value added services	624	599	447	287	191
Pay-TV, number of subscribers in Nordic, end of year					
Cable-TV	561,000	357,000	282,000	270,000	244,000
Smaller, closed Cable-TV	1,105,000	1,086,000	937,000	686,000	0
Satellite to residentials	657,000	506,000	405,000	352,000	251,000
Total	2,323,000	1,949,000	1,624,000	1,308,000	495,000
Internet, end of year					
Subscribers/registered users, Norway	831,000	625,000	400,000	260,000	165,000
Churn rate (subscriptions)	20 %	25.5 %	14 %	11.7 %	
Nextra business subscriptions, Norway	16,000	13,000	8,000	4,000	2,000
Nextra subscriptions, outside Norway	106,000	104,000	57,000	0	0
Work force, man years, end of year	21,000	20,150	21,968	20,226	19,598

Table 2 Trends in Telenor's operation (from annual report 2001)

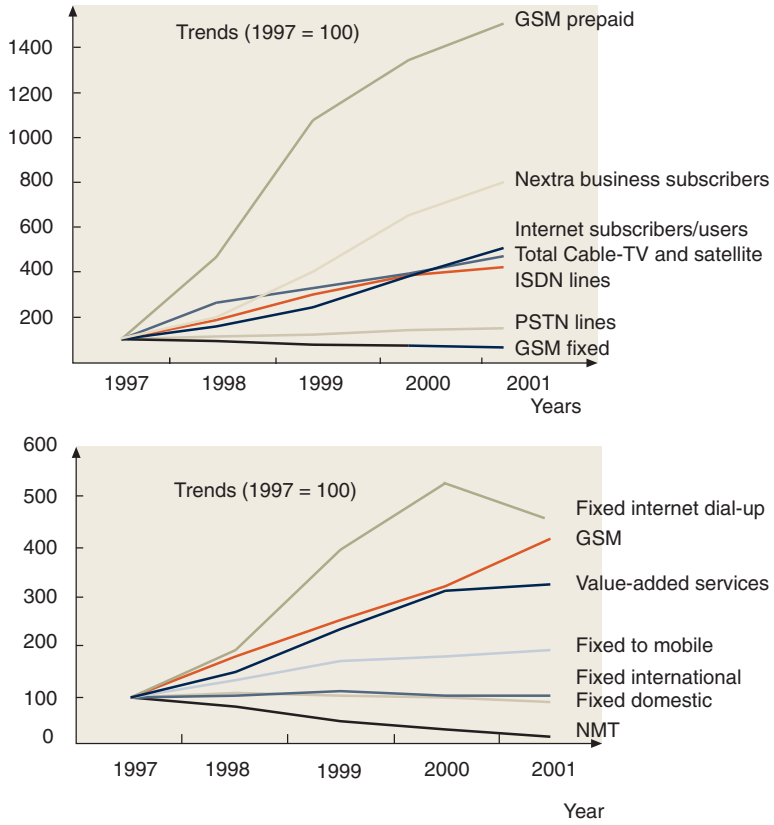


Figure 9 Trends in key indicators, referring to year 1997 as basis (100 – level)

ed (future) services. Hence a spiral effect can be seen, although new technical solutions (and commercial concerns, billing, etc.) may introduce more abrupt changes. Turning it around, in certain areas, this may also be seen as the "chicken and egg" problem.

5.2 Status – Background for Trends

A sound basis for discussing trends of a company is to assess the actual status and concise historic numbers. A few illustrations are given in the following, obtained from Telenor's annual report for year 2001. As a perspective on the dimensions behind different areas of Telenor, Table 1 shows turnover/revenue (million NOK) in 2001 compared to two previous years.

Figure 10 Trends in key indicators, referring to year 2001 as basis (100 – level)

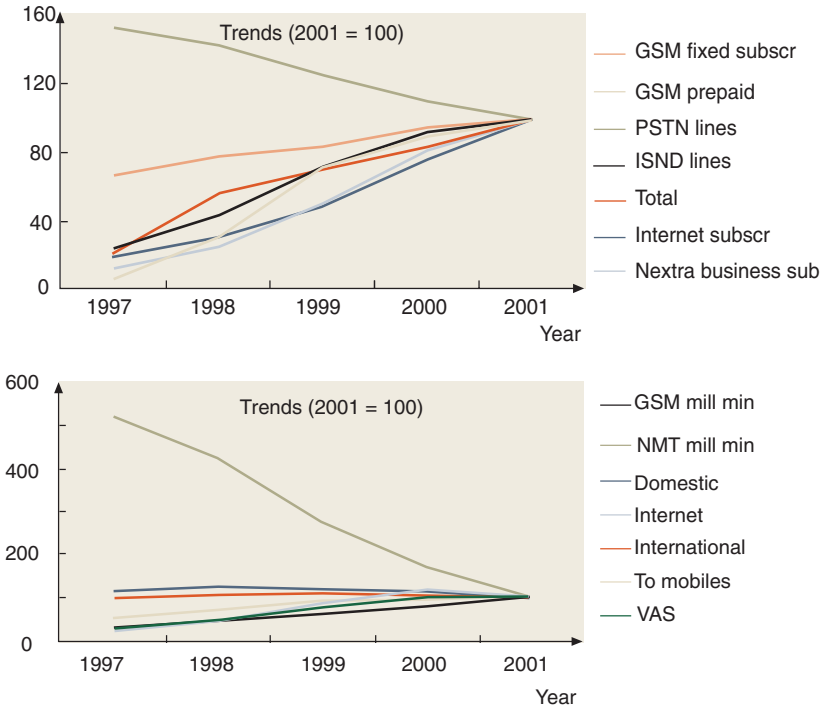


Table 2 gives more technical details, in particular for mobile systems (GSM, NMT), ISDN, PSTN, cable-TV and satellite, and Internet/IP-related operation. Numbers from the years 1997 – 2000 are included to indicate trends.

The numbers clearly point out areas of growth and reductions. In brief, analogue PSTN lines have a fairly drastic decrease, compensated for by the increase of ISDN lines. However, the overall voice-related traffic in PSTN/ISDN has decreased (about 10 % in 2001). Areas of growth are GSM, Internet and cable-TV/satellite in addition to ISDN. As expected, a significant decrease is also seen for NMT (to be phased out a few years after 2001).

The trends are shown in Figures 9 and 10, having year 1997 and year 2001 as basis, respectively. The former emphasises growth areas, while the latter clearer shows the areas in decline.

5.3 System Life-cycle

The different systems have different maturity levels and also address different levels of user needs. Some networks and associated technologies are still in the emerging phase, e.g. ADSL, while others are quite mature, such as PSTN. Traditionally, a system has gone through a competitive growth phase gaining as many customers/coverage as possible by customer acquisition and network infrastructure roll-out. Following that phase, more emphasis is put on revenues per customer, margins, churn rates and OPEX/CAPEX per customer. As a system eventually enters a decreasing phase, the provider's attention is gradually shifted towards customer retention and system consolidation. This is illustrated in Figure 11, also sketching a shift in positions in a future period.

The systems may not follow the same tempo through the different phases. That is, a system might well pass another system, as illustrated, comparing "now" with "*Y* years ahead". During the different phases it is natural that the network development is carried out according to different motivations. However, it is important that the overall life span is considered in the overall network planning avoiding that solutions are chosen that makes the operation very complex or expensive at a later stage.

Another factor is that the decreasing phase should be observed carefully, also reducing the

investment levels correspondingly, in order to steer clear of having much surplus equipment towards the end of system life.

Establishing such system life-cycles, scenario exercises are central. The observations made from scenarios will assist also when estimating the tempo of the systems during the different phases. It should also be noted that not every system has to follow the phases, in particular systems that are considered as flops may have difficulty reaching a growing – and hence a mature and declining phase.

6 Scenario Work – Qualitative

One should bear in mind that an approach based on scenario is commonly considered subjective. Hence, several of the steps and choices made may not be uniquely inferred from the previous steps. Still, the scenario approach can be applied to capture a possible future in order to identify the choices that one should prepare for.

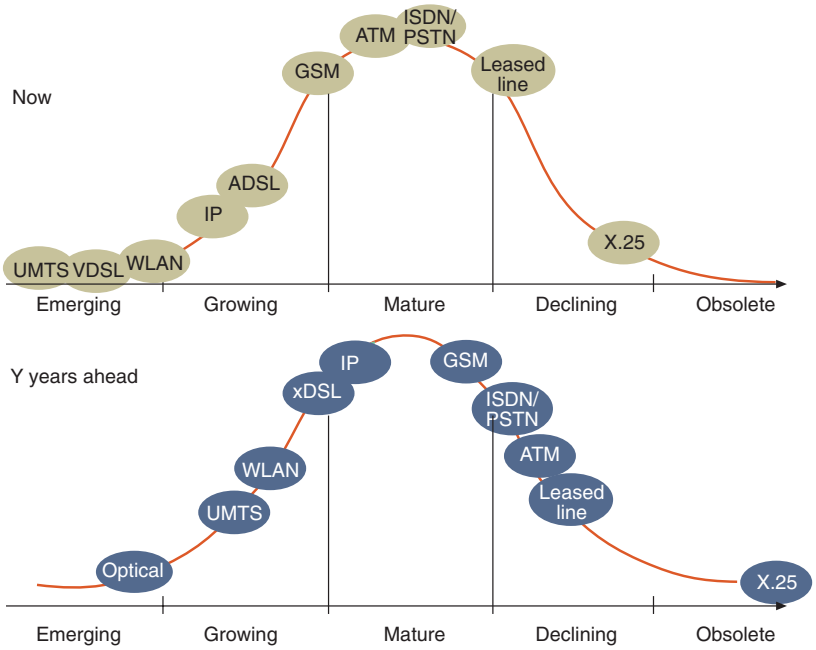


Figure 11 Illustration of system life-cycles; state of the systems today and a possible shift some years ahead

An overall illustration of an approach is given in Figure 12. As shown, the goal is to arrive at network roadmaps. Here the network roadmap shows how the networks in the portfolio should be developed in the time period considered. Hence, for each scenario, a network roadmap is derived. Deriving the network roadmaps, the evaluation criteria and technology time lines have to be taken into account.

Using the sample of network roadmaps corresponding to each scenario, the roadmaps are revisited to find candidates. As expected, fewer sets of candidates are seen than there are scenarios. Hence, these candidates are basis for elaborating a decision tree and the migration plans.

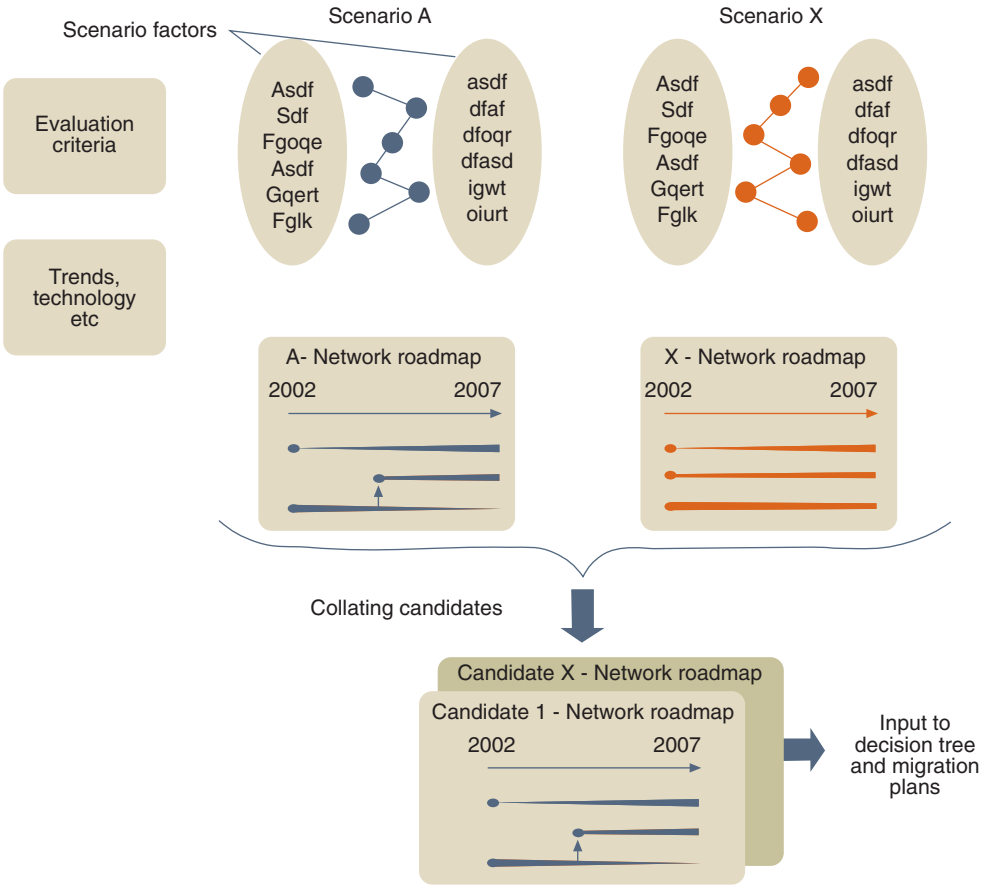


Figure 12 Overall approach followed for deriving candidates for network roadmaps and mapping of product categories (could be iterative)

6.1 Arriving at Scenarios

Deriving a set of scenarios, one faces the trade-off between finding as few scenarios as possible at the same time as the complete feasibility space is covered. Here, the feasibility space refers to all possible situations which may arise in the future. Naturally, a rather limited number of situations have to be selected in order to have a tractable task. Still these situations should include most of the challenges that can emerge. Deriving an adequate set of scenarios may be seen as an art, as several subjectively determined conditions have to be included. However, when open for iterations, one learns for each step, incorporating these lessons when the scenarios are adapted.

In order to describe a scenario, a number of scenario factors have to be defined. These factors will also to some extent assist when exploring whether or not the set of scenarios covers the feasibility space. Again, trade-offs are seen; the factors should be as few as possible (to be tractable) and still capture the main aspects. A high level of subjective judgement is involved when expressing the scenarios by the scenario factors.

A brief scenario description of a scenario definition would then contain:

- List of key characteristics
- Short story of what happened/is happening
- Short description of operator's situation
- Grading of the scenario factors

Even though one would start with a set of scenarios and carry on deriving network roadmaps, it is likely that the scenarios would be adjusted based on the lessons learned. Hence, a few iterations might take place until the scenarios seem to address the most likely and central issues. An alternative would be to derive the scenarios from the scenario factors. Although these factors may not be independent, quite a few combinations will appear, leading to an intractable number of scenarios. On the other hand, these could still be used as a starting point for identifying the scenarios.

6.2 Deriving Network Roadmaps from Scenarios

Given a scenario, certain key characteristics and main trends are included. The idea is then that these would motivate for a corresponding migration of the network portfolio. Again, subjective considerations are behind the story of expressing a network roadmap corresponding to the scenario.

A network roadmap shows how the different network solutions (e.g. ADSL, VDSL, Cable network, WLAN for access network) will migrate and be applied. That is, when new functionality will be introduced (and in what year), and is the specific solution expected to increase, be stable, or decrease in level of importance (number of subscribers, traffic volume, etc.). Moreover, a network roadmap also explains how the different solutions relate to one another. For instance, an IP-based network might be carried directly on an optical network and use ADSL, SHDSL, VDSL, Ethernet and WLAN as access forms.

When devising the network roadmaps, information from the expected trends/time lines as well as criteria is used. For example, the technology trends express views on when certain functions will be available including statements on when the solutions will be applied. This input is a natural starting point to describe the migration, and, considering the scenario characteristics, issues from the technology time lines can be selected accordingly.

6.3 Collating Network Roadmap Candidates

For each of the scenarios, a network roadmap (and product mapping) has to be derived. It is likely that fewer different network roadmaps will appear than the number of scenarios. Therefore, a "reverse processing" could be applied, meaning that the resulting roadmaps are put together. Given that these have the same starting point, they will differ by taking certain (different) steps at certain instances in time. In principle, this can be thought of as a decision tree; facing an instance when two scenarios differ expresses that a decision has to be made. The decisions may also be more related to certain events and less strictly to time instances.

Correlating these network roadmaps with the operator's situation, more figures will be introduced as the method so far has been followed mostly in a qualitative manner. This is done as part of elaborating network migration plans.

To speed up this process, it may be more efficient to start by raising the main questions relating to network migration. That is, devising these questions and then adapting the scenarios correspondingly.

6.4 Scenario Factors

Working on scenarios, a set of factors has to be devised. These factors apply when the scenarios are described, i.e. how the scenarios are placed into the "space" spanned by the factors. In principle, one could use the set of factors to identify potential scenarios. However, considering the set of factors the number of potential scenarios might then become too large.

The work on scenarios would likely be carried out iteratively, meaning that lessons learned from describing scenarios and belonging network roadmaps are used to return to the set of factors and potentially revise them. Note, however, that the scenarios as such are not the main outcome of the network planning, but are mostly used when dealing with the strategic perspective. Still, observations made could well be forwarded to other processes within the operator's sphere.

The scenario factors express the uncertainty attached to a future evolution. Hence, assigning values/grades to a factor "fixes" how that aspect evolves. The factors selected have been formulated as questions; like *A* versus *B*, where *A* and *B* are "opposite" choices on a scale. It is also implied that the scenario description refers to the "environment" as observed by an operator (Figure 13).

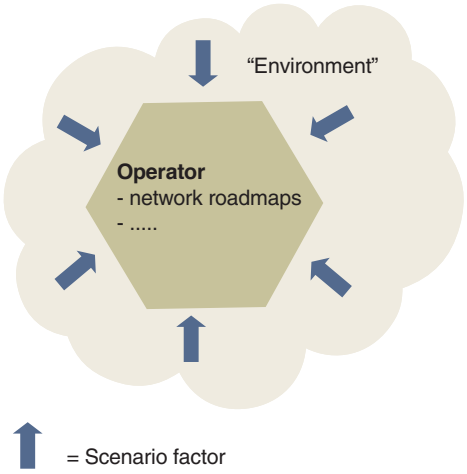


Figure 13 Scenario factors express how uncertain aspects evolve as seen by an operator

In some of the cases the different factors may not be orthogonal (or independent). In these cases, further work may be done to revise the set of factors based on the knowledge gained and the results one wants to look more closely into. Again, however, the scenarios themselves are not considered the main results in the network planning; rather these results are the potential and possibly recommended roadmap of the network portfolio.

Examples of scenario factors and corresponding scenarios are given in the following section.

7 Two Scenario Examples

7.1 Scenario Space With Two Axes

Assuming that one main question is whether or not a common packet-based (core) network should be used to carry the traffic, one should bear this in mind when deriving the scenarios. Here it is assumed that the current situation is to have a TDM-oriented common carrier. These technical combinations are depicted in Figure 14. The blue boxes on top refer to client systems, e.g. ISDN exchanges, ATM switches, IP routers, customer accesses.

Then we assume that there are two main uncertainties; i) the technical feasibility of supporting all the traffic on the packet-based network, and ii) the market demand for packet-based services versus more TDM-oriented services. The scenario space could then be made as in Figure 15.

Illustratively the four scenarios can be described as:

- Scenario I: Packet services have taken a significant market share at the same time as packet technology is able to service the demands in a common implementation. Hence, a common packet-based carrier network could be realised, transporting most of the client traffic flows.
- Scenario II: TDM-based services dominate the market although packet technology is capable of transporting the different traffic flows. This could for example happen when the TDM equipment is much cheaper than the packet equipment.
- Scenario III: TDM-based services dominate the market at the same time as the packet tech-

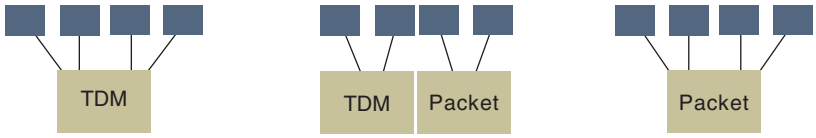


Figure 14 TDM or packet as carrier; left – TDM only, middle – TDM and packet in combination, right – packet only

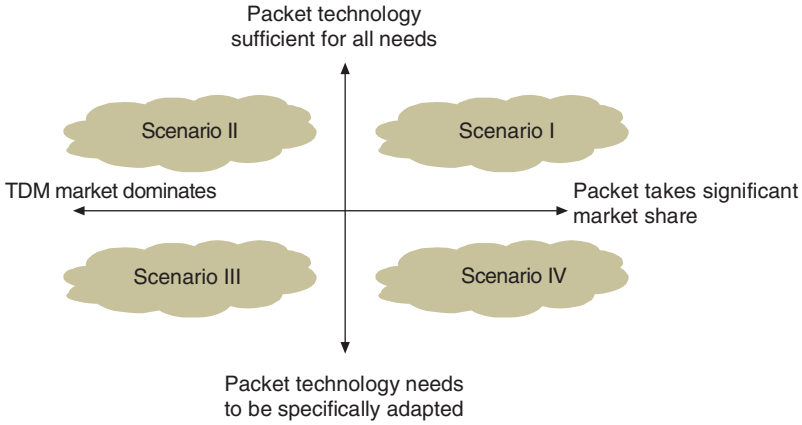


Figure 15 Two axes to illustrate the scenario space

Figure 16 Referring network configurations to scenarios (T = TDM-based, P = packet-based)

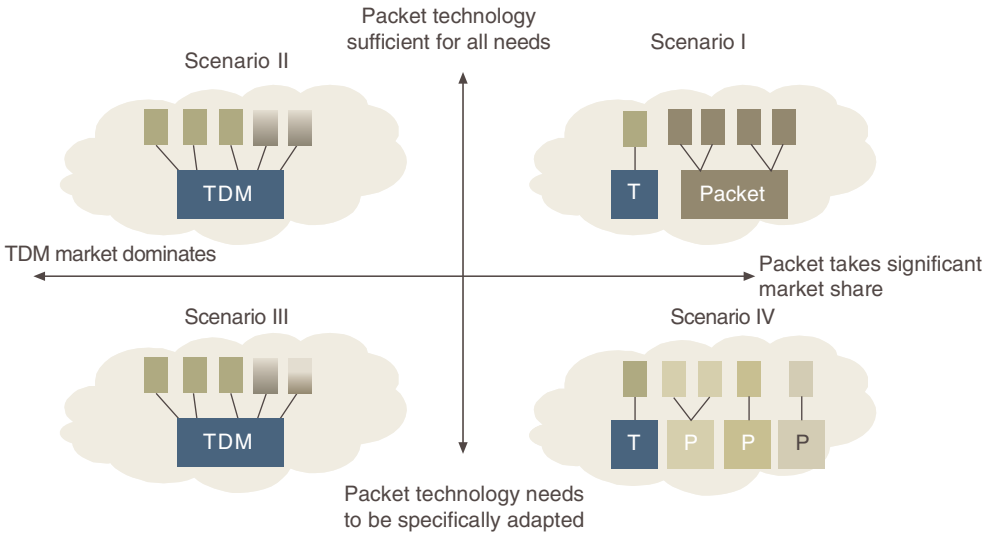
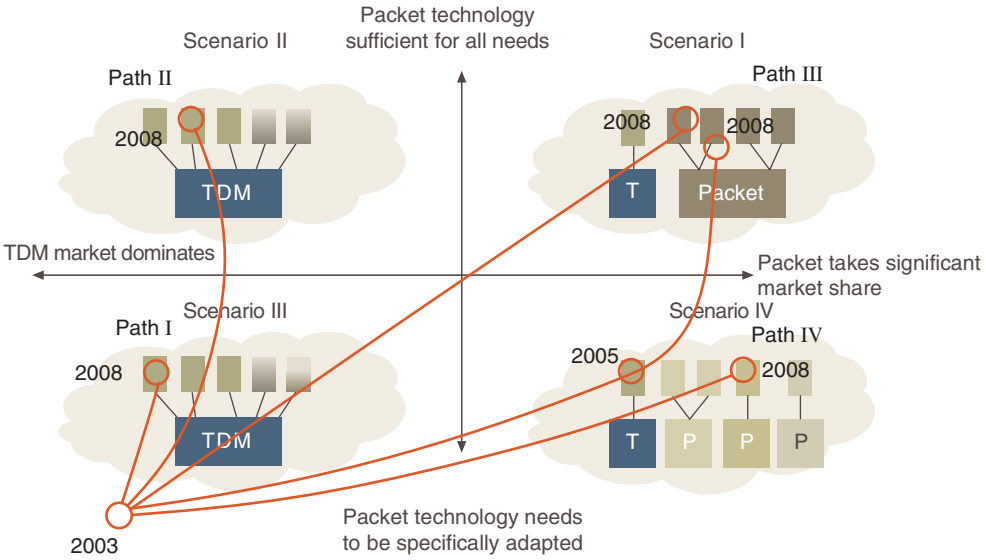


Figure 17 Describing potential network roadmaps referring to the scenarios



nology needs to be specialised to its applications. For example, common packet networks for different market segments cannot be realised.

- Scenario IV: Packet services take on a significant market share, although the packet technology needs to be adapted to its use. This could for example happen if the packet implementation is cheaper than the TDM-oriented way of constructing a network.

The four scenarios can be looked upon as extremes to illustrate the host of possibilities that exist. Practical and optimised solutions from a technical and commercial viewpoint may however show that intermediate solutions between these four are to be preferred.

The network configurations relating to the four scenarios are depicted in Figure 16. Having described these configurations, potential migration paths could be introduced as well. For example, assuming that the current configuration is at the lower left square, theoretically four paths (and hence target states) could be identified, each path reflecting a scenario. A more involved case could be present as well, such as when some of these configurations could be a state on the way towards a target in another square (see path V in Figure 17). Another variation is that the same path is followed in the early migration phases, although one does not reach as far along this path as believed; that is, the target state may be plotted at a closer stage along the migration.

7.2 Applying More Scenario Factors

For a more elaborative scenario description, a number of factors are commonly defined. One example is given in Table 3. Again, it has to be remembered that the most significant factors should be described, which both are chosen based on subjective evaluations and decided by the main network-related questions faced.

Utilising these scenario factors, a number of scenarios can be devised – two examples are illustrated in Figure 18. In addition to grading the scenario factors, a story has to accompany each of the scenarios. This should be a likely explanation how that scenario may occur. Naturally, on the management level, one would also like to know the distribution of market power between the actors present in a market and the roles in the value chain. An illustration of this is given in

Drivers for evolution	Scenario span	
User behaviour	Best-effort data, voice moving to mobile networks	Reach real-time media, seamless mobility
Customer ownership	Network operators	Value-added service providers
Business model	Integrated carriers	Fragmented value chain
Regional balance	US stays leading market	Asia-Pacific drives innovations
Network-or terminal-centric	Intelligence in terminal	Intelligence in network
Regulatory impact	Laissez faire	Strict regulation
Rise of global operators	Fragmented, national	Coordinated, global

Table 3 Example of scenario factors

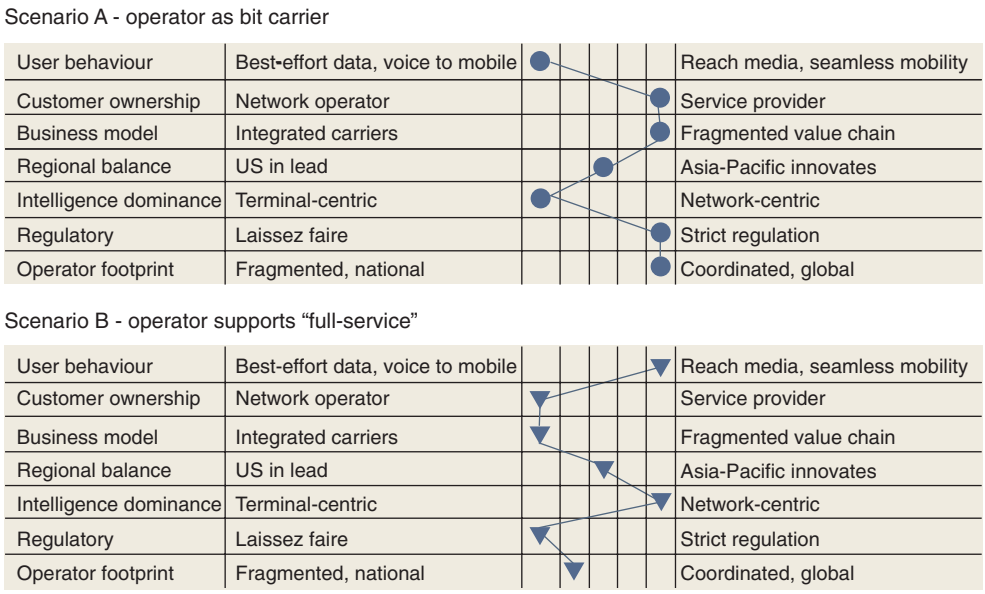


Figure 18 Samples of scenarios defined by the scenario factors

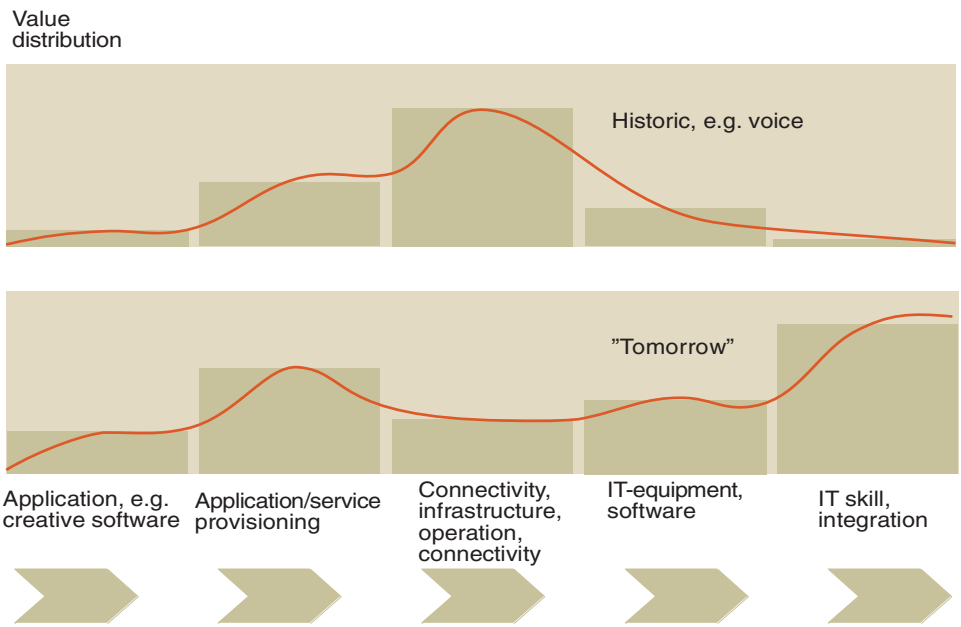


Figure 19 Shift in value chain (although depending on market/competition situation)

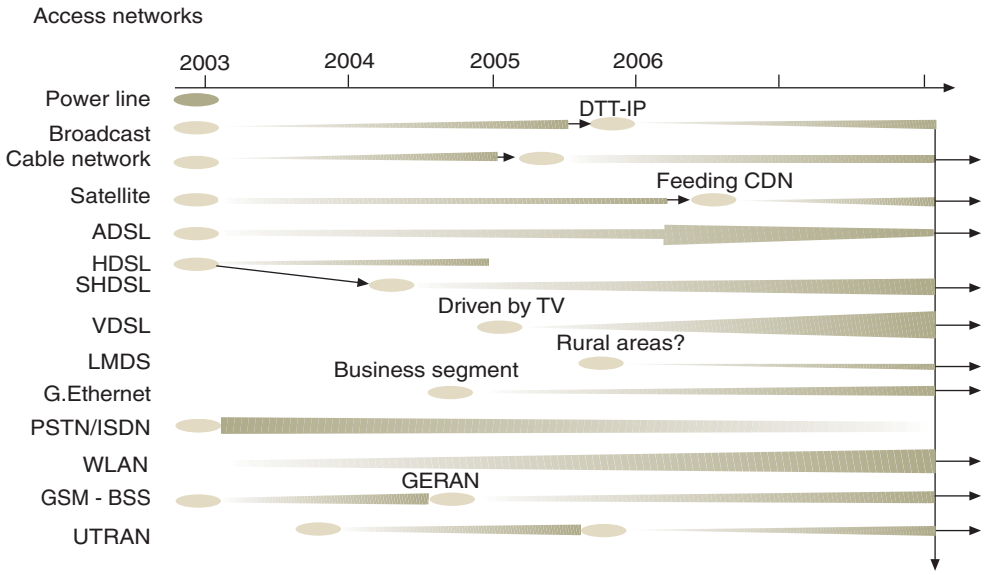


Figure 20 Qualitative illustration of system trends for different access networks related to a set of scenarios

Figure 19, showing a potential shift in value distribution related to a set of scenarios.

After describing the scenarios, actions and consequences on the network portfolio must be derived. Figure 20 shows an example of how a number of access systems could evolve related to a set of scenarios. As described above, each of the scenarios should have a network roadmap associated. Then, these resulting network roadmaps are collated to identify in what manner they differ and the phenomena that influence the decisions to be made for the different systems.

The scenario-based approach provides insight into the factors that influence the system evolution. This is commonly done on a qualitative level. The approach has to be complemented with calculations as described in the following section. The qualitative observations, however, can limit the number of candidates that should be input for the calculations. Moreover, they also try to systematise the uncertainties and risk factors to be followed.

The qualitative evaluations may also reveal a number of "winning systems" that seem to be safe to instal or enhance. This is also a valuable result that should be considered in the following work.

8 Qualitative Evaluations – Calculating Cost of Network Roadmaps

A complete calculation of network roadmaps should ultimately fill requirements for business case studies. However, there is often much uncertainty, particularly related to revenues. The following sections only include the investment aspects. To capture the total cost picture operational expenses must also be considered, given by staffing, license agreements, and so forth.

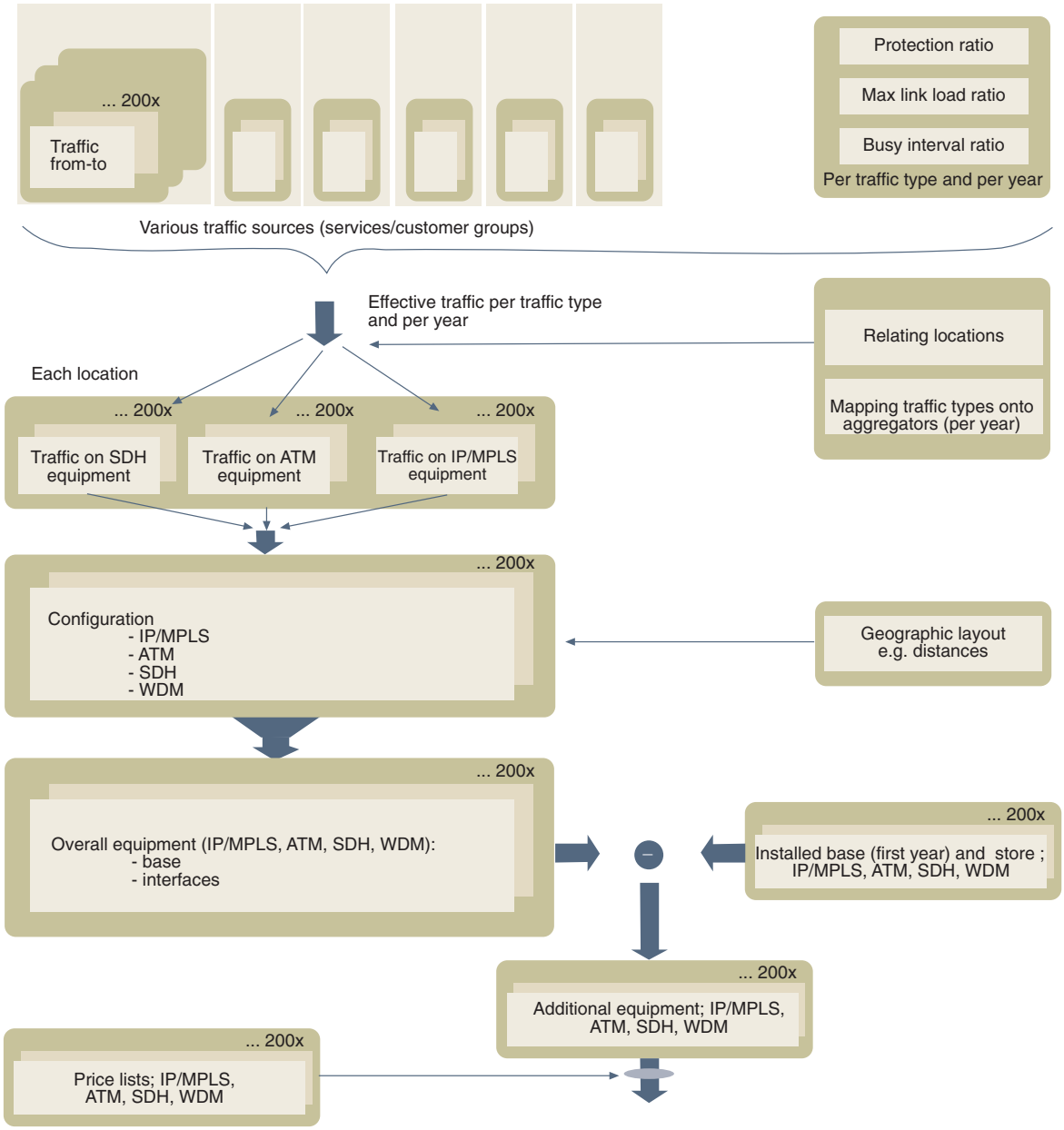
The description is based on an example where the question is how to carry traffic flows for a number of client systems between a number of locations in a core network. However, the same principles apply to other areas as well. Three types are assumed, SDH, ATM and IP/MPLS. These are also referred to as aggregators in the following.

Figure 21 depicts the overall procedure followed when carrying out the calculations. An end-result is the investments needed in order to carry the traffic loads. A study period of a number of years is given, hence, all the major input data must be specified for each of the corresponding years. This also allows for an evolution of the input data, which is essential at least for traffic demands and for component prices.

A major bulk of input data is composed of the traffic matrices specified as traffic load (e.g. in Mbit/s) between the locations (all locations as specified for the network, also giving the amount of traffic to/from abroad). A number of traffic types are specified.

A number of parameters give the dimensioning requirements for each of the traffic types. These parameters are:

- The ratio of traffic that should be kept during a failure situation (in the range 0 .. 1),
- The maximum link load that can be realised (in the range 0 .. 1),



Investment/ economics

Case	Investment in Mill. NOK								
Investment	2002	2003	2004	2005	2006	2007	Spare equip.	Invest -Spare	Sum 02-07
IP-related	9.6	2.9	4.7	6.5	11.3	8.6	4.9	38.7	43.6
SDH-related	3.2	0.7	2.3	0.0	0.9	1.5	0.0	8.7	8.7
ATM-related	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
WDM-related	1.1	0.5	1.1	3.2	3.4	5.2	0.0	14.6	14.6
sum	14.0	4.0	8.2	9.7	15.7	15.3	4.9	62.0	66.9

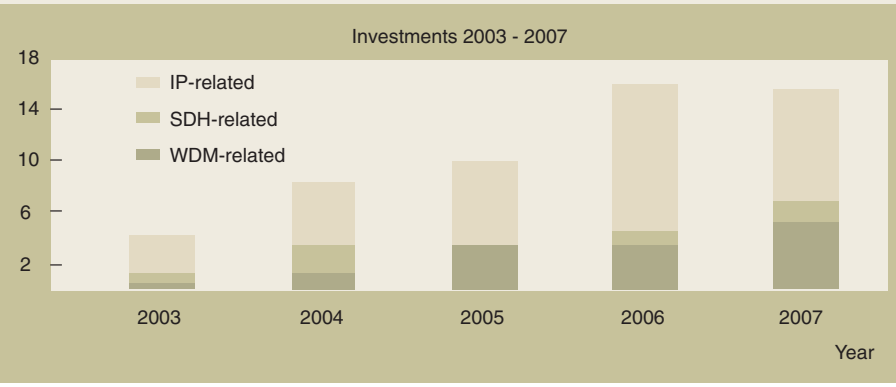


Figure 21 Schematic flowchart for calculations

- The relative portion of the peak traffic load that is present in the dimensioning interval, i.e. a busy interval indicator (in the range 0 ..1).

In combination, the two former indicate the traffic situation during a failure situation. For instance when a failure occurs, a ratio of the traffic (on the failed link) is carried by an alternative link and the maximum load on that link is given by the "maximum link load" parameter.

The result of taking these parameters into account is called the effective traffic load. This traffic load is given for each traffic type and each year. A step further is to consider the relations between the locations. This means to consider where links are placed (present between certain locations). Relating the effective traffic loads, mapping of traffic types onto aggregators and presence of links decide the traffic load on each aggregator type at each location.

The traffic loads on the aggregators give the configurations of each aggregator. "Directions" must also be included (that is where links are going) as well as the distances between the locations (in particular for WDM equipment).

Adding equipment types on all locations gives the overall needed equipment (for base configuration, interfaces, etc.). In order to consider actual deployment and equipment reuse between different years, an "equipment store" is present. Hence, only when needed equipment of a certain type in a year exceeds the sum of that equipment type in the previous year and the number in the "store" is additional equipment bought.

The result after looking at installed equipment, "stored" equipment and needed equipment is a list of additional equipment that needs investment funds. Considering the price of that equipment for the corresponding year gives an investment level. These investments can be aggregated in different ways, for instance per equipment class, per year, for the whole period, etc.

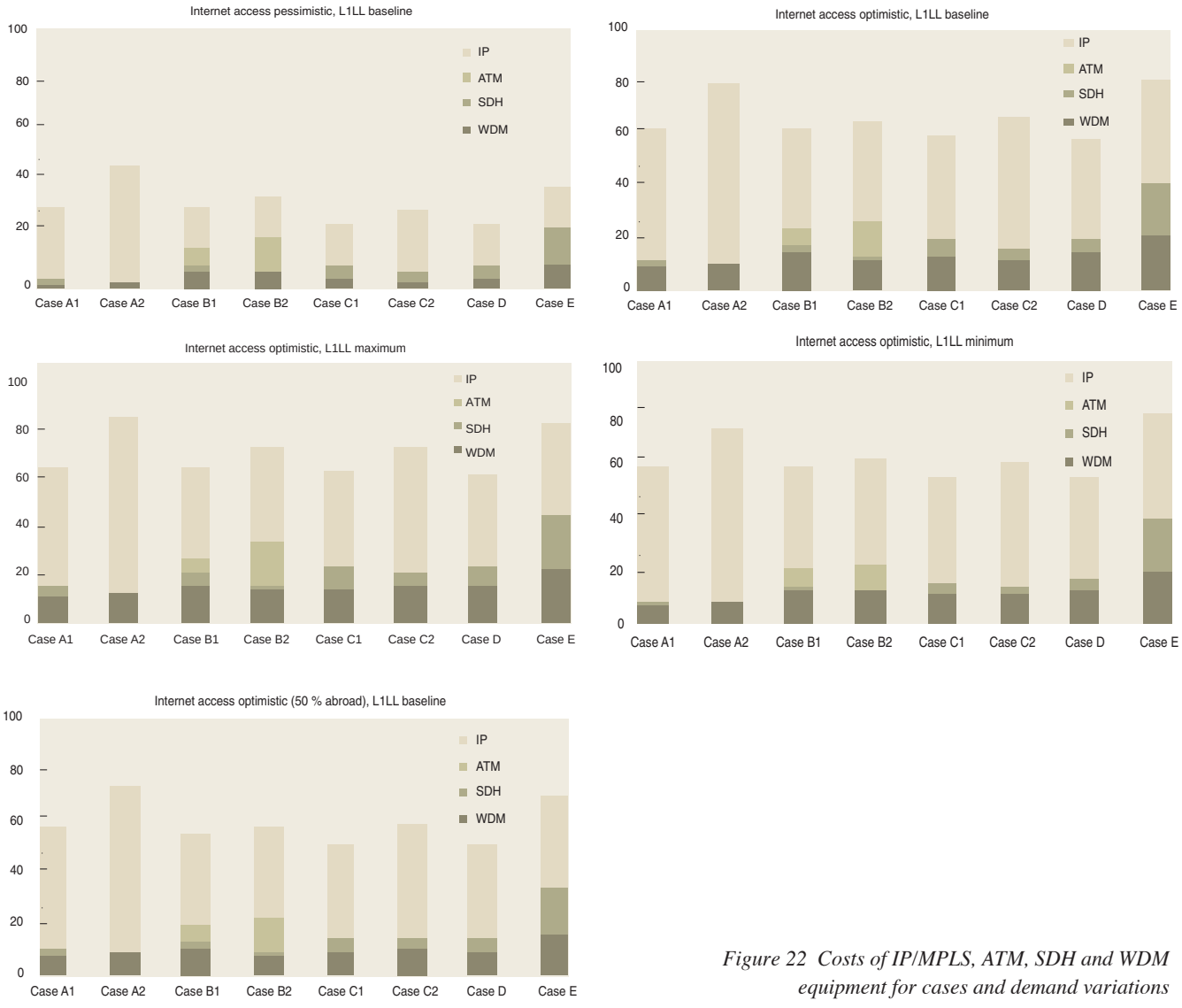


Figure 22 Costs of IP/MPLS, ATM, SDH and WDM equipment for cases and demand variations

In addition to economic output, traffic and capacity results can also be recorded.

Typically, calculations are carried out for a number of cases and the results are compared. These cases can be specified by the technology used, years for introducing functionality, variation in traffic demands, changes in equipment prices, and so forth. In addition different network architectures would be examined. An intention is to find out the more critical parameters, that is, those significantly influencing the results and the network roadmaps that should be preferred. At the end of the calculation period, the installed equipment could be assigned a value ("terminal value") depending on how investment levels should be estimated.

For most transmission systems the capacity often comes in certain granularities, e.g. 155 Mbit/s, 622 Mbit/s for SDH. This has to be included in addition to any consequent costs such as management systems and installation work.

An example of investment results is given in Figure 22. These are shown for five different combinations of traffic demands and where eight cases are examined (A1, .., E). One reason behind breaking up the investment cost showing network types is to give ideas on where more investment savings could be obtained.

An immediate observation from the results above is the fairly high ratio of costs relating to IP/MPLS equipment. In order to look more closely into the effect of varying costs of this equipment the following calculations were conducted; all IP/MPLS equipment costs were changed by *x* % (in the interval from –75 % to +50 %). Only cases A1, A2 and D are shown in Figure 23 as the objective was to estimate points of intersections. Two demand variants are included: Internet access pessimistic + L1LL baseline and Internet access optimistic + L1LL maximum.

For the variant 'Internet access optimistic + L1LL maximum' cases A1 and D come out with equal overall investments when the IP/MPLS cost is reduced by 25 %. For other combinations the IP/MPLS costs have to be reduced by about 75 % in order for the lines to intersect.

9 Risk Analysis and Exit Strategies

As described earlier, a number of decisions will commonly be revealed during the network strategy evaluations. For each step, a certain risk level is associated. In this section a general introduction to incorporate risks into the evaluation given.

9.1 Risk – General Introduction

Considering the constant change and uncertainty, an integrated risk management practice within an organisation is required to strategically deal with uncertainty and thereby capitalise on opportunities. The stakeholders in the decision process will also be involved in order to ensure better decisions in the future.

A risk management arrangement should cover all types of risks that face the organisation, including policy, operational, human resources, financial, legal, health and safety, environment, reputation. Hence, it is essential to integrate risk management into strategic decision-making. By establishing a risk management framework, a mechanism will be in place allowing to discuss, compare and evaluate different risk types.

Deploying a risk management regime also allows for a so-called risk-smart workforce that supports innovative and responsible risk-taking while ensuring legitimate precautions.

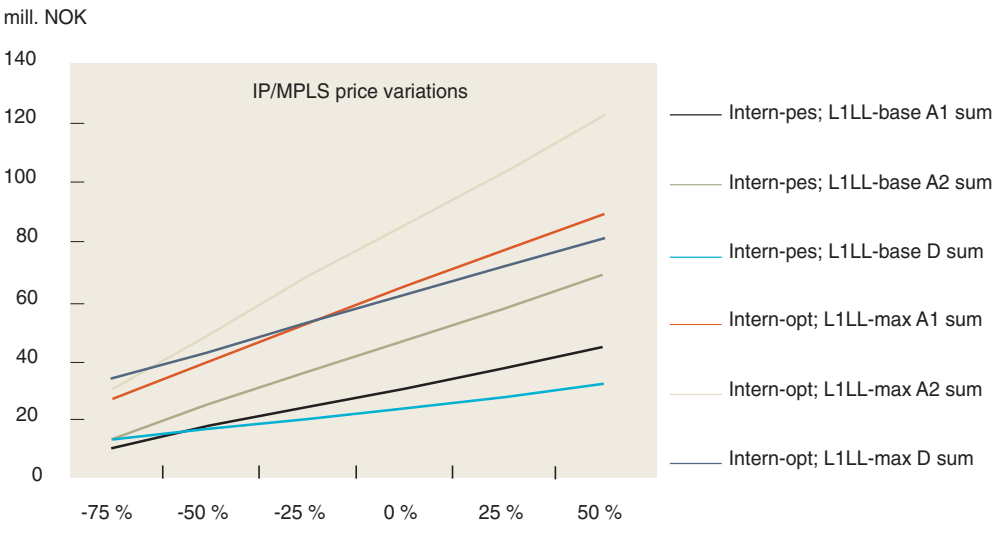


Figure 23 Impact on IP/MPLS cost variations for Cases A1, A2 and D

Risk is unavoidable and present in almost every situation in every-day life, for private as well as business roles. A number of definitions of risk are in use. However, a common concept in all definitions is that a certain level of uncertainty of the outcome is involved. One way the definitions differ is the characterisation of the outcomes; some describe risk as having only adverse consequences, while others are neutral. An ongoing discussion takes place arriving at an acceptable (common) generic definition of risk that recognizes the fact that when assessed and managed properly, risk management can lead to innovation and opportunity. This appears even more prevalent when dealing with operational risks and in the context of technological risks. One definition found is *the expression of likelihood and impact of an event with the potential to influence the achievement of an actor's objective*. Hence, risk refers to uncertainty around future events and outcome.

