26
Performance issues in Intelligent Networks

T. Jensen
Telenor Research and Development,
P.O.Box 83, N-2007 Kjeller, Norway
Tel. +47 63 84 88 39
Fax. +47 63 81 00 76
E-mail: terje.jensen@fou.telenor.no

Abstract

Intelligent Networks used as basis for an increasing number of services places more emphasis on the corresponding performance issues. Although ensuring sufficient service quality has been essential for most network operators, diversified customers and services together with growing competition request for effective utilisation of the network elements. A number of aspects have to be considered in order to maintain an effective operational network. In this paper, issues related to deployment of Intelligent Networks, including relevant services and sizing of network elements, characterisation of services and interconnect are treated.

Keywords

Intelligent Network, performance, dimensioning

Intelligent Networks and Intelligence in Networks D. Gaiti (Ed.)
Published by Chapman & Hall © 1997 IFIP

1 INTRODUCTION

Variants and usage of services based on Intelligent Network (IN) solutions grow steadily. The complexity, measured in terms of number of processing steps and devices involved for handling a call, does also seem to increase. Although one of the arguments for describing the IN concept was to ease service administration, e.g. (Q.12xx), the additional processing, signalling and usage of devices, may lead to that bottlenecks arise. In addition, patterns of service usage could further result in potential problems in certain portions of an IN.

Utilisation of network elements and resulting service quality would depend on the principles applied for deploying an IN. Two examples are overlay networks and integrated solutions. Although applying different philosophies for rolling out the network elements, similar questions with respect to performance are met. Such questions are given for a holistic view as well as for more specific aspects. A number of publications have been issued on these questions, like (Ramaswami, 1995) and (Pandya, 1994) for overall descriptions. Response times and related performance issues for querying data bases and personal communications have also been treated in papers, like (Demounem, 1992), (Saito, 1994), (Kwiatkowski, 1995). Corresponding analyses of the signalling network have also been carried out, e.g. (Bafutto, 1994). Other issues have also been examined. From an operational point of view, the relevant network elements and service logic/data must work together in a holistic sense. The presence of a number of actors involved in service handling could clutter the picture of determining the better ways of implementing services. In particular, when the different actors are not co-ordinated, appropriate mechanisms should be incorporated in the network solution.

One of the main objectives of this paper is to describe performance issues to be considered in relation to deployment of INs. Several questions arise during the belonging activities. The nature of these questions depends on the environment in which an actor is situated. As the resulting performance is tightly coupled with the dimensioning process, input data and scopes for carrying out dimensioning are treated in Section 2. Some specific issues of network elements are described in Section 3 outlining potential bottlenecks and studies to be undertaken. However, observing the performance from the users' point of view, a complete implementation should be examined. This also includes characterising services and service demands as presented in Section 4. Topics resulting from the presence of multiple actors are treated in Section 5.

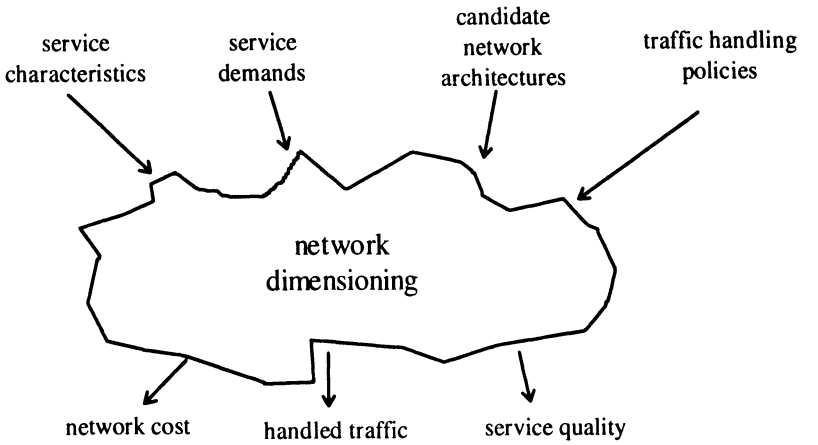

Figure 1 Potential input and output data for a dimensioning process

2 DIMENSIONING PROCESSES

The scopes for dimensioning an IN varies from establishing a network in a greenfield area to re-calculating the concerned parameters after changing the service demands. In addition, management activities are included in the performance studies. Commonly, different approaches are defined for the different scopes, natural as the possible means that can be undertaken vary. For a greenfield area, an optimisation problem can be formulated, like to minimise the cost of deploying a network when handling a set of service demands while meeting a set of requirements. In principle, the process can be illustrated as in Figure 1.