Other definitions found are e.g.:

- Combination of the probability of an event and its consequences. Note – in some situations, risk is a deviation from the expected,
- The chance of something happening that will have an impact on objectives. It is measured in terms of consequences and likelihood,
- The chance of injury or loss defined as a measure of the probability and severity of an adverse effect to health, property, the environment or other things of value,
- The possibility that one or more individuals or organizations will experience adverse consequences from an event or circumstance.

A risk measure can be formalised as

risk =

$$\sum_{event\_i} probability(event\_i) \cdot consequence(event\_i)$$

The summation is taken over all events considered and the product of each event's probability of occurring and consequence is added.

In order to take advantage of the present risk factors a management framework should be introduced. Risk management is defined as a systematic approach to setting the best course of action under uncertainty by identifying, assessing, understanding, acting on and communicating risk issues. As risk management is directed at uncertainty related to future events and outcomes, it is implied that all planning exercises encompass some form of risk management. Risk management is also impacting people at all levels as it concerns making decisions that contribute to the achievement of an actor's objectives. Hence, integrated risk management is a continuous, proactive and systematic process to understand, manage and communicate risk from an actor-wide perspective. It is about making strategic decisions that contribute to the achievement of an actor's overall corporate objectives. Integrated risk management requires an ongoing assessment of potential risks. Hence, it should be embedded in the actor's strategy and risk management culture. As stated above it is not limited to minimising the risks, but rather to foster innovation in order to achieve greatest returns with acceptable results, costs and risks. Hence, an optimal balance is strived for at the corporate level.

As pointed out in [Cari01], four key elements are included in an integrated risk management framework:

1. Develop the corporate risk profile – considering objectives and available resources
   a. Identify risks; resulting in description of threats and opportunities, i) type of risk – technological, financial, human resources, health, ii) source of risk – external, internal, iii) what is at risk – area of impact/type of exposure, iv) level of ability to control the risk – high (operational), moderate (reputation), low (natural disasters).

   b. Assess current risk management status; resulting in descriptions of challenges/ opportunities, capacity, practices.

   c. Identify risk profile; description of key risk areas, risk tolerance, ability and capacity to mitigate and needs for learning. In general, there seems to be lower risk tolerance for the unknown, where impacts are new, unobservable or delayed. In addition a higher risk tolerance is observed where people feel more in control (e.g. for car travel compared to air travel).

2. Establish an integrated risk management function
   a. Communicate, understand and apply management direction on risk management; risk management needs to be aligned with an actor's overall objectives, corporate focus, strategic direction, operating practices and internal culture. When a strategic risk management direction is set up, both internal and external concerns, perceptions and risk tolerances are taken into account. It is imperative to identify acceptable risk tolerance levels so the unfavourable outcomes can be remedied promptly and effectively.

b. Implement operational integrated risk management through existing decision-making and reporting structures. Integrating the risk management function into existing strategic management and operational processes ensures that risk management is an integral part of day-to-day activities.

c. Build capacity through development of learning plans and tools. Building risk management capacity is an ongoing challenge even after integrated risk management has become firmly entrenched. Environmental scanning will continue to identify new areas and activities that require attention, as well as the risk management skills, processes, and practices that need to be developed.

3. Practise integrated risk management
   a. Consistent application of common risk management at all levels. A common, continuous risk management process assists an actor in understanding, managing and communicating risk.

   b. Integrate results of risk management practices at all levels into informed decision-making and priority setting

   c. Apply tools and methods

   d. Consult and communicate with stakeholders

4. Ensure continuous risk management learning
   a. Establish supportive work environment where learning from experience is valued, and lessons are shared

   b. Build learning plans into actor's risk management practices

   c. Evaluate results of risk management to support innovation, learning and continuous improvement

   d. Share experiences and best practices

A potential risk management process wheel is illustrated in Figure 24. Internal and external communication and continuous learning improve understanding and skills for risk management practice at all levels of an organisation, from corporate through to front-line operations. The following steps may be included in a risk management process (from [Cari01]):

A. Risk Identification

1. Identify issues, set context:
   - Define the problems or opportunities, scope, context (social, cultural, scientific evidence, etc.) and associated risk issues.
   - Decide on necessary people, expertise, tools and techniques (e.g. scenarios, brainstorming, checklists).
   - Perform a stakeholder analysis (determine risk tolerances, stakeholder position, attitudes).

B. Risk assessment

2. Assess key risk areas:
   - Analyse context/results of environmental scan and define types/categories of risk to be addressed, significant organisation-wide issues, and vital local issues.

3. Measure likelihood and impact:
   - Determine degree of exposure, expressed as likelihood and impact, of assessed risks, choose tools.
   - Consider both the empirical/scientific evidence and public context.

4. Rank risks:
   - Rank risks, consider risk tolerance, use existing or developing criteria and tools.

C. Respond to risk

5. Set desired results:
   - Define objectives and expected outcome for ranked risks, short/long term.

6. Develope options:
   - Identify and analyse options – ways to minimise threats and maximise opportunities – approaches, tools.

Figure 24 A potential risk management process (adapted from [Cari01])

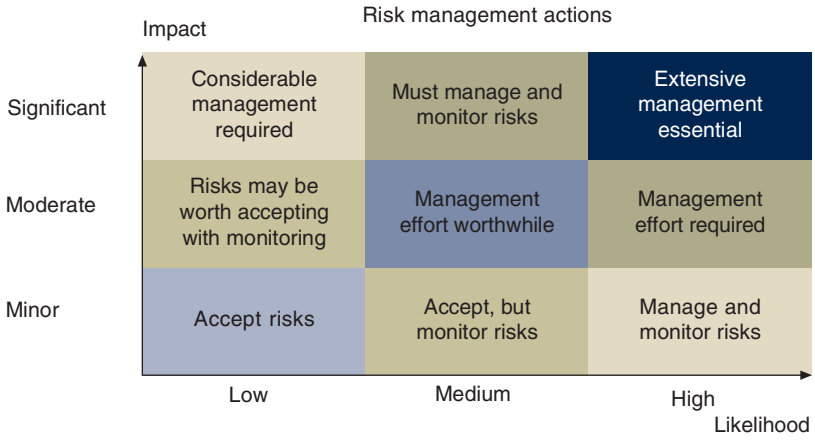


Figure 25 A potential risk management model

7. Select a strategy:
   - Choose a strategy, apply decision criteria – result-oriented, problem/opportunity driven
   - Apply, where appropriate, the precautionary approach/principle as a means of managing risks of serious or irreversible harm in situations of scientific uncertainty.
8. Implement the strategy:
   - Develop and implement a plan.

D. Monitor and evaluate

9. Monitor, evaluate and adjust:
   - Learn, improve the decision-making/risk management process locally and organization-wide, use effectiveness criteria, report on performance and results.

Each function or activity considered has to be examined from three perspectives:

- Its purpose: risk management would look at decision-making, planning, and accountability processes as well as opportunities for innovation.
- Its level: different approaches are required based on whether a function or activity is strategic, managerial or operational.
- The relevant discipline: the risks involved with technology, finance, human resources, and those regarding legal, scientific, regulatory, and/or health and safety issues.

A number of techniques can be used to assist in risk management, including:

- Risk maps: summary charts and diagrams that help organisations identify, discuss, understand and address risks by portraying sources and types of risks and disciplines involved/needed.
- Modelling tools: such as scenario analysis and forecasting models to show the range of possibilities and to build scenarios into contingency plans.
- Framework on the precautionary approach: a principle-based framework that provides guidance on the precautionary approach in order to improve the predictability, credibility and consistency of its application across several units.
- Qualitative techniques: such as workshops, questionnaires and self-assessment to identify and assess risks. Internet and organisational intranets: promote risk awareness and management by sharing information internally and externally.

Each assess risk can be illustrated according to its likelihood and impact as depicted in Figure 25.

9.2 Risk Allocation – Quantitative Aspects

In principle, two risk types can be defined, depending on the source of the risk:

- Systematic risk – coming from influences on the "market" in general. An example is change of interest rate.
- Non-systematic risk – coming from other sources than those affecting the "market" in general.

Considering a portfolio, the risk can be described by its exposure to the systematic factors, the volatility of those systematic factors, and the residual (or non-systematic risk). Regarding the systematic factors, an actor can gain passive exposure to these with little effort (or cost).

The non-systematic risk, in contrast, comes from the on-going decisions actors make in managing portfolios over time. When the actor is actively revising portfolio exposure to sources of systematic risk, these decisions generate non-systematic risk in a portfolio in addition to any systematic risk embedded.

The pay-off for bearing risk is commonly scaled according to the level of risk assumed. For systematic risk, the ratio involves dividing the risk premium by the level of systematic risk:

$$Sharpe\ ratio = \frac{Risk\ premium}{Systematic\ risk}$$

For non-systematic risk, the ratio involves dividing the non-systematic return by the level of non-systematic risk. This ratio is often referred to as the Information Ratio (IR):

$$IR = \frac{Non - systematic\ return}{Non - systematic\ risk}$$

For a portfolio, the return (pay-back) effects are additive between systematic and non-systematic risk. Hence, in variance terms:

$$\sigma^2_P = \sigma^2_S + \sigma^2_N + 2Cov_{SN},$$

where σ^2 indicates the variance (subscripts: P – total portfolio excess return, S – systematic excess return, N – non-systematic return) and Cov_{SN} indicates the covariance between the systematic and non-systematic returns.

For a typical well-diversified portfolio, the non-systematic risk makes up a relatively small portion of the total variance of the portfolio's return.

For an actor to allocate total risk in a portfolio between systematic and non-systematic sources, it is important to know the sources of risk for each alternative strategy that is considered as a candidate for inclusion in the portfolio. An important characteristic of most alternative strategies is that a greater portion of the total risk comes from non-systematic risk instead of from systematic risks.

An actor has the decision of how to i) allocate the risk budget across systematic factors, and ii) establish trade-off between systematic risk and non-systematic risk. This is the case even if the investor is not considering investing in any non-traditional strategies.

In these discussions it is essential to estimate the expected return from active management of each strategy (to exploit the non-systematic factors). One source of the expectations is the historical performance of different types of actively managed strategies. In some segments, possible gains may be high, while in other segments, there is less opportunity for additional gains.

An expected information ratio may support the understanding of actively managed portfolio. This can be illustrated as:

$$IR = \frac{E[\alpha]}{\sigma_\alpha} = IC \cdot TC \cdot \sqrt{N},$$

where:

IR	= information ratio
$E[\alpha]$	= expected non-systematic return
σ_α	= non-systematic risk
IC	= information coefficient
TC	= transfer coefficient
N	= number of decision points (often referred to as the breadth)

The information ratio is commonly a measure of how much non-systematic return that is expected relative to the amount of non-systematic risk. In general it is more preferable to obtain a higher information ratio:

- The information coefficient indicates how accurate the actor is in forecasting future returns. This is commonly estimated by computing the correlation between forecast returns and subsequent non-systematic returns. Different options with a high level of predictability have higher ICs than those with lower level of predictability.

  When an actor wants to increase the information ratio of an active strategy, this is to increase the information coefficient, either by finding better predictors of future relative returns or by recruiting individuals with better insight than others. Given the competitive nature of the market it is not that easy to increase the information coefficient.

- The transfer coefficient indicates how efficiently the actor's information is used in forming portfolio positions. The more constraints that are placed on the portfolio, the lower the transfer coefficient usually becomes. In a general case, constraints are commonly placed on the size of individual positions or combinations of positions (e.g. engagements in different countries, obligations to provide services, etc.). Relaxing any of these constraints tends to increase the transfer coefficient and improve the information ratio.

- The breadth indicates how many opportunities the actor has in applying the information gained. Hence, this shows the number of decisions that can be made during a process. This number is a function of both the number of elements in the portfolio and the frequency of decision-making. On the other hand, there is often a cost side of increasing the frequency of decision (such as preparing decision basis).

Considering these factors, in terms of risk budgeting, it is observed that a greater amount of non-systematic risk is allocated to the strategies that: i) focus on more inefficient markets (higher expected information coefficient), ii) are subject to fewer restrictive constraints (higher transfer coefficient), iii) have greater breadth (combination of a large universe to choose from and high frequency of portfolio management decisions).

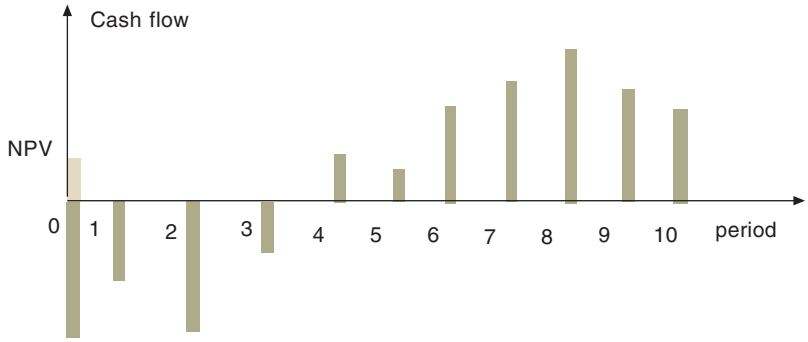


Figure 26 Net Present Value capturing the cash flow of an activity for a number of periods

9.3 Associating Gains with Options/Choices

On the one hand, relating to uncertain factors represents a risk to be included in the evaluation. On the other hand, in case a final decision can be postponed until more information is available, one would also consider a future option to decide as a positive effect. Hence, the latter should be included in the evaluation of a portfolio consideration where all future decisions could be added to the overall calculations.

A fairly standard way of evaluating an activity is to apply the Net Present Value (NPV). This measure collapses an activity's monetary timely schema into a single value. This is done by adjusting future cash flows and reflecting them into a current time value, see Figure 26.

Noting the cash flow in period *i* by CF_i and the periodic rate by *r*, a common way of expressing the *NPV* is:

$$NPV = \sum_i \frac{CF_i}{(1+r)^i}$$

By simply using this measure as a decision foundation, the activity with the highest *NPV* is chosen. Naturally, uncertainties are attached both to the cash flow in a certain period and to the set of rates (which might vary between periods).

Basing decision on a single parameter like *NPV*, a separation between the investment and finance matters has been assumed. That is, the investment decision is separate from the financing decision. An effect of this is that only the resulting *NPV* is looked at without considering the (cumulative) cash flows in selected periods. The cumulative cash flows indicate the amount of money needed to finance an activity.

Comparing *NPV* to 0 (zero) means that the decision rule is that investments could be increased as long as the marginal return on investment is greater than the cost of capital (determined by the finance market).

Performing the *NPV* calculations, using a proper value for the periodic (discount) rate, *r*, is essential. This rate must reflect the cost of capital, that is an investor's rate of return when making an investment. The discount rate should then both cover a compensation for the time-value of money as well as a risk premium as seen by the investor.

The Capital Asset Pricing Model (CAPM) is a frequently applied approach for assessing the discount rate. This postulates that the cost of capital, *r*, equals the return on risk-free securities, ϕ, plus the market price of the risk (market premium), $(\mu - \phi)$, multiplied by the company's systematic risk, β. Hence,

$$r = \phi + (\mu - \phi) \bullet \beta, \ where \ \beta = \frac{\sigma_a}{\sigma}.$$

Here the systematic risk, β, is expressed as the ratio between the covariance of the activity, σ_a, and the variance of the market, σ. The risk-free return, ϕ, is the rate one should obtain for an investment without any risk and can be approximated with government securities.

Then, the market price of risk becomes the difference between the expected rate of return and the risk-free rate; $E[\mu] - \phi$.

As seen from above, the systematic risk is the correlation of the activity return with the market divided by the variance of the market return. This may be a challenging task, allowing for subjective decisions by selecting activities belonging to the "market" (i.e. a portfolio of comparable activities).

The CAPM approach is applied to estimate the cost of capital and thereby the discount rate to apply in the *NPV* analysis. Introducing uncertainties in the *NPV* approach may be done in several ways; One way is to use simulations for "generating" cash flows in each period. The result is a distribution for the *NPV* which then can be used to deduce observations with respect to whether or not to go for an investment. An example of such a measure is the probability of having an *NPV* below a certain threshold.

One of the drawbacks of the traditional *NPV* calculation is that the future cash flows are often modelled as deterministic. Moreover, once an activity is initiated, it can simply be "passively managed". In practise, however, the cash flows are uncertain and the activity may commonly be adjusted/changed by more active management. Often traditional *NPV* can be referred to as static *NPV* calculations.

In trying to capture the uncertainties, a set of scenarios can be devised. These scenarios could

be identified in a number of ways; i) defining a set of uncertain factors and applying combinations of these to derive the scenarios, ii) defining a most likely outcome, together with a worst and a best outcome, iii) defining a set of scenarios depending on the outcome of (external) major factors.

The *NPV* can then be calculated for each of the scenarios. As described earlier, the scenarios help understand the uncertainties and derive the major factors to be observed and trigger actions. On the other hand, a number of drawbacks can be claimed, mainly due to i) the *NPV* calculation for each scenario is based on the same assumptions as the static *NPV* (deterministic cash flows and passive management), and ii) not capturing the option of jump between different scenarios during the course.

9.4 Decision Tree Analysis

The decision tree analysis is one approach for including the options revealed during the course of the activity. Setting up an event tree does this where the branches represent a cash flow outcome attached with the probability for that event to take place. The decision nodes are added to the tree where the management may choose to change the activity in order to respond to (external) factors or results of the activity. Such a method could for example be applied when analysing a complex sequential investment scheme where several decision points can be identified at discrete points of time, see Figure 27. Note the resemblance between decision tree and decision map in Figure 1. For a static *NPV* calculation, all uncertainty is presumably captured by the discount rate, the decision tree adds more understanding of the options during the activity's course and their probabilities. By modelling the investment as a decision tree, different actions in different scenarios or nodes in the tree can be introduced to incorporate the value of flexibility.

The tree is solved backwards from the leaf nodes and the beginning of the tree by discounting the relevant cash flows. The result is the value of the activity with flexibility as modelled in the tree.

This approach is intuitive and simple to comprehend, although it may have a number of limitations: i) the number of events may grow in a way such that analysing the tree is intractable, ii) the same discount rate is commonly used throughout the tree. Frequently, the risk differs at different nodes/branches inspiring for different discount rates. Commonly the discount rate applied is for an activity without flexibility.

9.5 Real Options

A net-present value (*NPV*) factor is usually derived for an activity and decisions made on this basis. A traditional *NPV* forces a decision to be made on the current set of information about the future. On the other hand, option valuation allows for additional flexibility of main decisions in the future, as more information will be available. This option of postponing a decision is captured in the option method that results in two additional types of values due to the simple fact that one will always want to postpone a final decision as long as possible. Firstly, one can earn the time value of money (e.g. interest). Secondly, as time passes until the decision has to be made, more information would be available such that a firmer foundation is achieved for making a decision.

Both arguments favour deferring a decision as long as possible. The traditional *NPV* criterion misses such an aspect. The real option methodology presumes the ability to postpone a decision and provides a way to quantify the value of deferring.

Using real option pricing, the value of an activity will always be greater than or equal to the value of a project using *NPV*. In case the identified flexibility by real option is unlikely to be used, the difference in pricing would be small. This is also likely to happen when the *NPV* is very high or very low (much negative). The greatest difference is likely to be seen when *NPV* is close to zero, hence referring to a project whose activation is questionable.

The higher the level of uncertainty, the higher the option value because flexibility allows for gains in the upside and minimise the downside potential. Often the total discounted operation

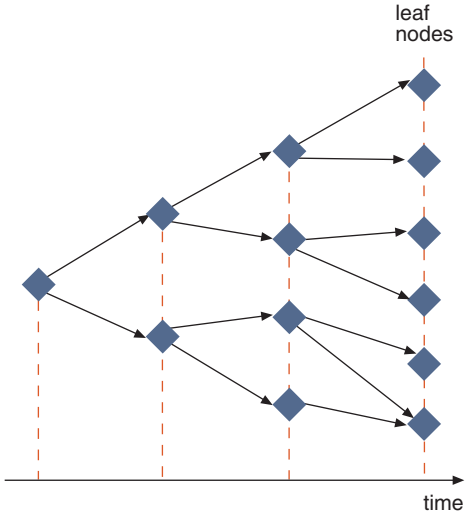


Figure 27 Decision tree analysis

[1] *Operating cash flow = revenue – all OAM, sales and other costs – variable follow investments (line cards, etc.)*

Investment option	Variable	Call option
Present value of a project's operating assets to be acquired	S	Stock price
Expenditure to acquire the project's assets	X	Exercise price
Length of time the decision may be deferred	T	Time to expiration
Time value of money	r_f	Risk-free rate of return
Riskiness of project assets	$\sigma\sqrt{T}$	Cumulative volatility

Table 4 Relating investment option and stock call option

cash flow[1)] is modelled by a lognormal stochastic variable. The uncertainty is modelled by a volatility commonly used for financial call options.

Utilising real options, theory from financial analysis is applied to evaluate the value of physical or real assets. It is claimed that real options have been very useful in assisting top management of companies facing a significant managerial flexibility and amount of uncertainty.

There are two types of financial options: i) call options, and ii) put options. The former is a right to buy, while the latter is a right to sell (note: these are options and not obligations). Referring to an activity and the corresponding managerial flexibility, some relevant options are:

- Abandonment option (a put option): when the complete activity is stopped.
- Option to contract (a put option): when part of the activity may be stopped, outsourced, sold, etc.
- Option to expand (a call option): when the activity is enhanced.
- Option to defer: when a decision is postponed.

Combinations of options may also be identified. Two cases are:

- Compound option: an option is an option of the value of another option. For example, one may have the possibility to shrink or expand an activity, but the amount (e.g. how many parts) may be flexible. Another example is sequential option plan (e.g. for a phased activity plan) where an option may be a potential follow-on of another option.
- Switching options: reflecting the possibility of switch between two operation modes. Two examples are to: i) exit and re-enter by a modified activity plan, ii) change from delivering a certain result to another result. When the switching cost is greater than zero, situations may occur when one is still carrying on with the activity even though the result would be better if changes were incorporated. This is due to the high cost of changing the activity plan.

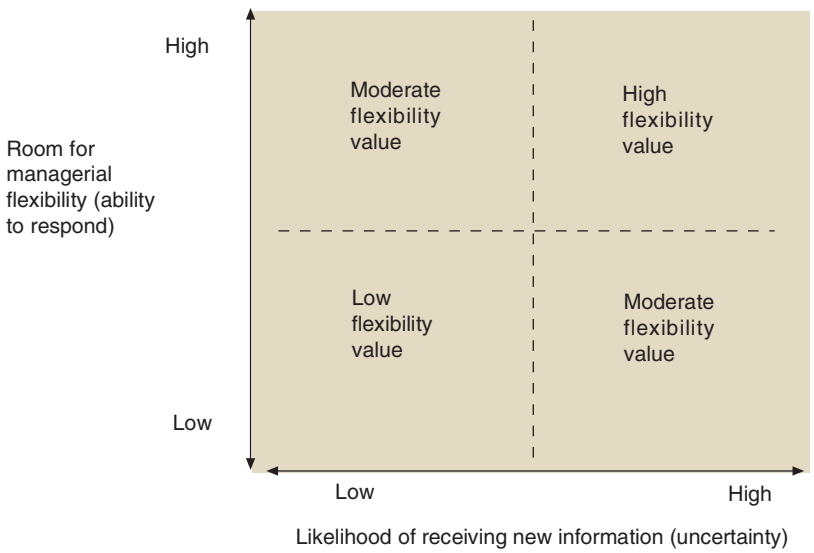
An important difference between real and financial options is that management can affect the value of the underlying risky assets as the actual project is under the management's control. Third parties control the financial options. This can be illustrated by mapping an investment question into a call options as shown in Table 4.

A number of observations can be made, see Figure 28:

- An increase in the present value of the project will increase the *NPV* (without flexibility) and therefore the option value will also increase.
- A higher investment cost will reduce *NPV* (without flexibility) and therefore reduce option value.
- A longer time to expiration allows learning more about the uncertainty and therefore will increase the option value.
- An increase in the risk-free rate of return will increase option value since it will increase the time value of money advantage in deferring the investment cost.
- In an environment with managerial flexibility an increase in uncertainty will increase option value.

In every case where the case value without flexibility is close to break even, the flexibility will have most relative value. Hence, the flexibility value is greatest when:

Figure 28 Estimating when managerial flexibility is greatest valued



- Great uncertainty about the future. Very likely to receive new information over time.

- Much room for managerial flexibility. Allows management to respond appropriately to this new information.

- *NPV* without flexibility near zero. If the activity is neither obviously good nor obviously bad, flexibility to change course is more likely to be used and is therefore more valuable.

In the financial world, the Black-Scholes formula is well-applied for valuing real options. This is an analytical solution to a differential equation describing the value of a European call option. The underlying risky asset is assumed to follow a geometric Brownian Motion with a Markov-Wiener stochastic process. When S is the value, the percent change is given by:

$$\frac{\delta S}{S} = \mu \cdot \delta t + \sigma \cdot \varepsilon \cdot \sqrt{\delta t}$$

where

μ is a drift term or growth parameter that increases at a factor of time steps δt.

δ is the volatility parameter, growing at a rate of the square root of time.

ε is simulated, usually following a normal distribution with a mean of zero and variance of one.

The first term is the deterministic part and the second term is the stochastic part.

A number of limitations on this formula are: there is only one source of uncertainty, no compound options are included, the underlying stochastic process is assumed to be known, the variance of return is constant through time, the underlying risky asset follows a Geometric Brownian Motion with a Markov-Wiener stochastic process.

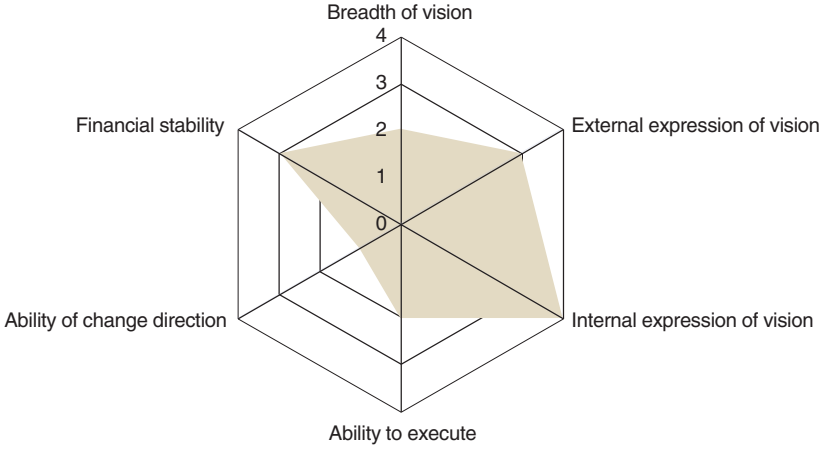
One way of modelling real options is to apply binomial lattices. Market-replicating portfolios and risk-neutral approach could for instance be used when calculating the values of the option sets. The basic idea is to represent the evolving uncertainty of the value of a risky asset by a binominal tree. The risk itself may or may not increase with time, but uncertainty increases with time. This can be considered as an uncertainty cone (hence growing as time passes). In the same way as the Black-Scholes formula this relies on a Geometric Brownian Motion following a Markov-Wiener process. As for any tree, time is discretised, and hence it may become quite demanding to compute every combination. As a limit, the value of an option calculated according to the binomial lattice approaches the time continuous solution of Black-Scholes as the number of time-steps in the binomial tree gets larger.

The calculations are started from the leaf nodes of the tree (similar to Figure 27) where the question of whether to exercise the option or not is a trivial one (being the end-points). The calculations are run backwards to the beginning of the tree, choosing the best decision of whether to exercise or not at each branch/node of the tree. At each branch/ node the market-replicating portfolios or the risk-neutral approach can be applied, both being equivalent to each other. The former refers to finding an equivalent combination of risky and risk-free assets, while the latter refers to adjusting the cash flows in the *NPV* calculations.

Naturally, one may raise the question of how to find an equivalent portfolio when referring to an activity "never-seen-before" for a certain company. On the one hand, it can be claimed that these are all models used as background when making the actual decisions. On the other hand, the static *NPV* of an activity could be used as a best-guess on the unbiased estimate of the market value of the activity. Considering the value including flexibility, one could consider this as placing one part of the money on the activity and another part in bonds.

The binomial event tree describes the evolving uncertainty of the underlying risky asset and the stochastic process is based on a single volatility. Monte Carlo simulation is commonly used to produce the volatility for the return of activity. This is however based on the assumption that the uncertainty is resolved continuously over time. There are often different causes for uncertainties, such as market and technology. These may well be resolved differently with time; an example being that technological uncertainties decrease with time while market uncertainties increase with time.

Figure 29 Potential categorisation of vendor's strength



Real Options – Referring to Stocks

One model for stock behaviour is a so-called geometric Brownian motion. The time discrete version is expressed as:

$$\frac{\Delta S(t)}{S(t)} = r_f \Delta t + \sigma \varepsilon \sqrt{\Delta t}$$

where:

$S(t)$ – stock price at time t

$\Delta S(t)$ – change in stock price during a short time interval, Δt

r_f – drift rate (risk-free rate of return)

σ – volatility

ε – random number drawn from a standard normal distribution (~N(0,1)).

Now it can be shown that $S(t)$ is log normally distributed with mean value given as

$$\mu_t = \ln(S(0)) + \left(r_f - \frac{\sigma^2}{2}\right) t$$

and standard deviation of $\sigma_t = \sigma\sqrt{t}$.

The measure under consideration, here the total discounted value of the operating cash flow of a project, T, is estimated at $S(0)$ with the current information. Up to decision time, T, the management receives information of relevant value and the measure value will therefore move up or down with accumulated volatility $\sigma\sqrt{t}$.

The project investment is only incurred if $S(T) > X'' = X + (1 + r_D)^2 V$, where X is the initial investment at time T and V is the NPV of the second phase of the project if the uncertain action is not carried out. r_D is the discount rate.

The *NPV* at time T is found as:

$$NPV = E\left[\max(S(T) - X,\ (1 + r_D)^2 V)\right]$$

$$= \int_{X''}^{\infty} (S(T) - X) p(S(T)) dS(T) + \int_0^{X''} (1 + r_D)^2 V p(S(T)) dS(T)$$

where

$$p(S(T)) = \frac{1}{\sigma_T \sqrt{2\pi S(T)}} e^{\frac{-(\ln(S(T)) - \mu_T)^2}{2\sigma_T^2}}$$

After some reordering and variable substitution (e.g. $w = \ln(S(t))$ -> $S(T) = e^w$, $dS(t) = e^w dw$), one gets

$NPV = e^{-rfT} (S\ \mathrm{N}(d_1) e^{rfT} - X\ \mathrm{N}(d_2)) + V\ (1 - \mathrm{N}(d_2))$

where

$$d_1 = \frac{\ln\left(\frac{S}{X''}\right) + \left(r_f + \frac{\sigma^2}{2}\right) T}{\sigma\sqrt{T}} \quad \text{and} \quad d_2 = \frac{\ln\left(\frac{S}{X''}\right) + \left(r_f - \frac{\sigma^2}{2}\right) T}{\sigma\sqrt{T}} = d_1 - \sigma\sqrt{T}.$$

N(.) is the cumulative standard normal distribution (~N(0,1)).

9.6 Stochastic Optimisation

Stochastic optimisation is an operational research technique for handling uncertainty. This is applied in order to optimise simulation models by defining a stochastic model and accompanying decision variables. Iterations are carried out between drawing input parameters from selected distributions and optimising the decision variables for each set of draws. In some sense this is similar to solving the task of finding the best path in a decision tree.

9.7 Evaluating Vendors

One element in the risk assessment is evaluating the situation of the current vendors. Both technical and financial criteria have to be considered.

Having an assessment of the key vendors for an operator is also central. Several criteria could be derived, of which some are (from [Heav03]):

- Breadth of vision
- External expression of vision
- Internal expression of vision

- Ability to execute
- Ability to change direction
- Financial stability

This inspires for illustrative factor diagrams as shown in Figure 29. However, it is important to be aware that the sequence of factors influences the area of the diagram.

10 Overall Results

Regarding the overall objective of network strategy studies, more than one target state may very well be described within each study. A natural justification for this is that a future time frame is commonly attached with uncertainty, leaving it unlikely that every factor influencing the solutions will be known. Hence, instead of looking for the ultimate solution, a decision tree approach could be strived for. That is, paths of migration are put together to choose which branch to follow if the tree is attached with a set of conditions. These conditions may be related to time and events. On a more abstract level every branching would actually be event-related, although some indication of timing will be of interest. This is schematically illustrated in Figure 30. The network naming has been chosen simply for illustration. Moreover, more than two branches might appear from one state. Following a line from a network without any merging or splitting into branches indicates an upgrade of network capabilities.

Following a branch in the tree both the consequences and the risks must be assessed. Consequences reflect costs, revenue side, product portfolio, organisation and so on. Here, the total portfolio has to be taken into account, possibly leading to investments in a certain system resulting in overall reduced operational expenses because a less efficient system could be phased out.

It is essential to have captured the relevant risk factors when choosing the right implementation track for core network migration. In other words it is not sufficient for an implementation track to be technically feasible and having the smallest cost among the various implementation options, the option must also have a sufficiently low risk to be considered for actual implementation. Risk factors can be grouped into technical, financial and business/strategic. The first refers to factors such as whether the technical solutions are actually able to deliver the services as promised, compatibilities between versions/vendors, etc. Financial issues come from the income and the expense side, where none of these are fixed in the period. Business and strategic issues refer to relations with regulatory bodies, organisation, communication with customers/competitors, and so forth.

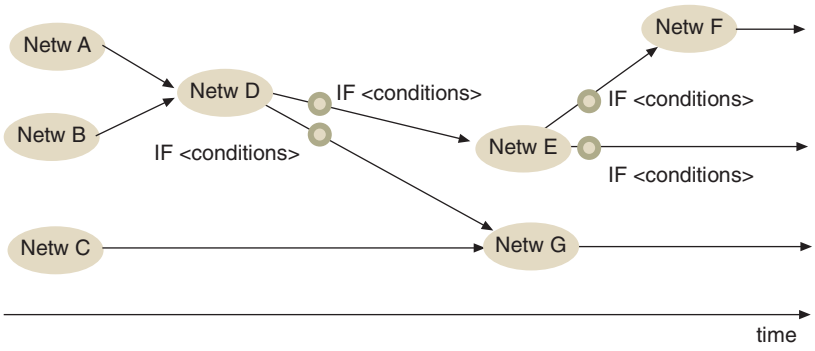


Figure 30 Illustration of network portfolio migration states with decision points

Production means

	Netw A	Netw B	Netw C	.....
Product z	V		V	
Product x		V	V	
Product w		V		
....				

Year 200y

Figure 31 Relating products and production means for each time period

Naturally, any modification of an existing operating network implies certain risks. In particular there are risk factors related to upgrading nodes related to cost, functionality and capacity. This also applies to any work of integrating networks, and other steps taken in case an additional aggregator function is to be introduced (as assumed for some of the cases/tracks). On the other hand, there may also be risks involved when no changes are carried out. This could come because the system portfolio may gradually become less efficient than what competitors manage and too slow to deploy new services.

As mentioned in the beginning it is essential that the network strategy elaborated is related to the current situation and thereby possibly leading to a change of the decisions to be made in the shorter term.

The products to be supported have to be related to the network portfolio illustrated in Figure 30. This may be thought of as a 'product – production means' matrix, as shown in Figure 31. Here the main product groups are given along one axis and the production means on the other axis. Such a matrix must be defined for each period/year looked at. In a period a product may be provided in a number of ways, partly competing and partly complementing. Doing the exercise of

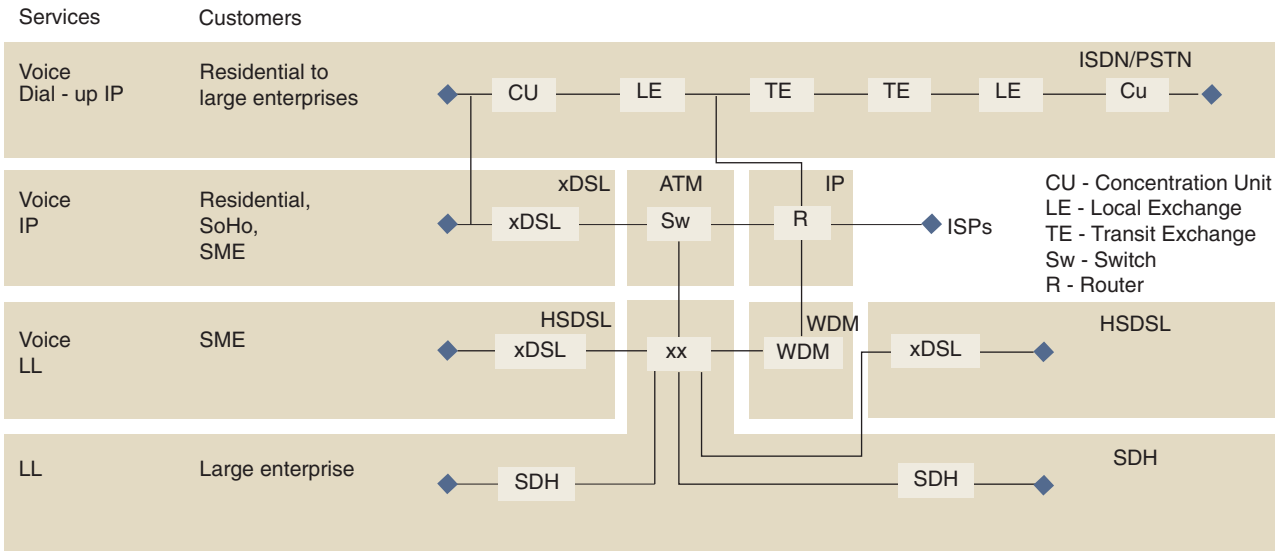


Figure 32 Example of relating services/customer segments to production means (Note: not all relevant systems are included, e.g. the mobile systems)

relating products to production means will reveal further areas of potential gains.

To some extent products to be supported have been described earlier. A few basic trends have also been outlined earlier. Regarding the demand patterns, most sources indicate a growth in most issues, leading to rather questionable aggregated demands. However, trends showing main growth areas within mobile, IP and xDLS are repeatedly observed.

Defining how to efficiently support the different products belongs to the network portfolio management. A basic challenge is to relate the services and market segments. An example of how this may be carried out is depicted in Figure 32.

11 Concluding Remarks

A network strategy is needed for every sound network operation. The strategy must also be related to decisions and actions in near-time – hence, the strategy must be made operational. A main goal of having a strategy is to be prepared for chances that can be revealed during the course of events. That is, the strategy is likely to assist when detecting business opportunities. The total network portfolio must be included, also seeing synergies between the different systems in the shorter and longer terms.

In elaborating the strategy a number of methods can be applied, including scenarios, cost/benefit calculations and risk assessment. A description of various aspects is outlined in this article.

Acknowledgement

Although the material in this text is the sole responsibility of the author, fruitful discussions with all colleagues involved in Telenor strategy work in later years are appreciated.

References

[Cari01] Canadian Treasury Board. *Integrated Risk Management Framework*. 2001. Online: http://www.tbs-sct.gc.ca

[Heav03] *Heavy Reading: Incumbents' wireline strategies*, 1 (3), 2003.

Portfolio Evaluation Under Uncertainty

RALPH LORENTZEN

Ralph Lorentzen (67) is daily leader of the company Lorentzen LP Service which does consulting work within mathematical programming. He retired July 1, 2003 from his job as senior scientist in Telenor R&D. Ralph Lorentzen graduated in mathematics at the University of Oslo, where he worked for some time as assistant professor giving lectures in statistics and operations research. He worked as distribution planner in Norske Esso, as principal scientist at Shape Technical Centre, as systems engineer in IBM, and as chief consultant in Control Data Norway, before joining Telenor R&D in 1985.

ralph.lorentzen@tiscali.no

1 Introduction

Principles for evaluation of the risk associated with a portfolio of investment projects within a company are based on models that describe potential changes in the factors that influence the profit of the individual projects and the interaction between these. Many different models are described in the literature. Some approaches are simple and may border towards the trivial, whilst others are more involved and rely on advanced stochastic models. During the last years it has become popular to apply principles and theory from the mathematics of finance (see e.g. [1] and [2]). Analogy is drawn between investment in securities on the one hand and investment in projects on the other. Both situations are characterised by the investment of an amount of money at one or several points in time with the hope of receiving larger sums of money at one or more future points in time. Such analogy has its strong and weak sides.

In this paper we describe an approach based on what we call *Risk-Adjusted Expected Net Present Value (RAENPV)* and contrast it to the more common *Net Present Value (NPV) with Risk Adjusted Discount Rate*. We also propose a way of describing the uncertainty associated with the cash flow of each individual candidate project in a portfolio and how it interacts with each of the other candidate projects. Our guiding principle has been to establish an approach, which on the one hand is not too narrow but on the other hand avoids relying on complicated models that cannot be justified because of lack of knowledge about the future.

2 Evaluation of a Single Project

2.1 Risk-adjusted Discount Rate

We shall first briefly comment in general terms on an approach which is often used where a project is rated according to its NPV calculated using a risk-adjusted discount rate that reflects the uncertainties in the cash flow. The approach is usually based on variants of the *Capital Asset Pricing Model (CAPM)*. A thorough description of CAPM may be found in [1]. A feature of this approach is that high uncertainty in the cash flow leads to a high risk-adjusted discount rate. The approach is not without problems, even for projects with a lifetime of one period only. We mention some of the conditions that must be satisfied:

- Investors have expectations about asset returns that are normally distributed.
- Investors may borrow or lend unlimited amounts of a risk-free asset.
- All projects (and parts of projects) can be bought/sold in a perfect market.

Even stronger are the conditions that must be satisfied if we want to use a risk-adjusted discount rate for projects with a lifetime of more than one period. Then it must be assumed that the cash flow in one period is a deterministic function of the cash flow in the previous period plus a random variable that is independent of what has happened in all previous periods.

2.2 Problems with the Use of a Risk-adjusted Discount Rate – Examples

In many situations it is not reasonable to assume that the conditions mentioned in the previous section are satisfied. We shall show by three project examples what happens when these conditions are not fulfilled. The examples are stylised to visualise the effect of risk-adjusting the discount rate. However, their basic features are not unrealistic.

The cost of capital is assumed to be 12 %, and we assume that the relevant risk-adjusted discount rate has been found (using for example CAPM type arguments) to be 13 % for all three projects. All amounts are in millions of dollars.

Example 1:
We invest 98 in year 0 in building a VDSL access network. There is uncertainty as to whether the revenue from subscribers will commence in year 1 or 2. With probability 1/2 we receive a net income of 13 in year 1. In the years 2, 3, … we receive with probability 1 a yearly net income of 13. The uncertainty is thus whether the income stream starts in year 1 or 2 (see Figure 1 where the income that occurs with probability 1/2 is hatched).

Situation A – Net income 13 in year 1:
The NPV becomes (13/1.12) / (1 – 1/1.12) – 98 = 10.3.

Situation B – Net income 0 in year 1:
The NPV becomes $(13/1.12^2)$ / (1 – 1/1.12) – 98 = –1.3.

Expected NPV: 1/2 × 10.3 + 1/2 × (–1.3) = 4.5

Expected NPV with risk-adjusted discount rate: $1/2 \times (13/1.13) / (1 - 1/1.13) + 1/2 \times (13/1.13^2) / (1 - 1/1.13) - 98 = -3.8$.

The method thus indicates that the project should be rejected.

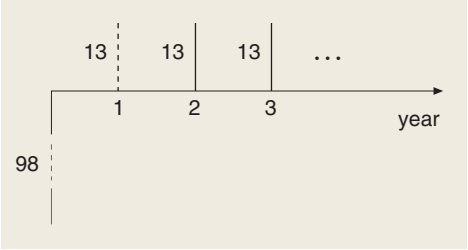


Figure 1 Cash flow in Example 1

Example 2:
We invest 63 in building an ADSL access network in year 0. There is uncertainty as to whether the ADSL technology will generate revenue for 4 or 5 years. With probability 1 we receive a net income of 20 in the years 1, ..., 4. With probability 1/2 we receive a net income of 20 in year 5. The uncertainty is thus whether we receive an income in year 5 (see Figure 2 where again the income that occurs with probability 1/2 is hatched).

Situation A – Net income 20 in year 5:
The NPV is $(20/1.12) \times (1 - (1 - 1/1.12)^5 / (1 - 1/1.12) - 63 = 9.1$.

Situation B – Net income 0 in year 5:
The NPV is $(20/1.12) \times (1 - (1 - 1/1.12)^4 / (1 - 1/1.12) - 63 = -2.3$.

Expected NPV:
1/2 × 9.1 + 1/2 × (–2.3) = 3.4

Expected NPV with risk-adjusted discount rate: $1/2 \times (20/1.13) \times (1 - (1 - 1/1.13)^5 / (1 \times 1/1.13) + 1/2 \times (20/1.13) \times (1 - (1 - 1/1.13)^4 / (1 - 1/1.13) - 63 = 1.9$.

The method thus indicates that Example 2 should be accepted.

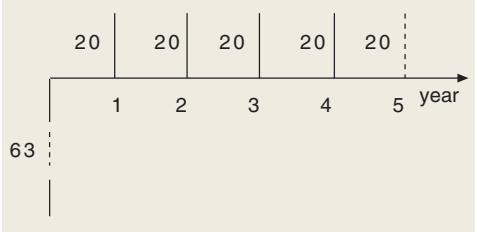


Figure 2 Cash flow in Example 2

Example 3:
We invest 98 in a VDSL access network in year 0. There is uncertainty as to whether the revenue from subscribers will commence in year 1 or 2. With probability 1/2 we receive a net income of 13 in year 1. In year 2 we sell the network to another company for 13 / (1 – 1/1.12) = 121.3 (which is the NPV at capital cost of the incomes received with certainty in Example 1 from year 2 and onwards; see Figure 3). We see that Example 3 is financially equivalent to Example 1. We have only replaced the income in year 2 with the NPV at capital cost of the incomes we receive with certainty from year 2 and onwards.

Situation A – Net income 13 in year 1:
The NPV becomes $13/1.12 + 121.3/1.12^2 - 98 = 10.3$.

Situation B – Net income 0 in year 1:
The NPV becomes $121.3/1.12^2 - 98 = -1.3$.

Expected NPV:
1/2 × 10.3 + 1/2 × (–1.3) = 4.5.

Expected NPV with risk-adjusted discount rate: $1/2 \times 13/1.13 + 121.3/1.13^2 - 98 = 2.8$.

The method thus indicates that Example 3 should be accepted.

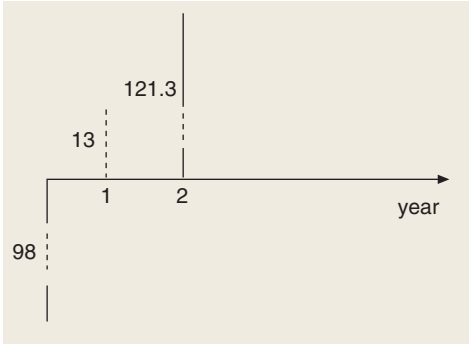


Figure 3 Cash flow in Example 3

These results are not reasonable. We see that the projects in Example 1 and Example 3 are financially equivalent. The criterion for whether they should be accepted or rejected should therefore be the same. Either should both be accepted or both be rejected. However, the expected NPV with risk-adjusted discount rate is negative for Example 1 and positive for Example 3. The NPV method with risk-adjusted discount rate thus implies that the project in Example 1 should be rejected whilst the project in Example 3 should be accepted. Furthermore, when the discount rate is set to capital cost, the project in Example 1 gives a higher NPV than the project in Example 2 for all statistical outcomes. Common sense thus implies that if the project in Example 2 is accepted, then the project in Example 1 should be accepted. The NPV method with risk-adjusted discount rate implies, however, that the project in Example 1 should be rejected

whilst the project in Example 2 should be accepted.

The examples demonstrate that any method based on risk-adjusting the discount rate is unable to differentiate correctly between different risk structures. By risk-adjusting the discount rate one tries to satisfy two masters. The discount rate should reflect

- Capital cost (i.e. the rate of return which can be achieved by alternative application of the money which are tied up/released in the project over time). This has nothing to do with the risk of the project under consideration;
- The risk associated with the project

These two requirements are fundamentally different, and it is difficult to satisfy both by adjusting the discount rate. The main reason to look for an alternative approach is to be able to differentiate between capital cost and risk.

2.3 Risk-adjusted Expected Net Present Value

We shall now describe our suggested approach. The expected NPV is calculated using capital cost w as discount rate. The capital cost should reflect the average expected rate of return for alternative projects in the company and has nothing to do with the risk associated with the project under evaluation.

Then an amount is subtracted from the expected NPV. This amount adjusts the Expected NPV because of the risk of the project. The suggested measure of utility of a project is

$$\text{Expected NPV} - \alpha \times \text{risk} \qquad (2.3.1)$$

where α is a parameter to be determined. In financial theory the risk is traditionally expressed through the standard deviation of the NPV. This is a reasonable measure for portfolios of securities. Such portfolios consist of a number of different securities, and the return is the sum of the returns of the individual securities. The central limit theorem in probability theory indicates that this sum is approximately normally distributed. Risk should reflect the down side of the distribution of the NPV. However, the normal distribution is symmetric, so the standard deviation expresses the down side as well as the up side of the distribution. For individual investment projects in a company the assumption that the NPV has a symmetric probability distribution is obviously unreasonable. The NPV will often have a skewed distribution. As the standard deviation also reflects the up side of the distribution, we may have a large standard deviation even if the risk of the project is small.

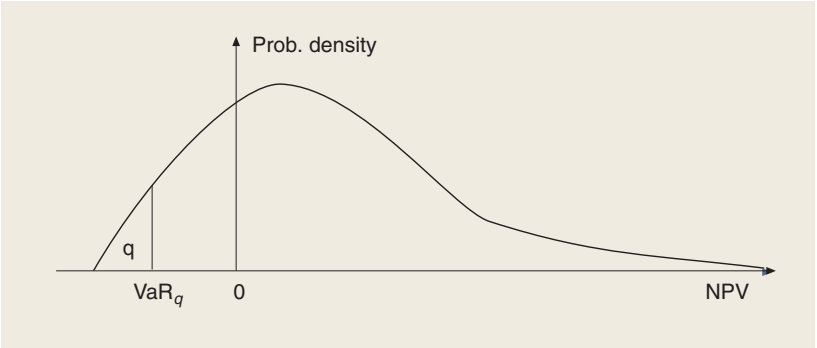


Figure 4 Value-at-Risk

Therefore we choose to measure risk by using the well-known concept *Value-at-Risk at q (VaR$_q$)*. *VaR$_q$* is defined as the q-point in the probability distribution of the NPV (see Figure 4).

q is usually chosen to be 1 % or 5 %. The suggested RAENPV then becomes

$$\text{Expected NPV} - \alpha \times (-\text{VaR}_q) \qquad (2.3.2)$$

where the NPV is calculated using the capital cost w as discount rate and where the parameter α reflects how the company (or its shareholders) weigh expected profit versus risk. We shall see later that the term $-\alpha \times (-\text{VaR}_q)$ can be interpreted as the insurance premium the company is willing to pay in order to be guaranteed against a negative NPV.

We shall now show how we determine α. For a perfect security market which contains a risk-free asset it is established knowledge that the substitution rate γ (price of risk) between expected return and risk (measured by the standard deviation of the return) in equilibrium is the same for all investors independent of their willingness to take risk (see e.g. [1] chapter 6). The value of γ can be obtained from stock exchange data and is not specific for a particular industrial sector. So for any investor an increment $\Delta\sigma$ in risk needs to be compensated by an increment of $\Delta\sigma/\gamma$ in expected return. This is shown in [1] for investments in a portfolio of securities where the portfolio is sold after one period, but it seems reasonable to use the same trade-off for multi-period projects. It is reasonable to measure the utility of an investment in a portfolio of securities by $E - \gamma\sigma$ where E is expected return and σ is the standard deviation of the return.

Figure 5 shows the classical picture where the dots represent securities in a diagram with the standard deviation of the return along the horizontal axis and the expected return along the vertical axis. The lower curve – the opportunity set border curve – limits all possible portfolios of securities. The straight line which starts at the risk-free asset and is tangent to the opportunity set border curve is the so-called *capital budget line*.

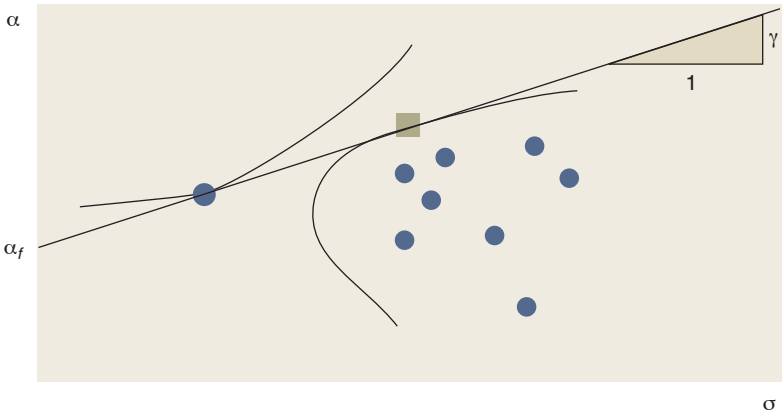


Figure 5 The capital budget line

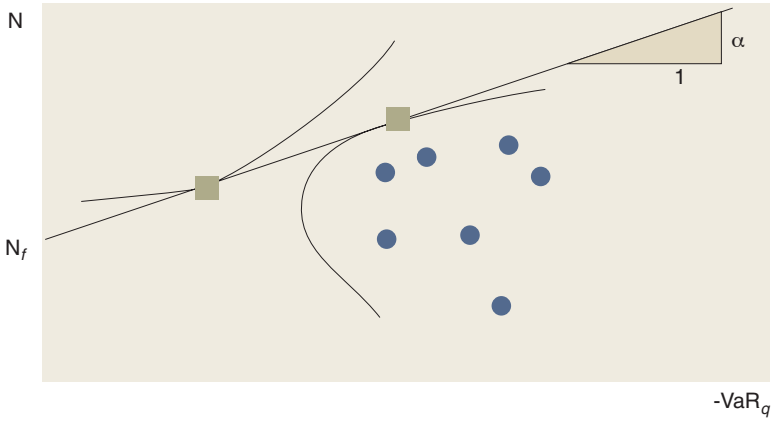
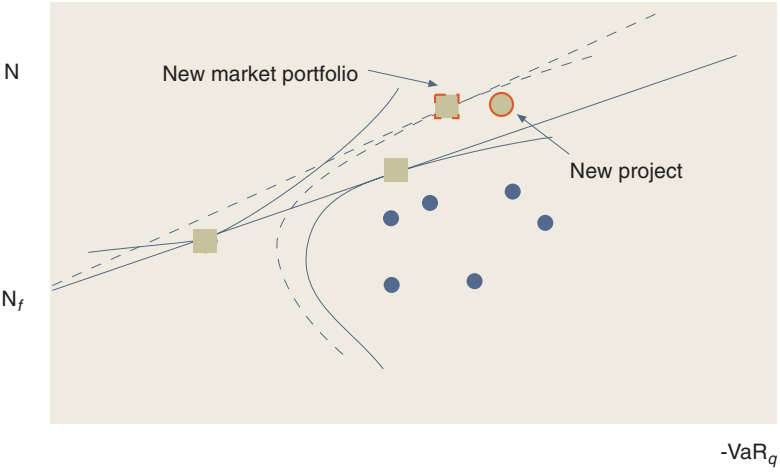


Figure 6 Transformed capital budget line

The point where the capital budget line touches the opportunity set border curve represents the market portfolio and is marked by a small square in Figure 5. The upper curve in Figure 5 represents an indifference curve for an investor. The point where this curve touches the capital budget line, and which is marked by a small circle, represents the combination of the market portfolio and the risk-free asset which is considered to be optimal for this particular investor.

Figure 7 Project with higher risk-adjusted NPV than the market portfolio

Figure 5 can be transformed to Figure 6 where we have replaced the standard deviation of the return by –Value-at-Risk for the NPV based on a discount rate r and the expected return by the expected NPV.



The rate of increase γ in Figure 5 has here been transformed to the rate of increase α in Figure 6. It can be shown (see the Appendix) that α **is independent of which discount rate** r **we apply**. We note further that the indifference curve for any investor has rate of increase equal to α at the point where it touches the transformed capital budget line. That is, α represents the optimal weighting between expected NPV and –Value-at-Risk. As mentioned earlier, the rate of increase γ can be found from capital market data. The formula for calculating α from γ is derived in the Appendix and is

$$\alpha = \frac{\gamma}{n_q - \gamma} \tag{2.3.3}$$

where n_q is the $(1 - q)$-point in the standardised normal distribution.

Since the market portfolio can be regarded as a yardstick, it is of interest to calculate the RAENPV for the market portfolio. In the Appendix it is shown that it can be expressed as

$$\frac{i - r}{(1 + r)(1 - \gamma/n_q)} \tag{2.3.4}$$

where i is the internal rate of return for the market portfolio.

A project should then be accepted if the RAENPV is greater than this. That will be in the interest of the shareholders. The project can then be represented by a point in Figure 6 that lies above the transformed capital market line. If the project is considered as a new investment possibility in the capital market, a larger opportunity set, a new transformed capital market line, and hence an improved market portfolio would arise. In Figure 7 the new opportunity set, capital market line, and market portfolio are indicated by broken lines. The new market portfolio would necessarily include the project.

An investment possibility represented by a point underneath the transformed capital market line might still improve the revised market portfolio and thus be a profitable investment. This would, however, be rather difficult to ascertain in general. We therefore propose to be conservative and leave to the planner the decision whether a project with a negative RAENPV should be accepted or not.

2.4 Analysis of the Examples Using Risk-adjusted Net Present Value

We shall now apply our criterion on the three examples described earlier where we assume that a has been estimated to be 0.2 and that the RAENPV of the market portfolio is very small.

In all three examples VaR_q at capital cost is the same whether q is 1 % or 5 % since the probability distributions of the NPV are different from zero for two values only, and for those values the probabilities are equal to 50 %.

Example 1:
The expected NPV and VaR_q with discount rate equal to capital cost are 4.5 and –1.3 respectively. The RAENPV becomes 4.5 – 0.2 × 1.3 = 4.24. The project should therefore be accepted.

Example 2:
The expected NPV and VaR_q with discount rate equal to capital cost are 3.4 and –2.3 respectively. The RAENPV becomes 3.4 – 0.2 × 2.3 = 2.94. The project should therefore be accepted. We see that the RAENPV is less than for the project in Example 1, which is reasonable since the project in Example 1 has a higher NPV at capital cost than the project in Example 2 under both statistical outcomes.

Example 3:
The expected NPV and VaR_q with discount rate equal to capital cost are 4.5 and –1.3 respectively. The RAENPV becomes 4.5 – 0.2 × 1.3 = 4.24. The project should therefore be accepted. The RAENPV is the same as for the project in Example 1, which is logical since the two projects are financially equivalent.

2.5 Establishing the necessary Data for Evaluation of a Single Project

The cash flow of a project is considered as a stochastic process. For every realisation of the process the NPV with capital cost as the discount rate can be calculated. This NPV is thus a random variable, and we propose the following procedure for estimating its probability distribution.

1 Establish a finite set of alternative future scenarios and assign a subjective probability of occurrence for each scenario.

2 For each scenario establish a pessimistic, an optimistic and a most probable cash flow for the project and calculate the NPV for each of those three realisations.

3 For each scenario specify an interval around the most probable realisation in which the NPV supposedly will lie with probability 1/2.

4 Fit a Beta distribution to the NPV for each scenario. Based on the mixture of these Beta distributions using the scenario probabilities, calculate the RAENPV.

Based on the resulting probability distribution of the NPV N one can calculate project characteristics like:

- Expected NPV ($\mathrm{E}N$)
- Value at risk at q ($\mathrm{VaR}_q(N)$)
- Standard deviation ($\sigma(N)$)
- Risk-adjusted ENPV ($\mathrm{E}N - \alpha(-\mathrm{VaR}_p(N))$)

Example 4:
We assume a single scenario. Based on a detailed analysis of the elements contributing to the cash flow the pessimistic and the optimistic estimates of the NPV for this scenario have been evaluated to be –400 and 300 respectively, and the most probable NPV is evaluated to be –100. The NPV is assumed to lie in the interval (-200, 0) with probability 1/2. (All figures are in mill. USD.)

The Beta distribution fit gives an expected NPV equal to –82.5 with standard deviation 134.4. The probability that the NPV is negative becomes 0.72 and $\mathrm{VaR}_{5\%}$ becomes –296. If α = .2, the RAENPV becomes –82.5 – .2 × 296 = –141.7.

3 Evaluation of Project Portfolios

3.1 Project Portfolio Alternatives

We shall consider the situation where we have a series of mutually exclusive portfolio alternatives. If we have n projects, there are in principle up to 2^n portfolio alternatives. In practice the number of alternatives will be much lower. Typically, some projects can be started only if some other projects are also started. Similarly there may be projects that cannot be part of the same portfolio. The planner must specify such restrictions, and a planning tool must facilitate the specification of such restrictions.

3.2 Scenarios

We can specify scenarios for project portfolios in two alternative ways:

- We can specify scenario alternatives across all projects.

- We can specify scenario alternatives for each individual project.

The two ways are in principle equivalent. If we for each project p specify the mutually exclusive scenarios $S_p^1, \mathrm{K}, S_p^{k(p)}$, we can define $\prod k(p)$ scenarios of the form $(S_1^{i_1} \cap S_2^{i_2} \cap \mathrm{L} \cap S_n^{i_n})$ which run across projects. We shall in the sequel assume that the scenarios are defined across the projects.

3.3 Evaluation of a Portfolio of Projects

The evaluation of a portfolio is done by a straightforward generalisation of the approach for a single project. Based on the probability distributions of the NPVs of the individual projects and the correlations between them we can calculate the same characteristics as for a single project such as expected NPV, value-at-risk, standard deviation, and RAENPV.

In general one proceeds as follows:

- Establish a set of mutually exclusive scenario alternatives. Which of these scenarios that occurs should have a significant influence on the NPV of the portfolio. Each scenario is assigned a subjective probability of occurrence.
- For each scenario establish a pessimistic, an optimistic and a most probable realization of each project in the portfolio and calculate the NPV for each of these realizations.
- For each scenario and each project in the portfolio specify an interval around the most probable realisation where it is assumed that the NPV will lie with probability 1/2.
- For each scenario specify a matrix of correlation coefficients between the NPVs of the individual projects in the portfolio.
- For each scenario calculate the expectation and the variance of the NPV of the portfolio, and approximate the distribution of the NPV by a suitable Beta distribution.
- Based on the mix of these Beta distributions over the scenarios, calculate the wanted characteristics of the portfolio's NPV.

3.4 Establishing Covariances

3.4.1 General

Establishing correlation coefficients (or equivalently, covariances) between the NPVs of all pairs of projects in a portfolio is not trivial. CAPM (see [1]) has an elegant solution to this for portfolios of securities which are registered on a stock exchange. Here it is only necessary to establish the covariance between the portfolio and the market portfolio. Unfortunately, the necessary conditions are in general unrealistic, even for portfolios of securities. For project portfolios these conditions cannot be justified at all. However, in this case the number of covariances to be estimated is considerably lower. In a security market the investors can choose between hundreds of securities so that there will be tens of thousands of covariances. A portfolio of investment projects in a company will normally consist of between 5 and 10 projects so that the number of covariances will be less than 50. We shall briefly discuss two approaches to estimation of the covariance between two NPVs, namely *direct estimation* and the *use of independent explanatory variables*.

However, we should make note of the fact that the RAENPV measure is approximately *additive* in the sense that the RAENPV of a union of two projects is approximately equal to the sum of the RAENPV of the individual projects. The expected NPV is obviously additive. The same is true for VaR_q when $q = 0$ %, and the two projects are not totally negatively correlated. So for $q = 1$ %, the approximation may not be too misleading. For preliminary analyses we may thus skip the correlation calculations altogether.

3.4.2 Direct Estimation

Here the planner estimates the covariances, or rather the correlation coefficients, based on experience and sound judgement. In order to facilitate this a planning tool should enable the planner to choose from a selection of preset correlation coefficients as suggested in Table 1.

Example 5

Suppose that we have three projects A, B and C with NPVs NA, NB and NC. We estimate the correlation coefficients for all pairs of NPVs as given in Table 2.

Not every table set up in this manner gives rise to valid correlation matrices. A necessary, but not sufficient, condition is that the matrix is positive semidefinite. Based on the distributions of the NPVs it is also possible to establish lower and upper bounds on the individual correlation coefficients. This, however, is computationally demanding and probably not worth the effort. These bounds will very seldom lie in the interval [–3/4,3/4], and total correlations (±T) will only

Table 1 Correlation codes

Description	Total negative	Large negative	Medium negative	Small negative	None	Small positive	Medium positive	Large positive	Total positive
Code	–T	–L	–M	–S	N	S	M	L	T
Correlation coefficient	–1	–3/4	–1/2	–1/4	0	1/4	1/2	3/4	1

be used when the NPVs are linearly dependent. Then the bounds will be –1 and +1 and thus be without value.

If the planner has set up a correlation matrix which is not positive semidefinite, the planning tool should be able to suggest a modified correlation matrix which is positive semidefinite.

3.4.3 Use of Independent Explanatory Variables

We assume that we can write the NPVs N_1 and N_2 for two investment projects 1 and 2 as functions of F independent random variables V_1, V_2, ..., V_F with known distributions:

$$N_p = N_p(V_1, V_2, ..., V_F) \text{ for } p = 1,2. \qquad (3.4.1)$$

The V-s are explanatory variables which determine e.g. equipment costs, total markets, and market shares. Furthermore, we assume that the functions N_p can be approximated by generalised polynomials:

$$N_p(V_1, V_2, \mathrm{K}, V_F) \approx \sum_{i_1,\mathrm{K},i_F} \alpha^p_{i_1\mathrm{K}i_F} V_1^{i_1}\mathrm{K}V_F^{i_F} \qquad (3.4.2)$$

where i_1, ..., i_F may take a finite number of integer values which are not necessarily positive. We are then able to calculate the covariance between N_1 and N_2:

$$\mathrm{cov}(N_1, N_2) = \mathrm{E}(N_1N_2) - \mathrm{E}N_1\mathrm{E}N_2 \qquad (3.4.3)$$

where

$$\mathrm{E}(N_1N_2) = \sum_{\substack{i_1,\mathrm{K},i_F \\ j_1,\mathrm{K},j_F}} \alpha^1_{i_1\mathrm{K}i_F}\alpha^2_{j_1\mathrm{K}j_F}\mathrm{E}V_1^{i_1+j_1}\mathrm{K}\mathrm{E}V_F^{i_F+j_F} \qquad (3.4.4)$$

and

$$\mathrm{E}N_1\mathrm{E}N_2 = \sum_{\substack{i_1,\mathrm{K},i_F \\ j_1,\mathrm{K},j_F}} \alpha^1_{i_1\mathrm{K}i_F}\alpha^2_{j_1\mathrm{K}j_F}\mathrm{E}V_1^{i_1}\mathrm{E}V_1^{j_1}\mathrm{K}\mathrm{E}V_F^{i_F}\mathrm{E}V_F^{j_F} \qquad (3.4.5)$$

such that

$$\mathrm{cov}(N_1, N_2) = \sum_{\substack{i_1,\mathrm{K},i_F \\ j_1,\mathrm{K},j_F}} \alpha^1_{i_1\mathrm{K}i_F}\alpha^2_{j_1\mathrm{K}j_F} (\mathrm{E}V_1^{i_1+j_1}\mathrm{K}\mathrm{E}V_F^{i_F+j_F} - \mathrm{E}V_1^{i_1}\mathrm{E}V_1^{j_1}\mathrm{K}\mathrm{E}V_F^{i_F}\mathrm{E}V_F^{j_F}) \qquad (3.4.6)$$

We assume furthermore that V_f is normalized to lie between 0 and 1 and is Beta distributed with parameters μ_f and ν_f. Then

$$\mathrm{E}V_f^i = \frac{\Gamma(\mu_f + \nu_f)\Gamma(\mu_f + i)}{\Gamma(\nu_f)\Gamma(\mu_f + \nu_f + i)} \qquad (3.4.7)$$

	N_B	N_C
N_A	M	–L
N_B		–S

Table 2 Specification of correlations

Using (3.4.6) and (3.4.7) we can find an expression for $\mathrm{cov}(N_1, N_2)$. Analogously we may find formulas for var N_p for relevant portfolios p.

Example 6

$N_1 = 10V_1 + V_2$
$N_2 = 2V_1 - 3V_2$

where V_1 is Beta distributed over (0,1) with parameters $\mu_{n_1} = 3$ and $\nu_{n_1} = 6$, whilst V_2 is Beta distributed over (0,1) with parameters $\mu_{n_2} = 4$ and $\nu_{n_2} = 2$, and where V_1 and V_2 are independent.

Then

$$\mathrm{E}V_1 = \frac{6}{3+6} = \frac{2}{3}, \mathrm{E}V_2 = \frac{2}{2+4} = \frac{1}{3}$$

$$\mathrm{var}V_1 = \frac{3\times 6}{(3+6)^2(3+6+1)} = \frac{1}{45},$$
$$\mathrm{var}V_2 = \frac{4\times 2}{(4+2)^2(4+2+1)} = \frac{2}{63},$$

$$\mathrm{cov}(N_1, N_2) = \mathrm{cov}(10V_1 + V_2, 2V_1 - 3V_2)$$
$$= 20\mathrm{var}V_1 - 3\mathrm{var}V_2 = \frac{22}{63}$$

$$\mathrm{E}N_1 = 10\mathrm{E}V_1 + \mathrm{E}V_2 = 7,$$
$$\mathrm{E}N_2 = 2\mathrm{E}V_1 - 3\mathrm{E}V_2 = \frac{1}{3}$$

$$\mathrm{var}N_1 = 100\mathrm{var}V_1 + \mathrm{var}V_2 = \frac{142}{63},$$
$$\mathrm{var}N_2 = 4\mathrm{var}V_1 + 9\mathrm{var}V_2 = \frac{118}{315}.$$

We may approximate the distributions of N_1 and N_2 by Beta distributions. The intervals are determined by:

$a_{N_1} = 0$, $b_{N_1} = 11$, $a_{N_2} = -3$, $b_{N_2} = 2$.

In order to find the parameters we normalize N_1 and N_2:

$$\tilde{N}_1 = \frac{N_1}{11}, \mathrm{E}\tilde{N}_1 = \frac{7}{11}, \mathrm{var}\tilde{N}_1 = \frac{142}{7623},$$

$$\tilde{N}_2 = \frac{N_2 + 3}{5}, \mathrm{E}\tilde{N}_2 = \frac{2}{3}, \mathrm{var}\tilde{N}_2 = \frac{118}{7875}.$$

This gives

$$\mu_{N1} = \frac{(7/11)(1-7/11)^2}{142/7623} - (1-7/11) = 4.15,$$

$$\nu_{N1} = \frac{(7/11)^2(1-7/11)}{142/7623} - 7/11 = 7.27$$

$$\mu_{N2} = \frac{(2/3)(1-2/3)^2}{118/7875} - (1-2/3) = 4.61,$$

$$\nu_{N2} = \frac{(2/3)^2(1-2/3)}{118/7875} - 1/3 = 9.55.$$

4 Extension to Options

4.1 A Single Project

We shall first focus on the analysis of a single project. We consider a Project *A* that we may start at some future time *T*. Between current time T_0 and time *T* we may receive information which can influence our decision whether to launch Project *A* or not. If, for example, this project competes with an alternative Project *B* with launching time T_0, we must consider the value of the option of not starting Project *A* that we will have at time *T*. If the information we receive during the time interval $[T_0,T]$ implies that the probability that the project will be profitable is unacceptably low we will at time *T* exercise the option of not starting the project. This should in general give a higher utility of the project than what we would get if we base the analysis merely on estimates of the cash flow of the project as seen at time T_0. What is required, however, is the probability distribution of the additional information as seen at time T_0.

Here we shall limit ourselves to describing a discrete options model where the additional information that we receive in $[T_0,T]$ relevant for the project in question is an adjustment of the probability of occurrence of the individual scenarios. We thus assume that we do not receive any additional information that is sufficiently significant to warrant a change in the Beta distributions for the individual scenarios. This may seem to be a severe limitation. It is, however, mitigated somewhat by the possibility of splitting a scenario where we would like to change the NPV distribution into two or more subscenarios.

We define λ^m to be the probability at time T_0 that we at time *T* are in *situation m* amongst *M* alternative situations. Situation *m* is characterized by a probability distribution over the different scenarios. The RAENPV for scenario *i* is n^i. All n^i are assumed to be known. So the uncertainty at time *T* is reduced to the question as to which situation that will occur. Let $\underline{n}$ be the RAENPV of the market portfolio. The overall RAENPV *n* for the project becomes

$$n = \sum_m \lambda^m n^m. \tag{4.1.2}$$

4.2 Portfolios

We now proceed to the analysis of portfolios of projects where each project has a starting date some time in the future and where information received before the starting date may indicate that it is not profitable to start the project.

Again we shall limit ourselves to describing a discrete options model where the additional information that we receive before the possible start of a project in the portfolio is an adjustment of the probability of occurrence of the individual scenarios. We enumerate all the projects according to increasing starting time: P_1, P_2, ..., and we want to decide whether to start project P_1. We shall approach this problem using dynamic programming. We let D_p denote a partial portfolio of projects amongst P_1, ..., P_p, and we denote by *m'* a situation at the starting time for project P_{p+1}. We assume that we for all such partial portfolios D_p and all *m'* know the partial portfolio D_p'(*m'*) amongst the projects P_{p+1}, P_{p+2} ... which together with D_p yield the portfolio $D_p \cup D_p$'(*m'*) with highest RAENPV.

We now assume that we have decided on a partial portfolio $D_{p\text{-}1}$ amongst the projects P_1, ..., $P_{p\text{-}1}$, that we are in situation *m*, and that we face the decision whether to start project P_p. We now consider two cases[1)]:

(i) We start project P_p. Then we form the partial portfolio $D_p = D_{p\text{-}1} \cup P_p$. The probability of being in situation *m'* at the starting time for project P_{p+1} is assumed to be known and equal to $\lambda_p^{mm'}$. The portfolio $D_p \cup D_p$'(*m'*) with the highest RAENPV $n(D_p, m')$ in situation *m'* is known, and the RAENPV becomes

$$\sum_{m'} \lambda_p^{mm'} n(D_p, m') \tag{4.2.1}$$

(ii) We do not start project P_p. The portfolio $D_{p\text{-}1} \cup D_p$'(*m'*) with highest RAENPV $n(D_{p\text{-}1}, m')$ in situation *m'* is known, and the RAENPV becomes

$$\sum_{m'} \lambda_p^{mm'} n(D_{p-1}, m') \tag{4.2.2}$$

---

1) *Note that in many situations we cannot choose whether we should start project P_p or not. It may happen that P_p either is not compatible with $D_{p\text{-}1}$ or that P_p is a necessary consequence of $D_{p\text{-}1}$. This contributes to a reduction of the number of possible portfolios and thus the time required for the necessary calculations.*

If (4.2.1) is larger than (4.2.2), we choose to start project P_p. If (4.2.1) is less than or equal to (4.2.2) we choose not to start project P_p. For every partial portfolio D_{p-1} of the projects P_1, ..., P_{p-1}, and every situation m at the start time for project P_p, we thus know the partial portfolio $D_{p-1}'(m)$ of the projects P_p, P_{p+1} ..., which together with D_{p-1} yield the portfolio $D_{p-1} \cup D_{p-1}'(m)$ with highest RAENPV $n(D_{p-1}, m)$.

This then forms the basis for a dynamic programming algorithm for finding the project portfolio with the highest RAENPV. The input we need in addition to what we have discussed earlier is the Markov chain matrices $\Lambda_p = (\lambda_p^{mm'})$ where $\lambda_p^{mm'}$ is the probability of being in situation m' at the starting time of project P_{p+1} given that we are in situation m at the starting time of project P_p.

To get going we start with the project P_n with the latest starting time and then work our way backwards in time:

For every partial portfolio D_{n-1} of projects from P_1, ..., P_{n-1}, and every situation m at the starting time for project P_n we consider two cases:

(i) We start project P_n, form the portfolio $D_n = D_{n-1} \cup P_n$ and calculate the RAENPV $n(D_n)$.

(ii) We do not start project P_n, put $D_n = D_{n-1}$ and calculate the RAENPV $n(D_n)$.

If the RAENPV we calculated in (i) is greater than the RAENPV we calculated in (ii), we choose to start project P_n. If the RAENPV we calculated in (i) is less than or equal to the RAENPV we calculated in (ii), we choose not to start project P_n. For every partial portfolio D_{n-1} from the projects P_1, ..., P_{n-1}, and every situation m at the starting time for project P_n we thus know whether it is profitable or not to start project P_n and can therefore for every situation m establish the extension of the partial portfolio D_{n-1} to a complete portfolio with the highest RAENPV. The set of relevant situations may of course vary from one project starting point to the next.

Example 7

We assume that we have five situations S_1, ..., S_5, and three projects P_1, P_2 and P_3 with startup times $t_1 < t_2 < t_3$. At t_1 we are in situation S_1. We assume that the transitions between the situations are as depicted in Figure 8.

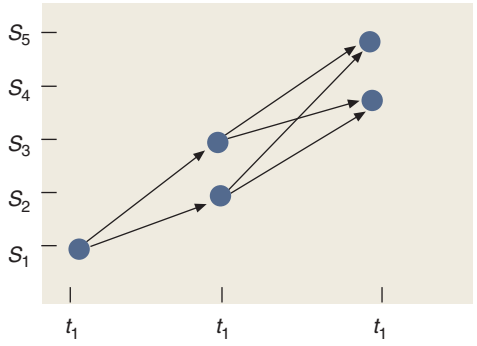


Figure 8 Situation transitions

The transition probabilities are given in Table 3.

Table 3 Transition probabilities

	S_2	S_3	S_4	S_5
S_1	1/2	1/2		
S_2			2/3	1/3
S_3			1/3	2/3

The RANPVs of the three projects in the relevant situations are given in Table 4.

Table 4 RANPVs for the individual projects

	P_1	P_2	P_3
S_1	1		
S_2		–5/4	
S_3		1/4	
S_4			5/2
S_5			–1/2

We assume further that P_3 cannot be started if P_2 is not started. Thus the possible portfolios are Ø, P_1, P_2, $P_1 \cup P_2$, $P_2 \cup P_3$, $P_1 \cup P_2 \cup P_3$. We want to know whether we shall start project P_1 at time t_1. We also make use of the approximate additivity of RAENPV mentioned in 3.4.1. We shall fill in Table 5 where the entries are to be interpreted as RAENPVs disregarding contributions from projects started earlier in time, and optimal decisions are made at later points in time.

From Table 4 we see that we at time t_3 start project P_3 in situation S_4 whilst we do not start project P_3 in situation S_5. Thus the rightmost column in Table 5 is easily established.

If we do not start P_2, the tail RAENPV becomes 0.

If we start P_2 in S_2, the expected tail RAENPV becomes $-5/4 + 2/3 \times 5/2 = 5/12$.

If we start P_2 in S_3, the expected tail RAENPV becomes $1/4 + 1/3 \times 5/2 = 13/12$.

If we do not start P_1, the expected RAENPV becomes $1/2 \times 5/12 + 1/2 \times 13/12 = 3/4$.

If we start P_1, the expected RAENPV becomes 1 + 1/2 × 5/12 + 1/2 × 13/12 = 7/4.

Thus we conclude that project P_1 should be started.

Table 5 RANPVs for the tail portfolios

	P_1	P_2	P_3
S_1	7/4		
S_2		5/12	
S_3		13/12	
S_4			5/2
S_5			0

5 Conclusion

We have argued that it makes more sense to evaluate portfolios based on a utility measure that balances expected net present value at capital cost against value-at-risk, rather than to use a risk-adjusted discount rate. The concepts used are fairly simple, and the approach can readily be implemented in a spreadsheet tool that calculates the total utility of all relevant portfolio candidates and ranks them according to decreasing total utility. We believe that the output of such a tool will be useful for the planners responsible for portfolio analysis and investment recommendations.

Appendix: Risk-Adjusted Expected NPV – Detailed Analysis

A1 The Market Portfolio

We consider a suitable market portfolio *M* and operate with a fixed period length which we for simplicity set equal to one year. At the beginning of the year we buy for one dollar worth of the market portfolio and sell it at the end of the year for *W* dollars. The rate of return *A* is then given by

$$A = \frac{W-1}{1} = W - 1. \tag{A1}$$

The NPV N_r^M with discount factor *r* then becomes

$$N_r^M = \frac{W}{1+r} - 1 = \frac{A-r}{1+r}. \tag{A2}$$

with expected value

$$\mathrm{E}N_r^M = \frac{\mathrm{E}A - r}{1+r}. \tag{A3}$$

We thus have

$$\mathrm{E}A = (1+r)\mathrm{E}N_r^M + r. \tag{A4}$$

Let a_q be the q-point in the distribution of *A*. Then

$$q = \mathrm{P}(A \le a_q) = \mathrm{P}\left(\frac{A-r}{1+r} \le \frac{a_q - r}{1+r}\right) = \mathrm{P}\left(N_r^M \le \frac{a_q - r}{1+r}\right) \tag{A5}$$

such that

$$\mathrm{VaR}_q(N_r^M) = \frac{a_q - r}{1+r}. \tag{A6}$$

We can estimate a_q based on the empirical distribution of *A* by using the q-point in this distribution. Since the return of the market portfolio is a weighted sum of the returns of the individual securities that constitute the portfolio we may alternatively permit ourselves to assume that *A* is normally distributed N(E*A*, σ_A). Then we can establish the following relation

$$q = \mathrm{P}(A \text{ " } a_q) = \mathrm{P}\left(\frac{A - \mathrm{E}A}{\sigma_A} \le \frac{a_q - \mathrm{E}A}{\sigma_A}\right) \tag{A7}$$

which gives

$$\frac{a_q - \mathrm{E}A}{\sigma_A} = -n_q \tag{A8}$$

where $-n_q$ is the q-point in the standardized normal distribution. (A8) can be written as

$$a_q = \mathrm{E}A - n_q\sigma_A. \tag{A9}$$

This gives further

$$\mathrm{VaR}_q(N_r^M) = \frac{\mathrm{E}A - n_q\sigma_A - r}{1+r} = \mathrm{E}N_r^M - \frac{n_q\sigma_A}{1+r} \tag{A10}$$

which gives the following expression for σ_A:

$$\sigma_A = \frac{1+r}{n_q}\left(\mathrm{E}N_r^M - \mathrm{VaR}_q(N_r^M)\right) \tag{A11}$$

The capital market line is given by:

$$\mathrm{E}A = f + \gamma\sigma_A \tag{A12}$$

where *f* is the risk free interest rate. γ reflects how the market portfolio balances expected return against risk[2)].

[2)] *A new investment possibility that lies above the capital market line will be included in the market portfolio and thus alter it. The investment will therefore be considered profitable in a perfect capital market. (An investment possibility below the capital market line could also be profitably included in the market portfolio, but this is considerably more difficult to ascertain.)*

From (A9) we see that

$$\mathrm{E}A = a_q + n_q\sigma_A \tag{A13}$$

We assume that the market portfolio is not risk free, in other words that $a_q < f$. (A12) og (A13) then imply that the coefficient of substitution γ between return and risk is less than n_q. (A12) gives the following expression for γ:

$$\gamma = \frac{\mathrm{E}A - f}{\sigma_A}. \tag{A14}$$

By substituting (A4) and (A11) into (A12) we get an analogous 'capital market line' which ties together $\mathrm{E}N_r^M$ and $\mathrm{VaR}_q(N_r^M)$:

$$(1+r)\mathrm{E}N_r^M + r = f + \gamma\frac{1+r}{n_q}\left(\mathrm{E}N_r^m - \mathrm{VaR}_q(N_r^M)\right) \tag{A15}$$

or

$$\mathrm{E}N_r^M = \frac{(f-r)n_q}{(1+r)(n_q-\gamma)} + \alpha\left[-\mathrm{VaR}_q(N_r^M)\right] \tag{A16}$$

where $\alpha = \dfrac{\gamma}{n_q - \gamma}$. This is meaningful since $\gamma < n_q$.

α represents the balancing of expected NPV against risk. We observe that α is independent of r.

A2 The Profitability of a Project

We propose to measure the profitability of a project by balancing the expected NPV against the value-at-risk where α is used as the weighting coefficient. Thus we choose to let the company's willingness to take risk be the same as for the market portfolio. The expected NPV and the value at risk depend on the choice of discount rate. We suggest the use of the company's estimated average capital cost as a value for r. It represents the average return on investment within the company and is thus a good estimate of what the company may obtain through an alternative use of the money. (We are not dogmatic here. Other considerations may cause the choice of another discount factor.) Thus the profitability of a project P is measured by the risk-adjusted expected NPV (RAENPV) given by

$$\mathrm{E}N_r - \alpha\,(-\mathrm{VaR}_q(N_r)). \tag{A17}$$

The definition (2.3.2) together with (A16) gives that the RAENPV of the market portfolio is

$$\frac{f-r}{(1+r)(1-\gamma/n_q)}. \tag{A18}$$

References

1 Copeland, T E, Weston, J F. *Financial Theory and Corporate Policy*. Boston, Mass, Addison Wesley, 1992.

2 Mina, J, Xiao, J Y. *Return to Risk Metrics: The Evolution of a Standard*. New York, RiskMetrics, 2001.

Forecasting – An Important Factor for Network Planning

KJELL STORDAHL

The paper gives an overview of forecasts and forecast methodology used for network planning. Specific attention is given to how forecasts are applied for development of strategies and planning. An extensive list of references is annexed for more detailed studies.

Kjell Stordahl (58) received his M.S. in statistics from Oslo University in 1972. He worked with Telenor R&D for 15 years and with Telenor Networks for 15 years, mainly as manager of Planning Department Region Oslo and then Market analysis. Since 1992 he has participated in various techno-economic EU projects (TITAN, OPTIMUM, TERA, TONIC) analysing rollout of different broadband technologies. Kjell Stordahl has been responsible for working packages on broadband demand, forecasting and risk analysis in these projects. He has published more than 140 papers in international journals and conferences.

kjell.stordahl@telenor.com

1 Is Forecasting Necessary for Network Planning?

Network planning is an important activity for the operators. In order to utilise the resources and investment means in the best possible way, it is of crucial importance to have insight in future telecommunication demand. A professional forecasting process will show the expected evolution of the telecommunication demand.

Questions to be solved in the network planning process are:

- Choice of technology and network components
- Design of network structure
- Routing principles and redundancy in the network
- Dimensioning of nodes and routes in the network
- Timing of network expansion
- Implementation of additional functionality
- Introduction of new and enhanced services
- Integration of functionality on various OSI levels
- Replacement strategies for old network components/old technology
- Long-term strategy planning for network evolution

Important forecasts to support the planning process are:

- Forecasts for service demand
- Forecasts for enhanced services demand
- Identification of demand for new services
- Subscription/access forecasts for the services
- Traffic volume forecasts for services and applications
- Busy hour traffic forecasts for services and applications
- Forecasts based on market segmentation
- Forecasts taking into account competition and market share
- Forecasts separated into national level, regional level and local level
- Forecasts allocated to the various networks
- Forecasts allocated to transport network, regional network and access network

Specific forecasts are defined by combining the items on the list.

For example, Telenor SHDSL subscription forecasts for the business market for one specific local access area are based on the following:

- Estimation of potential DSL accesses for different market segments, especially type of industry (SIC code) and size in the specific area
- Forecasts of the DSL business penetration in the area
- Forecasts of the evolution of demand of symmetric business access demand in the specific area
- Prediction of expected market share in the specific area

There are different players in the telecommunication market. Because of open network provisioning additional operators have been established in the market. The access network is available based on Local Loop UnBundling and several operators hire the copper pair from the incumbents. In addition wholesale has been introduced to create a more open and competitive telecommunication market with new service providers. The telecommunication market is more complex and there are even more need for forecast modelling taking into account the new environments.

There are always economic risks related to network planning. The forecasting process should generate reduced risks by identifying the new and enhanced applications and make predictions of access and traffic demand generated by the new and traditional services in environments where the competition and evolution of expected market share are described.

Of course there are significant uncertainties related to the forecasting process. Hence, an important part of the process is to describe the uncertainty and try to incorporate the uncertainty evaluations into the network planning process.

2 Network Evolution and the Forecasts

The circuit switched telephone network has evolved through certain milestones from being an analogue network to a network consisting of only digital exchanges. The next step was the

introduction of ISDN. Now, Telenor has the highest ISDN penetration in the world. During the last few years the IP based Internet has evolved significantly. In Norway the Internet penetration is more than 60 %. Before the residential broadband take off, the main capacity in the transport network was made up by leased lines and PSTN/ISDN traffic. A significant part of PSTN/ISDN traffic has been narrowband Internet traffic.

Now we see conversion of narrowband Internet traffic to broadband ADSL traffic. New services and applications are established and new demand generated. The broadband platform will be enhanced either through ADSL2+ or VDSL, which in turn include broadband entertainment services in the fixed network. In parallel HFC networks from the cable operators UPC and Telenor Avidi have been deployed and Fixed broadband radio systems like LMDS are established.

Today there is a significant conversion of voice traffic from the PSTN/ISDN to mobile networks. The mobile platform is evolving from GSM to GPRS, possibly to EDGE and to UMTS. In addition WLAN hot spots are deployed.

The business market has evolved through data line switched and data packet switched network without considerable success, to leased lines, Internet and DSL. Now, new services like symmetric DSL (SHDSL), fast- and GB Ethernet are introduced.

The market, the established services, the enhanced and new services and applications create telecommunication demand. Market forecasts are important and necessary for taking the right decisions for network evolution. The forecasts help to get the right timing from introduction of new network platforms and services. More detailed description of the forecasts are shown in [1, 3, 5, 7-8, 11-13, 26-27, 30-31, 34-35, 37, 51, 55, 57, 63, 65, 74, 82-83, 85-87].

3 Techno-Economic Tool for Strategic Evaluations of New Technology and Network Structures

To be able to make the right decision for rolling out a new network platform, a comprehensive techno-economic analysis has to be carried out.

Within the European programs RACE, ACTS and IST, the projects RACE 2087/TITAN, AC 226/OPTIMUM, AC364/TERA and IST-2000-25172 TONIC have developed a methodology and a tool for calculation of the overall financial budget of any access architecture. The tool handles the discounted system costs, operations, maintenance costs, life cycle costs, net present value (NPV) and internal rate of return (IRR). The tool has the ability to combine low level, detailed network parameters of significant strategic relevance with high level, overall strategic parameters for performing evaluation of various network architectures [40–46, 58–59, 66, 71–72].

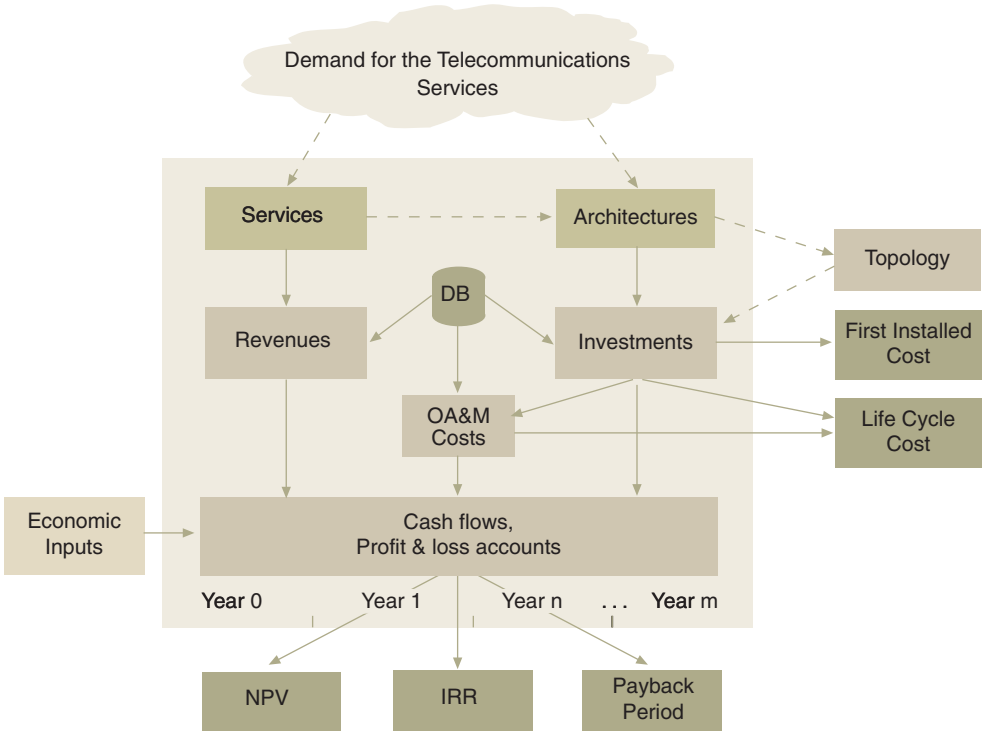


Figure 1 TONIC techno-economic tool for investment projects analysis

Figure 1 shows a techno-economic tool used in different international EU projects.

The figure shows that telecommunication demand is an input to the tool and to the techno-economic calculations. The TONIC tool is widely used to analyse economic consequences by implementing new network platforms. Important parts of the tool are:

- Service definitions
- Subscription and traffic forecasts
- Service tariff predictions
- Revenue model
- A topology model mapping geographic areas with given penetrations into the tool
- Network component cost data base including more than 300 network components
- Network component cost prediction model
- Investment model
- Operation and maintenance model
- Model for economic calculations
- Risk analysis model

The following steps are needed in the techno-economic evaluations of the network solutions:

The services to be provided must be specified. The market penetration of these services over the study period will be defined. The services have associated tariffs, i.e. the part of the tariff that is attributed to the network under study. From the combination of yearly market penetration and yearly tariff information TONIC calculates the revenues for each year for the selected service set.

Next, the architecture scenarios to provide the selected service set must be defined. This needs network design expertise and is mostly outside of the framework of TONIC methodology. However, TONIC includes several geometric models, which facilitate the network design by automatically calculating lengths for cables and ducting. These geometric models are optional parts of the methodology and TONIC can be used without them. The result of an architecture scenario definition is a so-called shopping list. This list indicates the volumes of all network cost elements (equipment, cables, cabinets, ducting, installation etc.) for each year of the study period and the location of these network components in different flexibility points and link levels.

The costs of the network components are calculated using an integrated cost database developed within the TONIC project, containing data gathered from many European sources. Architecture scenarios together with the cost database give investments for each year.

The OA&M costs are divided into different components like cost of repair parts including civil work and operations and administration costs. Typically the OA&M costs are being driven by services, say by number of customers and number of critical network elements.

Investment costs together with the OA&M costs give the life-cycle cost for the selected architecture scenario. Finally, by combining service revenues, investments, operating costs and general economic inputs (e.g. discount rate, tax rate) OPTIMUM, cash flows and other economic factors (NPV, IRR, Payback period etc) are calculated.

4 Strategic Evaluations of New Technology and Network Structures

The possibility to introduce new services, enhanced services, new applications, additional traffic growth and generation of additional revenue is important for network operators. Strategies for introduction of new technologies and network platforms open the possibilities for generation of additional revenue. Some examples of introduction of new technology during the last 20 years are:

- Establishment of NMT
- Deployment of optical fibre and fibre technology in the network
- Digitalisation of the PSTN
- Introduction of ISDN
- Establishment of GSM
- Introduction of Internet
- Establishment of ATM network
- Establishment of IP networks
- Establishment of SDH transmission technology
- Introduction of ADSL

The next step is enhancement of the ADSL platform to ADSL2+ or VDSL and for the mobile network UMTS, WLAN and 4G. Important input for making the right decisions of introducing new technology has been demand forecasts for new and enhanced services. The forecasts and tariff predictions give revenue forecasts, which are used together with techno-economic analysis to support decisions of introduction of new technology.

The traffic forecasts are also fundamental for dimensioning the networks and establishing optimal network structures.

Techno-economic evaluations of new network technology is analysed in [2, 4, 14–15, 18–25, 28–29, 33, 36, 38–39, 48, 50, 52–54, 60, 68, 70, 76, 81, 84, 88–89].

5 Subscriber/Access Forecasting

Network planning of the access network depends heavily on subscriber forecasts. Until recently, substantial investments in the local telephone exchanges have been based on forecasts of the subscriber growth in the local area. The subscriber forecasts indicated time for expansion of the telephone exchange. In addition the subscriber forecasts were used together with queuing models to estimate the busy hour traffic volume in the exchange. When the data traffic was limited, Erlang's dimensioning rule was used for traffic dimensioning taking into account the distribution of residential and business customers. 30 years ago the business customer traffic was dominating and the busy hour was around lunch. Now, traffic from residential customers is dominating on the main part of the Norwegian exchanges. During the last few years the busy hour has been moved from after the main television news programme to the 21.00–22.00 period, mainly because of the Internet traffic.

ISDN was introduced in the Norwegian market 10 years ago. There have been significant substitution effects between ISDN and PSTN in this period. The ISDN forecasts consist of: 2B+D residential, 2B+D business and 30B+D business. Also multiple ISDN line forecasts will be implemented.

The ISDN forecasts have been important for planning and providing new network components.

Because of competition and substitution effects between services, the penetration of PSTN and ISDN have saturated the Norwegian network. Now, the new PSTN and ISDN forecasts reflect a smaller number of subscribers. There are two main reasons for the decrease:

- Some residential customers substitute their main telephone subscriptions with mobile subscriptions
- Some residential customers substitute their PSTN/ISDN Internet connections with ADSL.

These forecasts are crucial for network planning. The forecasts show as a function of time the spare capacity in the network regarding traffic capacity and number of connections.

The subscriber forecasts have been important also for restructuring the access network. The forecasts are used for planning and establishing service connection points in the access network and for dimensioning the access lines and fibre rings between the service connection points.

Subscription forecasts have been developed for (narrowband) Internet, ATM, ADSL, SHDSL, SHDSL point-to-point, Fast Ethernet and GB Ethernet and leased lines. The subscription forecasts give valuable information for planning, dimensioning and expanding the network structure.