Characterising the services includes describing the network resources used by a service invocation. That is, load implied on the circuit switched connections, signalling links, processing elements and other relevant devices. The adequate service quality requirements are to be stated as well. Demand patterns for the different service must be given. Typically, these are related to a number of reference periods, leading to some mixtures of service usage. As the different customer groups often use different service variants, these variants could have peak demands at noncoinciding time periods. When the service variants are invoked and the user behaviour influence the load on network resources, identifying the time period resulting in highest load on a set of resources may be involved. Carrying out the examinations for several time periods are therefore important.

Candidate network structures have to be specified. For several scopes, however, the network structure is given (one candidate only). For studies of greenfield

areas, candidate locations of network elements including links will be specified. At some locations network elements could be already installed. Such aspects must be incorporated in the procedure allowing for flexibility in the configurations that can be considered. The candidate locations have to be given for all types of network elements considered, like SSPs, IPs, STPs, SCPs and SDPs. In addition, a number of combinations of the functional entities as well as equipment from different vendors could be taken into account.

Traffic handling candidates include routing and load control policies. Different sets of candidates could be given for the different service variants. These could allow for introducing priorities for some services and customers. Potential policies for traffic handling differ for the different portions on an IN. Although each of these could be studied in detail, reaching holistic profitable solutions are requested.

Output from the activity is a description of the network solution. The results include the cost of the network, traffic handled and the corresponding service quality. In addition, more specific data can be given, like utilisation of certain network elements and requirements for available storage devices.

Depending on the flexibility and current equipment in place, different scopes could be relevant as seen from a network operator's point of view, ref. Figure 2. Naturally, all of these scopes may not be of interest for all operators. The axis named time scale/flexibility indicates how much of the network is assumed to be given. Often, there is a correspondence between the time scale and the flexibility. For instance, in a long term solution, more possible candidates could be allowed. The axis named accuracy indicates the level of detail usually considered during the evaluations.

As indicated in Figure 2, when the flexibility increases, less accuracy may be considered. One aspect of this is that measurements and detailed information can be obtained for an existing network. For a greenfield study, however, more coarse descriptions are usually considered.

Dimensioning a network, the topology could be given. Then, finding capacities of the network elements and the relevant links is requested. All the input and output data outlined above may not be relevant for every case that is faced and approach that is used.

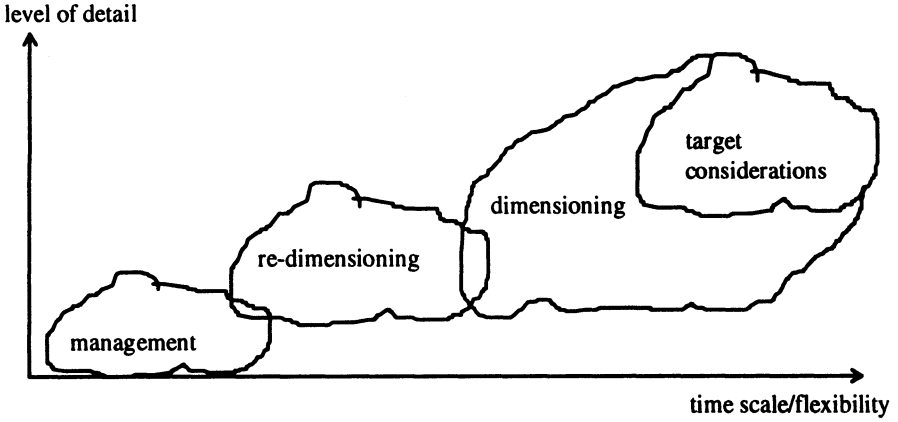

Figure 2 Scopes of network view

The IN-related part of the network can be divided into at least two portions: the circuit switched network and the signalling network. For dimensioning the circuit switched part, traditional methods can be applied. However, for certain service implementations special phenomena should be considered. A number of legs have been identified, e.g. involving the use of IP. Different holding times for the different portions could also be present. In addition, for some service calls, only a connection between a user and an SSP/IP is established. That is, a second user may not be involved. Typically, these aspects are considered when the traffic matrices are established.