However, it is important to take into account significant substitution effects between the services. The ISDN and PSTN forecasts are mutually dependent. The ADSL forecasts influence significantly the narrowband Internet forecasts and ISDN and PSTN forecasts. Lost PSTN and ISDN market shares for the incumbent increase the leased line demand forecasts. Subscription forecasts for 2.5G and 3G increase the leased lines forecasts. Increased ADSL subscriber forecasts and demand for LLUB increase the leased line forecasts because of expansion of the traffic capacity between DSLAM and broadband access point, and because of additional traffic in the transport network. Subscriber demand forecasts for SHDSL and SHDSL point-to-point and point-to-multipoint will lower the leased lines forecasts, since the new services are leased lines substitutes with cheaper tariffs and a degraded service quality.

To be able to control the dependencies, forecasting models for the different services are linked together.

6 Traffic Forecasting

An important part of network planning is design of network structures. The PSTN/ISDN consists of the access network including local exchanges, region networks including group exchanges and the long distance network including long distance exchanges and national exchanges. To maintain a high degree of reliability, independent routes are established between local exchanges and group exchanges on the one hand and between group exchanges and long distance exchanges on the other hand. Between the long distance exchanges there is a logical mesh network. However, even when there is a mesh network there is a simpler physical network taking into account the logical mesh network abilities. The physical network is based on deployed fibre rings and SDH transmission equipment.

Network planning designs optimal structure of the network with the nodes (exchanges) and the routes by minimising the costs for a given redundancy. Important factors are the exchange sites, the route length and the capacity on the routes. Specific network planning tools have also been developed to minimise the investments. Different tools have been developed for the access network and for the transport network. The tools can be applied for establishing new network structures or for expanding the network.

The traffic forecasts are important input for designing the network and for dimensioning the exchanges and capacity on the different routes and as input to the network planning tools.

The busy hour traffic forecasts quantify increases/decreases in the traffic. The dimensioning calculations based on queuing theory add some extra capacity to control random variation in the traffic load during the busy hours. When the traffic on a route or in an exchange approaches the capacity, an expansion will be carried out. The expansion of the capacity will be based on the forecasts for a given planning period.

The forecasting procedure described is used not only for the PSTN/ISDN but for other networks like PSDN, Frame relay/ATM, Digital Cross Connect, various IP networks, networks for mobile operators and leased lines. There are different possibilities for making traffic forecasts. One possibility is to measure the traffic and use the traffic measurements as a historic database for the forecasts. One problem is rerouting of the traffic based on new network structures or because of change of interconnection sites in the network. Such changes affect the traffic measurements, which have to be adjusted before the forecasts can be developed. Another possibility is to analyse the total traffic growth in the given network and make the forecasts on the aggregated level.

A detailed description of the traffic forecasting procedure for the transport network is described in [1]. The traffic forecasting procedure for the transport network is rather complex since the network carries business and residential traffic from all the other networks.

7 Capacity Forecasts

Models for making traffic forecasts during the busy hour have been described. However, the capacity needed to carry the traffic is higher. Additional capacity has to be dimensioned taking into account stochastic variations around the mean traffic in the busy hour. Usually Erlang's blocking formula is applied for voice traffic, while Lindberger's approximation is useful for dimensioning the data traffic capacity [90]. The traffic will probably be transported on SDH systems where packet overhead is added. In addition the systems' average load factor is less than the maximum capacity. Finally there will in general be some extra dimensioning since the expansion of the network is planned and deployed stepwise. A planning period is defined as the time between two expansions. The duration of the planning period is found by taking into account the repeating costs by performing the expansion and the unused investment means during part of the period. The optimal planning period is found by minimising the total costs. The estimation of additional capacity depends on the planning process and the technical systems and will vary from one incumbent to another.

8 Forecasting Methodologies

A variety of forecasting models are used for predicting the evolution of the telecommunication market. An important task is collection and analysis of statistics – historical data. A lot of resources are used to maintain rather large customer based systems. An extraction of data from these systems is used to create subscription statistics, which is a base for subscriber forecasts. In addition specific measurement systems are established to extract traffic measurements on specific points in the networks and also on the aggregated level.

The forecasting models take into account historical data. The most common models for *access forecasts* are:

- Diffusion models
- Regression models
- Econometric models
- ARIMA models
- Kalman filter models
- Smoothing models

The access forecasting models have to take into account substitution effects between services as pointed out in chapter 5. Even the competition between the incumbent and the other operators influences not only the access forecasts but also the BOT and leased lines forecasts. Telenor is using a rather comprehensive composite model where the substitution effects between PSTN, ISDN, Leased lines, ADSL, VDSL/ADSL2+, HFC, FTTH, FWA are included. Important factors in the model are the predicted technology coverage for the coming years, the predicted market shares etc.

The most common models for traffic forecasts are:

- Regression models
- Econometric models
- ARIMA models
- Kalman filter models
- Smoothing models

In addition specific forecasting procedures can be used. The traffic is generated by sources and moves in a network from source to a set of sinks. To dimension the network, it is important to know the traffic behaviour from the edge routers or the local exchanges. The traffic streams are mapped in a traffic matrix. Each element in the matrix denotes the traffic between two exchanges. Suppose the number of exchanges is

N. Then an *N*x*N* matrix describes the traffic between all exchanges. Specific forecasting procedures can be used to forecast the traffic in the network or the matrix. The procedures are used when the network structure is stable during a certain period.

The best known traffic matrix forecasting procedure is *Kruithof's method* [103]. The method makes traffic matrix forecasts based on the observed traffic matrix, let us say at time 0, and forecasts of the outgoing traffic and forecasts of the incoming traffic from an exchange, let us say at time *t*. The outgoing traffic corresponds to the row sums in the traffic matrix and the incoming traffic corresponds to the column sums. There will be inconsistency between the row sums at time 0 and the forecast row sums at time *t* and in the same manner for the column sums. To reach consistency, Kruithof's method tries to take into account both the traffic structure at time 0 and the row and column forecasts in the best possible way. Hence Kruithof's method uses an iteration procedure to get a compromise between the traffic matrix structure at time 0 and the forecasts. The iteration procedure upgrades the rows in the traffic matrix to correspond with the row sum forecasts. Then the iteration procedure upgrades the columns in the traffic matrix to correspond to the column forecasts. After some iterations the adjusted traffic matrix elements correspond to the row and column forecasts, which gives a traffic matrix forecast at time *t*.

An *extension of Kruithof's method* can be performed by making point-to-point forecasts for each element in the traffic matrix together with forecasts of the outgoing traffic and incoming traffic from each exchange and then adjust the traffic elements in the matrix based on the exchange forecasts using the same iteration procedure.

A further extension of Kruithof's traffic matrix forecasting procedure has been developed. *The weighted least squares method* is based on the fact that the relative uncertainties in the traffic element forecasts are larger than the row and column sum forecasts. The method calculates the adjusted traffic element forecasts by taking into account the relative uncertainty in the traffic element forecasts and the row and column sum forecasts. The different forecasts are found by weighting the forecasts according to their uncertainty using weighting least square method.

Extended least square method takes also into account the total traffic forecasts for the whole traffic matrix. The total traffic is defined as the sum of the row sums or the sum of the column sums in the matrix, which of course are identical. The method is based on the statistical principle to use as much information as possible. Hence, forecasts of the traffic elements, the row sums, the column sums and the total traffic are used. The model is described as follows:

- C_{ijt} is the traffic between i and j at the time t
- $C_{i.t}$ is the traffic from i at time t
- $C_{.jt}$ is the traffic from j at time t
- $C_{..t}$ is the total traffic at time t

The forecasts at time t is given by C_{ijt}, C_{ijt}, C_{ijt} and $C_{..t}$ respectively. The extended weighted leased square forecasts are denoted by $\{E\}$ and found by solving the following minimisation problem:

The square sum Q is defined by:

$$Q = \Sigma\, \alpha_{ij} \cdot (E_{ijt} - C_{ijt}) + \Sigma\, \beta_i\, (E_{i.t} - C_{i.t}) + \Sigma\, \gamma_j\, (E_{.jt} - C_{.jt}) + \Sigma\, \delta\, (E_{..t} - C_{..t})$$

where $\{\alpha_{ij}\}$, $\{\beta_i\}$, $\{\gamma_j\}$ and δ are given constants. The square sum is minimised given that adjusted forecasts $\{E\}$ are consistent, i.e. satisfy the condition that the row sum is equal to the row sum forecasts of each element and the column sum is equal to the column sum forecasts of each element and the sum of the elements in the whole matrix is equal to the total forecasts. A natural choice of the weights is the inverse of the variance of forecast uncertainty of the forecasts. One way to find estimates of the forecasts' uncertainty is to perform ex-post forecasts and then calculate the mean square error. The solution of the minimisation problem is found by using Lagrange's multiplicator method adding the constraints based on consistency in the adjusted forecasts. By using the method we get $N(N + 3) + 2$ equations when N is the dimension of the traffic matrix. The solution of the system of equations gives the traffic matrix forecasts.

The forecasts methodologies are described in more detail in [91–103].

9 Forecasts Uncertainty and the Risk

There are always uncertainties connected to forecasts. Network planning decisions have to be made based on expectations of future evolution. The uncertainty in the evolution can be classified in the following main groups:

- Uncertainty in market (penetration and market share) forecasts
- Uncertainty in tariff forecasts (predictions)
- Uncertainty in network component cost predictions.

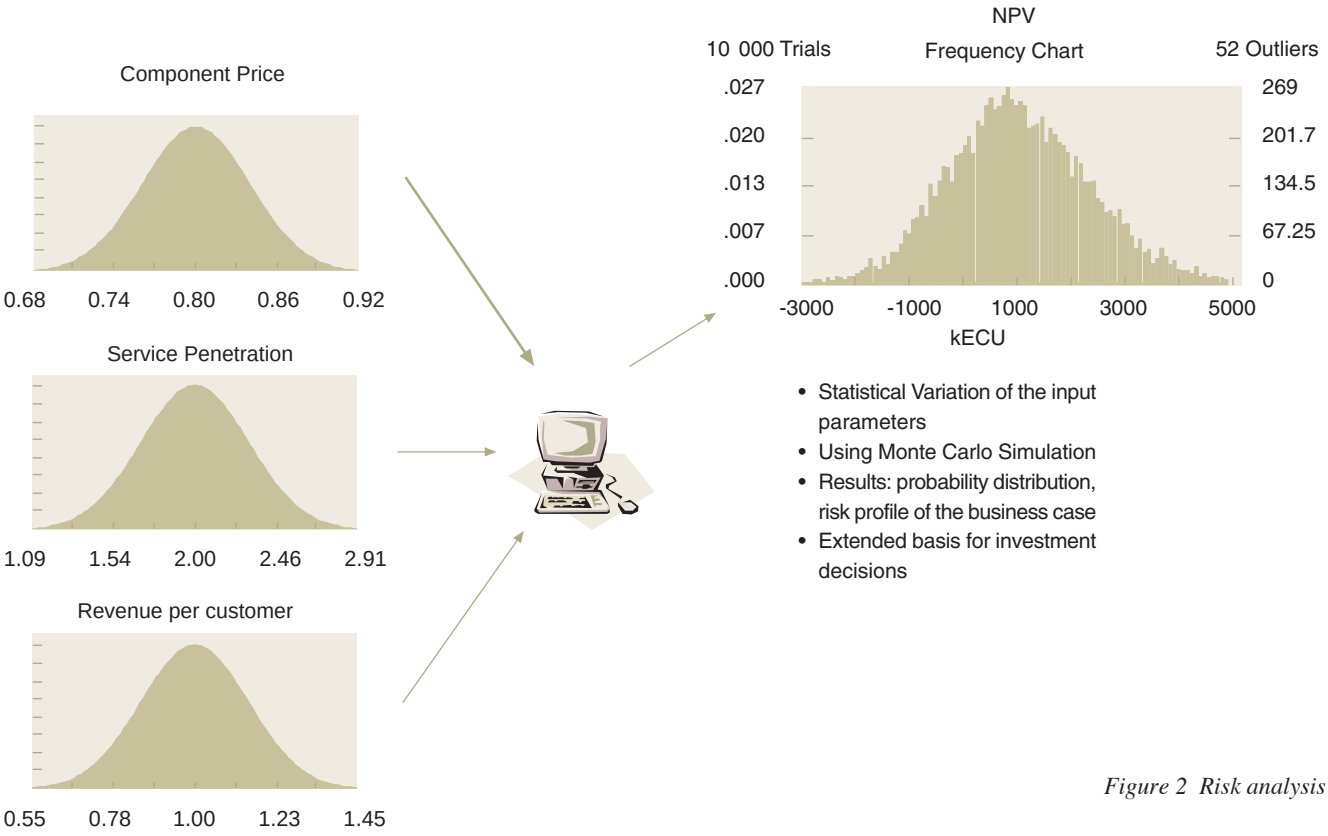


Figure 2 Risk analysis

The effect of uncertainty in the predictions should be quantified to examine the consequences. The first step is to find the most critical variables in the project. The next step is to perform calculations based on variation in the critical variables to identify the consequences.

One possibility is to apply sensitivity analysis. Another and more advanced method is to use the risk analysis. The output in an economic risk analysis is net present value (NPV), internal rate of return (IIR) and pay back period. One of the outputs or all of them simultaneously can be used in risk analysis.

Instead of using the expected forecast, say at time *t*, a probability distribution of the forecast at time *t* has to be used. Similar probability distributions have to be constructed for all critical variables. Usually truncated Normal distributions or Beta distributions are used.

A Monte Carlo simulation with 1000 – 5000 trials is performed. In each trial, a random number is picked from the predefined probability distributions; one for each of the critical variables. The simulation gives as output the cumulative distribution of NPV, IIR and pay back period and the ranking of the critical variables according to their impact. This uncertainty ranking is based on the percentage to the contribution to the variance of the NPV or the rank correlation with the NPV.

A commercial spreadsheet application Crystal Ball has been integrated with the techno-economic tool. A graphical interface in Crystal Ball makes it possible to specify the distribution functions directly from a palette type of menu inside the techno-economic model.

The risk analysis has been applied on a set of different cases studying strategies for rolling out new network technology [6, 9–10, 16–17, 32, 37, 47, 49, 56, 61–64, 67, 69, 73, 75, 77, 78–80].

10 Conclusions

The paper documents that forecasts and forecast methodology are important factors in the network planning process and strategy process for rolling out new technology platforms. Since the forecasts are uncertain, it is recommended to use risk analysis to evaluate the risk generated by forecasts and predictions of other critical variables. For more detailed studies the papers in the reference list should be studied.

References

1 Stordahl, K. Traffic forecasting models for the incumbent based on new drivers in the market. *Telektronikk*, 99 (3/4), 122–127, 2003 (this issue).

2 Elnegaard, N K, Olsen, B T, Stordahl, K. Broadband deployment in rural and non competitive areas: The European perspective. In: *Proc FITCE 2003*, Berlin, Germany, September 1–4, 2003.

3 Stordahl, K, Kalhagen K O. Broadband forecast modelling. Evaluation of methodology and results. *International Symposium on Forecasting 2003*, Merida, Mexico, June 16–19, 2003.

4 Welling, I et al. Techno-Economic Evaluation 3G & WLAN Business Case Feasibility Under Varying Conditions. In: *Proc 10th International Conference on Telecommunications*, Tahiti, French Polynesia, February 23 – March 1, 2003.

5 Stordahl, K, Kalhagen, K O. Broadband Access Forecasts for the European Market. *Telektronikk*, 98 (2/3), 21–32, 2002.

6 Elnegaard, N, Stordahl, K. Deciding on optimal timing of VDSL and ADSL roll-outs for incumbents. *Telektronikk*, 98 (2/3), 63–70, 2002.

7 Kalhagen, K O, Elnegaard, N K. Assessing Broadband Investment Risk Trough Option Theory. *Telektronikk*, 98 (2/3), 51–62, 2002.

8 Stordahl, K. Developing demand forecasts and tariff predictions for different telecom services both for fixed and mobile based on results from the IST TONIC project. *IIR Telecoms Market Forecasting 2002*, Prague, Czech Republic, October 7–10, 2002.

9 Kalhagen, K O, Elnegaard, N K. The Economics and Risks of 3rd Generation Mobile Service Deployment. *Launching Appealing Mobile Services*, September 18–19, London, UK.

10 Elnegaard, N K. How to incorporate the value of flexibility in broadband access network rollout investment projects. *41st European Telecommunications Congress (FITCE)*, Genova, Italy, September 4–7, 2002.

11 Stordahl, K, Kalhagen, K O, Olsen, B T. Broadband technology demand in Europe. *2002 International Communications Forecasting Conference*, San Francisco, USA, June 25–28, 2002.

12 Hjelkrem, C, Stordahl, K, Bøe, J. Forecasting residential broadband demand with limited information. A long term supply and demand model. *2002 International Communications Forecasting Conference*, San Francisco, USA, June 25–28, 2002.

13 Olsen, B T, Stordahl, K. Traffic forecast models for the transport network. In: *Proc Networks 2002*, Munich, Germany, June 23–27, 2002.

14 Monath, T et al. Economics of Ethernet versus ATM-based access networks for broadband IP services. In: *Proc Networks 2002*, Munich, Germany, June 23–27, 2002.

15 Varoutas D et al. Economic viability of 3G Mobile Virtual Network Operators. In: *Proc 3G Wireless 2002*, San Francisco, USA, May 28–31, 2002.

16 Kalhagen, K O, Elnegaard, N K. The Economics and Risks of 3rd Generation Mobile Service Deployment. In: *Proc 3G 2002*, Amsterdam, Netherlands, May 20–23, 2002.

17 Stordahl, K, Elnegaard, N K. Battle between cable operator and incumbent: Optimal ADSL/VDSL rollout for the incumbent. In: *Proc. XIV International Symposium on Services in the Local access, ISSLS 2002*, Seoul, Korea, April 14–18, 2002.

18 Olsen, B T, Kalhagen, K O, Stordahl, K. Provisioning of broadband services in non-competitive areas. In: *Proc. XIV International Symposium on Services in the Local access, ISSLS 2002*, Seoul, Korea, April 14–18, 2002.

19 Elnegaard, N K, Stordahl, K. Deciding on the right timing of VDSL rollouts: A real options approach. In: *Proc. XIV International Symposium on Services in the Local access, ISSLS 2002*, Seoul, Korea, April 14–18, 2002.

20 Monath, T et al. Economics of Fixed Access Networks for broadband IP services (in German). *3. ITG-Fachtagung Netze und Anwendungen*, Duisburg, Germany, February 28 – March 1, 2002.

21 Katsianis, D et al. The financial perspective of the mobile networks in Europe. *IEEE Personal Communications Magazine*, 8 (6), 58–64, 2001.

22 Ims, L A et al. Building a solid business plan for developing fibre in the access network. In: *Proc. Optical Access Networks*, London, UK, Oct 3–5, 2001.

23 Ims, L A et al. End-to-end-solutions for Telcos – is it a Broadcaster's World? In: *Proc. Interactive Convergence*, London, UK, Sep 3–5, 2001.

24 Ims, L A et al. From a large scale VDSL market trial towards commercial services – key issues. In: *Proc. DSLCon Asia*, Hong Kong, China, Aug 13–16, 2001.

25 Elnegaard, N K et al. From VDSL market trials to commercial launch – The key issues on content, services, technical platform and the business case. In: *Proc. XDSL Summit*, Geneva, Switzerland, July 11–13, 2001.

26 Hjelkrem, C. Forecasting with limited information: A study of the Norwegian ISDN access market. *Journal of Business Forecasting*, 20 (3), 2001.

27 Stordahl, K, Elnegaard, N K, Olsen, B T. Broadband access rollout strategies in a competitive environment. In: *Proc. Optical Hybrid Access Network/Full Service Access Network workshop*, Yokohama, Japan, April 3–5, 2001.

28 Ims, L A et al. Market driven deployment of next generation networks – increasing the service footprint by hybrid broadband access. In: *Proc. Hybrid Optical Access Network 2001 (HOAN 2001)*, Yokohama, Japan, April 3–5, 2001.

29 Elnegaard, N K, Stordahl, K. Broadband rollout strategies. In: *Proc. 17th European Network Planning Workshop*, Les Arcs, France, March 19–23, 2001.

30 Stordahl, K, Elnegaard, N K. Broadband market evolution in Europe and the upgrade risks. *2000 International Communications Forecasting Conference*, Seattle, USA, September 26–29, 2000.

31 Stordahl, K, Moe, M, Ims, L A. Broadband market – The driver for network evolution. In: *Proc. Networks 2000*, Toronto, Canada, September 10–16, 2000.

32 Elnegaard, N K, Stordahl, K, Ims, L A. The risks of DSL access network rollouts. In: *Proc. Networks 2000*, Toronto, Canada, September 10–16, 2000.

33 Stordahl, K et al. Optimal roll out strategies for XDSL broadband upgrades of the access network. In: *Proc. International Telecommunication Society 2000*, Buenos Aires, Argentina, July 2–5, 2000.

34 Stordahl, K, Ims, L A, Moe, M. Forecasts of broadband market evolution in Europe. In: *Proc. International Telecommunication Society 2000*, Buenos Aires, Argentina, July 2–5, 2000.

35 Hjelkrem, C. Forecasting access demand when information is scarce : A case study of the Norwegian ISDN market. In: *Proc. International Telecommunication Society 2000*, Buenos Aires, Argentina, July 2–5, 2000.

36 Cerboni, A et al. Evaluation of demand for next generation mobile services. In: *Proc. International Telecommunication Society 2000*, Buenos Aires, Argentina, July 2–5, 2000.

37 Stordahl, K et al. Broadband upgrading and the risk of lost market shares Under Increasing Competition. In: *Proc. ISSLS 2000*, Stockholm, Sweden, June 19–23, 2000.

38 Elnegaard, N K, Stordahl, K, Ims, L A. Roll-Out Strategies for the Copper Access Network: Evolution towards a Full Service Access Network. In: *Proc. ISSLS 2000*, Stockholm, Sweden, June 19–23, 2000.

39 Katsianis, D et al. The Economics of Next Generation Mobile Systems. *ISSLS 2000*, Stockholm, Sweden, June 19–23, 2000.

40 Olsen, B T. Introduction to telecom investments and techno-economics. SAS session. *Techno-economics of multimedia services and networks, ICC 2000*, New Orleans, USA, June 18–22, 2000.

41 Stordahl, K. Broadband market evolution – demand forecasting and tariffs. SAS session. *Techno-economic methodologies and case studies, ICC 2000*, New Orleans, June 18–22, 2000.

42 Ims, L A. Designing case studies. SAS session. *Techno-economics of multimedia services and networks, ICC 2000*, New Orleans, USA, June 18–22, 2000.

43 Olsen, B T. The TERA project – main objectives and results. SAS session. *Techno-economics of multimedia services and networks. ICC 2000*, New Orleans, USA, June 18–22, 2000.

44 Stordahl, K. Overview of risks in multimedia networks. SAS session. *Techno-economic methodologies and case studies, ICC 2000*, New Orleans, June 18–22, 2000.

45 Ims, L A. Case studies on broadband access: HFC, LMDS, VDSL, ADSL. SAS session. *Techno-economics of multimedia services and networks, ICC 2000*, New Orleans, USA, June 18–22, 2000.

46 Olsen, B T. The challenge of predicting cost evolution. SAS session. *Techno-economics of multimedia services and networks, ICC 2000*, New Orleans, USA, June 18–22, 2000.

47 Stordahl, K. Case studies on broadband access: Market risk analysis for fiber roll-out. SAS session. *Techno-economics of multimedia services and networks, ICC 2000*, New Orleans, USA, June 18–22, 2000.

48 Olsen, B T, Stordahl, K, Ims, L A. Evolution of network architectures. In: *Proc. 16th European Network Planning Workshop*, Les Arcs, France, March 20–24, 2000.

49 Stordahl, K et al. Overview of risks in Multimedia Broadband Upgrades. In: *Proc. Globecom '99*, Rio de Janeiro, Brazil, December 5–10, 1999.

50 Ims, L A et al. The economics of broadband infrastructure upgrade investments in various European markets : The EURESCOM view. *Telecom 99 Forum*, Geneva, Switzerland, October 9–17, 1999.

51 Stordahl, K. Identifying the key risks in network development, taking into account uncertainties in market forecasts and competition. In: *Proc. Future proof network planning strategies, Vision in Business*, Berlin, Germany, September 27–28, 1999.

52 Ims, L A. Introduction to broadband access networks. *Telektronikk*, 95 (2/3), 2–22, 1999.

53 Ims, L A. Wireline broadband access networks. *Telektronikk*, 95 (2/3), 73–87, 1999.

54 Ims, L A. Design of access network case studies. *Telektronikk*, 95 (2/3), 251–253, 1999.

55 Stordahl K, Rand, L. Long term forecasts for broadband demand. *Telektronikk*, 95 (2/3), 34–44, 1999.

56 Stordahl, K, Ims, L A, Olsen, B T. Risk methodology for evaluating broadband access network architectures. *Telektronikk*, 95 (2/3), 273–285, 1999.

57 Istad, S, Stordahl, K. Broadband demand survey in the residential and SOHO market in Norway. *Telektronikk*, 95 (2/3), 45–49, 1999.

58 Olsen, B T. OPTIMUM – a techno-economic tool. *Telektronikk*, 95 (2/3), 239–250, 1999.

59 Tahkokorpi, M, Lähteenoja, M. Techno-Economic Guidelines for Telecommunication Networks and Services. *Telektronikk*, 95 (2/3), 236–238, 1999.

60 Ims, L A et al. Towards broadband access in Norway – the view from Telenor. *Telektronikk*, 95 (2/3), 191–201, 1999.

61 Elnegaard, N K, Ims, L A, Stordahl, K. Techno-economic risk assessment of PNO access network evolutionary paths. *Telektronikk*, 95 (2/3), 286–290, 1999.

62 Stordahl, K et al. Risk methodology for multimedia projects assessments. In: *Proc. ECMAST 1999*, Madrid, Spain, May 26–28, 1999.

63 Stordahl, K. Evaluation of key risks in broadband network development taking into account uncertainties in market forecasts and competition. In: *Proc. Future Network Planning Strategies, Vision in Business*, London, March 2–4, 1999.

64 Elnegaard, N K, Stordahl, K. Technology choice risk analysis and its impact on your business development and market positioning. In: *Proc. Future Proof Network Planning Strategies, Vision in Business*, London, UK, March 2–4, 1999.

65 Hjelkrem, C. Forecasting the demand for ISDN accesses when the information is scarce : A case study from the Norwegian market. *International Business Forecasting 99*, Las Vegas, USA, February 26, 1999.

66 Budry, L et al. The economics of broadband service introduction : Area specific investments for broadband upgrade. In: *Proc. TRIBAN*, Bern, Switzerland, November 17–19, 1998.

67 Stordahl, K et al. Broadband access network competition – analysis of technology and market risks. In: *Proc. Globecom '98*, Sydney, November 8–12, 1998.

68 Ims, L A et al. Economics of broadband access network upgrade strategies: The European perspective. In: *Proc. Globecom '98 Access Networks Mini Conference*, Sydney, November 8–12, 1998.

69 Stordahl, K et al. Evaluating broadband strategies in a competitive market using risk analysis. In: *Proc. Networks 98*, Sorrento, October 18–23, 1998.

70 Ims, L A, Myhre, D, Olsen, B T. Investment costs of broadband capacity upgrade strategies in residential areas. In: *Proc. Networks 98*, Sorrento, Italy, October 18–23, 1998.

71 Lähteenoja, M et al. Broadband access networks: techno-economic methodology and tool application results. *IEEE symposium on planning and design of broadband Networks*, Montebello, Quebec, October 8–11, 1998.

72 Ims, L A (ed). *Broadband Access networks : Introduction Strategies and Techno-economic Evaluation*. London, Chapman & Hall, 1998. (ISBN 0 412 82820 0)

73 Stordahl, K, Ims, L A, Olsen, B T. Risk analysis of residential broadband upgrades based on market evolution and competition. In: *Proc. Supercom ICC '98*, Atlanta, June 7–11, 1998.

74 Stordahl, K. Forecasting long term demand for broadband services. In: *Proc. Market forecasts in the telecoms industry*, Institute of International Research (IIR), Hong Kong, May 25–27, 1998.

75 Olsen, B T et al. Multimedia business cases in a deregulated environment : Opportunities and risks. *ISSLS 1998*, Venice, March 22–27, 1998.

76 Ims, L A et al. Key factors influencing investment strategies of broadband access network upgrades. In: *Proc. ISSLS '98*, Venice, Italy, March 22–27, 1998.

77 Stordahl, K. Evaluating broadband upgrade strategies in a competitive market using risk analysis. Institute of International Research (IIR). In: *Proc. Evaluating Broadband Service Strategies*, London, UK, February 18–20, 1998.

78 Ims, L A, Stordahl, K, Olsen, B T. Risk analysis of residential broadband upgrade in a competitive environment. *IEEE Communication Magazine*, June, 1997.

79 Ims, L A, Olsen, B T, Myhre, D. Economics of residential broadband access network technologies and strategies. *IEEE Networks*, 11 (1), 58–64, 1997.

80 Stordahl, K. Risk assessments of broadband upgrade strategies. In: *Proc. Broadband strategies – the battle for the customer*, IBC conference, London, December 4–5, 1997.

81 Ims, L A et al. Evolution of technologies and architectures to a full services network. In: *Proc. Broadband strategies – the battle for the customer*, IBC conference London, December 4–5, 1997.

82 Stordahl, K. Forecasting long term demand for broadband services and their influence on broadband evolution. Institute of International Research (IIR). In: *Proc. Market Forecasting in the Telecoms Industry*, London, UK, December 1–4, 1997.

83 Stordahl, K, Ims, L A, Olsen, B T. Forecasts and risk analysis of PNO and Cable operators by introducing broadband upgrades in the access network. *1997 International Communication Forecasting Conference*, San Francisco, June 24–27, 1997.

84 Ims, L A et al. Key factors influencing the overall costs and project values of broadband access network upgrades. *NOC '97*, Antwerp, Belgium, June 17–20, 1997.

85 Stordahl, K. Forecasting long term demand for broadband services in the residential market: Devising a method for developing forecasts for new services. Institute of International Research (IIR). In: *Proc. Market Forecasting in the Telecoms Industry*, London, UK, May 19–21, 1997.

86 Hjelkrem, C. Forecasting models for ISDN accesses in the Norwegian market and their influences on the investments strategies. Institute of International Research (IIR). In: *Proc. Market Forecasting in the Telecoms Industry*, London, UK, May 19–21, 1997.

87 Stordahl, K. Techno-economic analysis, market forecasts and risk analysis of broadband access network upgrading. In: *Proc. 13th European Network Planning Workshop*, Les Arcs, France, March 9–15, 1997.

88 Olsen, B T. Access Network Evolution. In: *Proc. 13th European Network Planning Workshop*, Les Arcs, France, March 9–15, 1997.

89 Olsen, B T et al. Techno-economic evaluation of narrowband and broadband access network alternatives and evolution scenario assessment. *IEEE Journal of Selected Areas in Communications*, 14 (8), 1996.

90 Lindberger, K. Analytical methods for the traffical problems with statistical multiplexing in ATM network. In: *Proc. 13th International Teletraffic Congress*, Copenhagen, June 19–26, 1991.

91 Stordahl, K. Methods for traffic matrix forecasting. In: *Proc. 12th International Teletraffic Congress*, Torino, Italy, June 1–8, 1988.

92 Stordahl, K. Extended weighted minimum squared method for forecasting traffic matrices. In: *Proc. VII Nordic Teletraffic Seminar*, Lund, Sweden, August 25–27, 1987.

93 Stordahl, K. Forecasting models for network planning. In: *Proc. VI Nordic Teletraffic Seminar*, Copenhagen, Denmark, August 28–30, 1986.

94 Stordahl, K, Damsleth, E. Traffic forecasting methods for network planning. In: *Proc. 6th International Conference on Forecasting and analysis for business in the information age*, Tokyo, Japan, August 1986.

95 ITU. *Procedures for traffic matrix forecasting*. Geneva, ITU, 1986. CCITT Com II 50-E.

96 Stordahl, K. Prognoser som grunnlag for planlegging. ("Forecasts as a base for planning"). *Telektronikk*, 82 (3), 1986.

97 Stordahl, K, Holden, L. Traffic forecasting models based on top down and bottom up procedures. In: *Proc. 11th International Teletrafic Congress*, Kyoto, Japan, September 1985.

98 Stordahl, K. *Routing algorithms based on traffic forecast modelling*. Time series analyses – Theory and practice. O D Anderson (ed.). Amsterdam, North Holland Publishing Company, 1985.

99 Stordahl, K. Forecasting models for traffic in the future broadband network. In: *Proc. V Nordic Teletraffic Seminar*, Trondheim, Norway, June 5–7, 1984.

100 ITU. *Forecasting of traffic*. Geneva, ITU, 1984. CCITT Red Book Rec E.506.

101 Stordahl, K. Routing algorithms based on traffic forecast modelling. *Time series analyses – Theory and practice*. Toronto, Canada, August 18–21, 1983.

102 Stordahl, K. Centralized routing based on forecasts of the telephone traffic. In: *Proc. 10th International Teletraffic Congress*, Montreal, Canada, June 1983.

103 Kruithof, J. Telefoonverkeerrsrekening. *De Ingenieur*, 52 (8), 1937.

Traffic Forecasting Models for the Incumbent Based on New Drivers in the Market

KJELL STORDAHL

Kjell Stordahl (58) received his M.S. in statistics from Oslo University in 1972. He worked with Telenor R&D for 15 years and with Telenor Networks for 15 years, mainly as manager of Planning Department Region Oslo and then Market analysis. Since 1992 he has participated in various techno-economic EU projects (TITAN, OPTIMUM, TERA, TONIC) analysing rollout of different broadband technologies. Kjell Stordahl has been responsible for working packages on broadband demand, forecasting and risk analysis in these projects. He has published more than 140 papers in international journals and conferences.

kjell.stordahl@telenor.com

1 Introduction

The circuit switched voice traffic has traditionally been a significant part of the traffic in the transport network. However, during recent years Leased lines ordered by different operators and also the Leased lines ordered from the business market have expanded the transport network capacity. A new capacity wave has also started: the data traffic moving from narrowband to broadband. The data traffic is increasing exponentially and will for some years be the dominating traffic in the transport network. Important traffic drivers are broadband applications carried by HFC, ADSL, VDSL, LMDS, UMTS and WLAN.

This paper analyses the traffic and capacity evolution of the transport network of an incumbent operator having the possibility to integrate different type of traffic into the network. A traffic volume indicator is developed for traffic increase in the transport network. The access forecast modelling has been developed based on parts of the results from the projects ACT 384 TERA and IST-2000-25172 project TONIC [1–16].

2 Market Segments

Services

Traffic from the services is transported on different network platforms or on leased lines. Important services for the transport network are: PSTN/ISDN, Internet, Leased lines, PSDN (packet switched data network), Frame relay, ATM, IP Virtual private network (VPN), ADSL/SDSL, VDSL/LMDS, Fast Ethernet, Gigabit Ethernet, Lamda wavelength, Dynamic bandwidth allocation.

Market Segments

An incumbent operator leases transport capacity to other operators. In addition the incumbent offers transport capacity to the residential and the business market. The incumbent operator offers transport capacity either via own service provider or as wholesale. A segmentation of the market will be:

Residential market: PSTN, ISDN, Internet, ADSL, VDSL, LMDS, HFC.

Business market: PSTN, ISDN, IP VPN, Internet, PSDN, Frame Relay, ATM, ADSL, SDSL, VDSL, LMDS, Leased lines, Fast Ethernet, Gigabit Ethernet, Lamda wavelength, Dynamic bandwidth allocation.

Operators: ADSL, SDSL, VDSL, LMDS, Leased lines, Fast Ethernet, Gigabit Ethernet IP-VPN, Lamda wavelength, Dynamic bandwidth allocation. (mobile operators, ISPs, other operators).

Network Platforms

Relevant network platforms are: PSTN/ISDN, PSDN, Frame Relay/ATM, Digital Cross Connect, Various IP Networks including IP networks for mobile operators, Leased lines. PSTN/ISDN is a circuit switched network. Leased lines have no concentration effect at all, while the other network platforms also have packet switched concentration.

3 Traffic From the Residential Market

The residential market generates different types of traffic: Voice traffic, Dialled Internet traffic, ADSL traffic, VDSL/LMDS traffic.

The *voice traffic* has nearly reached saturation. During the next years, the circuit switched voice traffic will be rather stable before parts of the circuit switched voice traffic are substituted by IP voice. The *dialled Internet traffic* will reach maximum within a few years. Then the dialled Internet traffic will continuously be substituted by broadband traffic. A battle has already started between broadband operators to capture parts of the broadband market.

Broadband access forecasts have been developed in the IST-2000-25172 project TONIC. Figure 1 shows the market share evolution of ADSL, VDSL, FWA (Fixed wireless broadband access) and HFC/cable modem for West European countries.

Figure 2 shows total broadband penetration. The total broadband penetration forecasts are adjusted compared to the forecasts developed in the TONIC projects. A combination of the figures gives the broadband penetration for each technology.

Now, the question is which traffic is carried in the incumbent's transport network. Usually the cable TV/HFC traffic is carried outside the transport network, while a part (market share) of the

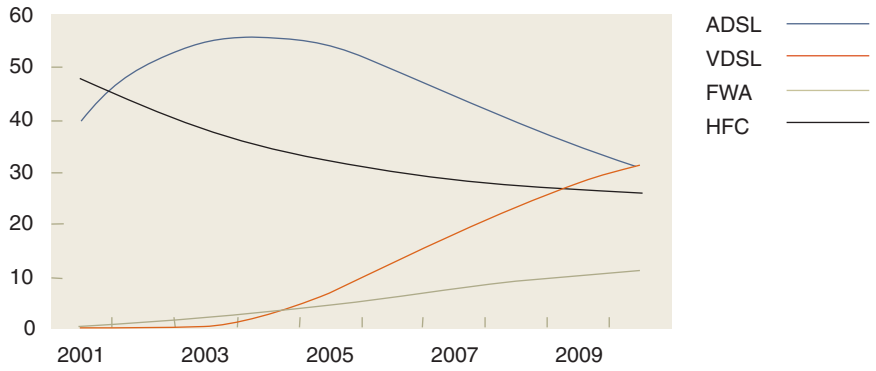


Figure 1 Market share distribution and prediction of ADSL, VDSL, FWA and HFC (cable modem) for West European countries

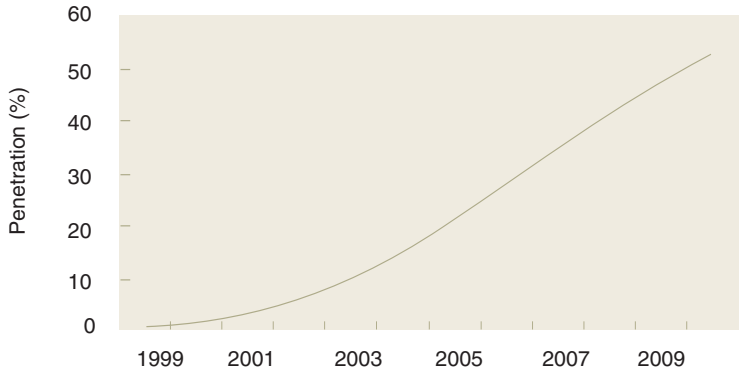


Figure 2 Broadband penetration forecasts for the European residential market

penetration of each of the other technologies will be carried in the incumbent's transport network.

A traffic volume forecast indicator, $V_R(t)$, for the residential busy hour traffic into the transport network is given by:

$$V_R(t) = N_t \Sigma_{i=1,2,3,4,5}\, b_{it}\, u_{it}\, A_{it}\, C_{it}\, M_{it}\, p_{it} \qquad (1)$$

where

$i = 1$ denotes voice traffic
$i = 2$ denotes Dialled Internet
$i = 3$ denotes ADSL
$i = 4$ denotes VDSL
$i = 5$ denotes FWA
N_t is the number of households in year t
b_{it} is busy hour concentration factor for technology i in year t
u_{it} is packet switching concentration factor for technology i in year t
A_{it} is the access capacity utilisation for technology i in year t
C_{it} is mean downstream access capacity for technology i in year t
M_{it} is incumbent's access market share for technology i in year t
p_{it} is the access penetration forecasts (%) for technology i in year t

Market Share and Access Penetration

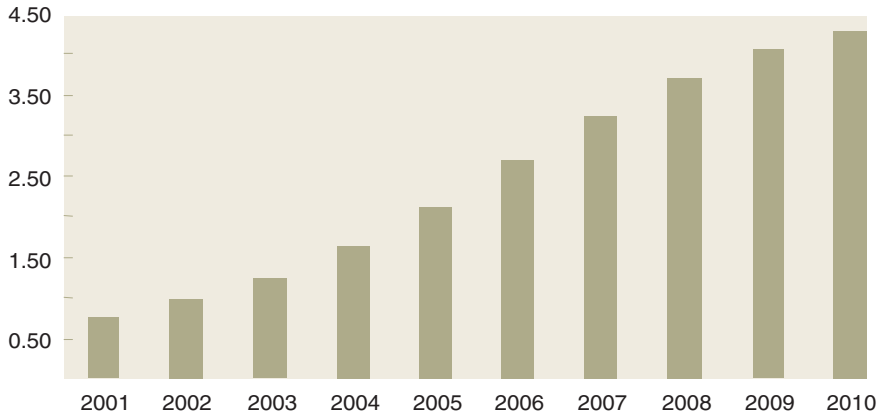
The factor $N_t\, M_{it}\, p_{it}$ represents a forecast of number of households connected to the incumbent's transport network in year t using technology i. Suppose the incumbent operator has 40 % of the ADSL market share and expects to be in the same position the next years, then $M_{3t} = 0.40$ for $t = 2001, 2002, \ldots$.

Mean Downstream Access Capacity and the Technology Evolution

The downstream access capacity, C_{1t}, for voice traffic is 64 kbs and will be the same the next years. However, the downstream access capacity C_{3t} for ADSL changes. Figure 3 shows how the mean downstream access capacity increases. Now, the operators offer a set of different access capacities. There will be an evolution from low access capacities to higher access capacities especially because of new and enhanced applications. In addition, new functionality such as bandwidth on demand will be introduced by the operators.

So far, ADSL based on reasonable low downstream capacity is introduced. The next step will be to enhance the ADSL capacity in the range 6 – 8 Mbs. However, the capacity size offered will be dependent on the twisted copper pair length. The capacity offered will decrease with the copper pair length.

Figure 3 Evolution of predicted average downstream capacity in Mbs (Tonic 2002)



An extension to the VDSL platform will increase the offered capacity further. New services like VoD and peer-to-peer applications based on downloading and exchange of films will increase the capacity demand. The access capacity for VDSL will be in the range 15 – 24 Mbs depending on the copper pair length.

ADSL2+ is a supplement to VDSL and ADSL. While VDSL covers subscribers up to 1.5 km from the exchange or the last fibre drop, ADSL2+ has the possibility to offer 10 Mbs up to 2.0 km and probably within a few years 2.5 km distance from the exchange/last fibre drop.

ADSL2+ and VDSL have enough capacity to offer Internet and digital TV while also offering the possibility to use interactive TV. The number of independent TV streams will be dependent on the selected solutions. It will also be possible to transfer traditional TV programs on the copper pairs like the ones seen in cable TV networks. However, only part of this capacity will be individual, while the capacity for dedicated TV channels will be a common resource for all households in the transport network.

Dynamic Spectrum Management (DSM) will enhance the copper line capacity further. The methodology is based on dynamic regulation of the frequencies on the copper line to reduce the noise and cross talk.

The market share evolution for different ADSL access capacities is described in the IST-2000-25172 project TONIC. Mean access capacity is calculated according to the distribution of different access capacities for each year. The results are shown in Figure 3.

The capacity predictions are based on the assumption of flat rate principle for the first years. However, there may be a delay in the demand for capacity if and when traffic charges are introduced by the operators.

Access Capacity Utilisation

A broadband customer is not utilising the maximum capacity all the time. The access capacity utilisation factor indicates average capacity use taking into account the proportion of time during the conversation for downloading and the proportion of time for uploading. The factor also reflects the degree of using the specified bandwidth.

Busy Hour Concentration

The busy hour concentration effect is well known. Usually about 10 % ($b_{1t} = 0.1$) of the customers make phone calls in the busy hour. Traditionally Erlang's blocking formula (assuming exponential interarrival time and holding time) is used for dimensioning. The busy hour concentration factor is increasing because of the Internet. For broadband connections the busy hour concentration factor is significantly higher because of heavy users, longer holding times, flat rate and evolution of new applications.

The busy hour for residential narrowband and broadband traffic is in the evening.

Packet Switching Concentration

The circuit switched services PSTN and ISDN have no packet switching concentration ($u_{1t} = 1$). The other services have significant concentrations. Internet use consists of sessions based downloading, thinking and uploading. Traffic will be packet according to use. Traditional Internet use gives low packet switching concentration. Applications like music on demand and video on demand generate high packet switching concentration. The evolution of the packet switching concentration factor is complex.

Uncertainty in the Concentration Factors

Figure 4 shows possible evolutions of combinations between busy hour concentrations and packet switching concentrations. There are significant uncertainties in the evolution. The basis

for the predictions in Figure 4 is 0.15 (15 %) busy hour concentration and 0.20 packet switching concentration in 2001. *Four alternatives* are defined having a linear yearly increase:

Busy hour concentration: 0.15 in 2001 to respectively 0.195 – 0.24 – 0.285 – 0.33 in 2010.

Packet switching concentration: 0.20 in 2001 to respectively 0.335 – 0.47 – 0.605 – 0.74 in 2010.

Figures 2–4 show a nearly exponential evolution of broadband penetration, capacity increase and traffic concentration in the coming years. The most probable prediction will be between alternative 2 and 3. The traffic volume indicator described in equation (1) has a much stronger exponential evolution because of a multiplicative effect of the same factors.

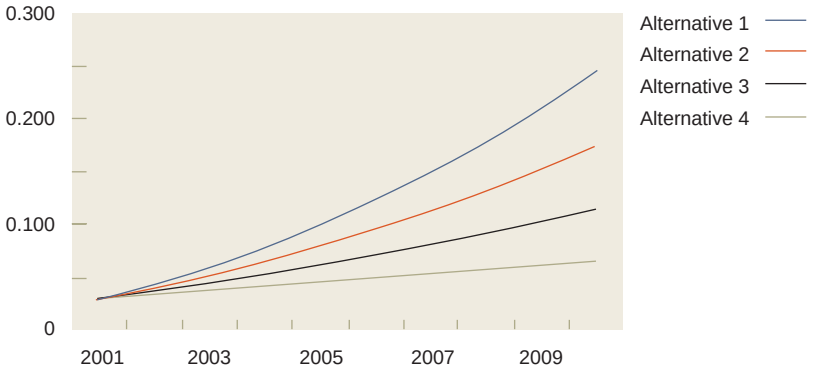


Figure 4 Description of possible evolutions of concentrations of ADSL traffic as a function of busy hour and packet switching concentration

4 Traffic from the Business Market

The business market generates the following types of traffic/capacity: Voice traffic, Dialled Internet traffic, PSDN, ATM, Frame Relay, DSL traffic, IP Virtual Private Networks (IP VPN) traffic, Leased lines, Fast and Gigabit Ethernet.

There are significant substitution effects between DSL, IP VPN, Leased lines, Fast and Gigabit Ethernet, which have to be taken into account in the forecasting process. Leased lines are used to establish fixed connections between sites often based on head office and branch offices or between different enterprises. The established network forms a local network with high service quality. There are no busy hour concentration or packet switching concentration. Leased lines constitute a significant part of the transport network capacity. Some parts of the leased lines capacity will be transferred to IP VPN or DSL because of cheaper tariffs and in spite of reduced service quality/SLA.

A traffic volume forecast indicator $V_B(t)$ for the business market busy hour traffic is given by:

$$V_B(t) = N_t \Sigma_i b_{it} u_{it} A_{it} C_{it} M_{it} p_{it} \qquad (2)$$

where the different traffic/capacity types i are defined in the first paragraph of the chapter.

5 Traffic Generated by Other Operators

Different operators like mobile operators, ISPs and other fixed network operators lease necessary capacity in the transport network. The capacity demand depends on type of services offered and the market share to the operators and of course the probability to use the transport network of the incumbent.

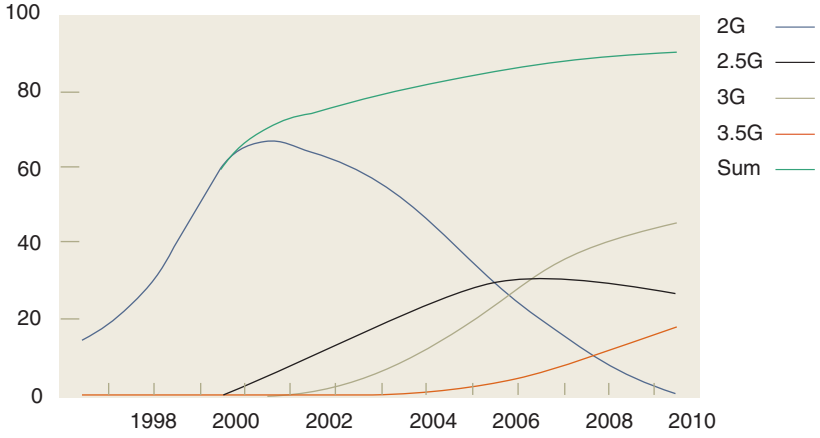
The mobile operators are important transport network customers using leased lines in the transport network. In the coming years these operators will generate the following traffic:

- 2G traffic (Digital mobile systems such as GSM)
- 2.5G traffic (HSCSD, GPRS, EDGE)
- 3G traffic (UMTS)
- 3.5G traffic (Ubiquitous roaming among 3G and WLAN systems)

The access forecasts for the different systems have been developed in the IST-2000-25172 project Tonic.

Subscription forecasts for the mobile systems, allowing the subscribers to have more than one subscription have also been developed. The traffic capacity for data applications per user increases from 40 kbs to 1.92 Mbs from 2G to 3G, while WLAN offers up to 54 Mbs. The traffic forecast model for mobile traffic is similar to the traffic in the fixed network. The busy hour concentration factor and the packet switching concentration factor are extremely important for the dimensioning. The services/applications are defined in different service classes: conversation, streaming, best effort, and the capacity is reserved according to service level agreements.

Figure 5 Subscriber penetration forecasts for different mobile systems



The traffic forecasting model $V_M(t)$ for mobile operators is given by:

$$V_M(t) = N_t \Sigma_i b_{it} u_{it} A_{it} C_{it} M_{it} p_{it} \quad (3)$$

Here N_t is number of persons, not number of households. M_{it} denotes the market share to the operator and i = 1, 2, 3, 4 the system generations 2G, 2.5G, 3G and 3.5G. The penetration forecasts p_{it} are shown in Figure 5. The capacity C_{it} is a mean capacity. A 3.5G subscriber has the possibility to use a data rate up to 144 kbs using UMTS but a significantly higher capacity in WLAN hot spot. The number of hot spots available and the proportion of time the subscriber uses WLAN compared with UMTS give the mean capacity C_{4t}. The factor A_{it} indicates the real utilisation of the capacity.

Let $V_O(t)$ be the busy hour traffic forecasts for the other operators, then the total of busy hour traffic forecasts $V(t)$ is given by:

$$V(t) = V_R(t) + V_B(t) + V_M(t) + V_O(t) \quad (4)$$

There are definitely possibilities to reduce $V(t)$ since the business traffic has busy hour before/after lunch, while the residential traffic has busy hour in the evening. Since the operators use leased lines, day and night capacity is equal. Let $V_{BL}(t)$ be the leased line capacity and $V_{BP}(t)$ be the packet switched traffic in the business market. Then $V_B(t) = V_{BL}(t) + V_{BP}(t)$. The residential broadband traffic is larger than the packet switched business traffic in the busy hour. Suppose that $\Delta = 0.2$ (20 %) of the packet switched business busy hour traffic is transferred during the residential busy hour the adjusted busy hour traffic forecast $V^*(t)$ is:

$$V^*(t) = V_R(t) + \Delta V_{BP}(t) + V_{BL}(t) + V_M(t) + V_O(t) \quad (5)$$

6 Capacity Forecasts

Models for making traffic forecasts during busy hour have been described. However, the capacity needed to carry the traffic is higher. Additional capacity has to be dimensioned taking into account stochastic variations around the mean traffic in the busy hour. Usually Erlang's blocking formula is applied for voice traffic, while Lindberger's approximation is useful for dimensioning the data traffic capacity [12]. The traffic will probably be transported on SDH systems where packet overhead is added. In addition the system's load average factor is less than the maximum capacity, and finally there will in general be some extra dimensioning since the expansion of the network is planned and deployed stepwise. The estimation of additional capacity depends on the planning process and the technical systems and will vary from one incumbent to another.

7 Conclusions

A traffic volume indicator has been developed to estimate busy hour traffic increase in the transport network. The indicator estimates the traffic entering the transport network. The traffic volume indicator does not include redundancy and protection capacity in the core network. Telenor uses the traffic volume indicator forecasts as input to the transport network planning process – evaluation of new network structures, expansion of the network and introduction of new core network technology.

The indicator depends on the application evolution and also the tariff regime for broadband services. So far, most countries use a flat rate for broadband traffic. However, specific applications will overload the transport network heavily if no actions are performed. There exist no incitements to control the size of the traffic in the transport network. A probable solution will be to introduce a tariff on high capacity traffic and on bandwidth on demand and specific applications. The traffic forecasts used assume that a new tariff regime for broadband services will be implemented within the next two years.

References

1 Stordahl, K, Kalhagen, K O. Broadband Access Forecasts for the European Market. *Telektronikk*, 98 (2/3), 21–32, 2002.

2 Elnegaard, N, Stordahl, K. Deciding on optimal timing of VDSL and ADSL roll-outs for incumbents. *Telektronikk*, 98 (2/3), 63–70, 2002.

3 Stordahl, K, Kalhagen, K O, Olsen, B T. Broadband technology demand in Europe. *ICFC 2002*, San Francisco, USA, June 25–28, 2002.

4 Stordahl, K, Elnegaard, N K. Battle between cable operator and incumbent : Optimal ADSL/VDSL rollout for the incumbent. In: *Proc. XIV International Symposium on Services in the Local Access – ISSLS 2002*, Seoul, Korea, April 14–18, 2002.

5 Monath, T et al. Economics of Ethernet based Access Networks for broadband IP Services. *Proc. ISSLS 2002*, Seoul, Korea, April 14–18, 2002.

6 Stordahl, K, Elnegaard, N K. Risk analysis of broadband access rollout strategies in a competitive environment. In: *Proc. Optical Hybrid Access Network/Full Service Access Network workshop*, Yokohama, Japan, April 4–6, 2001.

7 Stordahl, K, Elnegaard, N K. Broadband market evolution in Europe and the upgrade risks. *ICFC*, Seattle, September 26–29, 2000.

8 Stordahl, K et al. Broadband market, the driver for network evolution. In: *Proc. Networks 2000*, Toronto, Canada, September 10–15, 2000.

9 Elnegaard, N K, Ims, L A, Stordahl, K. The risks of DSL access network rollouts. In: *Proc. Networks 2000*, Toronto, Canada, September 10–15, 2000.

10 Stordahl, K, Ims, L A, Moe, M. Forecasts of broadband market evolution in Europe. In: *Proc. ITS 2000*, Buenos Aires, Argentina, July 2–5, 2000.

11 Stordahl, K. Broadband market evolution – demand forecasting and tariffs. *ICC 2000*, SAS session. Techno-economics of multimedia services and networks. New Orleans, USA, June 18–22, 2000.

12 Stordahl, K et al. Broadband upgrading and the risk of lost market share under increasing competition. In: *Proc. ISSLS 2000*, Stockholm, Sweden, June 19–23, 2000.

13 Elnegaard, N K, Ims, L A, Stordahl, K. Rollout strategies for the copper access network – evolution towards a full service access network. In: *Proc. ISSLS 2000*, Stockholm, Sweden, June 19–23, 2000.

14 Stordahl, K et al. Overview of risks in Multimedia Broadband Upgrades. In: Proc. *Globecom '99*, Rio de Janeiro, Brazil, Dec 5–10, 1999.

15 Stordahl, K et al. Evaluating broadband strategies in a competitive market using risk analysis. In: *Proc. Networks 98*, Sorrento, Italy, Oct 18–23, 1998.

16 Ims, L A (ed.). *Broadband Access Networks – Introduction strategies and techno-economic analysis*. Chapman-Hall, 1998.

17 Lindberger, K. Dimensioning and design methods for integrated ATM networks. In: *Proc. 14th ITC*, Antibes, France, June, 1994.

Planning Dependable Network for IP/MPLS Over Optics

TERJE JENSEN

Dr. Terje Jensen (41) is Research Manager 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 on network planning, performance modelling/analyses and dimensioning.

terje.jensen1@telenor.com

It is always challenging for an operator to find a sufficient level of functionality in the different network layers, also in the IP and optics areas as discussed in this paper. With the growing traffic demand, steadily more emphasis is placed on finding efficient network solutions, also considering resilience options. A number of options are described in this article.

1 Introduction

The ongoing traffic growth, mainly related to IP-based services requires a steady improvement in network efficiency. This is to carry the traffic while still reducing the infrastructure cost, and, hence, allowing for lowering the prices to customers.

Many people question to what extent the traffic load growth in the core of IP-based networks is hindered by the access link capacity (e.g. dial-up, ISDN, GSM). Introducing access links with higher capacity, such as DSL and Ethernet/fibre, the traffic loads in the core networks may grow even more drastically. This is one argument for investigating the use of optics in closer connection with IP (although the general traffic growth and price trends for optics also advocate this).

One of the means to step up the traffic handling capability of the IP network is to develop routers with higher throughput. Some means to be undertaken are:

- Separate forwarding and route determination and make the routing software leaner;
- Introduce interfaces with higher transfer rates, increased switching speed;
- Introduce hardware adapted solutions (e.g. through application-specific integrated circuits, ASICs).

Making leaner software solutions, in some respect, may be contrasting the functionality for traffic handling according to the Traffic Engineering mechanisms. Reducing the number of layers is one step to reduce the overhead. Hence, IP "direct" over optics has become a subject gaining more interest.

Basically, there will be a number of "clients" to an optical network sublayer. However, in this article the interplay between IP and optics is looked at, expected to become the dominating traffic volume to be carried in the long run.

Models for interconnecting the IP and the optical layers are described in Chapter 2, including inputs from sources such as the Optical Interconnection Forum, IETF and ITU. The Generalised Multi-Protocol Label Switching (GMPLS) has been proposed as a means for controlling the resources (mainly seen from the IP layer, but also from other protocol layers). A brief description is given in Chapter 3, together with MPLS recognized as a starting point for the GMPLS specification. Resilience can be provided in different ways and supported at different layers, as reflected in a number of options treated in Chapters 4–10.

2 Various Models for Interconnecting IP and Optics

2.1 Principles

In order for the resources to be utilised efficiently, the optical networks must be survivable, flexible and controllable. Some have suggested introducing more «intelligence» in the control plane for the optical networks in order to achieve this. Hence, utilising similar mechanisms as found for the IP-based networks is looked at as IP is seen as being a common protocol for much of the traffic carried by the optical network.

A first issue is the adaptation and reuse of IP control plane protocols for the control plane in optical networks. These are to be used no matter which traffic flows (IP or non-IP) are carried. A second issue is how IP traffic can be carried where joint control and co-ordination between the IP and optical layer are utilised.

This is illustrated in Figure 1: An optical subnetwork may consist of all-Optical Cross-Connects (OXCs) or some nodes where optical-electrical-optical conversion is applied. The switching function within an OXC is controlled by setting the parameters of the cross-connect unit. This can be considered as configuring a cross-connect table that specifies which input port is connected to which output port. Here, a port can indicate fibre as well as wavelength.

Two types of control interface are indicated; User-Network Interface (UNI) between the clients and the optical network, and, Network-

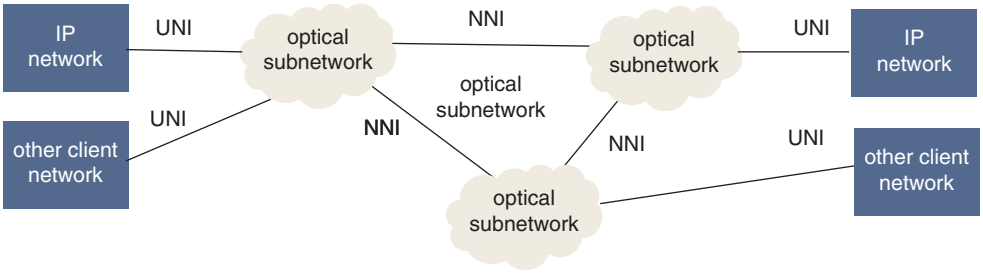


Figure 1 Schematic illustration of optical network with client networks

Network Interface (NNI) between optical subnetworks. The control flow across the UNI would naturally depend on the services offered to the client. As the NNI control flow would be derived from IP control, similarities between NNI and UNI may well exist when an IP network is the client. In addition to these two interface types, interface within in an optical subnetwork also needs to be defined, such as the interface between two OXCs. Some refer to this interface as an internal NNI.

Between administrative domains it is essential to consider issues like security, scalability, stability and information hiding. In principle, the UNI and the NNI could be implemented in the same way. However, one commonly seeks to limit the information needed to be transferred across interfaces, thereby motivating for a separation of UNI and NNI, allowing for a specialisation of their implementations.

The UNI can be regarded as a client-server interface; for example, the IP layer is the client, while the optical layer is the server. The client roles would then request a service connection and the server role establishes the connection to meet the request when all admission control conditions are fulfilled. The physical implementation of the UNI may vary, like

- direct interface with an in-band or out-of-band IP control channel. This channel is used to exchange signalling and routing messages between the router and the OXC (like a peering arrangement);
- indirect interface with out-of-band IP control channel. The channel may run between management systems or servers;
- provisioned interface involving manual operations.

Two service models, in principle applicable both for UNI and NNI, are outlined in [ID_ipofw]:

- Domain service model where the optical network primarily offers high bandwidth connectivity (services like light-path creation, deletion, modification and status enquiry);
- Unified service model where the IP and the optical network are treated together, as seen from the control plane perspective, as an integrated network. Then the OXCs will be treated like any router as seen from the control plane. No distinction is then made between UNI, NNI and any other router-to-router interface. Such an interface is assumed to be based on extensions for MPLS/GMPLS.

It is important to make a separation between the control plane and the data plane over the UNI. As mentioned, the optical network basically provides services to clients in the form of transport capacities (by light-paths). IP routers at the edge of the optical network must establish such paths before the communication at the IP layer can start. Therefore, the IP data plane over optical network is done over an underlying network of optical paths. On the other hand, for the control plane, the IP routers and the OXCs can have peering relations, in particular for routing information exchanges. Various degrees of loose or tight coupling between the IP and the optical network may be used. The coupling is given by the details of topology and routing information exchanged, level of control that IP routers can exercise on selecting specific paths, and policies regarding dynamic provisioning of optical paths between routers (including access control, accounting and security).

Three interconnection models are sketched:

- *Overlay* model: The routing, topology distribution, signalling protocols are independent for the IP/MPLS and the optical network.
- *Augmented* model: Routing instances in the IP layer and the optical network are separated by information exchanged (e.g. IP addresses are known to the optical routing protocols).
- *Peer* model: The IP/MPLS layers act as peers to the optical network. Then a single routing protocol instance can be used for the IP/MPLS network and the optical network.

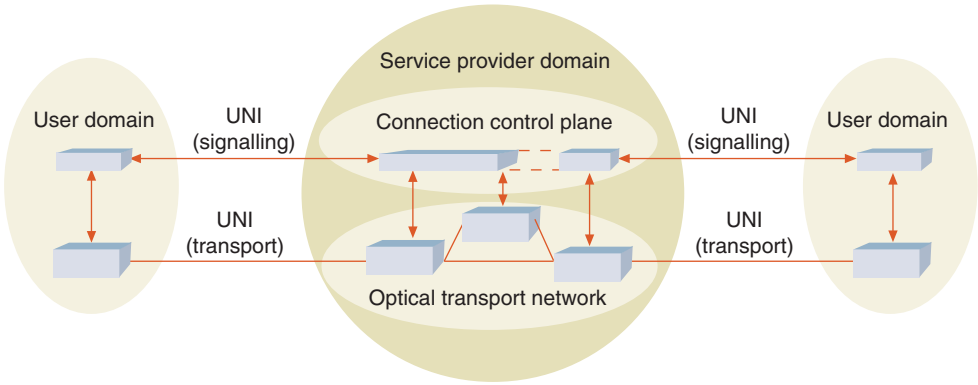


Figure 2 UNI used between service provider and user domains

These models give a certain degree of implementation complexity; the overlay being the least complex one for near-term deployment and the peer model the most complex one. As each of the models has its advantages, an evolution path for IP over optical network may be seen.

A possible migration path is to start with the simpler functionality, meaning the domain service model with overlay interconnection and no routing exchange between the IP and the optical network. A provisioned interface would be expected. The next phase of the migration path is to exchange reachability information between IP and the optical network. This may allow for light-paths to be established in conjunction with setting up Label Switched Paths (LSPs). The third phase of the migration might be supporting the peer model.

Applying a common signalling framework from the start would assist the migration. For the domain service model, implementation agreement based on Generalised MPLS (GMPLS) UNI signalling is being developed by the Optical Interworking Forum (OIF). This is intended for near-term deployment, although helping in the migration toward the peer model. This is said to support incremental development as the interconnection model increases in complexity.

2.2 User Network Interface – OIF

Utilising WDM as an intelligent transport service allows its clients such as IP routers to interconnect and relate to optical interfaces. The increasing bandwidth need places more weight on the capability to manage and control optical networks. Currently, optical networks are provisioned through element management systems, implying that end-to-end connections across equipment from different manufacturers are likely to involve several incompatible management systems. This, again, will likely result in longer provisioning times and manual effort required.

The Optical Interconnection Forum (OIF) has addressed this issue by adopting MPLS-based control intelligence for use within optical networks. A standard signalling interface between client and optical network has also been defined to support dynamic provisioning requests.

Version 1.0 of OIF UNI allows clients to establish optical connections dynamically by applying signalling procedures compatible with Generalized MPLS (GMPLS), see Figure 2.

In addition to signalling, the IOF UNI specification also describes a neighbour discovery mechanism and a service discovery mechanism. The former allows nodes at both ends of a fibre link to identify each other (e.g. reducing the manual effort needed to build data bases of network inventory). The service discovery mechanism allows clients to determine the services that are supported by the optical network, also any new services introduced.

An application of the capabilities is to support dynamic traffic engineering controlled by the IP router traffics. An example, as shown in Figure 3, is to establish or release bandwidth depending on the traffic level – for example estimated by traffic monitoring.

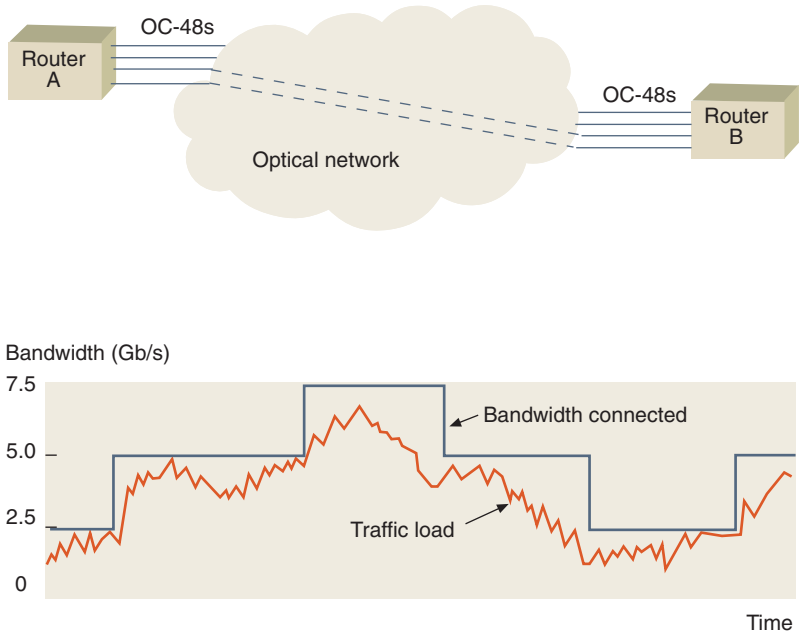


Figure 3 Illustration of UNI signalling used to establish and release bandwidth between routers

2.3 Requirements on Optical Network Services

Some of the challenges for network operators are efficient bandwidth management and fast service provisioning in a multi-technology and possibly multi-vendor networking environment. ITU has started work on Automatically Switched Optical Network (ASON) recommendations aiming to providing optical networks with intelligent network functions and capabilities in its control plan. This enables fast connection establishment, dynamic rerouting and multiplexing and switching at different granularity levels (including fibre, wavelength and TDM channel). The following benefits are strived for:

- Automated discovery (network inventory, topology and resource)
- Rapid circuit provisioning
- Enhanced protection and restoration
- Service flexibility (protocol and bit-rate transparent and bandwidth-on-demand)
- Enhanced interoperability

As described in [ID-iopreq] a number of additional benefits may also be offered by ASON:

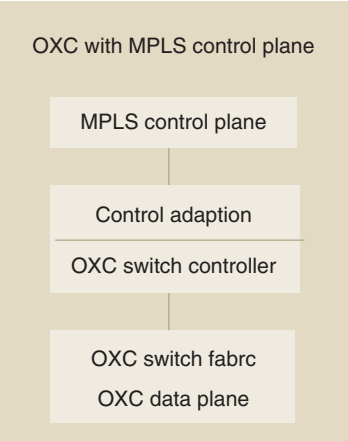
- Reactive traffic engineering at optical layer that allows network resources to be dynamically allocated to traffic flow.
- Reduce the need for service providers to develop new operational support systems software for the network control and new service provisioning on the optical network, thus speeding up the deployment of the optical network technology and reducing the software development and maintenance cost.
- Potential development of a unified control plane that can be used for different transport technologies, including OTN, SDH, ATM and PDH.

Services are commonly described by their topologies; bi-directional or uni-directional and point-to-(multi-)point. In addition, persistency may be given, such as permanent, switched and soft permanent. The difference between permanent and soft permanent is that only end-points in a domain are specified for the latter, leaving it to the network to establish a connection between these points.

A number of requirements, based on the description in [ID-iopreq], can be categorised as:

- General requirements:
  - Separation of networking functions – grouping functions into control plane, data plane (transport plane) and management plane.
  - Separation of call and connection control – call relates to end-to-end, admission control, while connection control relates to establishing and releasing the connections for a call.
  - Network and service scalability and response times to be supported, enabling several thousands of termination ids per OXCs, set-up time should preferably be less than a number of seconds.
  - Services offered over different underlying types of transport technologies – including optical and TDM.
  - Well-defined service building blocks.
- Service user requirements:
  - The common services must be supported with the corresponding interfaces (SDH, wavelength, Ethernet, Storage Area Networks) with various bit rates, Service Level Agreement (SLA) conditions, service classes and routing options.
  - Service invocation may be initiated by the provider (permanent and soft permanent) or by the user side (switched). Interplay between control and management must take place to allow for different initiation modes.
- Service provider requirements:
  - Supporting different modes for connecting users; direct connection, remote access via sub-networks, dual homing).
  - Different inter-domain connectivity schemes must be supported. In general one seeks autonomy between different domain signalling/control, both for domain within an operator and between operators. Several points of interconnection between domains must be possible.
  - Different name and address arrangements must be allowed, depending on the clients (e.g. IP addresses for IP routers). Address resolution and translation functions should be present.
  - Policy-based service management must be supported to enable flexible service provisioning.

Figure 4 Potential OXC system architecture



- Control plane requirements (in combination with management plane):
  - Basic capabilities required are network resource discovery, address assignment and resolution, routing information propagation and dissemination, path calculation and selection, connection management.
  - A signalling network can be established independently of the data transport plane topology. Three types of relations can be found:
    i) in-band signalling: signalling messages are carried in a logical channel imbedded in the data-carrying link;
    ii) in-fibre, out-of-band: signalling messages are carried in a dedicated communication channel separated from the data-carrying channels, but within the same fibre;
    iii) out-of-fibre: signalling messages are carried over a different fibre than the data-carrying links.
  - Interface to data plane to be standardised allowing for the control plane to configure switching fabrics and port functions (via management plane) and receive failure and degradation status information.
  - The control plane is considered a managed entity and, hence, interacts with the management plane for configuration, status reporting and so forth.
- Requirements on signalling, routing and discovery:
  - Signalling must support both strict and loose routing and allow individual as well as groups of connections. Fault notifications must also be supported. Crank-back and rerouting shall be supported for inter-domain signalling.
  - Routing includes information on reachability, topology/resource and path computation. Routing mechanisms to support are hop-by-hop, source-routing and hierarchical routing. In the latter case, information aggregation would be needed. Both scalability and stability aspects are to be considered. Exchange of routing information may be trigger-based and timeout-based.
  - Path selection must allow for different algorithms, including shortest path and constraint-based routing (constraints such as cost, link utilization, diversity, service class).
  - Both manual and automatic discovery shall be supported for neighbour (physical and logical), resource and service.
- Requirements on resiliency for service and control plane:
  - Different service levels are to be offered, accompanied by SLA conditions.
  - Priorities for signalling messages should be implemented in order to allow for faster restoration.

2.4 Optical Transport Networks

A high level architectural model has been worked on in IETF, grouping the modelling aspects into a horizontal dimension and a vertical dimension. The horizontal dimension refers to special requirements for an Optical Transport Network (OTN) including considerations like:

- Type of OTN state information should be discovered and disseminated to support path selection for optical channels (e.g. attenuation, dispersion).
- Infrastructure used for propagating the control information.
- Computing constrained paths fulfilling performance and policy requirements.
- Domain specific requirements for establishing optical channels and enhancements for MPLS signalling protocols for addressing these requirements.

The vertical dimension includes concerns when porting MPLS control plane software onto an OXC. A potential architecture of an OXC is illustrated in Figure 4.

Looking closer into an OTN as described by ITU-T, it should be noted that it is itself divided into layers, including

- an optical channel (OCh) layer network;
- an optical multiplex section (OMS) layer network;

Box A – OTNT: ITU-T Recommendations on the OTN Transport Plane

Framework for Recommendations

G.871/Y.1301 Framework for Optical Transport Network Recommendations 10/00

Architectural Aspects

G.872 Architecture of Optical Transport Networks (Revised, 11/01 pre-publ.) 2001

Structures & Mapping

G.709/Y.1331 Network node interface for the optical transport network (OTN) 02/01

G.975 Forward Error Correction 10/00

Functional characteristics

G.681 Functional characteristics of interoffice long-haul line systems using optical amplifiers, including optical multiplexing 10/96

G.798 Characteristics of optical transport network (OTN) equipment functional blocks 11/01, corr. 2002

G.806 Characteristics of transport equipment – Description Methodology and Generic Functionality 10/00

G.7710/Y.1701 Common Equipment Management Requirements 11/01

Protection Switching

G.841.x Protection Switching in the OTN 2002

G.gps Generic Protection Switching 2002

Management Aspects

G.874 Management aspects of the optical transport network element 11/01

G.874.1 Optical Transport Network (OTN) Protocol-Neutral Management Information Model For The Network Element View 01/02

G.875 Optical Transport Network (OTN) management information model for the network element view

Data Communications network (DCN)

G.7712/Y.1703 Architecture and specification of data communication network 11/01

G.dcn living list

Error Performance

G.optperf Error and availability performance parameters and objectives for international paths within the Optical Transport Network (OTN) 2003

M.24otn Error Performance Objectives and Procedures for Bringing-Into-Service and Maintenance of Optical Transport Networks 2003

Jitter & Wander Performance

G.8251(G.otnjit) The control of jitter and wander within the optical transport network (OTN) 11/01, amend/corr. 2002

Physical-Layer Aspects

G.691 Optical Interfaces for single-channel SDH systems with Optical Amplifiers, and STM-64 and STM-256 systems 10/00

G.692 Optical Interfaces for Multichannel Systems with Optical Amplifiers 10/98

G.694.1 Spectral grids for WDM applications: DWDM frequency grid 2002

G.694.2 Spectral grids for WDM applications: CWDM wavelength grid 2002

G.664 General Automatic Power Shut-Down Procedures for Optical Transport Systems 06/99

G.959.1 Optical Transport Networking Physical Layer Interfaces 02/01

G.693 Optical interfaces for intra-office systems 11/01

Sup.dsn Optical System Design 2003

Fibres

G.651 Characteristics of a 50/125 μm multimode graded index optical fibre cable 02/98

G.652 Characteristics of a single-mode optical fibre cable 10/00

G.653 Characteristics of a dispersion-shifted single mode optical fibre cable 10/00

G.654 Characteristics of a cut-off shifted single-mode fibre cable 10/00

G.655 Characteristics of a non-zero dispersion shifted single-mode optical fibre cable 10/00

Components & Subsystems

G.661 Definition and test methods for the relevant generic parameters of optical amplifier devices and subsystems 10/98

G.662 Generic characteristics of optical fibre amplifier devices and subsystems 10/98

G.663 Application related aspects of optical fibre amplifier devices and sub-systems 04/00

G.671 Transmission characteristics of passive optical components 02/01

Figure 5 Boundary of an optical transport network and client-server relationship

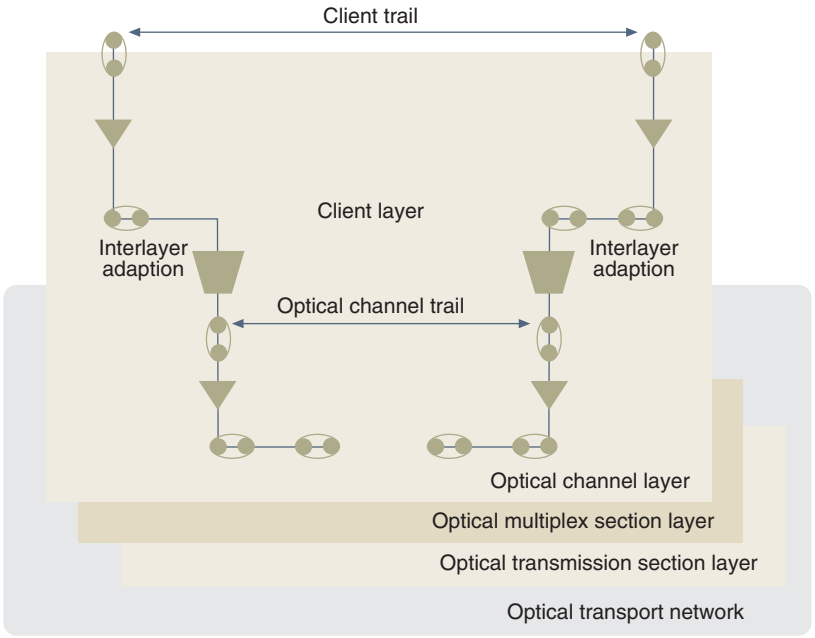
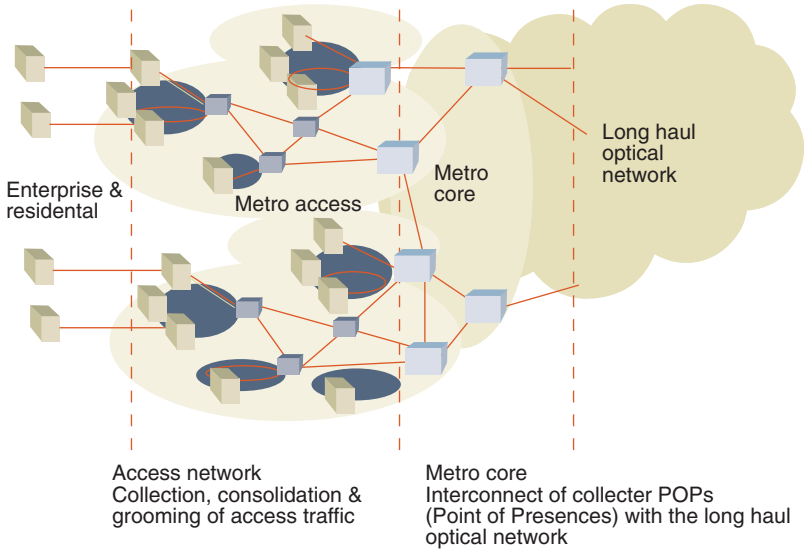


Figure 6 Possible relationship of MON and LHON



Box B – OTNT: ITU-T Recommendations on the ASTN/ASON Control Plane

Requirements

G.807/Y.1302 Requirements for the Automatic Switched Transport Network (ASTN) 07/01

Architecture

G.8080/Y.1304 Architecture for the Automatic Switched Optical Network (ASON) 11/01

G.ason living list

Protocol Neutral Specifications for key signalling elements

G.7713/Y.1704 Generalised Distributed Connection Management (draft version 0.3, 06/01) 10/01

G.7713.1/Y.1704 Distributed Call and Connection Management – PNNI Implementation

G.7713.2/Y.1704 Distributed Call and Connection Management – GMPLS RSVP-TE Implementation

G.7713.3/Y.1704 Distributed Call and Connection Management – GMPLS CR-LDP Implementation

G.7714/Y.1705 Generalised automatic discovery techniques 10/01

G.7715/Y.1706 Architecture and requirements for routing in automatically switched optical networks 2002

G.7716/Y.1707 [ASTN link connection status]

G.7717/Y.1708 Connection Admission Control Specific Protocols to realise the signalling elements

Data Communication Network (DCN)

G. 7712/Y.1703 Data Communication Network (Draft, 06/01) 10/01

G.dcn living list version 02/01

- an optical transmission section (OTS) layer network.

Within ITU-T the optical transport network (OTN) is defined as *transmission of information over optical media in a systematic manner*. An OTN is composed of a set of optical network elements connected by optical fibre links, able to provide the functionality of transport, multiplexing, routing, management, supervision and survivability of optical channels carrying client signals. The OTN is able to transport any digital signal independent of client-specific aspects. An illustration is given in Figure 5.

The term metropolitan optical network (MON) has entered the stage in recent years. MONs are said to have a role distinct from the long-haul network as well as the enterprise and access networks. Due to the increasing traffic demands optical solutions are increasingly promoted in the aggregation and region networks. A few characteristics do however motivate for treating MON differently than long-haul networks, see Figure 6:

- MONs are inherently designed for short to medium distances, say less than 200 km. Hence, topics like signal regeneration, amplification and error correction are of less importance.
- Lower cost is more emphasized in combination with wide coverage.
- Service developments and fast provisioning might be more pronounced in MONs, such as bandwidth-on-demand and awareness of service classes.

Box A and Box B list ITU-T recommendations related to OTN transport plane and ASON/ASTN control plane, respectively.

3 Generalised Multi-Protocol Label Switching

The GMPLS is described in [RFC3471]. This basically contains extensions to signalling for MPLS, needed to include time-division, wavelength and spatial switched/divided systems, see illustration in Figure 7. Hence, this chapter starts with a description of MPLS before briefly addressing the more generalised version.

3.1 MPLS – the Starting Point

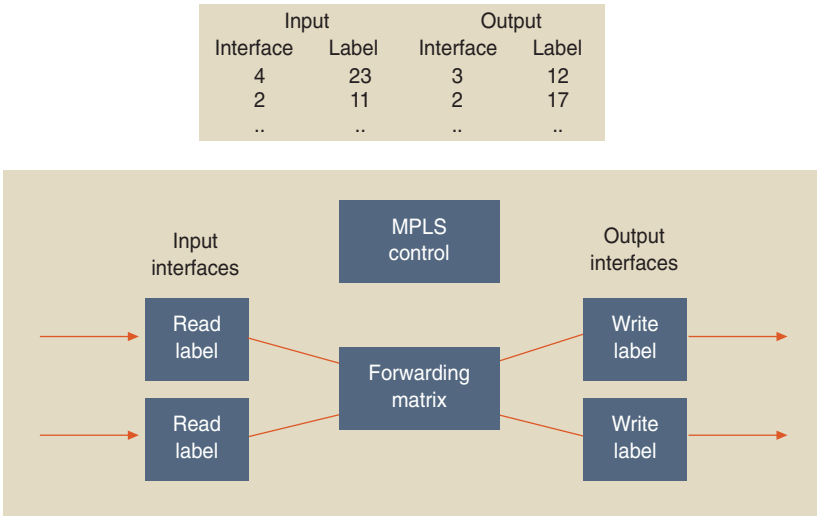
3.1.1 Principles and Format

At the IP layer (layer 3) a router makes forwarding decisions for a packet based on information in the IP header. The analysis of the packet header is performed and an algorithm is executed in each router to decide upon further treatment. This can be viewed as a two-step process, ref [RFC3031]: i) The packets are classified into a set of *Forwarding Equivalence Classes (FECs)*; ii) Each FEC is mapped to a next hop.

In contrast to legacy IP networks, when a packet enters an MPLS network it is assigned a label. At subsequent hops the label is used as an index into a table that specifies the packet's next hop and new label. The old label is swapped with the new label, and the packet is forwarded to its next hop. The path the packet traverses is therefore called a Label Switched Path (LSP). A set of LSPs can be merged at a specific node if packets from these LSPs are forwarded in the same manner (e.g. over the same downstream path, with the same forwarding treatment). This is called label merging.

A key feature of MPLS is that once the labels required for an LSP have been assigned by LSP

Figure 7 Similarities between MPLS and GMPLS for a Label Switched Router and an Optical Cross-Connect



set up or label distribution protocols, intermediate Label Switched Routers (LSRs) transmitted by the LSP do not need to examine the content of the data packets flowing on the LSP. Several labels can be placed on a packet to form a label stack that allows multiple LSPs to be tunnelled, one within the other. The outermost LSP therefore becomes a tunnel that makes inner LSPs transparent to the intermediate LSRs, and therefore simplifies the forwarding tables at these LSRs. This feature is critical for the local repair approach as described later. An explicitly routed LSP is an LSP whose path is established by means other than normal IP routing. This requires i.a. a management system representation of the LSPs.

The following advantages of MPLS are listed in [RFC3031]:

- MPLS forwarding can be done by nodes not capable of analysing the IP packet headers, or not capable of analysing these headers with sufficiently high speed.
- Assigning a packet to an FEC, the ingress router may use information about the packet that goes beyond the content of the packet header, like the interface. Hence, assignment to FECs can be a more involved process, without impacting all routers in the network.
- Forwarding decisions within a network may be made depending on which ingress router a packet used. Then a packet may be forced to follow a particular route explicitly chosen, circumventing the ordinary routing.

To some extent, the use of LSPs can be considered as introducing tunnelling as seen from the IP layer. That is, when an LSP is introduced an intermediate node would not examine the IP header information in order to decide upon the proper handling of the packets arriving in that LSP. That is, with MPLS the classification of packets into FECs is only performed at the ingress to the MPLS domain. The packet is then mapped to an LSP by encapsulation of an MPLS header. The LSP is identified locally by the header, see Figure 8. In successive routers within the MPLS domain the label is swapped (therefore it can have only local significance) and the packet is mapped to the next hop.

An LSP can be considered as a path created by concatenation of one or more hops, allowing a packet to be forwarded by swapping labels from an incoming to an outgoing side of the MPLS node. An MPLS path is frequently referred to as layer 2 1/2 in the OSI model. That is, it may be considered as a tunnel. In order to introduce a tunnel, a "header" is attached to the IP packet as shown in Figure 8 for the Point to Point Protocol (PPP) case. The MPLS architecture is described in [RFC3031].

Referring to Figure 8, the fields in the MPLS header can be used as follows:

- Label – contains a 20 bit tag identifying an LSP.
- Exp – contains 3 bits (originally not allocated, intended for experimentation) which can refer to a certain service class, e.g. in analogy to the DiffServ classes.
- S – 1 bit indicates end of label stacking as several labels may be stacked.
- TTL – 8 bit giving the Time To Live information.

When an MPLS packet enters a Label Switching Router (LSR) a table containing information – Label Information Base (LIB) – on further treatment of the packet is looked up. This base may also be referred to as the Next Hop Label Forwarding Entry (NHLFE), typically, containing the following information (ref. [RFC3031]):

- next hop of the packet;
- operation to perform on the packet's label stack (replace the label at the top with another label, pop the label stack; or, replace the label at the top of the stack with a new label, at the same time push one or more new labels onto the stack);
- data link encapsulation to use when transmitting the packet;

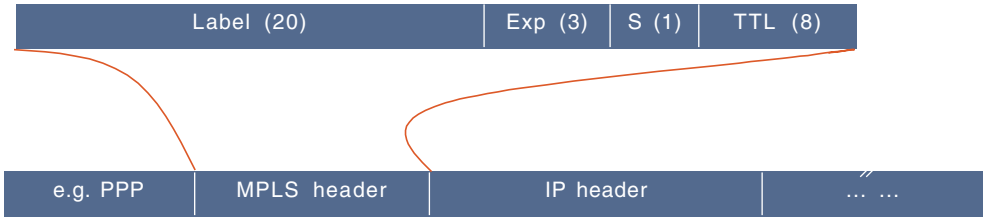


Figure 8 MPLS header and placement in layer "2 1/2"

- way of encoding the label stack when transmitting the packet;
- other information relevant for forwarding treatment.

In a given LSR, the "next hop LSR" may be the same LSR, implying that the top level label should be popped and the packet "forwarded" to itself, allowing for more forwarding decisions.

At the ingress of an MPLS domain an FEC-to-NHLFE mapping is needed; that is when packets arrive without an MPLS label.

Within an MPLS domain an incoming label mapping is executed, mapping the packet onto a set of NHLFEs.

MPLS can operate on a label stack. Operations on this stack are push, pop and swap. This can be used to merge and split traffic streams. The push operation adds a new label at the top of the stack and the pop operation removes one label from the stack. The MPLS stack functionality can be used to aggregate traffic trunks. A common label is added to the stack of labels. The result is an aggregated trunk. When this MPLS path is terminated the result will be a splitting (de-aggregation) of the aggregated trunk into its individual components. Two trunks can be aggregated in this way if they share a portion of their path. Hence, MPLS can provide hierarchical forwarding, potentially becoming an important feature. A consequence may be that the transit provider need not carry global routing information, thus making the MPLS network more stable and scalable than a full-blown routed network.

3.1.2 Traffic Engineering and MPLS

When utilising MPLS with Traffic Engineering, a number of mapping relations is asked for:

- Mapping packets onto FECs. An FEC composes a group of packets to be forwarded over the same path with the same forwarding treatment. In order to carry out this mapping fields in the IP packet are examined.
- Mapping FECs onto traffic trunks. A traffic trunk is an aggregation of traffic flows of the same class. A traffic trunk can again be routed (placed inside an LSP; i.e. a traffic trunk is only given for one LSP and not a sequence of LSPs).
- Mapping traffic trunks onto LSPs.
- Mapping LSPs onto links in the physical network.

In several sources, the terms traffic trunk and LSP are used synonymously. A fundamental difference between traffic trunk and LSP can be observed, though. That is, a traffic trunk is an abstract representation of traffic to which specific characteristics can be associated. An LSP is a description of a path in the network through which the traffic traverses.

Trunks having the same egress point may be merged into a common tree towards the egress. This may reduce the number of trees significantly. Trunks can also be aggregated by adding a new label to the stack for each trunk (that is, bundling the trunks into a single path/tunnel).

Designing an MPLS network "on top of" a physical network could be looked upon as relating two graphs to each other:

- Physical graph, $G = (V, E, c)$ is a capacitated graph depicting the physical topology of the network. V is the set of nodes in the network and E is the set of links. For v to w in V, (v, w) represents the link in E when v and w are directly connected under G. c indicates the set of capacity and other constraints associated with E and V.
- MPLS graph, $H = (U, F, d)$, where U is a subset of V representing the LSRs; that is the set of LSRs that are endpoint of at least one LSP. F is the set of LSPs. For x and y in U, (x, y) is in F if there is an LSP going from x to y. d represents the set of demands and restrictions associated with F.

The fundamental problem of designing an MPLS network is to relate the two graphs such that an objective function is optimised.

One of the requirements from Traffic Engineering is to be able to reroute an LSP under a number of conditions (failure, better route available, etc.). It is desirable that this is done without disturbing the traffic flows. This could be done by establishing the new LSP before the old/existing LSP is released, which is called make-before-break. In case the existing and new LSP compete for the same resources, particular concerns have to be made, also considered by the admission control.

In addition to attributes related to traffic trunks, see Box C, some attributes are also related to resources (frequently thought of as the links). These attributes are ([RFC2702]):

- Maximum allocation multiplier attribute. The value of this attribute tells what proportion of the link and buffer capacity is available. Then,

Box C – Attributes characterising traffic trunks are ([RFC2702]):

- Traffic parameter attributes. These are used to describe the traffic flows (the FECs) transported in the traffic trunk. Relevant parameters include peak rates, average rates, maximal burst size, etc. Possibly, equivalent measures could be applied, like the effective bandwidth.
- Explicit path specification attribute. An explicit path assignment for a traffic trunk is a path that is specified through "operator" action (e.g. management procedures). Such a path can be completely or partially specified. Path preference rules may be associated with explicit paths, telling whether the explicit path is "mandatory" (has to be followed) or "optional" (other paths could be selected in case sufficient resources are not available on the preferred path).
- Resource class affinity attribute. This attribute can be used to specify which resource types can be explicitly included or excluded from the path through which the traffic trunk is routed. If no affinity attribute is given a "don't care" condition is assumed. Routing traffic trunks onto resources has to take these attributes into account, matching the requirements.
- Adaptivity attribute. As network state and traffic state change over time, more optimal routes of traffic trunks could appear. Setting this attribute tells whether or not the route can be re-optimised for the traffic trunk. However, appropriate thresholds should be given avoiding too frequent changes of routing.
- Load distribution attribute. In case several traffic trunks are used between the pair of nodes, the load distribution attribute can tell whether or not the load (traffic trunk) can be distributed on these trunks. In general the packet order should be maintained, implying that packets belonging to the same traffic flow are transferred on the same traffic trunk.
- Priority attribute. This attribute gives the relative importance of the traffic trunk. The value can be used to determine the order in which trunks are assigned to paths under establishment and failure situations. Priorities will also be used together with preemption.
- Preemption attribute. The value of this attribute tells whether or not a traffic trunk can preempt another traffic trunk, and whether or not another traffic trunk can preempt a specific traffic trunk. This will assist to ensure that high priority traffic trunks are routed through even though the capacity is not sufficient to handle all traffic trunks.
- Resilience attribute. The resilience attribute gives the behaviour of a traffic trunk when faults occur along the path followed by a traffic trunk. In case of fault the traffic trunk could be rerouted or not depending on the value of this attribute. For rerouting, the constraints given (e.g. by affinity) could be observed or not.
- Policing attribute. The value of this attribute tells which actions to take when the traffic on the trunk is not complying (traffic is exceeding). Examples of actions are packet dropping (rate limiting), packet tagging and packet shaping.

"over-allocation" could be achieved (as well as "under-allocation").

- Resource class attributes. The attributes express the resource type (e.g. thought of as colours). These are matched with the traffic trunk affinity attribute when finding paths onto which the traffic trunks are routed.

Basic operations on traffic trunks are ([RFC2702]):

- Establish a traffic trunk
- Activate a traffic trunk, to start passing packets
- Deactivate a traffic trunk
- Modify attributes for a traffic trunk
- Reroute a traffic trunk
- Remove a traffic trunk.

In addition to these basic operations, a few more, like policing and shaping could also be defined.

A traffic trunk is defined as unidirectional. As a bidirectional transfer capability is commonly asked for, two traffic trunks having the same end-points but passing packets in opposite directions can be defined. In case these are always handled as a unit, it is called a bidirectional traffic trunk (they are established as an atomic operation and one may not exist without the other). If a trunk is routed through a different physical path from the corresponding trunk in the opposite direction, the bidirectional traffic trunk is called topological asymmetric. Otherwise, it is called topological symmetric.

3.2 GMPLS Multiplexing Hierarchy

MPLS uses labels to support forwarding of packets. Label Switching Routers (LSRs) have a forwarding table recognising the cells/frames with the labels, or the IP packet headers (at the border of the MPLS domain). This is extended in GMPLS where the following interfaces are given for an LSR:

- Interfaces that recognise packet/cell/frame boundaries and forward the data based on the content in the packet or label/cell header. This is referred to as *Packet-Switch Capable*. Examples are MPLS-capable routers and ATM switches.
- Interfaces that forward data based on time slot in a periodic cycle. This is referred to as *Time-Division Multiplex Capable*. SDH cross-connect is one example.
- Interfaces that forward data based on the wavelength. Such interfaces are referred to as *Lambda Switch Capable*. An optical cross-connect is an example.
- Interfaces that forward data based on physical space position of data. This is referred to as *Fibre-Switch Capable*. An optical cross-connect operating on a level of single or multiple fibres is an example.

These can be organised in a hierarchical manner as shown in Figure 9 and corresponding labels and Label Switched Paths (LSPs) defined. Then,

Figure 9 Organising "labels" hierarchically

an LSP that starts and ends on a packet-switch capable interface can be grouped together with other similar LSPs into a common LSP that starts and ends on a time-division multiplex interface. This LSP can be grouped together with other similar LSPs into an LSP that starts and ends on a lambda switch capable interface, and so forth. This is similar to a multiplexing hierarchy. Note that all these levels may not be present in all cases.

Compared with MPLS, the GMPLS introduces additional interface types. The formats of the labels on the interfaces are given in [RFC3471].

Motivations for combining solutions for MPLS, in particular related to Traffic Engineering, and mechanisms for control plane in OXCs are:

- to provide a framework for real-time provisioning of optical channels in automatically switched optical networks;
- to foster the expedited development and deployment of a new class of versatile OXCs;
- to allow the use of uniform semantics for network management and operation control and hybrid networks consisting of OXCs and label switching routers (LSRs).

A particular emphasis may be placed on support of various protection and restoration schemes.

3.3 GMPLS Signalling Functional Description

Extensions to MPLS signalling to support GMPLS are described in [RFC3471]. A generalized label contains sufficient information for a receiving node to program its cross connect fabric. Between nodes generalized labels are exchanged in order to allow these nodes to know how to handle the information attached with these labels in the next phase. Generalized label requests are used consisting of

- 8 bit LSP encoding type: gives the LSP types, such as 1 = packet, 2 = Ethernet, 5 = SDH, 8 = lambda, 9 = fibre
- 8 bit switching type: gives the type of switching (packet switch, layer 2 switch, TDM switch, lambda switch, fibre switch)
- 16 bit Generalized payload identifier: gives the type of client layer to the LSP (IP, MPLS, Ethernet, SDH, lambda, fibre)

The label itself depends on the link type it is referring to. For fibre and lambda a 32 bit label has been described being significant between two neighbour nodes. Allowing for a range of (close-by) lambdas to be switched along the same route a waveband label has been devised, giving the start and end label (and thereby the range of wavelengths to be routed together).

GMPLS allows for bi-directional LSPs to be established as an atomic activity (compared with having to set up two uni-directional LSPs for MPLS). This allows for reduced set-up times, although one label each has to be assigned in the different directions.

Link-related protection can also be carried by the signalling field called Protection Information. The use of this field is optional. The protection capabilities of a link may be advertised by the routing scheme. The Protection Information field also indicates whether the LSP is primary or secondary (backup). The resources of the secondary LSP may be used by other LSPs until the primary LSP fails over to the secondary LSP.

The Protection Information field consists of three units:

- 1 bit secondary flag – set to 1 for a secondary LSP
- 25 bit reserved (set to zero)
- 6 bit link flags – indicates protection type (e.g. 1 + 1, 1 : 1, shared, unprotected)

4 Single-Layer Survivability

Four options exist for a two-layer structure if each of the layers is considered in isolation: i) no survivability mechanism; ii) survivability mechanisms in the lower layer; iii) survivability mechanisms in the upper layer; and iv) survivability mechanisms both in the lower and upper layer, but without any coordination. The first option is of little relevance for this discussion. Factors to be considered for the others are (adapted from [Coll02]):

- *Lower layer survivability only:* A basic advantage for having recovery mechanisms in the lower layer is that simple root failure has to be treated and recovery actions are performed on the coarsest granularity. This results in the lowest number of required recovery actions. Moreover, failures do not need to propagate through other layers before triggering any recovery actions. A drawback, however, is that failure in upper layers may not be resolved. In addition, when a failure occurs in a lower layer node, this layer recovers affected traffic where this node is involved, possibly leaving upper layer nodes in isolation (see Figure 10).

Figure 10 Only recovery mechanisms in the lower layer may result in isolated upper layer nodes

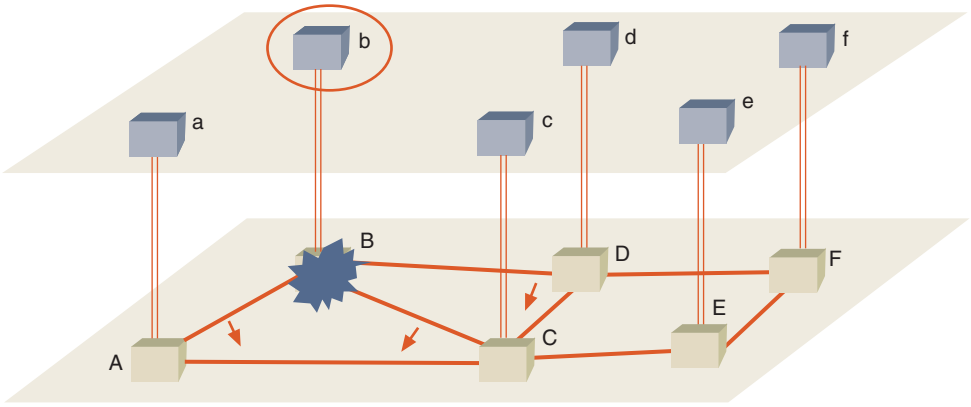
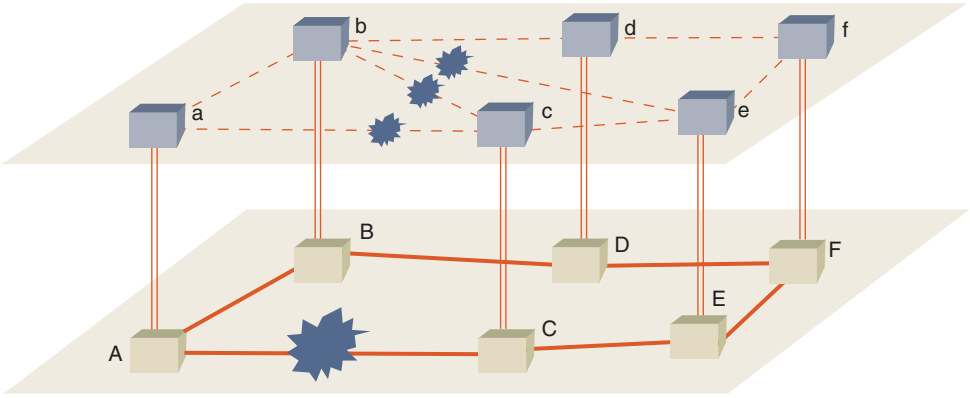


Figure 11 Single failure at lower layer may result in several failure indications at upper layer



- *Upper layer survivability only:* An advantage of upper layer survivability is that it can handle node or higher layer failures more easily. A disadvantage is that several recovery actions may be needed because of finer granularity of the traffic flows in the upper layer, see Figure 11. On the other hand, this finer granularity allows for a differentiation of traffic flows, both in speed of recovery and decision on whether or not recovery actions should be activated. In some cases finer granularity might also lead to more efficient capacity usage. One cause of this is that aggregated flows (at lower layer), poorly filled with working traffic flows, may have quite a lot of spare resources. A second cause is that finer granularity allows distributing flows on more alternative paths.

- *Single layer survivability combinations:* A number of variants considering recovery mechanisms in different layers can also be included under the single layer survivability. The argument is that for each failure scenario, the responsibility to recover traffic is the situation on only one layer. A variant to the above is *survivability at the lowest detecting layer*. That is in a multi-layer configuration, the (single) layer detecting the (root) failure is the only layer taking any recovery actions. Hence, a failure in a higher layer node will then be detected and dealt with in that layer. A situation depicted in Figure 10 may still not be handled. Another variant is *survivability at the highest possible layer*. This addresses the situations where traffic flows are injected at different layers (e.g. combination of optical channels, SDH-based leased lines and IP traffic flows). A recovery strategy could be to deal with failures at the layers where the different traffic flows are injected.

5 Multi-Layer Survivability

A common design goal for a network with multiple technological layers is to provide the desired level of service in the most cost-effective manner. Multi-layer survivability may allow the optimisation of spare resources through the improvement of resource utilization by sharing spare capacity across different layers. Coordination during recovery among different network layers might necessitate the development of a vertical hierarchy. The benefits of proving survivability mechanisms at multiple layers and the optimisation to the overall approach must be weighed with the associated cost and service impacts.

A default coordination mechanism for inter-layer interaction could be the use of nested timers and current SDH fault monitoring, as has been done traditionally for backward compatibility. Thus, when lower-layer recovery happens in a longer time period than higher-layer recovery, a hold-off timer is utilized to avoid contention between the different single-layer survivability schemes. In other words, multiplayer interaction is addressed by having successively higher multiplexing levels operate at a protection/restoration time scale greater than the net lowest layer. This can impact the overall time to restore service. For example, if SDH protection switching is used, MPLS recovery timers must wait until SDH has had time to switch. Setting such timers involves a trade-off between rapid recovery and creation of a race condition where multiple layers are responding to the same fault, potentially allocating resources in an inefficient manner.

In other configurations where the lower layer does not have a restoration capability or is not expected to protect, say an unprotected SDH linear circuit, then there must be a mechanism for the lower layer to trigger the higher layer to take recovery action immediately. This difference in network configuration means that implementations must allow for adjustment of hold-off timer values and/or a means for the lower layer to immediately indicate to a higher layer that the fault has occurred so that the higher layer can take restoration or protection actions.

Moreover, faults at higher layers should not trigger restoration or protection actions at lower layers.

The fact that there are a number of layers present raises the question of whether these could be utilised in a coordinated manner to improve the recovery schemes. Three main classes of schemes are:

- *Uncoordinated approach:* This may be considered as the simplest approach, that is to install recovery mechanisms in several layers, but without any coordination between them. An advantage is that it is simple to install and operate. A main disadvantage is that each recovery mechanism requires spare resources. This implies that spare resources are seen for each layer (and the overall situation could be fairly low resource utilisation). Moreover, more extra traffic (that is pre-emptable and unprotected) could be disrupted during a failure. An example (Figure 12) is where the lower layer switches traffic from a working link to a recovery path and hence pre-empts extra traffic on that link. However, the upper layer may gradually route the traffic on another path (possibly also pre-empting additional traffic along that path). For the example in Figure 12 link $A - D$ fails and the lower layer moves traffic to links $A - B - D$. However, the upper layer may find that another path is better so the traffic is effectively moved onto links $A - C - E$ (potentially resulting in poor utilisation of links $A - B - D$).

- *Sequential approach:* Realising that some coordination would likely result in improved utilisation, different options are investigated. The next level of complexity is seen where the responsibility for recovery is handed over to the next layer when it is clear that the current layer is not able to fulfil the recovery task. In principle any sequence of layers can be thought of, however, two obvious ones are *bottom-up* and *top-down*. For the former the lowest detecting layer starts the recovery and traffic that cannot be resolved by this layer (e.g. due to capacity shortage) will be restored by a higher layer. This has the advantage that recovery actions are taken at the appropriate granularity and more complex secondary failures are only treated when needed. For the *top-down* approach the upper layer starts the recovery action and only if that layer is unable to restore all traffic are actions at lower layers initiated. An advantage of this is that higher layers may differentiate traffic due to require-

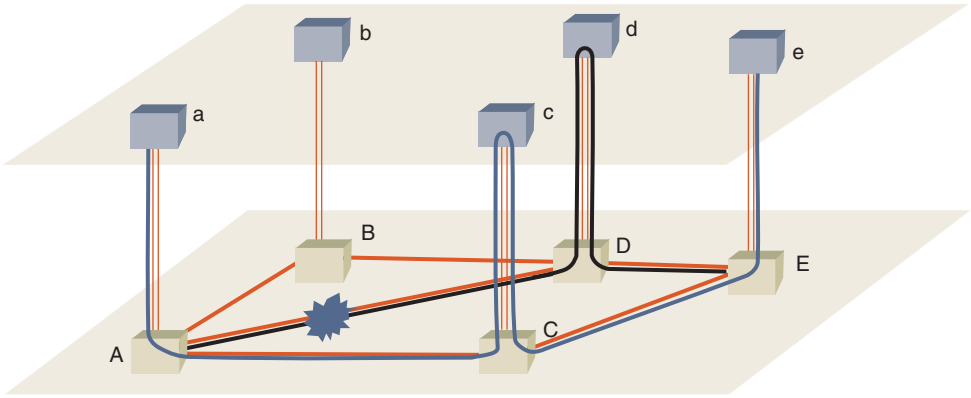


Figure 12 Illustration of uncoordinated recovery; lower layer routes on neighbour link, while upper layer may find that another path is better

Table 1 Overview of some recovery strategies (from [Coll02])

Factor	Survivability strategy				Preferred value
	Bottom layer	Bottom-up	Top layer	Integrated approach	
Switching granularity	Coarse	Coarse	Fine	Coarse	Coarse
Failure scenario	Simple	Simple	Complex	Simple	Simple
Recovery close to root	Yes	Yes	No	Yes	Yes
Capabilities, flexibility	Low	High	High	High	High
Failure coverage	Low	High	High	High	High
Coordination management	Low	High	Low	Low	Low
Resources	Low	High	Low	Low/High	Low

ments and hence restore high-priority traffic first. When coordinating multiple layers a few mechanisms have to be in place to avoid that the different layers take steps destroying for each other. One proposal is to introduce a *hold-off timer*. This timer is set when a layer starts to restore traffic. If the timer expires and the traffic is not adequately restored, the next layer initiates its actions. A disadvantage by this is that recovery actions at the next layer would always be delayed independent of the failure. To try to compensate for this an explicit *recovery token signal* can be exchanged between layers. Such a signal has to be included in the standards covering the interfaces between these layers.

- *Integrated approach:* This is based on a common integrated recovery scheme for the stack of layers. Hence a full overview of all the layers is needed in order to decide which layer and which actions to take. In principle, this approach is the more flexible one, however the algorithms have to be devised and implemented.

Table 1 summarises a number of recovery strategies by a number of selected factors. The column called preferred value indicates a typical (but not necessarily needed) value.

The aspects discussed above can be applied on the configuration of IP/MPLS over optical/SDH networks. In principle MPLS is able to provide fairly fast protection switching. Therefore, one may choose to promote recovery at the IP/MPLS layer (hence promoting survivability at the top layer). An argument backing this is that IP/MPLS may allow for less spare resources. This may come from packet switching being suitable for shaping spare capacity among pre-established backup paths. Then low-priority traffic can be dropped, e.g. by applying DiffServ.

In the case of peer models becoming a reality in the longer term, an integrated survivability approach may be likely. This is due to the single integrated control plane in the peer model.

Automation of wavelength establishment/release does not require a fixed logical IP/MPLS network design. Hence, the logical network design can be re-optimised during a failure situation. Then, if a router fails, an automatic reconfiguration of the logical IP/MPLS network could be undertaken to restore the traffic handling capability (instead of traditional rerouting or ensuring bi-connected IP/MPLS network designs).

A disadvantage of current IP/MPLS networks is that failure detection is based on the periodic exchange of *Hello* messages between adjacent routers. If no *Hello*s are received through an interface, the conclusion is that the opposite side of the interface in unreachable. This does not allow for a separation of route failures and link failures. Another concern with *Hello* messages is the detection time. Currently this may be in the range of 10 s and a failure is declared after missing 4 *Hello* messages. This would allow the optical/SDH to spend plenty of time restoring the traffic handling capability without even informing the IP/MPLS layer. As more capacity is introduced on the links (say 2.5 Gbit/s and 10 Gbit/s), more frequent *Hello* messages can be exchanged without stealing too much of the link capacity.

When discussing gains and drawbacks by utilising IP/MPLS and optics survivability mecha-

nisms, the observations would naturally depend on the network configuration looked at. Basically, using cost of links and nodes, it is expected that dedicated protection is more expensive than shared protection. Moreover, this difference may be bigger for local protection. This could be caused by the fact that sharing two (or more) backup paths using the same resource is only possible if the two working segments do not overlap (here a segment is a path in the case of path protection or a local loop-back of a link in the case of link protection and two links in case of node protection). For local protection, the working segments are in general shorter than for path protection or local loop-back implying lower probability of working segments' overlapping and hence a higher probability that sharing between the two backup paths is indeed allowed. This is comparing one or two links versus a complete path. Hence, the relative difference between dedicated and shared protection in terms of capacity requirements will be more substantial for local protection than for path protection or local loop-back.

It is also expected that the network topology has a significant impact, in particular for the local protection. That is, the topology with the smallest nodal degrees suffers the most from dedicated protection (node degree = number of links divided by number of nodes). This can be caused by the fact that as a topology becomes sparser, backup paths tend to become longer and more overlapping (the extreme is a ring topology). This explains why dedicated protection is relatively more costly in sparse networks than in dense networks.

An overall observation, from a capacity point of view, is that MPLS recovery may allow for fewer resources. On the other hand, the switchover delay would likely be longer. This may have a severe impact on the traffic flows. An example of this is when a number of TCP flows are carried by an MPLS path that is switched over and share underlying resources with another MPLS path. Just before the failure, the TCP sources on both MPLS paths may be transmitting at high rates. Then immediately after the switchover congestion may lead to longer queueing delays and therefore back-off by the TCP sources. Depending on the duration of the switchover time, a number of the TCP sources on the failed MPLS path may have backed off. This may lead to a more gentle traffic handling of the TCP sources running on the restoring link. An observation in [Coll02] is that when the switchover delays are in the range 0 – 250 ms, the differences in the goodput do not differ all that much, especially when the number of TCP flows is large.

Typically, a restoration time requirement of 50 – 60 ms is seen referring to SDH. This is advocated by the fact that voice calls (circuit switched) may potentially be dropped if restoration takes longer than this. As SDH may detect a transmission failure in the range of 2.3 – 100 μs, there may be little assistance in the optical layer supporting restoration faster than 1 ms. The SDH layer would already potentially detect the failure and any further actions by the SDH could be held back until the optical layer has finished its actions (as described for sequential and integrated approach above).

6 Optical Layer Survivability

A protection taxonomy referring to the optical layer is given in Table 2. The terms refer to SONET. A protection scheme either operates at i) an aggregate WDM level, called line layer or optical multiplexer section (OMS) layer; or, ii) at individual wavelength level, called path layer or optical channel (OCh). As shown, protection is either dedicated or shared. When rings are used, these are called dedicated protection rings (DPRings) and shared protection rings (SPRings), respectively. Again in SONET, DPRings are called optical unidirectional path-switched rings (OUPSRs) or optical unidirectional line-switched rings (OULSRs), depending on whether they operate at the path layer or the line layer. Shared protection rings are called optical bidirectional line/path switched rings (OBLSRs or OBPSRs, respectively).

Ring-based survivability involves the use of bi-directional line switched rings (BLSRs) or unidi-

Table 2 Protection taxonomy for optical layer survivability, referring to SONET terms (from [Gers00])

Network topology	Line layer (OMS)		Path layer (OCh)	
	Dedicated protection	Shared protection	Dedicated protection	Shared protection
Point-to-point	1+1 linear APS (OMS-DP)	1:1/1:N linear APS (OMS-SP)	1+1 linear APS (OCh-DP)	1:1/1:N linear APS (OCh-SP)
Ring	OULSR (OMS-DPRing)	OBLSR (OMS-SPRing)	OUPSR (OCh-DPRing)	OBPSR (OCh-SPRing)
Mesh	(no standard term)	(no standard term)	1+1 OSNCP	1:N mesh protection

rectional path-switched rings (UPSRs) as self-protecting transmission system overlaid on the network topology. More recent variants of UPSR- and BLSR-based transport networks are optical path protection ring (OPPR) and optical shared protection ring (OSPR) variants in a WDM context. An important factor is that rings may use simple switching mechanisms, which permits restoration in about 50 – 60 ms, although by nature they require 100 % redundancy. In particular, the BLSR uses a working-to-protection loop-back switching mechanisms at the two nodes adjacent to a failure. In essence, this is identical to the switching mechanism employed for *p*-cycles (in WDM or SDH). According to [Stam00], in conventional multi-ring network designs, where the working fibres of channel groups are not fully utilizable, effective protection-to-working capacity ratios (capacity redundancy) can be 200 – 300 %. This may show that rings allow fast restoration but may not be capacity-efficient.

Mesh-based survivability is more capacity-efficient due to the fact that each unit of spare capacity can be used in several manners, i.e. for several failures. Hence, overall performance close to the ideal maximum-flow routing-efficiency can be realised. Mesh restoration has traditionally been based on cross-connect systems embedded in a mesh-like set of point-to-point transmission systems. This is one factor for slower restoration compared with ring systems. In addition comes the more general nature of solving a capacitated multiple-path rerouting problem for mesh networks compared with rings.

Hence, mesh and ring both have their strengths and weaknesses. Mesh networks tend to be applied in the long-haul networks where capacity on transmission links is expensive. Rings tend to be more cost-efficient in the metro areas where cost is dominated by terminal equipment and not so much with the link and through-connect (due to short links). A method named *p*-cycle has been promoted trying to merge the advantages of ring and mesh. *p*-cycles are based on the formation of closed paths (elementary cycles in graph theoretic terms), called *p*-cycles, in the protection capacity of a mesh-restorable network. They are formed in advance of any failure, out of the previously unconnected protection capacity units of a restorable network. A *p*-cycle protects both on-cycle and straddling failures (the latter refers to paths not following the cycle during normal working situation, but may follow the cycle during failure of any link not being part of the cycle). This implies that protection capacity on a *p*-cycle is more widely accessible, i.e. more highly shared for restoration compared with traditional rings. Still, the cycle formation allows for fast restoration. Moreover, a few cases (e.g. ref. [Stam00]) show that *p*-cycles can be about as capacity-efficient as mesh networks. Compared to rings, *p*-cycles also de-couple the routing of working and protecting paths by the fact that the working paths may be routed along the shortest paths, while protecting paths follow the cycle.

The reason for *p*-cycles being as fast as rings is that only two traffic-substituting connections are made for any working path failure to be restored. The two end nodes perform the switching only by transmitting a control signal as they know in advance which working-to-protection functions are needed for any given failure. *p*-cycles can be introduced, in principle, on optical, SDH, and MPLS layers.

It is quite common to find mesh physical topologies in the backbone networks. The SDH layer placed on that, however, is in most cases protected in the form of rings. These rings are interconnected to provide an overall network connectivity and protection. It is a question whether these schemes are less bandwidth efficient to protect a physical meshed network. Efficiency improvements in the area 20 – 60 % have been demonstrated by introducing mesh protection schemes, although these observations naturally depend on the topology, traffic patterns and protection scheme. Taxonomy of mesh protection variations including trade-offs between them as depicted in Figure 13.

Deploying a robust mesh protection scheme, a number of challenges have to be met (adapted from [Gers00]):

- A fast signalling mechanism is needed to propagate information about failure events to relevant nodes so they can reconfigure the switches to accommodate the recovery path.

- The recovery routing tables become large if pre-computed routes are used to expedite the protection and full flexibility in the choice of routes for protection is requested (to maximise the bandwidth efficiency). In particular maintaining large sets of routing tables is a non-trivial task, e.g. with regard to ensuring consistency.

- Failure propagation may be easier in a mesh network than in a ring topology. Larger portions of a network may therefore be affected. This is essential to consider as the protection scheme in itself may be inadequate, e.g. not considering all failure configurations that can occur (such as multiple failures), operator errors, software bugs, and so forth.

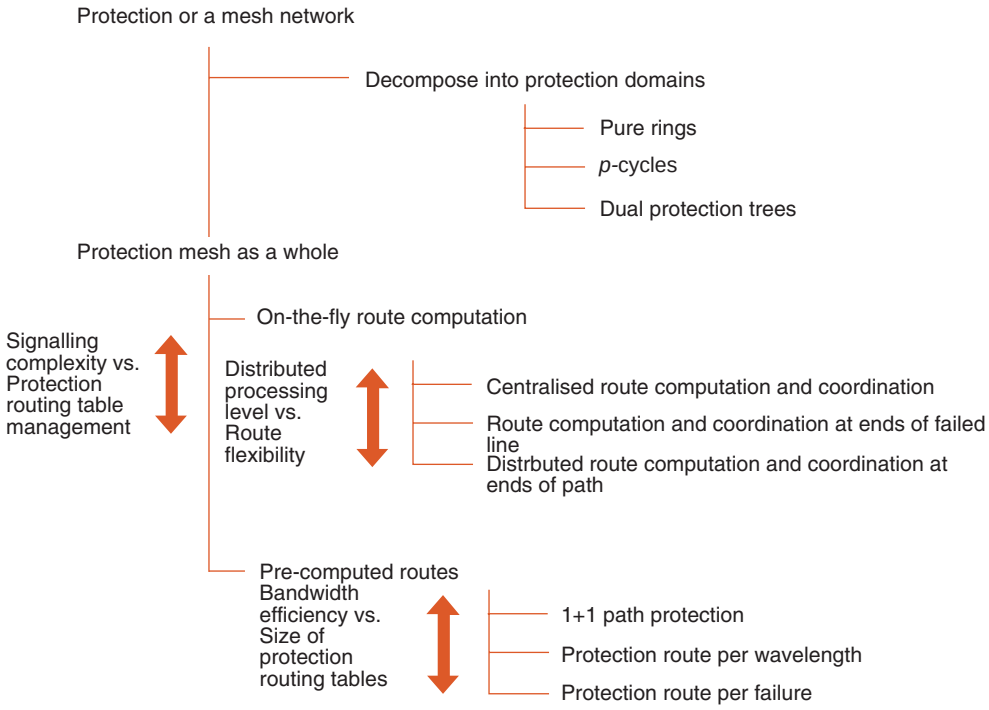


Figure 13 Taxonomy and trade-offs for mesh protection variants (from [Gers00])

- Making full use of the bandwidth efficiency may require sophisticated design algorithms and network planning tools, themselves adding to the complexity.

- Different variants of mesh protection schemes may imply that standardisation and interoperability are harder to achieve.

Still comparing mesh and ring topologies, a mesh scheme allows the entire network to be managed in an integrated manner. Moreover, interconnecting rings avoiding the single point of failure adds complexity to this scheme.

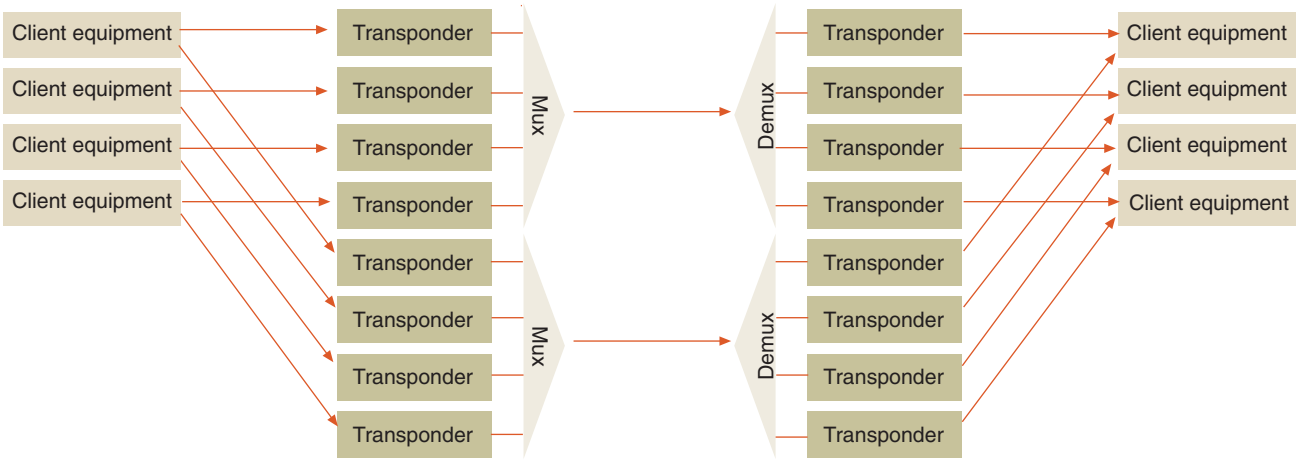
Introducing wavelength conversion may considerably improve the utilisation of available wavelengths. Avoiding full conversion, limited conversion capability is proposed to reduce the cost of optical nodes. However, this may increase the protection complexity. For example, without conversion in a ring, the same wavelength could simply be looped around the other way, while when wavelength conversion is introduced one has to check that the wavelength has not been used for other purposes on any hops.

7 Survivability Issues for Optical/WDM and Client Networks

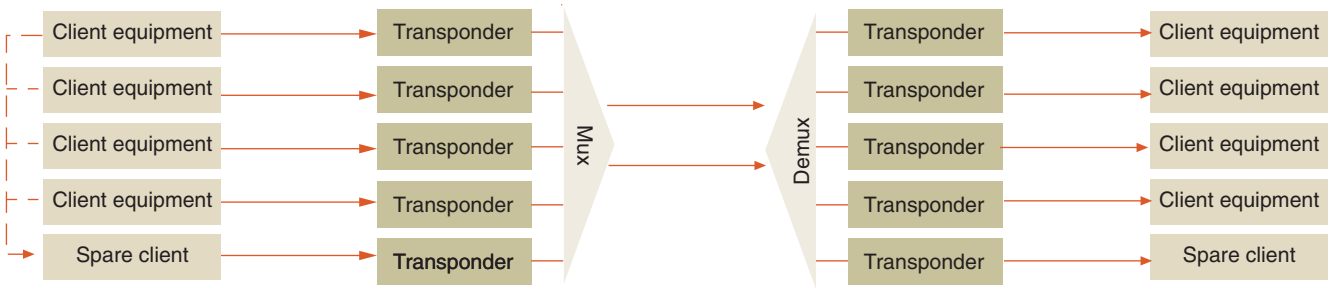
When discussing options for multi-layered networks, one may apply an approach of treating the neighbouring layers iteratively. Hence, one may start by examining the options for two layers. Here the lower layer can be considered as the optics, while the higher layer can be considered as the IP/MPLS combined layer. However, this description can also be further generalised.

Referring to configurations with clients over optical networks, a few protection schemes are depicted in Figure 14. In part a) clients have connections to duplicated transponder/optical systems. Introducing a spare client as shown in part b) allows for fewer transponders, although clients have to coordinate any switchover. The taxonomy given in Figure 14 does not concern itself with internal node implementation. Hence, there could be varying cost levels for implementing the different protection schemes. As transponders may be relatively expensive (most of the WDM terminal costs come from the transponders), there may be a benefit by having an arrangement with fewer transponders. This may favour a post-split arrangement versus a pre-split arrangement as shown in part d) and c), respectively.

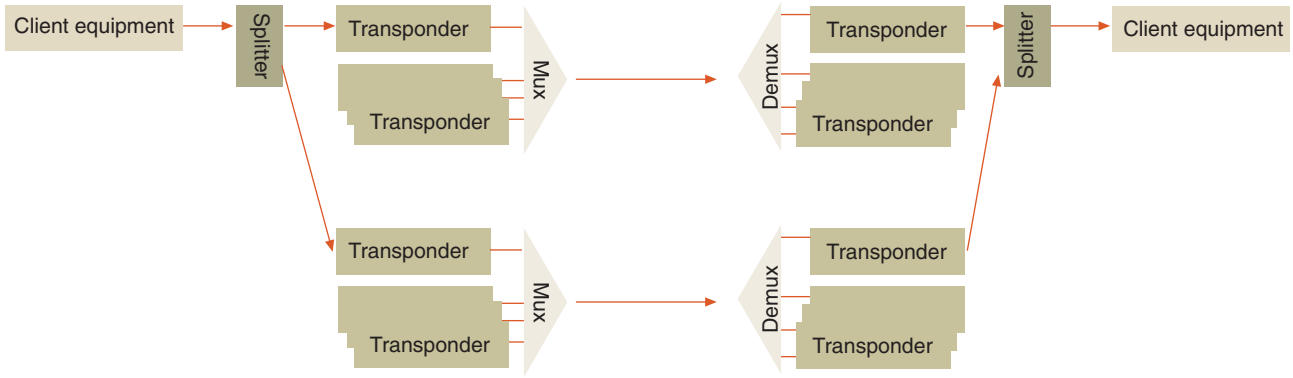
One approach to achieve fewer transponders is to introduce spare arrangements. Three cases are depicted in Figure 15. These are mostly for protecting failures in transponders. In part a) a transponder on a designated wavelength is used as spare. In case of failure of a working transponder, the signal link is routed to the spare transponder through its switch. The same operation takes place at the receiver end, where the signal is routed to the corresponding client equipment. When a tuneable laser/detector is used in the spare transponder no designated wavelength for spare is needed. For a failure, the signal from the client equipment is routed to the



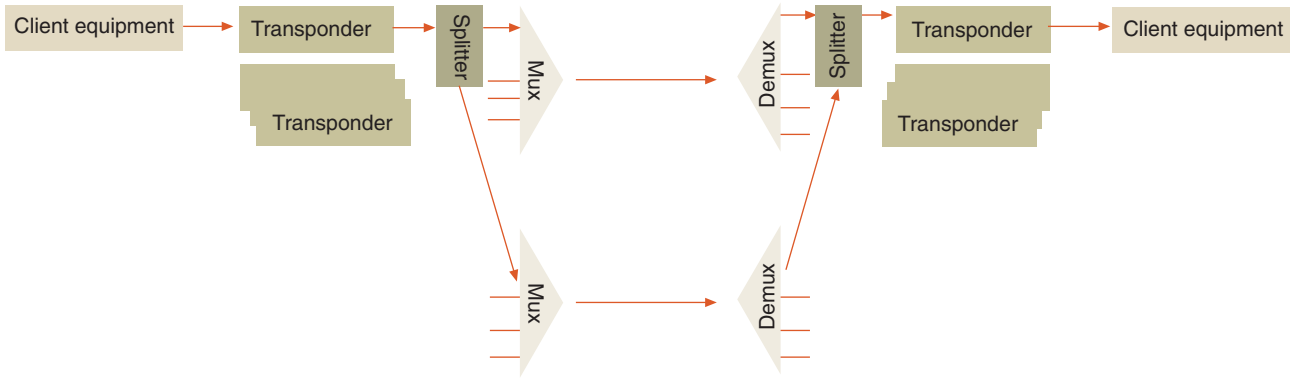
a) Protection using 1 + 1 client scheme



b) Protection by 1: N client protection scheme and 1 + 1 optical line protection

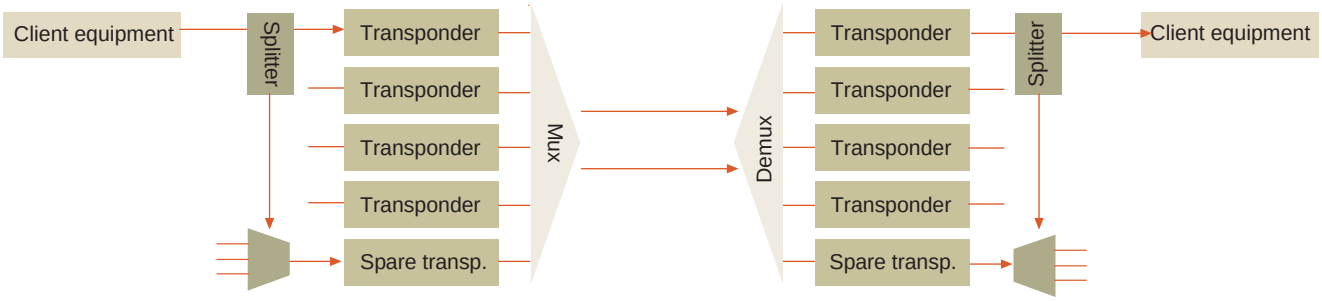


c) Protection by transponder pre-split

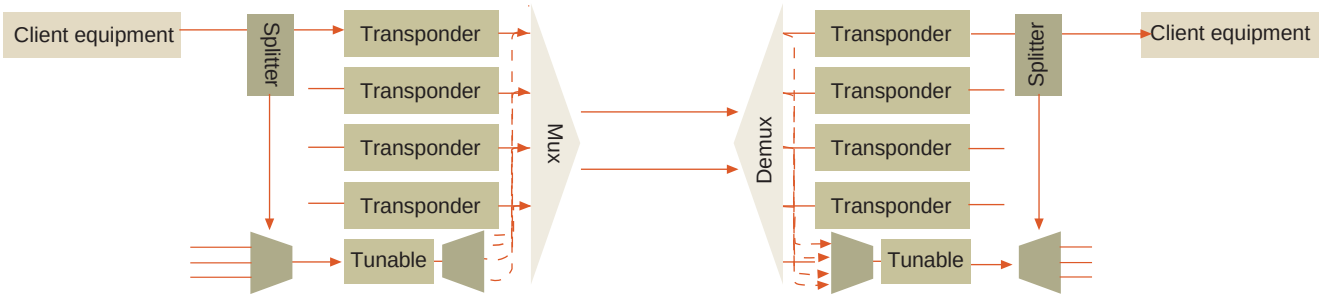


d) Protection by transponder post-split

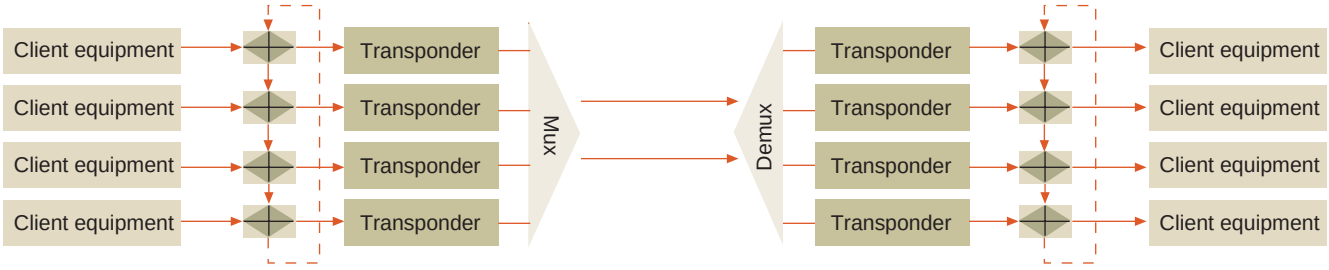
Figure 14 Illustrating a few protection arrangements for client and optical layers



a) Fixed spare



b) Tunable spare



c) Provisionable spare without additional transponder

Figure 15 Three transponder configurations

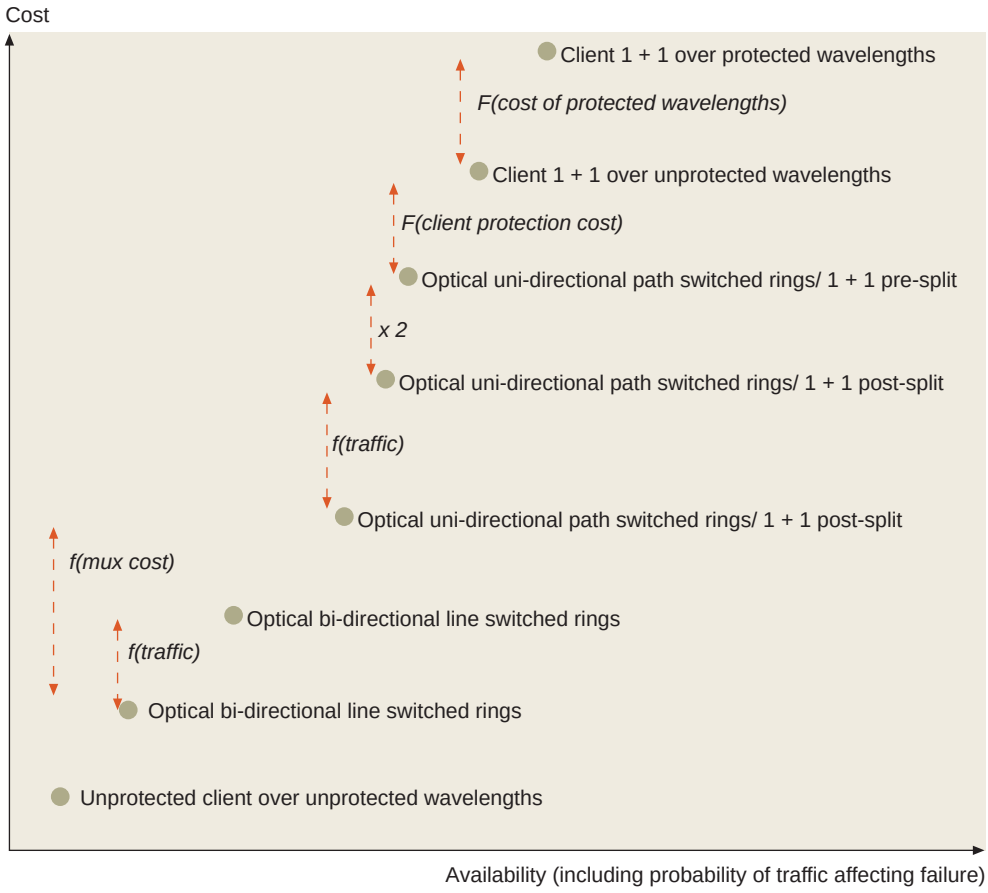
tuneable spare, which is then tuned to the wavelength of the previously used transponder. A similar operation may take place at the receiver end, although this operation is not mandatory, and may also allow for additional protection. In part c) a series of small optical switches is introduced. Under normal situation the switches connect the signal straight through, although in case of failure, the signal is routed to the next switching element until a spare transponder is found. Hence, no dedicated transponder is spare. However, for this part the overall traffic handling capability may be reduced during a failure situation.

The three schemes may be compared as follows: The fixed spare scheme is the least demanding scheme from the hardware and software point of view. A complicating factor is that when a failure occurs, both ends have to coordinate the switch to the protect wavelength using a fast signalling protocol, e.g. similar to Automatic Protection Switching (APS) in SDH. In a mesh network the coordination grows in complexity and may require a few additional wavelengths for protection only.

Tuneable spares do not need the coordination between the two ends because as the same wavelength is used, the receiver may see no changes. A downside is that tuneable lasers are considered expensive. Moreover, the scheme cannot handle other failures in the network related to wavelengths, such as failures in wavelength filters.

For the provisionable case, flexibility is allowed, e.g. by defining protection groups. Multiple transponder failures may also be supported, although then with further decrease in traffic handling capability or introducing more spare transponders. The small optical switches must be controlled by a signalling protocol, which may be adding complexity.

Figure 16 Qualitative illustration of different protection schemes (adapted from [Gers00])



Introducing more advanced protection schemes will also add to the cost. However, the cost of the implementations within a node has to be balanced with the transport cost, and the achieved availability. A qualitative comparison is shown in Figure 16.

A set of requirements on survivability and hierarchy, both current and near-term, are described in [RFC3386]. These are summarised as:

A. Survivability requirements:
- Need to define a small set of interoperable survivability approaches in packet and non-packet networks
- Suggested survivability mechanisms including
  - 1:1 path protection with pre-established backup capacity (non-shared)
  - 1:1 path protection with pre-planned backup capacity (shared)
  - local restoration with repairs in proximity to the network fault
  - path restoration through source-based rerouting
- Timing bounds for service restoration to support voice call cut-off (140 ms to 2 s), protocol timer requirements in premium data services, and missing critical applications
- Use of restoration priority for service differentiation

B. Hierarchy requirements

B.1 Horizontally oriented hierarchy (intra-domain)
- Ability to set up many LSPs in a service provider network with hierarchical IGP, for the support of layer 2 and layer 2 VPN services
- Requirements for multi-area traffic engineering need to be set up to provide guidance for any necessary protocol extensions.

B.2 Vertically oriented hierarchy; the following functionality for survivability is common on most routing equipment today:
- Near-term need in some loose form of coordination and communication based on the use of nested hold-off timers, instead of direct exchange of signalling and routing between vertical layers

- Means for an upper layer to immediately begin recovery actions in the event that a lower layer is not configured to perform recovery

C. Survivability requirements in horizontal hierarchy:
   - Protection of end-to-end connection is based on a concatenated set of connections, each protected within their area.
   - Mechanisms for connection routing may include: i) a network element that participates on both sides of boundary; ii) a route server.
   - Need for inter-area signalling of survivability information i) to enable a "least common denominator" survivability mechanism at the boundary; ii) to convey the success of failure of the service restoration action, e.g. if a part of a "connection" is down on one side of a boundary, there is no need for the other side to recover from failures.

In a survivable network design, spare capacity and diversity must be built into the network from the beginning to support some degree of self-healing whenever failures occur. A common strategy is to associate each working entity with a protection entity having either dedicated resources or shared resources that are pre-reserved or reserved-on-demand. According to the methods of setting up a protection entity, different approaches to providing survivability can be classified. Generally, protection techniques are based on having a dedicated protection entity set up prior to failure. Such is not the case in restoration techniques, which mainly rely on the use of spare capacity in the network. Hence, in terms of trade-offs, protection techniques usually offer fast recovery from failure with enhanced availability, while restoration techniques usually achieve better resource utilization.

A 1+1 protection architecture is rather expensive since resource duplication is required for the working and protection entities. It is generally used for specific services that need a very high availability.

A 1:1 architecture is inherently slower in recovering from failure than a 1+1 architecture since communication between both ends of the protection domain are required to perform the switch-over operation. An advantage is that the protection entity can optionally be used to carry low-priority extra traffic in normal operation, if traffic pre-emption is allowed. Packet networks can pre-establish a protection path for later use with pre-planned but not pre-reserved capacity. That is, if no packets are sent on to a protection path, no bandwidth is consumed. This is not the case in traditional transmission networks like most optical or TDM networks, where path establishment and resource reservation cannot be decoupled.

In the 1:n protection architecture, traffic is normally sent on the working entities. When multiple working entities have failed simultaneously, only one of them can be restored by the common protection entity. This contention could be resolved by assigning a different pre-emptive priority to each working entity. As in the 1:1 case, the protection entity can optionally be suited to carry pre-emptable traffic in normal operation.

While the m:n architecture can improve system availability with a small cost increase, it has rarely been implemented or standardised.

When compared with protection mechanisms, restoration mechanisms are generally more frugal as no resources are committed until after the fault occurs and the location of the failure is known. However, restoration mechanisms are inherently slower, since more must be done following the detection of a fault. Also, the time it takes for the dynamic selection and establishment of alternate paths may vary, depending on the amount of traffic and connections to be restored. It is also influenced by the network topology, technology employed, and the severity of the fault. As a result, restoration time tends to be more variable than the protection switch time needed with pre-selected protection entities. Hence, in using restoration mechanisms, it is essential to use restoration priority to ensure that service objectives are met cost-effectively.

Once the network routing algorithms have converged after fault, it may be preferable in some cases to re-optimise the network by performing a reroute based on the current state of the network and network policies.

[RFC3386] states some typical requirements on protection switch time or restoration time:

- Best effort data: recovery of network connectivity by rerouting at the IP layer would be sufficient
- Premium data service: need to meet TCP timeout or application protocol timer requirements
- Voice: call cut-off is in the range of 140 ms to 2 s (the time that a person waits after interruption of the speech path before hanging up or the time that a telephone switch will disconnect a call)

- Other real-time services (e.g. streaming, fax) where an interruption would cause the session to terminate

- Mission-critical applications that cannot tolerate even brief interruptions, for example real-time financial transactions.

It then goes on to propose some timing bound for different survivability mechanisms (times excluding signal propagation):

- 1:1 path protection with pre-established capacity 100 – 500 ms

- 1:1 path protection with pre-planned capacity 100 – 750 ms

- local restoration 50 ms

- path restoration 1 – 5 s

In order to ensure that requirements for different applications can be met, restoration priority should be implemented. This determines the order in which connections are restored.

8 IP over Optics

The continuing growth of IP-based traffic (both Internet-related and for other IP-based services), advocates more streamlining of the network layers. Hence, IP-over-optics with few intermediate network layers has been promoted by several parties. The fact that available capacity on a fibre cable increases, a single cable cut may affect huge traffic volumes. Therefore special attention is paid to the survivability of such networks.

As claimed in [Coll02] a starting point for several of the incumbent operators is an IP/ATM/SDH/WDM multi-layer configuration, as depicted in Figure 17. It should be noted that typically there is a range of "clients" to each of the layers. Considering the arrangement as seen from the IP point of view, [Coll02] states two main drawbacks by carrying IP over ATM:

- The ATM cell size; introducing the so-called cell tax (overhead of ATM cell header and adaptation layer), and that many ATM cells must be transported and processed to carry a single IP packet.

- The additional (ATM) layer to be maintained and managed.

Naturally, ATM also has a number of benefits; such as its connection orientation (when needed) providing extra support for traffic engineering, and decoupling of control plane (e.g. routing) and user plane (e.g. forwarding).

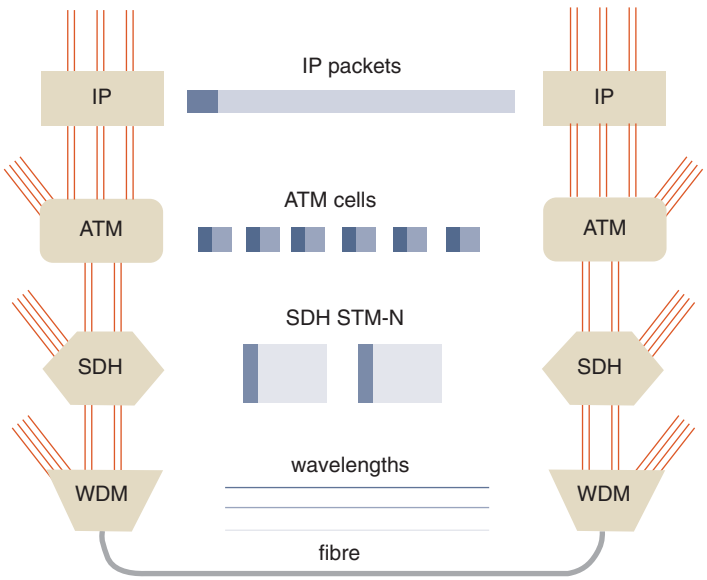
However, capabilities such as Traffic Engineering defined for IP imply that several of the arguments in favour of using ATM can be achieved by other means. Introducing MPLS has been promoted to cover several of the objectives.

Furthermore, as traffic intensities and router interface rates increase, dedicated wavelengths could be likely, allowing the IP/MPLS traffic flows to bypass both ATM and SDH networks. (A distinction should be made between the networks and the "framing" protocols, e.g. SDH framing may still be applied for a wavelength between two IP routers.) This also means that cross-connection functionality offered by the SDH network may not be utilised for such IP/MPLS traffic flows. On the other hand, optical network elements might be introduced to provide such functionality if needed.

Another factor is that most transport network configurations so far tend to be rather static. That is, the connections are commonly established/changed by manual operations through a management system. In case the IP/MPLS traffic flows show highly dynamic behaviour, more automatic procedures are asked for in order to adapt the transport capacity in the different directions to the variations in the IP/MPLS network. Much research effort is currently put into solving how this provisioning process can be automated. This would expectedly require some control plane functions.

Broadly speaking, two main (opposite) models have emerged for an Automatic Switched Optical Network (ASON). One is an overlay of an Automatic Switched (optical) Transport Network (ASTN) (ASTN can be seen as a generalisation of ASON). In the overlay model both the

Figure 17 Illustration of an IP/ATM/SDH/WDM configuration



transport network and its client networks have a separated and independent control plane. This is described in ITU-T recommendations.

The IETF seems to target a peer model with the Generalised MPLS (GMPLS) concept. This concept stems from MPλS, where the underlying idea was that a wavelength is a "label" like any other layer and therefore the MPLS concept can be adopted in the optical domain to serve the need for fast and automatic provisioning of wavelengths. GMPLS is more generic in the sense that it considers any type of label: a header string for packet switch capable LSR (PSC-LSR), a time slot for a TDM switch capable LSR (TSC-LSR, e.g. by SDH), a wavelength for a lambda-switch capable LSR (LSC-LSR, e.g. an optical cross-connect), or a fibre in a fibre-switch capable LSR (FSC-LSR). These are described in Section 3.2.

Although both client and transport networks may have their own separate and independent (G)MPLS control planes, an integration of these controls planes into a singe one seems tempting. This then results in the so-called peer model. The following advantages are seen for the peer model:

- Avoiding duplication of control plane functionality in distinct layers;
- Avoiding the requirements of standardisation of a User-Network Interface between IP/MPLS routers and optical network components.

A basic drawback, however, is the fact that integration and compatibility among multiple client type networks seem hard, as well as the requirement that all necessary information of the optical network (e.g. the topology) will be accessible in the client domain.

9 All Protection at IP Layer

Some people claim there is a trend that most of the protection is located at the IP/MPLS layer. This is backed by arguments such as:

- Protection at SDH is considered as expensive as seen by the IP/MPLS layer (one pays for a back-up capacity that is rarely used).
- Supporting differentiated protection schemes implies that such mechanisms must be located at the IP/MPLS layer.
- Several people would remove the SDH network as such for carrying the IP/MPLS traffic (although SDH framing may still be applied).

A consequence of this is that links between IP routers should be disjoint in order to handle failures. In this perspective, one can consider the IP network as consisting of a set of logical links that are mapped onto a set of physical links in the fibre/WDM network. In the WDM network the logical links are assigned a wavelength, or possibly a sequence of wavelengths if wavelength conversion is present. In this way several logical links could go on the same fibre (or conduit), which is not a preferred situation as a single cable cut could affect several links.

When a failure occurs in the optical equipment, one should still have a connected topology at the IP layer. Hence, for every possible failure, alternative operational logical links (or combinations of logical links) must exist so that alternative routes can be found by the IP routing protocol.

A requirement may then be that the alternative link offers about the same delay as the normal operating link. For example, a voice-over-IP application may suffer if the end-to-end delay increases by more than 50 ms after switching over to the alternative link. This comes in addition to the fact that the primary operating link usually follows the shortest path and thereby has the shortest delay.

Hence, designing the robust IP network, both disjoint links as well as delay concerns have to be included. The latter would naturally be related to the link loads. Having two different types of requirements commonly asks for a trade-off (or combination) to be included. The basic problem can be formulated as an Integer Programming problem (integer refers to the fact that one must choose from certain capacities and that logical links must be mapped onto physical links). Generally this is known as an NP-complete problem, where delay constraints, limited number of wavelengths and fibre/WDM network structure have to be considered.

10 MPLS-Based Protection

Several of today's networks are firstly built for connectivity, and secondly to support one service class. The migration of real-time and higher-priority traffic to IP networks means that modern IP networks increasingly carry mission-critical business data and must therefore provide reliable transmission. Current routing algorithms, despite being robust and survivable, can take a substantial amount of time to recover from a failure, which can be in the order of several seconds to minutes and can cause serious disruption of service in the interim. This is unacceptable for many applications that require a highly reliable service, and has motivated network providers to give serious consideration to the issues of network survivability.

Multi Protocol Label Switching (MPLS) can be utilised to integrate forwarding based on label-swapping of a local label and hence allows more advanced forwarding mechanisms to be applied. One area of these mechanisms is QoS, where MPLS can be used to transport data reliably and efficiently and also to support a range of service classes. One of the drawbacks with current IP routing algorithms is the relatively long time to recover from a fault, which is in the order of seconds or minutes. This may well cause a disruption of service as observed by the applications/users. As more business-critical applications/usages and real-time demanding applications are placed on IP networks, such long recovery times may be unacceptable. It is often seen that applications require recovery times down to tens of milliseconds; examples being virtual leased lines, voice traffic, real video transfers, stock exchange data. For several of these applications SLA conditions have been stated which do not allow for long recovery times.

Generally, operators would provide the fastest and best protection mechanisms that can be provided at a reasonable cost. However, the higher level of protection, the more resources consumed. Hence, it is naturally expected that operators will offer a range of service levels. MPLS-based recovery should give the flexibility to select the recovery mechanism, choose the granularity of which traffic to protect and also to choose the specific types of traffic that are protected in order to give operators more control over that trade-off.

Based on the above, one motivation for MPLS is the notion that the current IP routing algorithms have limited potential for improving the recovery times. Introducing recovery mechanisms on the MPLS level, or partly enhance the current routing algorithms with MPLS aspects may alleviate the underlying restriction.

Another motivation for introducing recovery mechanisms operating on the MPLS level is the presence of MPLS in several of the IP networks. Moreover, MPLS is commonly promoted as a base solution for IP-based transport networks. Hence, it is natural that MPLS provides protection and restoration of the traffic that has such requirements.

MPLS may facilitate the convergence of network functionality on a common control and management plane. Further, a protection priority could be used as a differentiating mechanism for premium services that require high reliability.

Terms

The intent of protection is to protect against any link or node fault on a path or a segment of a path. The protected path is commonly referred to as a *working path*. A set of paths used for carrying traffic on the working path in case of failure along it, is commonly referred to as *recovery paths*. An LSR that can select to use the working path or the recovery path is called a *Path Switch LSR* (PSL). The LSR where the working and recovery paths meet has to decide which packets that should be forwarded. This LSR is commonly called a *Path Merge LSR* (PML).

As LSPs are unidirectional and recovery may well require notification of faults being carried to the LSR responsible for switch-over to the recovery path, a mechanism must be provided for propagating fault and repair indications from the point of occurrence of the fault back to that LSR. These are called *Fault Indication Signal* (FIS) and *Fault Repair Signal* (FRS).

10.1 Recovery Principles

A nomenclature for recovery principles is also drafted in [Shar02]. Firstly it is noted that MPLS-based recovery refers to the ability to quickly and completely restore capabilities for serving traffic affected by a fault in an MPLS network. The fault may be detected at the IP layer or in lower layers. To some extent the different terms can be arranged in a “nomenclature tree” as depicted in Figure 18.

The issues and corresponding options are:

1. Configuration of recovery:
   - Default recovery: no MPLS-based recovery is enabled; hence traffic must be recovered using IP layer or lower layer mechanisms.
   - Recoverable: MPLS-based recovery is enabled; hence one or more MPLS recovery paths are identified for a working MPLS path.
2. Initiation of path set-up:
   - Pre-established: recovery paths are established prior to any failure of the working path. This is also called protection switching.
   - Pre-qualified: a path created for other purposes is also designated as a recovery path for a working path.
   - Established on-demand: a recovery path is established after a failure on its working path has been detected and notified to the Path Switched LSR (PSL). This is also called re-routing option.

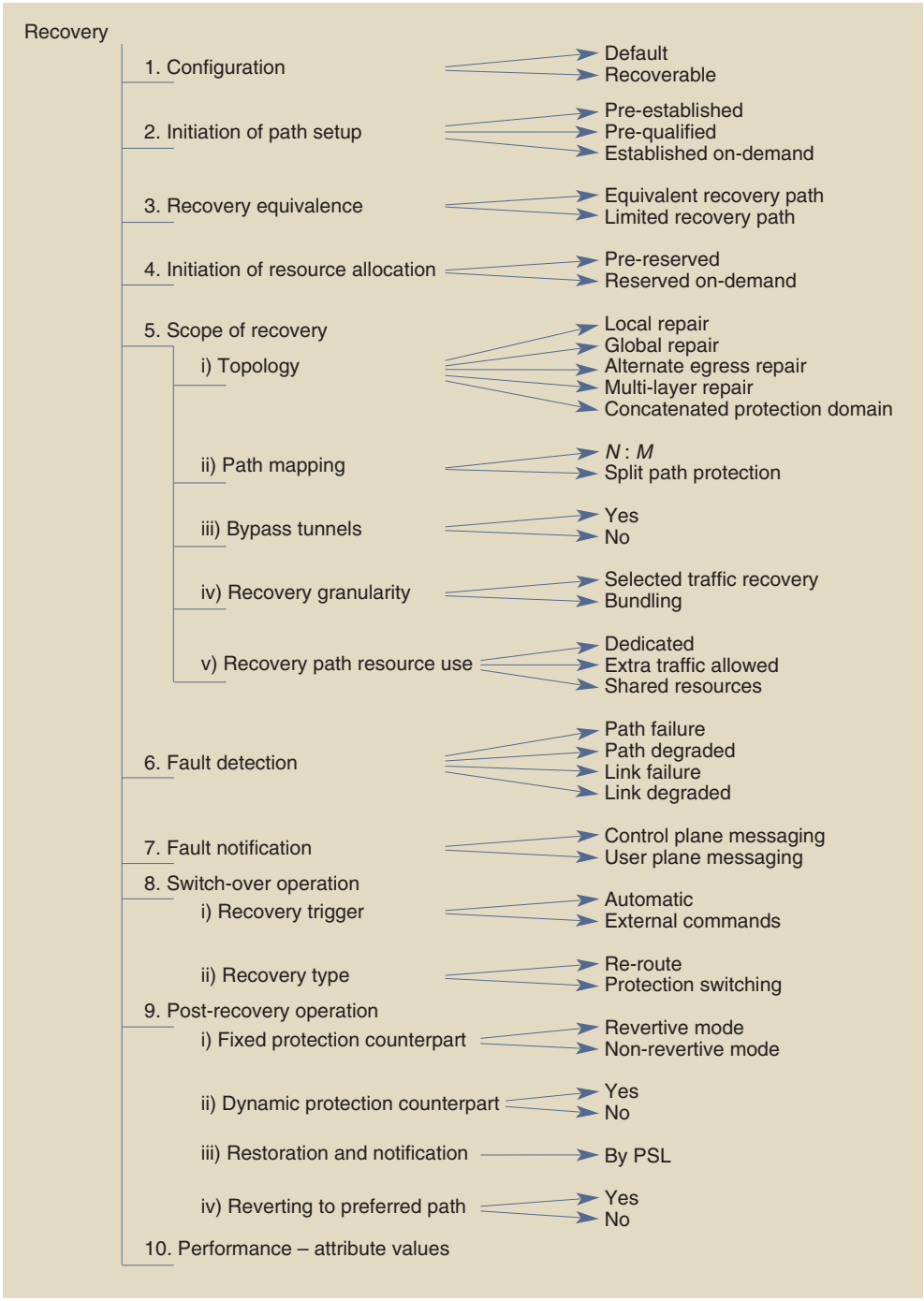


Figure 18 Recovery terms (adapted from [Shar02])

3. Recovery equivalence:
   - Equivalent recovery path: recovery paths are capable of replacing the working path without degrading the service levels.
   - Limited recovery path: recovery paths do not replace the working path without degrading the service levels.

4. Initiation of resource allocation:
   - Pre-reserved (applies only to protection switching): recovery path has reserved resources on all hops along its route.
   - Reserved on-demand: recovery path reserved the required resources after a failure on the working path has been detected and notified to the PSL and before the traffic is switched over to the recovery path.

5. Scope of recovery:
   i) Topology:
   - Local repair: node immediately upstream of the fault is the one to initiate recovery. This is either a) Link recovery/restoration/recovery path is configured to route around a certain link (working and recovery paths disjoint only at a certain link); or, b) Node recovery/restoration: recovery path routed around a neighbour node (working and recovery path disjoint at a certain node).

- Global repair: Point Of Repair (POR) is at the ingress of a path (segment), hence also covering end-to-end path recovery. In case the recovery path is completely disjoint from the working path, this protects against any link and node faults.

- Alternate egress repair: the recovery path may have a different egress point from the working path.

- Multi-layer repair: several network layers are managed together in order to achieve adequate protection of the traffic flows (ref. Chapter 5).

- Concatenated protection domains: recovery provided for traffic flows crossing a number of network domains. As these domains may apply different recovery mechanisms it is still important that the end-to-end protection is adequate.

ii) Path mapping (mapping of traffic from working path to set of recovering paths):
- *N:M* protection: up to *N* working paths are protected using *M* recovery paths. Common configurations are 1:1 and *N*:1 protection.

- Split path protection: multiple recovery paths carry the traffic of a working path according to configurable splitting ratios. Typically, this is used when no single recovery paths can carry the entire traffic load of the working path. This may require the Path Merge LSR, PML(s) to correlate the traffic arriving on multiple recovery paths with the working path. In principle a 1+1 configuration could be seen as a special case where the traffic is carried on two paths between the PSL and PML, and the PML correlates the packets arriving along the two paths.

iii) Bypass tunnels: creation of a tunnel (MPLS path) for a Path Protection Group (PPG) between a PSL and PML. Then one MPLS path – the tunnel – carries a number of paths by label stacking, thereby making all the "inner paths" transparent to intermediate LSRs.

iv) Recover granularity (refers to the ratio of traffic requiring protection (e.g. ranging from a fraction of a path to a bundle of paths):
- Selective traffic recovery: protection of a fraction of traffic within the same path, referred to a protection traffic portion (PTP).

- Bundling: grouping several working paths allowing for simultaneous recovery. The logical bundling of these working paths routing between the same PSL and PML is called a Protected Path Group (PPG). When a fault occurs on the working path carrying the PPG, the PPG as a whole can be protected.

v) Recovery path resource use (referring to optional use of pre-reserved recovery paths when that path is not in use – no fault on the working path):
- Dedicated resource: the recovery path resources may not be used by other traffic flows.

- Extra traffic allowed: the recovery path may carry other traffic flows when that capacity is not needed by traffic on the working path, e.g. when there is no fault on the working path. Extra traffic can be displaced whenever the recovery path resources are needed to carry the working path traffic.

- Shared resource: recovery path resources are used by several working paths (where all of the working paths are not assumed to fail simultaneously).

6. Fault detection
   - Path Failure (PF): an indication to the MPLS mechanisms that the connectivity of the path is lost. This could, for example, be identified by path continuity test or liveness messages/probes.

   - Path Degraded (PD): an indication to the MPLS mechanisms that the service level on the path is unacceptable. This may typically be done local to an LSR, e.g. by receiving many packets with incorrect labels or TTL.

   - Link Failure (LF): an indication from a lower layer to MPLS recovery mechanisms that the link has failed. This may for example be Loss Of Signal (LOS) from SDH.

   - Link Degraded (LD): an indication from a lower layer to MPLS recovery mechanisms that the link gives an unacceptable service level.

7. Fault notification (when a fault is severe enough to require path recovery, an informed node must take appropriate actions. In

case the node is not capable of initiated direct actions, notification of the faults should be sent to the POR):

- Notification by control plane messaging: relaying hop-by-hop along the path until a POR is reached.

- Notification by user plane messaging: typically sent downstream to the PML which then takes corrective action (e.g. as POR for 1+1) or communicate with a POR (by control plane messaging or user plane messaging for a bi-directional LSP).

8. Switch-over operation:
   i) Recovery trigger: MPLS protection switching may be initiated because of automatic inputs of external commands. Automatic inputs include fault notifications received (when several notifications are received for the same fault case, these should be correlated).

   ii) Recovery action: a POR executes the recovery actions, either re-routing or protection switching as described earlier.

9. Post-recovery operation (refers to decisions made after traffic flows on the recovery path):
   i) Fixed protection counterparts:
      - Revertive mode: when the fault on the (original) working path is cleared, the traffic is switched back to this path, from the (original) recovery path.

      - Non-revertive mode: a switch back to the (original) working path is not immediately carried out even after the fault has been cleared. When the fault in the (original) working path is cleared, that path may be used for other purposes, or be defined as a recovery path for the one where traffic has been switched on to.

   ii) Dynamic protection counterparts: when the network reaches a stable state after the fault, traffic on the recovery path may be switched over to a different preferred path (not the original working path).

   iii) Restoration and notification: reversion is performed by the PSL by receiving notification that the working path is repaired or that a new working path is established.

   iv) Reverting to preferred path (or controlled rearrangement): typically this involves a "make before break" switching.

10. Performance: the recovery paths will typically have a number of attributes, including i) resource class attributes (in case of equivalence these are the same as the working path), ii) path priority attributes, iii) path pre-emption attributes.

When several MPLS-based recovery schemes are compared, a set of criteria has to be defined. [Shar02] proposes the following criteria:

- Recovery time: this time can be defined as the interval from the fault occurred until a recovery path is carrying the relevant traffic. Hence, it is the sum of the fault detection time, hold-off time, notification time, recovery operation time and traffic restoration time.

- Full restoration time: this time is defined as the interval from the fault occurred until a permanent restoration is reached. A permanent restoration reflects the state when traffic flows on links capable of or having been engineered to handle traffic as expected. This time may differ from the recovery time depending on when limited recovery paths are used (as opposed to equivalent paths).

- Set up vulnerability: the amount of time a set of working paths is left unprotected during recovery (such as path computation and set-up).

- Back-up capacity: capacity needed during recovery (also depending on traffic characteristics as well as protection plan selection algorithms and signalling and re-routing methods.

- Additive latency: additional delays along the recovery path compared with the working path.

- Quality of protection: includes relative and absolute survivability. Relative means that the traffic flow in question has the same service class as other flows (contending for the same resources). Absolute means that explicit guarantees may be given to the protected traffic.

- Re-ordering: packets may arrive in a different order than sent during recovery (the same may happen during restoration)

- State overhead: the amount of information needed to maintain the paths

- Loss: a certain amount of packet loss may be introduced during switch-overs.

- Coverage: representing various types of failovers: i) Fault types (link faults, node faults, degraded service level, etc.), ii) Number of concurrent faults, iii) Number of recovery paths, iv) Percentage of coverage (certain percentage of fault types), and v) Number of pro-

tected paths compared with the total number of paths (e.g. given as n/N where n is the number of protected paths and N is the total number of paths).

10.2 MPLS Recovery Objectives

As stated in [Shar02], MPLS-based recovery is motivated by a number of arguments:

- Ability to increase network availability by enabling faster response to faults than possible with mechanisms only related to the IP layer.
- IP rerouting may often be too slow for a core MPLS network that needs to support recovery times smaller than convergence times of IP routing protocols.
- Utilising lower layers (e.g. optical channels – sometimes referred to as layer 0, and SDH – sometimes referred to as layer 1) may result in wasteful use of resources. That is, the granularity of the resources may be too coarse for the group of traffic flows that should be protected.
- IP-traffic can be put directly over WDM channels to provide recovery options without an intervening SDH layer (supports building IP-over-WDM networks). Commonly mechanisms on optical layer and SDH layer do not have visibility into higher layer operations. Hence, they may provide e.g. link protection by not easily providing node protection or protection of traffic transported at the IP layer. Moreover, this may prevent the lower layers from providing restoration based on the needs of the traffic flows. An example is that fast restoration could be provided only to traffic flows that need this, while slower restoration could be used for traffic flows that do not need fast restoration (slower restoration may also allow for time to find more optimal use of the resources). For cases where the latter traffic type is dominant, providing fast restoration to all traffic types may not be most effective from an operator's point of view.

MPLS Recovery Goals

[Shar02] also states a number of goals for MPLS-based recovery. These may be broadly divided into:

A. Resource-related:
   - Allow for better overall use of the network resources, e.g. subject to traffic engineering goals
   - Maximise network reliability and availability, e.g. by minimising number of single points of failure in the MPLS-protected domain
   - Recovery techniques may be applicable for an entire end-to-end path or for segments thereof.
   - MPLS-based recovery mechanisms should take into account recovery actions for lower layers. MPLS-based mechanisms should not trigger lower layer protection switching.
   - Minimise the state overhead incurred for each recovery path maintained.
   - Preserve the constraints on traffic after switch-over. That is, if desired, the recovery path should meet the resource requirements of, and achieve the same performance characteristics as the working path.

B. Traffic-related:
   - Facilitate restoration times that are sufficiently fast for the end-user application, i.e. matching the application's requirements.
   - Enhance the reliability of the protected traffic while minimally or predictably degrading the traffic carried by the diverted resources.
   - Protection of traffic at various granularities. For example to specify MPLS-based recovery for a portion of the traffic on an individual path, for all traffic on an individual path, or for all traffic on a group of paths.
   - Minimise loss of data and packet reordering during recovery operations.

Several of these goals may be in conflict with each other. Moreover, engineering compromises would be done based on a range of factors such as cost, application requirements, network efficiency, and revenue considerations.

Features

From an operational point of view, [Shar02] lists the following desirable features:

- MPLS recovery provides an option to identify protection groups (PPGs) and protection portions (PTPs).
- Each PSL is capable of performing MPLS recovery upon detection of impairments or receipt of notification of impairments.
- Manual protection switching commands are not precluded.

- A PSL is capable of performing a switch back to the original working path when the fault is cleared or a switch-over to a new working path.

- Path merging at intermediate LSRs must be possible. If a fault affects the merged segment all the paths sharing that merged segment should be able to recover. If a fault affects a non-merged segment only the path that is affected by the fault should be recovered.

10.3 Recovery and Protection Schemes

Basically, there are two models for path recovery: rerouting and protection switching. Recovery by rerouting is defined as establishing new paths or path segments on demand for restoring traffic *after* the occurrence of a fault. For protection switching mechanisms a recovery path (or path segment) is pre-established. The pre-established path may or may not be link and node disjoint with the working path. If it is not disjoint, the overall reliability is degraded.

As traffic dynamics grows, requirements on network survivability are not becoming more relaxed. During recent years, much work and interest have been attached to finding efficient ways of allocating spare capacity in survivable networks. Several approaches still utilise NP-hard optimisation process based on static working traffic demands [Ho02]. Most of the static schemes can be used for reallocating the spare capacity while the network is running. However, this is mostly after a time-consuming optimisation process. Hence the derived solution can be far from optimal as traffic rapidly changes. A conclusion in [Ho02] is that the static schemes are more suited for use in designing small-sized networks, or networks where demands are less dynamic.

One attempt to overcome the computational complexity is to introduce heuristic algorithms, introducing a compromise between performance (efficient resource utilisation) and computational efficiency. [Ho02] names this process a survivable routing. A survivable routing algorithm is used to dynamically allocate the current connection request into a network with protection service, while maximising the probability of successfully allocating subsequent connection requests in the network.

In contrast to dedicated protection such as 1+1, when a shared protection scheme is used, several working paths may use the same protection resource. However, when two working paths share the same risk of failure (e.g. caused by the same fault), they may not share the same protection resources. Obeying this, it is possible to guarantee the level of impact that a single fault (link or node) would have.

Basically, two shared protection schemes are:

- Path-based: the source node of a working path computes a protection path by ensuring that the protection path is diversely routed from the working path. If a fault occurs on the working path, the terminating node in its control plane detects the fault and sends a notification to the first node in the path to activate a switch-over. The source node switches over the traffic from the working path to the protection path. As the path length influences the restoration time, a short failure recovery time may be hard to reach.

- Link-based: fault localisation and restoration is done at upstream neighbour node of the failed link or node and the traffic merged back to the original working path at the downstream neighbour node. This may allow fast restoration due to rapid fault localisation.

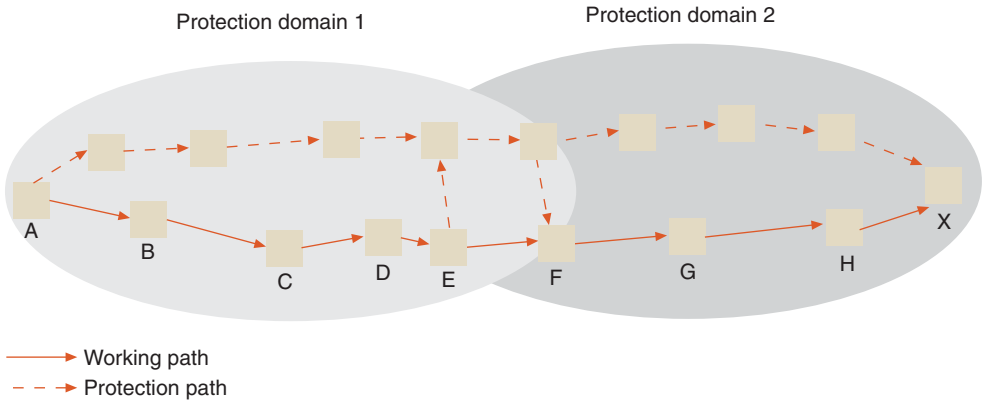
In addition to the short restoration times, fine service granularity and high network throughput are sought.

Protection domains are commonly introduced. Subsequently, a diversely routed protection path segment is sought for each working path segment within a protection domain. A domain diameter may then be introduced, defined as the shortest path (hop count) between a source path node and termination path node. Referring to this measure, link-based would have a diameter equal to one, while path-based would have a diameter equal to the working path length. Adjacent protection domains must overlap in order to protect against every single link/node failure along the working path (e.g. the node on the edge of a domain), see Figure 19.

When a fault occurs on a working path, the first node in the protection domain where the fault occurs is notified and activates a traffic switchover. For example, a fault on link $C - D$ is detected by node D and signalled to node A that switches the traffic to the protection path. A fault on link $G - H$ is detected by node H and signalled to node E that switches over the traffic.

As argued in [Ho02], introducing protection domains allows for shorter restoration times (only within a domain), less computational complexity (needs to find restoration within the domain only), and potentially higher resource utilisation (protection paths could be shared outside the domain). From a network management perspective, it allows a trade-off between restoration time and amount of protection resources consumed. A major drawback is the increase of signalling.

Figure 19 Dividing the paths into protection domains



Survivability mechanisms are available at several network layers, e.g. SDH, MPLS and IP. Moreover, these mechanisms may be in operation in multiple layers at the same time. Another issue is that a single failure at the physical layer would likely result in several failures detected at the IP layers (several routes unavailable as seen by the IP router).

Generally, recovery at lower layers has advantages in short recovery time. The recovery at IP or MPLS layer allows for better resource efficiency and recovery granularity. The latter is important when recovery/resilience classes are offered. An example of such classes is shown in Table 3.

As seen by the values, class 1 has the strictest requirements, while class 4 has no requirements. Hence, traffic of class 4 may be pre-empted if a failure occurs, even when this traffic is not directly affected by the failure itself. This is to release network resources for recovery of other classes.

A number of recovery options, matching the resilience schemes in Table 3 are given in Table 4.

Table 3 Example of service classes and resilience options (adapted from [Aute02])

Service class	**1**	**2**	**3**	**4**
Resilience requirements	High	Medium	Low	None
Recover time	10 – 100 ms	100 ms – 1 s	1 s – 10 s	n.a.
Resilience scheme	Protection	Restoration	Rerouting	Pre-emption
Recovery path set-up	Pre-established	On-demand immediate	On-demand delayed	None
Resource allocation	Pre-reserved	On-demand (assured)	On-demand (if available)	None
Service level after failure	Equivalent	May be temporarily reduced	May have reduced service level	None

Table 4 Examples of recovery options (from [Aute02])

Recovery model	Protection switching		Restoration (MPLS rerouting)		(IP) rerouting
Resource allocation	Pre-reserved			Reserved on-demand	
Resource use	Dedicated resources		Shared resources		Extra traffic allowed
Path setup	Pre-established		Pre-qualified		Established on-demand
Recovery scope	Local repair	Global repair	Alternate egress pair	Multi-layer repair	Conc. prot. domain
Recovery trigger	Automatic inputs (internal signals)			External commands (OAM signalling)	

The characteristics of the different recovery models are:

- Protection switching: the alternative path is pre-established and pre-reserved (pre-provisioned). Hence, the shortest traffic disruption is achieved. Two main groups are 1+1 and 1:1. In the former group packets are forwarded simultaneously on working and protection paths. When the working path fails, the downstream node simply selects packets from the alternative path. For 1:1, packets are forwarded on a predefined path in case of failure on the working path. When no failure is present, the alternative path may carry other traffic flows. However, these flows must be preemptible as they should be dropped if the alternative path is needed for the protected traffic.

- Restoration (MPLS rerouting): the recovery path is established on-demand after detecting a failure. As it takes a while to calculate new routes, signal them and configure the relevant mechanisms, this takes considerably longer than protection switching.

- IP rerouting: ordinary routing protocol and exchange principles are utilised to identify alternative paths.

A number of recovery scopes may also apply. A path is either switched at the ingress and egress router or locally at the router adjacent to the failure. A protection switching scheme where an alternative path is pre-established for each link is commonly called MPLS fast reroute. In this case no end-to-end failure notification and signalling are required. As described above, another method is to set up an alternative path to handle fast rerouting. One approach is to set up an alternative path for each working path as indicated from the last hop router in the reverse direction to the source of the working path and along a node-disjoint path to the destination router. When a failure is detected, the adjacent upstream router immediately switches the working path to the recovery path. Therefore, only a single protection path must be set up, and the rerouting may still be triggered based on a local decision in the router directly upstream of the failure. Hence, no recovery signalling is needed.

Protection switching schemes with global repair are commonly called path protection. For each protected path a protection path is established either between the ingress and egress router or between designated recovery-switching points (e.g. for segment protection/domain). Then the switching node must be notified in case a path fails in order to switch that path to the protection path.

10.4 Recovery Cycle and Time Description

Restoration time, T_R, can be decomposed as:

$$T_R = T_{detection} + T_{signalling} + T_{config}$$

Where $T_{detection}$ is the time for failure detection and localisation, $T_{signalling}$ is the time for signal propagation and node processing, and T_{config} is the time for configuring the network elements/ resources along the protection path.

Figure 20 [Shar02] describes three recovery cycles containing a set of events. During the first cycle, the fault is detected and traffic restored onto the MPLS recovery path. In case the recovery path is non-optimal, it may be followed by any of the two latter cycles to achieve an optimised network again. The reversing cycle applies for explicitly routed traffic that does not rely on any dynamic routing protocols to be converged. The dynamic re-routing cycle applies for traffic that is forwarded based on hop-by-hop routing.

The different time intervals are defined as follows:

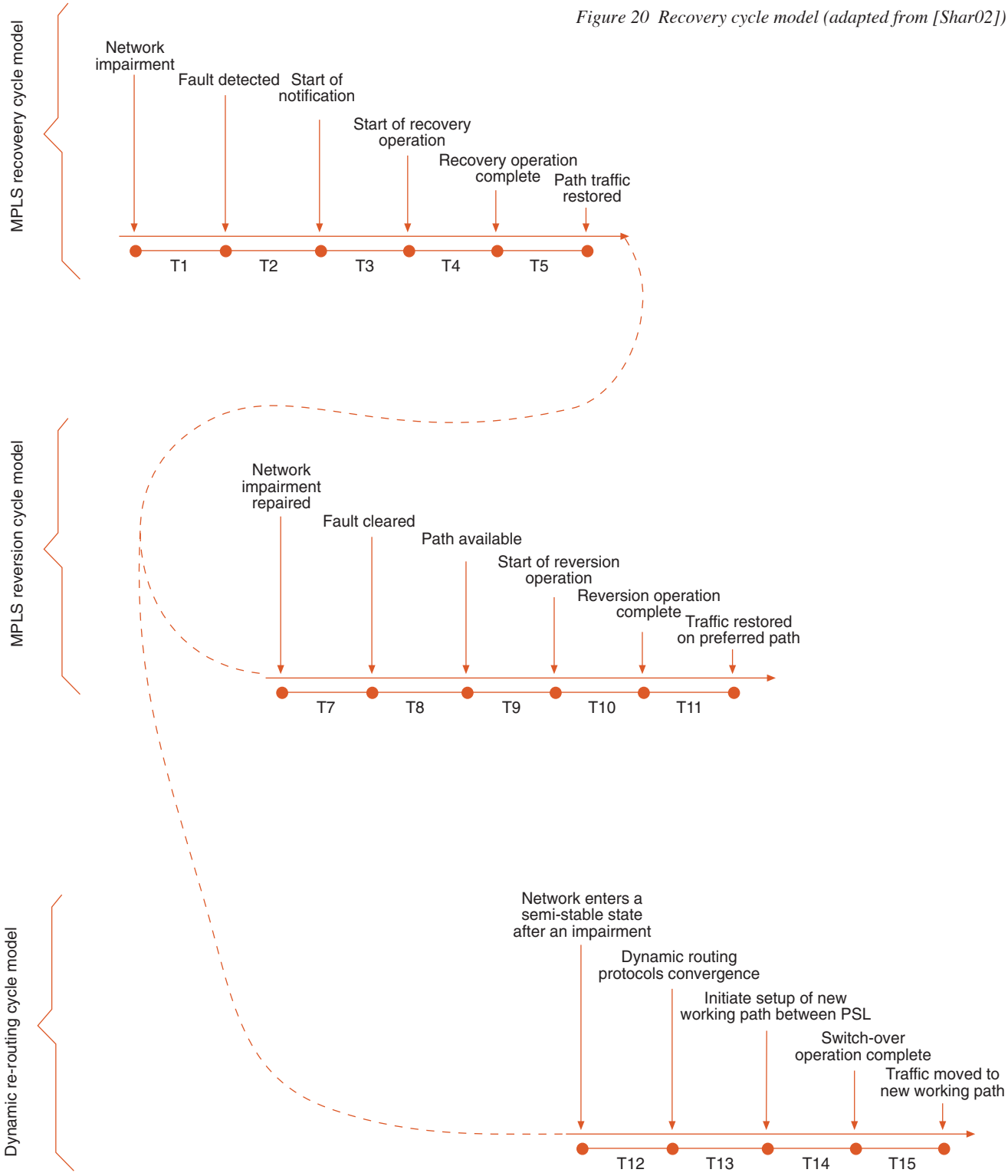
A. MPLS recovery cycle model:
   - T1 – Fault detection time: the time between the occurrence of an impairment and the moment the fault is detected by the MPLS-based recovery mechanisms. The length of this interval may be heavily dependent on lower layer protocols.

   - T2 – Hold-off time: the waiting time between the detection of a fault and taking MPLS-based recovery actions. This may be configured to allow time for lower layer protection to take effect (however, it may also be zero). The hold-off time may also occur after the notification time interval if the node responsible for the switch-over, the Path Switch LSR (PSL), rather than the detecting LRS, is configured to wait.

   - T3 – Notification time: the time between initiating a Fault Indication Signal (FIS) by the LSR detecting the fault and the instant at which the PSL begins the recovery operation. This is zero if the PSL detected the fault itself or infers a fault from such events as an adjacency failure. As noted above, if the PSL detected the fault itself, there may be a hold-off time period between detecting and starting the recovery operation.

   - T4 – Recovery operation time: the time between the first and the last recovery

Figure 20 Recovery cycle model (adapted from [Shar02])



action. This can include message exchanges between the PSL and PML to coordinate recovery actions.

- T5 – Traffic restoration time: the time between the last recovery action and the instant when the traffic is completely recovered. This interval is intended to account for the time required for traffic to once again arrive at the point of the network that experienced disrupted or degraded service due to the occurrence of the fault (the length of this interval depends on the location of the fault, the recovery mechanisms, and the propagation delay along the recovery paths.

- T6 – (not indicated): the interval when the impairment has not been repaired and traffic is sent on the recovery path.

B. MPLS reversion cycle model (executed to switch the traffic back to the preferred path after the fault on that path has been cleared):

- T7 – Fault clearing time: the time between the repair of an impairment and instant when the MPLS-based mechanisms learn that the fault has been cleared. This will likely be heavily dependent on lower layer protocols.
- T8 – Wait-to-restore time: the waiting time between the clearing of a fault and MPLS-based recovery actions. The waiting time may be configured to ensure that the path is stable and to avoid flapping in cases where a fault is intermitted. This time may also be set to zero. The waiting time may occur after the notification time interval if the PSL is configured to wait.
- T9 – Notification time: the time between initiating a Fault Recovery Signal (FRS) by the LSR clearing the fault and the time at which the PSL begins the reversion operation. This is zero if the PSL clears the fault itself. As noted above, there may still be a Wait-to-restore time period between fault clearing and the start of the reversion operation.
- T10 – Reversion operation time: the time between the first and last reversion actions. This may include message exchanges between the PSL and PML to coordinate reversion actions.
- T11 – Traffic restoration time: the time between the last reversion action and instant when traffic is completely restored on the preferred path. This interval is expected to be small since both paths are working and care may be taken to limit the traffic disruption.

In practise, the most interesting times are the Wait-to-restore time and the Traffic restoration time. As this cycle is completed when paths are in operation, there would likely be little need for a rapid operation. Hence, a well-controlled operation is sought with minimal disruption.

C. Dynamic re-routing cycle model (aims to bring the network to a stable state after impairment has occurred. Hence, a re-optimised network is sought after the routing protocols have converged and the traffic has been moved from a recovery path to a (possibly) new working path.

- T12 – Network route convergence time: the time taken for the routing protocols to converge and the network to reach a stable state.
- T13 – Hold-down time: the time for which a recovery path must be used. This is to prevent flapping of traffic between a working and a recovery path.
- T14 – Switch-over operation time: the time between the first and the last switch-over actions. This may include message exchanges between the PSL and PML to coordinate the switch-over actions.
- T15 – Traffic restoration time: the time when all traffic has been moved to working paths and a new stable network situation reached.

[Huan02] presents a protection mechanism that is built on the following principles:

- Tree structure for efficient distribution of fault and/or recovery information (called reverse notification tree).
- *Hello* packets to detect faults as complementary fault detection (in addition to any lower layer mechanism).
- Notification transport protocol, e.g. utilising UDP.

As pointed out in that article, one of the major considerations in a path protection mechanism is to control the delay that must be met by the notification message travelling from the fault detection node to the protection switching node, i.e. POR. This delay may cause additional packet loss and misordering.

A key aspect considered in that article is that LSPs may be merged; several working paths may converge to form a multipoint-to-point tree with the PSLs as leaves. The fault indications and repair notifications should then be sent along a reverse path of the working path to all the PSLs affected by the fault. This is in case an end-to-end recovery is to be initiated. If a single segment is to be recovered, the notification would be stopped in the PSL at the beginning of that segment.

Each LSR must be able to detect certain fault types, such as path failure, path degraded, link failure and link degraded, and send corresponding FIS message to the PSL. Hence, a node upstream of the fault must be able to detect or learn about the fault. This motivates for use of a *hello* protocol commonly using a much shorter timer than applied for most routing protocols.

11 Concluding Remarks

Considering the steadily increasing growth of traffic, in particular related to IP-based services, it is essential for an operator to configure an efficiently operating network. The interplay of IP and optics has been promoted as a strong candidate in the core network. This is also the main topic in this article, with particular emphasis on how to achieve resilience for the traffic flows that have such requirements.

Several studies and international projects have been carried out in this area, only a few of the results and questions are addressed in this article. Currently it can be observed that several of the vendors are taking positions in both the IP and the optics area, also proposing networking solutions to manage the topics raised. However, a final answer on how to take care of the resilience seems not to be found yet. Neither is the general model for IP and optics interplay settled, advocating for continuing effort.

References

[Aute02] Autenrieth, A, Kirstädter, A. Engineering End-to-End IP Resilience Using Resilience-Differentiated QoS. *IEEE Communications Magazine*, 4 (1), 50–57, 2002.

[Coll02] Colle, D et al. Data-Centric Optical Networks and Their Survivability. *IEEE Journal on Selected Areas in Communications*, 20 (1), 6–20, 2002.

[Ho02] Ho, P-H, Mouftah, H T. A Framework for Service-Guaranteed Shared Protection in WDM Mesh Networks. *IEEE Communications Magazine*, 40 (2), 97–103, 2000.

[Huan02] Huang, C et al. Building Reliable MPLS Networks Using a Path Protection Mechanism. *IEEE Communications Magazine*, 40 (3),156–162, 2002.

[ID_ipofw] Rajagopalan, B et al. *IP over Optical Networks: A Framework*. IETF draft. Draft-ietf-ipo-framework-05.txt. Oct.2003. Work in progress.

[ID-iporeq] Xue, Y. *Optical Network Service Requirements*. IETF draft. Draft-ietf-ipo-carrier-requirements-05.txt. Dec 2002. Work in progress.

[Gers00] Gerstel, O, Ramaswami, R. Optical Layer Survivability – An Implementation Perspective. *IEE Journal on Selected Areas in Communications*, 18 (10), 1885–1899, 2000.

[RFC3386] Lai, W, McDysan, D. *Network Hierarch and Multilayer Survivability*. IETF RFC 3386, Nov. 2002.

[RFC3471] Berger, L. *Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description*. IETF RFC 3471, Jan. 2003.

[Shar02] Sharma, V, Hellstrand, F. *Framework for MPLS-based Recovery*. IETF draft Draft-ietf-mpls-recovery-frmwrk-07.txt. Mar. 2002. Work in progress

[Stam00] Stamatelakis, D, Grover, W D. IP Layer Restoration and Network Planning Based on Virtual Protection Cycles. *IEEE Journal on Selected Areas in Communications*, 18 (10), 1938–1949, 2000.

Terms and Acronyms

Note: Several of the acronyms have been described in articles of this issue.

2.5 G Two and a half Generation (mobile system)	Commonly used term for enhancements in data capabilities and bit rates of second generation systems (GSM, TDMA, etc) like EDGE.
2B+D ISDN basic access	ISDN term: 2 channels 64 kbit/s (user traffic), 1 channel 16 kbit/s (signalling, maintenance and user traffic).
30B + D ISDN enhanced access	ISDN term: 30 channels 64 kbit/s (user traffic), 1 channel 64 kbit/s (signalling, maintenance and user traffic).
3G 3rd generation (mobile system)	The generation of mobile systems which is a descendant of today's widespread and voice centric second generation mobile systems like GSM. 3G systems are characterised by multimedia capability and data rates up to 2 Mbit/s. Ref. UMTS
4G 4th generation (mobile system)	Term often used to denote future broadband mobile communication systems or standards to follow third generation. Often referred to as "systems beyond 3G" (B3G).
AAA Authentication, Authorization and Accounting	Key functions to intelligently control access, enforce policies, audit usage and provide the necessary information for billing of Internet services.
Accounting	Tracking which services are used, by whom, when and for how long. Accounting is carried out by the logging of session statistics and usage information and is used for authorization control billing, trend analysis, resource utilization, and capacity planning activities.
ACTS Advanced Communications Technologies and Services	Thematic Programme of the Fourth Framework Research Programme funded by the European Union. It ran from 1994 to 1998 (the last projects ended in 2000) with a total funding from the EU of 681 million Euro. http://www.cordis.lu/acts
ADM Add-drop multiplexer	
Aggregator/ Aggregation broker	An entity providing access to several underlying networks/systems. This is commonly done by trading services (e.g. bandwidth) between member providers. Brokers typically provide centralized authentication services in order to compute and validate the broadband traffic.
AP Access Point	A point where users access the system/network, e.g. a base station in a wireless network.
APS Automatic Protection Switching	
ARIMA Autoregressive Integrated Moving Average	Time series modelling
ASIC Application-Specific Integrated Circuit	
ASON Automatically Switched Optical Network	
ASP Application Service Provider	
ASTN Automatically Switched Transport Network	
ATM Asynchronous Transfer Mode	
AUC Authentication Centre	
Authentication	The process of determining who the user is. It can take the form of ensuring that data has come from its claimed source, or of corroborating the claimed identity of a communication party.
Authentication server	The server that is responsible for authenticating the content user {source [BCD01a]}.
Authorization	The process of enforcing policies – of determining the types or qualities of activities, resource, or services a user is permitted. Usually, authorization is in the context of authentication; once you have authenticated a user, (s)he may be authorized for different types of access or activities.
Authorization server	Determines if an authenticated user is authorized to view content with a particular set of delivery parameters. May also bill the user for use and generate a billing record that is sent to the content provider. The content provider or a third-party could own this server {source [BCD01a]}.
B-ISDN Broadband Integrated Services Digital Network	
BA Behaviour Aggregate	A collection of packets with the same (DiffServ) code point crossing a link in a particular direction. {source [RFC3564]}
Baseline analysis	A study conducted to serve as a baseline for comparing to the actual behaviour of the network. {source [RFC3272]}
Billing cycle	The period between the creation of billing files. Typically, this is done once a day at a specific time. The billing file contains the traffic of all end users per roaming partner for one billing cycle. The shorter the billing cycle, the better the risk of fraud can be monitored in roaming environments.
BLSR Bi-directional line switched ring	

Bluetooth A short-range wireless specification that allows radio connection between devices within a 10-metre range of each other. Bluetooth is designed as a Personal Area Network (PAN) technology with a wide variety of theoretical uses.

BGP Border Gateway Protocol

BOT Build, operate and Telenor collocation — Telenor product offering to cover deployment, operation and maintenance of ICT equipment in addition to collocation of technical equipment.

Bottleneck A network element whose input traffic rate tends to be greater than its output rate. {source [RFC3272]}

Branch and bound A tree of linear programs is formed in which each problem has the same constraints and objectives except for some additional bounds on certain variables. At the root of this tree is the original problem without additional requirements. The solution to this root problem will not, in general, have all integer components. We now choose some non-integer solution component and define two variants by fixing one of the variables. This gives rise to two sub-problems. The left-child and right-child problems separate the solution space. This branching process can be carried out recursively; each of the two new problems will give rise to two more problems when we branch on one of the non-integer components of their solution. It follows from this construction that the tree is binary.
Eventually, after enough new bounds are placed on the variables, integer solutions are obtained. The value for the best integer solution found so far is retained and used as a basis for pruning the tree. If the continuous problem at one of the tree nodes has a final objective value different from the value obtained in a node, so have all of the node's descendants, since they have smaller feasible regions and hence even larger optimal objective values. The branch emanating from such a node cannot give rise to a better integer solution than the one obtained so far, so we consider it no further; that is, we prune it. Pruning also occurs when we have added so many new bounds to some continuous problem that its feasible region disappears altogether.
Two strategies for solving the node problems are depth-first and width-first, of which the former is frequently preferred.

BRAS Broadband Remote Access Server

BSC Base Station Controller — Network node in the GSM network controlling a number of BTSs. http://www.etsi.org

BTS Base Transceiver Station — The radio base station of a GSM network. It consists of one or more transmitter-receiver units, each serving one carrier frequency. http://www.etsi.org

Busy hour A one hour period within a specified interval of time (typically 24 hours) in which the traffic load in a network or sub-network is greatest. {source [RFC3272]}

BW Bandwidth

Bypass tunnel A path that serves to back up a set of working paths using the label stacking approach. The working paths and the bypass tunnel must all share the same PSL and PML {source [Shar02]}.

CAPEX Capital Expenses

CAPM Capital Asset Pricing Model

CaTV Cable TV

CDMA Code Division Multiple Access — Multiple access method used e.g. in UMTS. Allowing for multiple transmissions to be carried simultaneously on a single wireless channel (same time slot and carrier frequency). This is done by spreading the signal across the frequency band using one of a set of spreading codes such that one user's signal appears as noise/interference for other (non-intentional) receivers.

CDN Content Distribution/Delivery Network — A collection of network elements arranged for more effective delivery of content to clients. Typically a CDN consists of a request-routing system, surrogates (a content server other than the origin server), a distribution network, and an accounting system. {source [BCD01a] – from IETF content peering group}

Clearing house A provider of roaming billing data services. A clearing house typically provides post-connection services by handling the billing data, offering validation and rating services as well as settling the financial liabilities. A distinction should be made between data and financial clearing house. The latter is also referred to as settlement house. The former only delivers the data processing service whereas the latter also provides invoice and payment services.

Congestion A state of a network resource in which the traffic incident on the resource exceeds its output capacity over a time interval. {source [RFC3272]}

Congestion avoidance An approach to congestion management that attempts to obviate the occurrence of congestion. {source [RFC3272]}

Congestion control An approach to congestion management that attempts to remedy congestion problems that have already occurred. {source [RFC3272]}

Constraint based routing A class of routing protocols that take specified traffic attributes, network constraints, and policy constraints into account when making routing decisions. Constraint-based routing is applicable to traffic aggregates as well as flows. It is a generalization of QoS routing. {source [RFC3272]}

Content A conceptual identifier of a media object and its constituent digital bit streams. An identifier is a movie title, a song name, and – in the case of a live event – an event identifier. Note that content could be comprised of several different bit streams. For example, a particular movie could be encoded in three different qualities and have several language tracks. A live event could also be encoded in different qualities. {source: [BCD01a]}

Content server The server that is responsible for sending the stream to a target device over some network. This may be the original server or a surrogate server owned by the content network.

CPU Central Processing Unit

CT Class Type — The set of Traffic Trunks crossing a link that is governed by a specific set of bandwidth constraints. CT is used for the purposes of the link bandwidth allocation, constraint based routing and admission control. A given Traffic Trunk belongs to the same CT on all links. {source [RFC3564]}

CWDM
Coarse wavelength division multiplexing

A relatively low-cost WDM technology that uses wide spacing (about 10 nm) between wavelengths and un-cooled optics.

DACS
Digital access cross-connect system

DB
Database

DCF
Discounted Cash Flow

DCME
Digital Circuit Multiplication Equipment

Demand side congestion management

A congestion management scheme that addresses congestion problems by regulating or conditioning offered load. {source [RFC3272]}

Dijkstra's algorithm

An algorithm to find the shortest paths from a single source node to all other nodes in a weighted, directed graph.

DPP
Discounted Payback Period

DPRings
Dedicated protection rings

DRM
Digital Rights Management

DS
DiffServ – Differentiated Services

DSL
Digital Subscriber Line

DSLAM
Digital subscriber line access module

DSM
Dynamic spectrum management

DTH
Direct to home

Satellite access

DWDM
Dense wavelength division multiplexing

A carrier-class WDM technology that uses expensive, cooled optics and tight spacing between wavelengths of less than a nanometre based on the specifications of the International Telecommunications Union (ITU) DWDM wavelength grid. http//www.itu.int

E-LAN service

Ethernet LAN service provides multipoint connectivity by connecting two or more users. One or more of the connected users can receive data. At each site the user is connected to a multipoint Ethernet Virtual Connection (EVC). As new sites are added they are connected to the same multipoint EVC. From a user point of view the E-LAN service makes the carrier network (metro and wide areas) look like a switched LAN. For an easy comparison the Virtual Private Line Service (VPLS) can be seen as an E-LAN service that is centred on the use of MPLS labels in the forwarding plane as opposed to the use of CLAN tags in the Metro Ethernet Forum (MEF) definitions. Both E-LAN and VPLS focus on enabling carriers to deliver true VPN services to users. {source [MEFsite]}

E-Line service

Ethernet line service providing point-to-point Ethernet Virtual Connection (EVC) between two subscribers. The E-line service is used for Ethernet point-to-point connectivity. {source [MEFsite]}

EBIT
Earnings Before Interests and Income Taxes

EBITDA
Earnings Before Income Taxes, Depreciation and Amortisation

EDGE
Enhanced Data for GSM Evolution

A modulation method for GSM and IS-136 TDMA networks that allows for wireless data transfer up to 384 kbit/s. Standardized by ETSI. http://www.etsi.org

Edge server

A network element close to the edge of a backbone network that is able to perform content services. Service examples include caching, video/audio stream serving and splitting, insertion, transcoding, content listing, etc. {source [BCD01a]}

Effective bandwidth

The minimum amount of bandwidth that can be assigned to a flow or traffic aggregate in order to deliver 'acceptable service quality' to the flow or traffic aggregate.

Egress traffic

Traffic exiting a network or network element.

EIR
Equipment Identity Register

EPG
Electronic Programme Guide

Erlang's blocking formula

Well-applied formula for calculating the blocking probability (used for decades in telephony networks).

Ethernet

International standard networking technology for wired implementations. Basic, traditional networks offer a bandwidth of about 10 Mbit/s. Fast Ethernet (100 Mbit/s) and Gigabit Ethernet (1000 Mbit/s), as well as 10 Gigabit, are also available.

ETSI
European Telecommunications Standards Institute

A non-profit membership organisation founded in 1988. The aim is to produce telecommunications standards to be used throughout Europe. The efforts are coordinated with the ITU. Membership is open to any European organisation proving an interest in promoting European standards. It was responsible for the making of the GSM standard, among others. The headquarters are in Sophia Antipolis, France. http://www.etsi.org

Extra traffic
(also referred to as pre-emptable traffic)

Traffic carried over the protection entity while the working entity is active. Extra traffic is not protected, i.e. when the protection entity is required to protect the traffic that is being carried over the working entity, the extra traffic is pre-empted. {source [RFC3386]}

FEC
Forward Equivalence Class

FIS
Fault Indication Signal

A signal that indicates that a fault along a path has occurred. It is relayed by each intermediate LSR to its upstream or downstream neighbour until it reaches an LSR that is set up to perform MPLS recovery (the POR). The FIS is transmitted periodically by the node (closest to the point of failure) for some configurable length of time. {source [Shar02]}

Forecasting traffic

Statement in advance of traffic characteristics, including volume (e.g. measured in Mbit/s) and number of customers.

Flow

The term flow is commonly used to signify the smallest non-separable stream of data from the point of view of an end-point or termination point (source or destination points). {based on [ID-ipofw]}

Footprint

A network of access points referring to the overall coverage of the network.

FRS
Fault Recovery Signal

A signal that indicates that a fault along a working path has been repaired. Again, like the FIS, it is relayed by each intermediate LSR to its upstream or downstream neighbour until it reaches the LSR that performs recovery of the original path. The FRS is transmitted periodically by the node/nodes closest to the point of failure for some configurable length of time. {source [Shar02]}

FTP
File Transfer Protocol

FTTH
Fibre to the home

FWA
Fixed Wireless Access

Fixed Wireless Access consists of a radio link to the home or the office from a cell site or base station. It replaces the traditional wireless local loop, either if the wire based infrastructure is sparse or to gain rapid expansion in denser urban and suburban areas.

GB
Gigabit

'One billion bits'. Term used to denote the number 1 073 741 824 bits (2^{30}).

GMPLS
Generalized Multi-Protocol Label Switching

GoS
Grade of Service

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 kbit/s gross transfer rate. Practical rates are from 12–48 kbit/s. http://www.etsi.org

Greenfield provider

A service provider that introduces itself to a new market/region without prior engagements.

GSM
Global System for Mobile communication

A digital cellular phone technology that is the predominant system in Europe, but is also used around the world. Development started in 1982 by CEPT and was transferred to the new organisation ETSI in 1989. Originally, the acronym was the group in charge, 'Groupe Special Mobile', but later the group changed its name to SMG, and GSM became the name of the system. GSM was first deployed in seven European countries in 1992. It operates 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 2000 there were more than 250 million GSM users, which is more than half the world's mobile phone population. http://www.etsi.org

GSM Alliance

Consortium of North American GSM 1900 operators.

GSM Association

World's leading wireless industry representative body, consisting of more than 660 second- and third-generation wireless network operators and key manufacturers and suppliers to the wireless industry. www.gsmworld.com

HB@
Hybrid Broadband Access

Telenor project running from 2000–2002 with the aim of investigating and testing how different access technologies could be combined in order to develop a complete broadband network able to provide comparable broadband services both in urban, suburban and rural parts. The aim was to develop Telenor's superior strategy and specific solutions on the implementation and roll-out of a hybrid access infrastructure suitable for television and personal computer centric broadband services in the period 2001–2005.
Four different technical solutions were tested in Stavanger, Svolvær, Beito and Oslo providing a selection of next generation interactive broadband services. http://www.telenor.no/prosjekt/hba/indeks.shtml (in Norwegian)

HDTV
High Definition TV

HFC
Hybrid Fibre Coax

Hierarchy

A technique used to build scalable complex systems. It is based on an abstraction, at each level, of what is most significant regarding the details and internal structures of the levels further away. This approach makes use of a general property of all hierarchical systems composed of relevant subsystems that interactions between subsystems decrease as the level of communication between subsystems decreases.

HLR
Home Location Register

A database that holds subscription information about every subscriber in a mobile network. An HLR is a permanent SS7 database used in cellular networks. The HLR is located at the SCP (Signal Control Point) of the cellular provider of record, and is used to identify/verify a subscriber. It also contains subscriber data related to features and services. http://www.etsi.org

Home operator

The provider that owns the relationship with the user of the service by way of a subscription agreement allowing the provider to bill the user for services.

Horizontal hierarchy

The abstraction that allows a network at one technology layer, for instance a packet network, to scale. Examples of horizontal hierarchy include BGP confederations, separate autonomous systems and multi-area OSPF. In the horizontal hierarchy, a large network is partitioned into multiple smaller, non-overlapping subnetworks. The partitioning criteria can be based on topology, network function, administrative policy, or service domain demarcation. Two networks at the "same" hierarchical level, e.g. two autonomous systems in BGP, may share a peer relation with each other through some loose form of coupling. On the other hand, for routing in large networks using multi-area OSPF, abstraction through the aggregation of routing information is achieved through a hierarchical partitioning of the network. {source [RFC3386]}

Hot spot

- A venue where a commercial public WLAN service can be accessed. Hotspots are typically found in hotels, airports and cafés.
- A network element or subsystem, which is in a state of congestion. {source [RFC3272]}

HSCSD
High speed circuit switched data

An addition to GSM for adding faster data transmission. While GSM originally supports 9.6 kbit/s, HSCSD supports from 14.4 up to 57.6 kbit/s circuit switched data connections. http://www.etsi.org

HSSG
Higher speed study group

The group commissioned by the IEEE 802.3 Ethernet working group with drafting the technical scope and objectives of the standards effort.

Hub

A multi-port device used to connect PCs to a network, typically via Ethernet cabling or via wireless access (WLAN). Wired hubs can have numerous ports and can transmit data at speeds ranging from 10 Mbit/s to multi-gigabyte rates per second. A hub transmits the packets it receives to all the connected ports. A small wired hub may only connect four computers; a large hub can connect 48 or more, whereas a wireless hub can connect hundreds.

ICT
Information and communication technologies

IEC
International Electrotechnical Commission

An organization that sets international electrical and electronics standards founded in 1906. It is made up of national committees from over 40 countries. Headquarters in Geneva, Switzerland. http://www.iec.ch

IEC/ISO 11801

The international and European standard for building cabling.

IEEE
The Institute of Electrical and Electronics Engineers

USA based organisation open to engineers and researchers in the fields of electricity, electronics, computer science and telecommunications. Established in 1884. The aim is to promote research through journals and conferences and to produce standards in telecommunications and computer science. IEEE has produced more than 900 active standards and has more than 700 standards under development. Divided into different branches, or 'Societies'. Has daughter organisations, or 'chapters' in more than 175 countries worldwide. Headquarters in Piscataway, New Jersey, USA. http://www.ieee.org

IEEE 802.11

Refers to a family of specifications developed by the IEEE for wireless local area networks. It also refers to the "Wireless LAN working group" of the IEEE 802 project. 802.11 specifies an over-the-air interface between a wireless client and a base station, or between two wireless clients. The IEEE accepted the specification in 1997. There are several specifications in the 802.11 family, including i) 802.11 – provides 1 or 2 Mbit/s transmission in the 2.4 GHz band; ii) 802.11a – an extension that provides up to 54 Mbit/s in the 5 GHz band (it uses an orthogonal frequency division multiplexing encoding scheme rather than FHSS or DSSS); iii) 802.11b – provides 11 Mbit/s transmission in the 2.4 GHz band and was ratified in 1999 allowing wireless functionality comparable to Ethernet; iv) 802.11g provides 20+ Mbit/s in the 2.4 GHz band; v) 802.11z is a method for transporting an authentication protocol between the client and access point, and the Transport Layer Security (TLS) protocol. More variants are also under preparation, including support of 100 Mbit/s traffic flows. http://www.ieee802.org/11/

IEEE 802.3 Working Group

The permanent CSMA/CD (Ethernet) working group of the IEEE 802 project. http://www.ieee802.org/3/

IEEE 802.3ae Task Force

The Task Force chartered by the IEEE 802.3 working group with writing the specification for 10 Gigabit Ethernet. IEEE 802.3ae is also the name of the document produced by the Task Force. http://www.ieee802.org/3/ae/index.html

IETF
The Internet Engineering Task Force

The Internet Engineering Task Force is a large open international community of network designers, operators, vendors, and researchers concerned with the evolution of the Internet architecture and the smooth operation of the Internet. It is open to any interested individual. The actual technical work of the IETF is done in its working groups, which are organized by topic into several areas (e.g. routing, transport, security, etc.). The IETF working groups are grouped into areas, and managed by Area Directors (AD). The ADs are members of the Internet Engineering Steering Group (IESG). Providing architectural oversight is the Internet Architecture Board, (IAB). The IAB also adjudicates appeals when someone complains that the IESG has failed. The IAB and IESG are chartered by the Internet Society (ISOC) for these purposes. The General Area Director also serves as the chair of the IESG and of the IETF, and is an ex-officio member of the IAB. http://www.ietf.org

IGP
Internal Gateway Protocol

IM
Instant Message

IN
Intelligent Network

Industry organisation

A non-profit organisation that helps to create common technical or commercial recommendations and specifications that streamline the marketplace.

Ingress traffic

Traffic entering a network or network element. {source [RFC3272]}

Integer programming

Optimisation formulation with integer constraints on some variables.

Integrator

A company that offers consulting services to help put everything together.

Inter-domain traffic

Traffic that originates in one autonomous system and terminates in another. {source [RFC3272]}

Intermediate LSR

An LSR on a working or recovery path that is neither a PSL nor a PML for that path. {source [Shar02]}

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

IOT
Inter Operator Tariff

Rate that the settlement price between Wireless ISPs for roaming traffic is based on. The roaming agreement between the home and the visited WISPs regulates the rebate (if any) that the home operator is entitled to relative to the IOT. The IOT is often roughly equal to the normal per minute price for users in the operator's network.

IRR
Internal rate of return

ISDN
Integrated Services Digital Network

A digital telecommunication 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 multi-purpose user-network interfaces. The user is offered one or more 64 kbit/s channels. http://www.itu.int

ISO
International Organization for Standardization

The International Organization for Standardization (ISO) is a world-wide federation of national standards bodies from more than 140 countries, one from each country. ISO is a non-governmental organization established in 1947. The mission of ISO is to promote the development of standardization and related activities in the world with a view to facilitating the international exchange of goods and services, and to developing cooperation in the spheres of intellectual, scientific, technological and economic activity. http://www.iso.org/

ISP
Internet Service Provider

A vendor who provides access for customers to the Internet and the World Wide Web. The ISP also typically provides a core group of Internet utilities and services like e-mail and news group readers.

IST
Information Society Technologies

Thematic Programmes of the 5th and 6th Framework Research Programmes funded by the European Union. The first IST programme runs from 2000 to 2004, while the IST programme of the 6th Framework Research Programme started in 2003 and will run until 2006. http://www.cordis.lu/ist/

ITU
International Telecommunication Union

The ITU is a United Nations (UN) body with approx. 200 member countries. It is the top forum for discussion and management of technical and administrative aspects of international telecommunications. Its main objective is to promote connectivity and interoperability between its member countries and to foster the development of all kinds of telecommunications worldwide. It was established in 1865 as an intergovernmental organization with the aim of streamlining interconnection of the national telegraph networks. In 1947 it became a specialized agency of the UN. http://www.itu.int

ITV
Interactive TV

LD
Link Degraded

A lower layer indication to MPLS-based recovery mechanisms that the link is performing below an acceptable level. {source [Shar02]}

LEX
Local Exchange

LF
Link Failure

A lower layer fault indicating that link continuity is lost. This may be communicated to the MPLS-based recovery mechanisms by the lower layer. {source [Shar02]}

LIB
Label Information Base

Link-disjoint

Two connections are link-disjoint if they do not share any link along the path.

Liveness message

A message exchanged periodically between two adjacent LSRs that serves as a link proving mechanism. It provides an integrity check of the forward and backward directions of the link between the two LSRs as well as a check of neighbour aliveness. {source [Shar02]}

LL
Leased Line

LLUB
Local Loop Unbundling

Option to rent only the local loop (e.g. copper access line to a customer) by a non-incumbent operator.

LMDS
Local Multipoint Distribution System

System for Wireless Local Loop applications which can replace cable based access to private and corporate customers. Typically operates in the millimetre wave band. Can provide bandwidth of several Mbit/s to the end customer.

Loss network

A network that does not provide adequate buffering for traffic, so that traffic entering a busy resource within the network will be dropped rather than queued. {source [RFC3272]}

LP
Location Point

LS
Local switch

LSP
Label Switched Path

LSR
Label Switching Router

MAC
Media access control protocol

MAC Address
Media Access Control

A hardwire address that uniquely identifies each node of a network.

Management control/tactical planning

The process by which managers assure that resources are obtained and used effectively and efficiently in the accomplishment of the organisation's objective. {source [Anth65]}

Mbit/s
Megabit per second

'One million bits per second'. Term used to denote a data signalling rate of 1 048 576 b/s (2^{20}).

MD
Main distribution point

Measurement basis

Measurements can be classified on the basis of where and at which level of aggregation the traffic data are gathered. This refers to a set of packets, possibly relative to a particular pair of source and destination, for the purposes of defining performance parameters. Typical measurement bases are; flow-based, interface-based, link-based, node-based, node-pair-based, path-based. {source [ID_Temea03]}

Measurement entity

A measurement entity defines what is measured; it is a quantity for which data collection must be performed with a certain measurement. A measurement type can be specified by a (meaningful) combination of a measurement entity with the measurement basis. {source [ID_Temea03]}

Measurement interval

A measurement interval is the time interval over which measurements are taken. Some traffic data are collected continuously, while other data are collected by sampling, or on a scheduled basis. A measurement interval consists of a sequence of consecutive read-out periods. Summarisation is usually done by integrating the raw data over a pre-specified read-out period. The period must be short enough to capture – with acceptable accuracy – the bursty nature of the traffic, while not being so short that the volume of data produced gets too voluminous.

Measurement methodology — A repeatable measurement technique used to derive one or more metrics of interest. {source [RFC3272]}

Metric — A parameter defined in terms of standard units of measurement. {source [RFC3272]}

Mesh optical network — A topologically connected collection of optical subnetworks whose node degree may exceed 2. Such an optical network is assumed to be under purview of a single administrative entity. It is also possible to conceive a large scale global mesh optical network consisting of the voluntary interconnection of autonomous optical networks, each of which is owned and administered by an independent entity. In such an environment, abstraction can be used to hide the internal details of each autonomous optical cloud from external clouds. {based on [ID-ipofw]}

MON Metropolitan Optical Network

MPEG Moving Pictures Expert Group — MPEG is a committee of ISO/IEC that is open to experts duly accredited by an appropriate National Standards Body. It is in charge of the development of standards for coded representation of digital audio and video. Established in 1988, the group has produced MPEG-1, the standard on which such products as Video CD and MP3 are based, MPEG-2, the standard on which such products as Digital Television set top boxes and DVD are based, MPEG-4, the standard for multimedia for the fixed and mobile web and MPEG-7, the standard for description and search of audio and visual content. Work on the new standard MPEG-21 "Multimedia Framework" has started in June 2000. http://www.chiariglione.org/mpeg/

MPLS Multi Protocol Label Switching — An IETF standard intended for Internet application. A widely supported method of speeding up IP-based data communication. {RFC 2702}

MPLS domain — A continuous set of nodes which operate MPLS routing and forwarding and which are also in one routing of administrative domain. {from [RFC3031]}

MPLS edge node — An MPLS node that connects an MPLS domain with a node which is outside of the domain, either because it does not run MPLS and/or because it is in a different domain. Note that if an LSR has a neighbouring host which is not running MPLS, that LSR is an MPLS edge node. {from [RFC3031]}

MPLS egress node — An MPLS edge node in its role in handling traffic as it leaves an MPLS domain. {from [RFC3031]}

MPLS ingress node — An MPLS edge node in its role in handling traffic as it enters an MPLS domain. {from [RFC3031]}

MPLS label — A label which is carried in a packet header and which represents the packet's FEC. A short fixed length physically continuous identifier used to identify an FEC of local significance. {from [RFC3031]}

MPLS label merging — The replacement of multiple incoming labels for a particular FEC with a single outgoing label. {from [RFC3031]}

MPLS label stack — An ordered set of labels. {from [RFC3031]}

MPLS label swap — The basic forwarding operation consisting of looking up an incoming label to determine the outgoing label, encapsulating, port, and other data handling information. {from [RFC3031]}

MPLS label swapping — A forwarding paradigm allowing streamlined forwarding of data by using labels to identify classes of data packets which are treated indistinguishably when forwarding. {from [RFC3031]}

MPLS label switch hop — The hop between two MPLS nodes, on which forwarding is done by use of labels. {from [RFC3031]}

MPLS label switched path — The path through one of more Label Switched Routers (LSRs) at one level of the hierarchy followed by a packet of a particular FEC. {from [RFC3031]}

MPLS merge point — A node at which label merging is done. {from [RFC3031]}

MPLS node — A node which runs MPLS. An MPLS node will be aware of MPLS control protocols, will operate one or more layer 3 routing protocols, and will be capable of forwarding packets based on labels. {from [RFC3031]}

MPLS protection domain — The set of LSRs over which a working path and its corresponding recovery path are routed. {source [Shar02]}

MPLS protection plan — The set of all LSP protection paths and the mapping from working to protection paths deployed in an MPLS protection domain at a given time. {source [Shar02]}

MRA Multilateral Roaming Agreement — Cooperation agreement between a wireless ISP and a central legal entity where the former guarantees the provision of services to the need users of the members of the latter. The central legal entity guarantees the setting of minimum standards of service for existing and future members. The result of the cooperation agreement is that end users can roam in the network of the members of the platform.

MSC Mobile Switching Centre — The switching node in a GSM network. http://www.etsi.org

MTX Mobile Telephone Exchange — The switching node in an NMT network.

MVNO Mobile Virtual Network Operator — An actor offering mobile services not having its own mobile network, but relying on other actors to provide the network resources. (Ref. *Telektronikk,* 97 (4), 2001)

NB Narrowband

NCF Net Cash Flow

NE Network Element

Network handover — Handover of an uninterrupted session between different networks or access points.

Network hierarchy — An abstraction of part of a network's topology, routing and signalling mechanism. The abstraction may be used as a mechanism to build large networks or as a technique to enforce administrative, topological, or geographic boundaries. For example, network hierarchy might be used to separate the metropolitan and long-haul regions of a network, or to separate the regional and backbone sections of a network, or to interconnect service provider networks (with BGP which reduces a network to an autonomous system). Two perspectives may be taken on:

- Vertically oriented: between two network technology layers;
- Horizontally oriented: between two areas or administrative subdivisions within the same network technology layer. {source [RFC3386]}

Network survivability
The capability to provide a prescribed level of QoS for existing services after a given number of failures occur within the network. {source [RFC3272]}

Next-gen SDH
Including the Ethernet functionality in SDH and SONET equipment using GFP, LCAS and virtual concatenation. Next-gen SDH may also support virtual switching, including multipoint connections.

NHLFE
Next Hop Label Forwarding Entry

NMT
Nordic mobile telephony
Automatic mobile telephone system based on analogue transmission technology. NMT was developed by the Nordic public telephone administrations (Norway, Sweden, Denmark, Finland and Iceland) in the period 1969 to 1980. The first version operated in the 450 MHz band (NMT-450) and was launched in 1981/1982. Later the system was enhanced to operate in the 900 MHz band (NMT-900, 1986). The system offered voice telephony with international roaming. The technology used was narrowband frequency modulation (FM) with 25 kHz user channels. The NMT-900 service was discontinued to release the frequency resources in 2001 when the GSM coverage had reached a sufficiently high level. In Norway the NMT-450 system will be discontinued at the end of 2004. (Ref. *Telektronikk,* 91 (4), 1995)

NNI
Network-Network Interface

Node-disjoint
Two connections are node-disjoint if they do not share any node along the path except the ingress and egress nodes.

Non-revertive mode
Case where there is no preferred path or it may be desirable to minimize further disruption of the service brought on by a revertive switching operation. A switch-back to the original working path is not desired or not possible since the original path may no longer exist after the occurrence of a fault on that path. {source [RFC3386]}
A recovery mode in which traffic is not automatically switched back to the original working path after this path is restored to a fault-free condition. {source [Shar02]}

NOC
Network Operations Centre
The physical space from which a typically large telecommunication network is managed, monitored and supervised. The NOC coordinates network troubles, provides problem management and router configuration services, manages network changes, allocates and manages domain names and IP addresses, monitors routers, switches, hubs and UPS that keep the network operating smoothly, manages the distribution and updating of software and coordinates with affiliated networks.

Normalization
Sequence of events and actions taken by a network that returns the network to the preferred state upon completing repair of a failure. This could include the switching or rerouting of affected traffic to the original repaired working entities or new routes. Revertive mode refers to the case where traffic is automatically returned to a repaired working entity (also called switch-back). {source [RFC3386]}

NP
Network Performance

NPV
Net present value

NSP
Network Service Provider

NVoD
Near Video on Demand

OA
Ordered Aggregate
A set of BAs that share an ordering constraint. The set of PHBs that are applied to this set of Behaviour Aggregates constitutes a PHB scheduling class. {source [RFC3564]}

OA&M
Operation, Administration & Maintenance

OBLSR
Optical bi-directional line switched ring

OBPSR
Optical bi-directional path switched ring

OCh
Optical Channel

Offline traffic engineering
A traffic engineering system that exists outside of the network. {source [RFC3272]}

OIF
Optical Inter-working Forum

OMS
Optical Multiplex Section – layer network
The optical multiples section layer provides transport for the optical channels. The information contained in this layer is a data stream comprising a set of optical channels, which may have a defined aggregate bandwidth.

Online traffic engineering
A traffic engineering system that exists within the network, typically implemented on or as adjuncts to operational network elements. {source [RFC3272]}

Opaque vs. transparent optical networks
A transparent optical network is an optical network in which optical signals are transported from transmitter to receiver entirely in the optical domain without OEO conversion. Generally, intermediate switching nodes in a transparent optical network do not have access to the payload carried by the optical signals. Note that amplification of signals at transit nodes is permitted in transparent optical networks.
On the other hand, in opaque optical networks transit nodes may manipulate optical signals traversing through them. An example of such manipulation would be OEO conversion which may involve 3R operations (reshaping, retiming, regeneration, and perhaps amplification). {based on [ID-ipofw]}

Operational control
The process of assuring that specific tasks are carried out effectively and efficiently. {source [Anth65]}

OPEX
Operations Expenses

OPPR
Optical path protection ring

Optical channel trail of lightpath
An optical channel trail is a point-to-point optical layer connection between two access points in an optical network. {based on [ID-ipofw]}

Optical internetwork
An optical internetwork is a mesh-connected collection of optical networks. Each of these networks may be under a different administration. {based on [ID-ipofw]}

Optical mesh sub-network
An optical sub-network is a network of OXCs that supports end-to-end networking of optical channel trails providing functionality like routing, monitoring, grooming, and protection and restoration of optical channels. The interconnection of OXCs in this network can be based on a general mesh topology. The following sub-layers may be associated with this network.

OPTIMUM
Optimised Network Architectures for Multimedia Services
Research project in the European Union's 4th Framework Research Programme ACTS. The objective of the project was to establish guidelines for the introduction of advanced communications networks and multimedia services in a competitive multiservice environment. The project used techno-economic evaluation of different case studies derived from field trials and projects within and outside ACTS. Project duration was from 1996 to 1998.
http://www.cordis.lu/infowin/acts/rus/projects/ac226.htm, http://www.telenor.no/fou/prosjekter/optimum/

OSI
Open System Interconnection
Reference model from ISO for data communication divided into 7 main layers; 1 – physical, 2 – data link, 3 – network, 4 – transport, 5 – session, 6 – presentation, 7 – application.

OSPF
Open Shortest Path First

OSPR
Optical shared protection ring

OSS
Operations Support System

OTN
Optical Transport Network

OTNT
Optical Transport Networks & Technologies

OTS
Optical Transmission Section – layer network
This layer provides functionality for transmission of optical signals through different types of optical media. {based on [ID-ipofw]}

OULSR
Optical unidirectional line-switching ring

OUPSR
Optical unidirectional path-switched ring

OXC
Optical Cross-Connect
A space-division switch that can switch an optical data stream from an input port to an output port. Such a switch may utilise optical-electrical conversion at the input port and electrical-optical conversion at the output port, or it may be all-optical. An OXC is assumed to have a control plane processor that implements the signalling and routing protocols necessary for computing and instantiating optical channel connectivity in the optical domain. {based on [ID-ipofw]}

P2P
Peer-to-Peer

Path continuity test
A test that verifies the integrity and continuity of a path or path segment. {source [Shar02]}

PDH
Plesiochronous Digital Hierarchy

Peering
A collection of network elements supporting some form of interconnected operation among two or more entities owned by separate organizations. Examples include accounting peering, content list peering and distribution peering. {source [BCD01a]}

Performance management
A systematic approach to improving effectiveness in the accomplishment of specific networking goals related to performance improvement. {source [RFC3272]}

Performance measures
Metrics that provide quantitative or qualitative measures of the performance of systems or subsystems of interest. {source [RFC3272]}

Performance Metric
A performance parameter defined in terms of standard units of measurement. {source [RFC3272]}

PD
Path Degradation
Fault detected by MPLS-based recovery mechanisms indicating that the quality of the path is unacceptable. {source [Shar02]}

PDB
Per-Domain Behaviour
The expected treatment that an identifiable or target group of packets will receive from edge-to-edge of a DS domain. A particular PHB (or – if applicable – list of PHBs) and traffic conditioning requirements are associated with each PDB. {source [RFC3564]}

PF
Path Failure
Fault detected by MPLS-based recovery mechanisms that is defined as the failure of the liveness message test or a path continuity test, which indicates that the path connectivity is lost. {source [Shar02]}

PG
Path group
A logical bundling of multiple working path, each of which is routed identically between a PSL and a PML. {source [Shar02]}

PHB
Per-Hop-Behavior
The externally observable forwarding behaviour applied at a DiffServ-compliant node to a DiffServ behaviour aggregate. {source [RFC3564]}

PHY
Physical layer device
The Ethernet PHY at Layer 1 of the OSI model defines the electrical and optical signalling, line states, clocking guidelines, data encoding, and circuitry needed for data transmission and reception. Contained within the PHY are several sub-layers that perform these functions including the physical coding sub-layer (PCS) and the optical transceiver or physical media dependent (PMD) sub-layer for fibre media. The Ethernet PHY connects the media to the MAC (Layer 2).

PML
Path Merge LSR
An LSR that is responsible for receiving the recovery path traffic and either merges the traffic back on to the working path or, if it is itself the destination, passes the traffic on to the higher layer protocols. {source [Shar02]}

PNO
Public Network Operator

PoP
Point of Presence

POR
Point Of Repair

An LSR that is set up to perform MPLS recovery. In other words, an LSR that is responsible for carrying out the repair of an LSP. The POR can be a PSL or a PML depending on the recovery scheme employed. {source [Shar02]}

PPG
Protected path group

A path group that requires protection. {source [Shar02]}

Provisioning

The process of assigning or configuring network resources to meet certain requests.

PSC
PHB Scheduling Class

A PHB group for which a common constraint is that ordering of at least those packets belonging to the same micro-flow must be preserved. {source [RFC3564]}

PSDN
Public Switched Data Network

PSTN
Public Switched Telephony Network

Pre-emption priority

A method of determining which traffic can be disconnected in the event that not all traffic with higher restoration priority is restored after the occurrence of a failure. {source [RFC3386]}

Protection (also called protection switching)

Survivability technique based on predetermined failure recovery; as the working entity is established, a protection entity is also established. Protection techniques can be implemented by several architectures; 1+1, 1:1, 1:n and m:n. In the context of SDH they are referred to as Automatic Protection Switching (APS). {source [RFC3386]}

1+1 protection architecture

A protection entity is dedicated to each working entity. The dual-feed mechanisms is used whereby the working entity is permanently bridged onto the protection entity at the source of the protected domain. In normal operation mode, identical traffic is transmitted simultaneously on both the working and protection entities. At the other end (sink) of the protection domain, both feeds are monitored for alarms and maintenance signals. A selection between the working and protection entity is made based on some predetermined criteria, such as the transmission performance requirements or defect indication. {source [RFC3386]}

1:1 protection architecture

A protection entity is also dedicated to each working entity. The protection traffic is normally transmitted by the working entity. When the working entity fails, the protected traffic is switched to the protection entity. The two ends of the protected domain must signal detection of the fault and initiate the switchover. {source [RFC3386]}

1:*n* protection architecture

A dedicated protection entity is shared by *n* working entities. In this case, not all of the affected traffic may be protected. {source [RFC3386]}

***m*:*n* protection architecture**

A generalisation of the 1:*n* architecture. Typically $m < n$ where *m* dedicated protection entities are shared by *n* working entities. {source [RFC3386]}

Protection counterpart

The 'other' path when discussing pre-planned protection switching schemes. The protection counterpart for the working path is the recovery path and vice versa. {source [Shar02]}

Protection entity (also called backup entity or recovery entity)

Entity that is used to carry protected traffic in recovery operation mode, i.e. when the working entity is in error or has failed. {source [RF3386]}

Protection switching

A recovery mechanism in which the recovery path or path segments are created prior to the detection of a fault on the working path. Hence, the recovery path is pre-established. {source [Shar02]}

Protection switch time

Time interval from the occurrence of a network fault until the completion of the protection-switching operations. It includes the detection time necessary to initiate the protection switch, a hold-off time to allow for the interworking of protection schemes, and the switch completion time. {source [RFC3386]}

PSL
Path Switch LSR

An LSR that is responsible for switching or replicating the traffic between the working path and the recovery path. {source [Shar02]}

PTP
Protected Traffic Portion

The portion of the traffic on an individual path that requires protection. {source [Shar02]}

PVR
Personal Video Recorder

QoS routing

Class of routing systems that select paths to be used by a flow based on the QoS requirements of the flow {source [RFC3272]}.

RACE
Research in Advanced Communications in Europe

Name of the 2nd and 3rd Thematic Programmes in Information and Communication Technologies funded by the European Union. In the 1st Framework Research Programme (FP) the 'Community Action' RACE 0 was the definition phase from 1985–1986 and formulated a reference model for Integrated Broadband Communications (IBC). The first RACE programme (RACE 1) ran from 1987 to 1992 as a "Community programme in the field of telecommunications technologies". The 2nd RACE programme (RACE 2) ran from 1990 to1994. In the 4th FP it was followed by the ACTS Thematic Programme. http://www.cordis.lu

RADIUS
Remote Authentication Dial-In User Service

An authentication and accounting system used by many (W)ISPs. When logging into a public Internet service you must enter your user name and password. This information is passed to a RADIUS service, which checks that the information is correct, and then authorises access to the WISP. Though not an official standard, the RADIUS specification is maintained by a working group of the IETF. http://www.ietf.org/

RAENPV
Risk-adjusted expected net present value

Recovery

Sequence of events and actions taken by a network after the detection of a failure to maintain the required performance level for existing services (e.g. according to service level agreements) and to allow normalisation of the network. The actions include notification of the failure followed by two parallel processes: i) a repair process with fault isolation and repair of the failed components; and ii) a reconfigurations process using survivability mechanisms to maintain service continu-

	ity. In protection, reconfiguration involves switching the affected traffic from a working entity to a protection entity. In restoration, reconfiguration involves path selection and rerouting of the affected traffic. {source [RFC3386]}
Recovery path	The path by which traffic is restored after the occurrence of a fault. Hence, the path onto which the traffic is directed by the recovery mechanisms. Back-up path, alternative path and protection path are also seen as synonyms to recovery path. {source [Shar02]}
Request manager	The system that responds to the user's request to view a particular piece of content. This is usually determined by the URI of the content. It could also be a fixed proxy. This system could be owned by the content provider or content network. {source [BCD01a]}
Rerouting	A recovery mechanism where the recovery path or path segments are created dynamically after the detection of a fault on the working path. Hence, the recovery path is not pre-established. {source [Shar02]}
Restoration (also called recovery by rerouting)	A survivability technique that establishes new paths or path segments on demand, for restoring affected traffic after the occurrence of a fault. The resources in these alternate paths are the currently unassigned (unreserved resource in the same layer). Pre-emption of extra traffic may also be used if spare resources are not available to carry the higher-priority protected traffic. As initiated by detection of a fault on the working path the selection of a recovery path may be based on pre-planned configurations, network routing policies, or current network status such as network topology and fault information. Signalling is used to establish the new paths to bypass the fault. Thus, restoration involves a path selection process followed by rerouting of the affected traffic from a working entity to the recovery entity. {source [RFC3386]}
Restoration priority	Method of giving preference to protect higher-priority traffic ahead of lower-priority traffic. Its use is to help determine the order of restoring traffic after a failure has occurred. The purpose is to differentiate service restoration time as well as to control access to available spare capacity for different classes of traffic. {source [RFC3386]}
Restoration time	Time interval from the occurrence of a network fault to the instant when the affected traffic is either completely restored, or until spare resources are exhausted, and/or no more extra traffic exists that can be pre-empted to make room. {source [RFC3386]}
Revertive mode	Procedure in which revertive action, i.e. switch-back from the protection entity to the working entity, is taken once the failed working entity has been repaired. In non-revertive mode, such action is not taken. To minimize service interruption, switch-back in revertive mode should be performed at a time when there is least impact on the traffic concerned, or by using the make-before-break concept. {source [RFC3386]} A recovery mode in which traffic is automatically switched back from the recovery path to the original working path upon the restoration of the working path to a fault-free condition. This assumes a failed working path does not automatically surrender resources to the network. {source [Shar02]}
Roaming	The possibility of changing operator or end user. Roaming does not refer to technical handover between access points or different networks such as GPRS and WLAN.
RSS Remote subscriber stage	
RSU Remote subscriber unit	
SAP Service access point	
SCP Service Control Point	
SDH Synchronous Digital Hierarchy	
SIM Subscriber Identity Module	
SLA Service Level Agreement	A contract between a provider and a customer that guarantees specific levels of performance and reliability at a certain cost. {source [RFC3272]}
SMATV Satellite master antenna television	
SME Small Medium Enterprise	
SMP Service Management Point	
SMP Significant Market Power	
SMS-C Short Message Service Centre	
SNCP Subnetwork Connection Protection	SDH protection schema.
SOHO Small Office Home Office	
SONET Synchronous Optical NETwork	
SPRings Shared protection rings	
SRG Shared risk group	Set of network elements that are collectively impacted by a specific fault or fault type. For example, a shared risk link group (SRLG) is the union of all the links on those fibres that are routed in the same physical conduit in a fibre-span network. Besides shared conduit, this concept includes other types of compromise such as shared fibre cable, shared right of way, shared optical ring, shared office without power sharing, etc. The span of an SRG, such as the length of the sharing for compromised outside plant, needs to be considered on a per fault basis. The concept of SRG can be extended to represent a 'risk domain' and its associated capabilities and summarisation for traffic engineering. {source [RFC3386]}

	A group of links or nodes that share a common risk component, whose failure can potentially cause the failure of all the links or nodes in the group. When referring to link resources, Shared Risk Link Group (SRLG) applies – an example being fibre links in the same conduit (all links are affected when the conduit is broken).
SS Special subscriber	
SSno7 Signalling System No. 7	A signalling protocol defined by the International Telecommunication Union (ITU), basically as a way to control circuit switched connections in ISDN although used in several systems (including GSM, IN).
SSP Service Switching Point	
Stability	An operational state in which a network does not oscillate in a disruptive manner from one mode to another. {source [RFC3272]}
STM Synchronous transport module	
Strategic planning	The process of deciding on the objectives of the organisation, on changes in these objectives, on the resources used to attain these objectives, and on the policies that are to govern the acquisition, use and disposition of these resources. {source [Anth65]}
Supply side congestion management	A congestion management scheme that provisions additional network resources to address existing and/or anticipated congestion problems. {source [RFC3272]}
Survivability	The capability of a network to maintain service continuity in the presence of faults within the network. Survivability mechanisms such as protection and restoration are implemented either on a per-ink basis, on a per-path basis, or throughout an entire network to alleviate service disruption at affordable costs. The degree of survivability is determined by the network's capability to survive single failures, multiple failures, and equipment failures. {source [RFC3386]}
Switch-back	The process of returning the traffic from one or more recovery paths back to the working paths. {source [Shar02]}
Switch-over	The process of switching the traffic from the path that the traffic is flowing on onto one or more recovery paths. This may involve moving traffic from a working path onto recovery paths or may involve moving traffic from a recovery path onto more optimal working paths. {source [Shar02]}
TA Traffic Aggregate	A collection of packets with a code-point that maps to the same PHB, usually in a DS domain or some subset of a DS domain. A traffic aggregate marked for PHB x is referred to as the x traffic aggregate or x aggregate. This generalizes the concept of Behaviour Aggregate from a link to a network. {source [RFC3564]}
TAP Transferred Account Procedure	A standard maintained by the GSM organization, by which GSM operators exchange roaming billing information. This is how roaming partners are able to bill each other for the use of network and services through a standard process. The TAP files are generated and sent, at the latest 36 hours from call end time. This means that operators can send one or many TAP files per day. TAP files contain rated call information according to the operator's Inter-operator Tariff (IOT), plus any bilaterally agreed arrangements or discounting schemes. The transfer of TAP records between the visited and the home mobile networks may be performed directly, or more commonly, via clearing house. Invoicing between the operators then normally happens once per month.
TCBH Time-Consistent Busy Hour	
TCP Transmission Control Protocol	
TDM Time Division Multiplexing	
TE Traffic Engineering	
TERA Techno-Economic Results from ACTS	Research project in the ACTS programme in FP4. The main objective of this project is to support consolidation, condensing and rationalising of deployment guidelines for introduction of advanced communication services and networks. http://www.telenor.no/fou/prosjekter/tera/; http://www.cordis.lu/infowin/acts/rus/projects/ac364.htm
TEX Transit Exchange	
TIA 568	The U.S. standard for building cabling.
TITAN Tool for Introduction strategies and Techno-economic evaluation of Access Network	EU research project.
TONIC TechnO-ecoNomICs of IP optimised networks and services	Research project in the IST programme in FP5. It concentrates on techno-economic evaluation of new communication networks and services, in order to identify the economically viable solutions that can make the Information Society to really take place. http://www-nrc.nokia.com/tonic/
TP Transmission point	
Traffic characteristic	A description of the temporal behaviour or a description of the attributes of a given traffic flow or traffic aggregate. {source [RFC3272]}
Traffic engineering system	A collection of objects, mechanisms and protocols that are used conjunctively to accomplish traffic engineering objectives. {source [RFC3272]}
Traffic flow	A stream of packets between two end-points that can be characterized in a certain way. A micro-flow has a more specific definition: A micro-flow is a stream of packets with the same source and destination addresses, source and destination ports, and protocol ID. {source [RFC3272]}
Traffic intensity	A measure of traffic loading with respect to a resource capacity over a specified period of time. In classical telephony systems traffic intensity is measured in units of Erlang. {source [RFC3272]}

Traffic matrix A representation of the traffic demand between a set of originating and destination abstract nodes. An abstract node can consist of one or more network elements. {source [RFC3272]}

Traffic monitoring The process of observing traffic characteristics at a given point in a network and collecting the traffic information for analysis and further action. {source [RFC3272]}

Traffic trunk An aggregation of traffic flows of the same class which are placed inside a Label Switched Path. {source [RFC3564]}
A traffic trunk is an abstraction of traffic flow traversing the same path between two access points which allow some characteristics and attributes of the traffic to be parameterised. {based on [ID-ipofw]}
An aggregation of traffic flows belonging to the same class which are forwarded through a common path. A traffic trunk may be characterized by an ingress and egress node, and a set of attributes which determine its behavioural characteristics and requirements from the network. {source [RFC3272]}

Transit traffic Traffic whose origin and destination are both outside of the network under consideration. {source [RFC3272]}

Trust domain A trust domain is a network under a single technical administration in which adequate security measures are established to prevent unauthorized intrusion from outside the domain. Hence, it may be assumed that most nodes in the domain are deemed to be secured to trusted in some fashion. Generally, the rule of 'single' administrative control over a trust domain may be relaxed in practice if a set of administrative entities agree to trust one another to form an enlarged heterogeneous trust domain. However, within a trust domain, any subverted node can send control messages which can compromise the entire network. {based on [ID-ipofw]}

TSC Transit Switching Centre

UDP User Datagram Protocol

UMTS Universal Mobile Telecommunication System — The candidate for a 3rd generation system standardised and promoted by 3GPP. The standardisation work began in 1991 by ETSI but was transferred in 1998 to 3GPP as a cooperation between Japanese, Chinese, Korean and American organisations. It is based on the use of WCDMA technology and is currently deployed in several European countries. The first European services opened in 2003 and more are expected in 2004 and 2005. In Japan, NTT DoCoMo opened its 'pre-UMTS' service FOMA (Freedom Of Mobile multimedia Access) in 2000. The system operates in the 2.1 GHz band and is capable of carrying multimedia traffic. It offers a maximum user bit rate of 2 Mbit/s, however more realistic rates are 144 kbit/s and 384 kbit/s for data services. http://www.3gpp.org/

UPSR Uni-diretional path switched ring

UPT Universal Personal Telecommunications

URI Universal Resource Identity

USO Universal Service Obligation

Vertical hierarchy An abstraction, or reduction in information, which would be of benefit when communicating information across network technology layers, as in propagation information between optical and router networks.
In the vertical hierarchy the total network functions are partitioned into a series of functional or technological layers with clear logical, and maybe even physical separation between adjacent layers. Survivability mechanisms either currently exist or are being developed at multiple layers in networks. The optical layer is now becoming capable of providing dynamic ring and mesh restoration functionality, in addition to traditional 1+1 or 1:1 protection. The SDH layer provides survivability capability with automatic protection switching (APS), as well as self-healing ring and mesh restoration architectures. Similar functionality has been defined in the ATM layer with work ongoing to also provide such functionality using MPLS. At the IP layer rerouting is used to restore service continuity following link and node outages. Rerouting at the IP layer, however, occurs after a period of rerouting convergence, which may require from a few seconds to several minutes to complete. {source [RFC3386]}

Visited provider The provider that provides services to a user with whom the service providing provider does not have any legal ties but who is a subscriber of a provider with whom a roaming agreement exists.

VoD Video on demand

VoIP Voice-over-IP — Voice transmission using Internet Protocol to create digital packets.

VPN Virtual Private Network — A network that is constructed by using public wires to connect nodes. For example there are a number of systems that enable you to create networks using the Internet as the medium for transporting data. These systems use encryption and other security mechanisms to ensure that only authorized users can access the network and that the data cannot be intercepted.

Wavelength continuity property A lightpath is said to satisfy the wavelength continuity property if it is transported over the same wavelength end-to-end. Wavelength continuity is required in optical networks with no wavelength conversion feature. {based on [ID-ipofw]}

Wavelength path A lightpath that satisfies the wavelength continuity property is called a wavelength path. {based on [ID-ipofw]}

WDM Wavelength division multiplexing — A means of data transmission that uses optical multiplexing to enable two or more wavelengths, each with its own data source, to share a common fibre optic medium.
Wavelength Division Multiplexing is a technology that allows optical signals operating at different wavelengths to be multiplexed onto a single optical fibre and transported in parallel through the fibre. In general, each optical wavelength may carry digital client payloads at a different data rate and in a different format. For example, a mixture of SDH 2.5 Gbit/s and 10 Gbit/s can be carried over a single fibre. An optical system with WDM capability can achieve parallel transmission of multiple wavelengths gracefully while maintaining high system performance and reliability.

	In the near future, commercial dense WDM systems are expected to concurrently carry more than 160 wavelengths at rates 10 Gbit/s or above, resulting in more than 1.6 Tb/s. {based on [ID-ipofw]}
Wi-Fi alliance	A non-profit international association formed in 1999 to certify interoperability of wireless Local Area Network products based on IEEE 802.11 specification. Currently the Wi-Fi Alliance has 207 member companies from around the world, and over 1000 products have received Wi-Fi® certification since certification began in March 2000. http://www.wifialliance.org/
WISP Wireless Internet Service Provider	A commercial provider of WLAN services in public places.
WLAN Wireless Local Area Network	Usually used to define wireless access to local area networks based on IP. The term is often used specifically to denote the 802.11 standards, or Wi-Fi. A WLAN access point (AP) usually has a range of 20–300 m. A WLAN may consist of several APs and may or may not be connected to the Internet.
WLL Wireless Local Loop	A means of provisioning a local loop facility without wires. Usually employing low power radio systems running in the microwave range. Widely deployed in Asia and other developing countries where they offer the advantages of rapid deployment, and rapid configuration and reconfiguration, as well as avoidance of the cost of burying wires and cables. WLL is also attractive to new operators bypassing the incumbent PNOs in a deregulated, and competitive environment.
Working entity	Entity that is used to carry traffic in normal operation mode. Depending upon the context, an entity can be a channel or a transmission link in the physical layer, a Label Switched Path (LPS) in MPLS or a logical bundle of one or more LSPs. {source {RFC3386]}
Working path	The protected path that carries traffic before the occurrence of a fault. The working path exists between a PSL and PML. The path is a hop-by-hop routed path, a trunk, a link, an LSP or part of a multipoint-to-point LSP. Commonly primary path and active path are also seen as synonyms to working path. {source [Shar02]}
WTSA World Telecommunication Standardization Assembly	The World Telecommunication Standardization Assembly (WTSA) is held every four years by the International Telecommunication Union (ITU), a specialised agency of the United Nations, to define general policy for ITU's Telecommunication Standardization Sector (ITU-T Sector). WTSA adopts the work programme, sets up the necessary Study Groups (SGs), designates chairmen and vice-chairmen of the Study Groups, reviews the working methods and approves Recommendations. The WTSA is open for participation from both Member States and Sector Members (operators and industry) of the T-Sector. http://www.itu.int/
xDSL	Various configurations of digital subscriber line: X = ADSL – asymmetric, VDSL – very high speed, SHDSL – single pair high speed, SDSL – symmetric, HDSL – high speed.

References

[Anth65] Anthony, R N. *Planning and Control System; A Framework for Analysis*. Cambridge, MA, Harvard University Press, 1965.

[BCD01a] Broadband Content Delivery forum: *Requirements for end-to-end delivery of broadband content*. 2001.

[ID_ipofw] Rajagopalan, B et al. *IP over Optical Networks: A Framework. IETF draft*. Draft-ietf-ipo-framework-05.txt. Oct. 2003. Work in progress.

[ID-Temea03] Lai, W S, Tibbs, R W, Van den Berghe, S. *Requirements from Internet Traffic Engineering Measurement*. Draft-ietf-tewg-measure-06.txt. July 2003. Work in progress.

[MEFsite] *Metro Ethernet Forum*. www.metroethernetforum.com

[RFC3272] D. Awduche et al. *Overview and Principles of Internet Traffic Engineering*. IETF RFC 3272. May 2002.

[RFC3386] Lay, W, McDysan, D. *Network hierarchy and multiplayer survivability*. IETF RFC 3886. Nov. 2002.

[RFC3564] Le Faucheur, F, Lai, W. *Requirements for Support of Differentiated Services-aware MPLS Traffic Engineering*. IETF RFC 3564. July 2003.

[Shar02] Sharma, V, Hellstrand, F. *Framework for MPLS-based Recovery. IETF draft*. Draft-ietf-mpls-recovery-frmwrk-07.txt. Mar. 2002.