Similar comments can be attached when dimensioning the signalling network. The lengths of signalling messages can be different implying that allowed arrival rates may be lower compared to other applications of the signalling protocols. Specified mechanisms for load control could also influence the characteristics of service implementations. However, it is questionable whether the load control should be considered during dimensioning or if these mechanisms can be introduced afterwards to ensure that specific measures are reached.

When locations and capacities of these parts are considered as variables, an optimisation problem could be formulated. In case an estimate of the network cost is to be minimised, this could be calculated as the sum of costs for elements and connections when their corresponding capacities are considered. Main constraints to be fulfilled could be derived from the service quality requirements. The network dimensioning is usually performed for a large portion of a network where the capacities of most nodes and link/circuit sets are subject to changes.

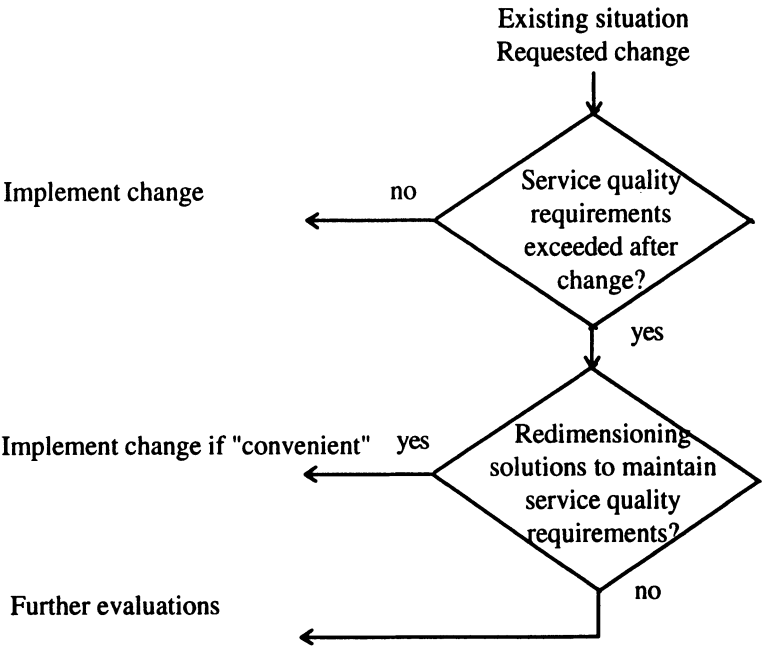

Figure 3 Redimensioning procedure

In a shorter time scope, like when a new IN-based service is introduced or the demand for an existing IN-based service is changed, another approach could be applied. In this procedure, the revised service demand can be added to the demand of an existing network in order to check whether or not this network is able to handle this demand satisfactorily. In case the answer is affirmative, the procedure is ended. Otherwise, a bottleneck is to be identified and measures undertaken to resolve this bottleneck. The procedure is iterated until no more bottlenecks are found. In some cases, major changes could be needed to resolve the congestion. Then, the dimensioning procedure could be started implying that more equipment may be needed, the change of service demand is postponed or service quality reduction could be expected. This procedure is illustrated in Figure 3, based on the description found in (E.734).

In case specific means have to be undertaken in order to maintain the service quality, the corresponding cost could be estimated in order to decide whether or not the activities should be carried out.

The management scope refers to mechanisms implemented which take care of traffic handling in the operational state. Examples of such mechanisms are load control and failure protection.

For all these scopes, elaborating adequate performance models and analyses are fundamental. In particular, performing sensitivity studies on selected groups of input data are requested in order to identify the critical factors. Then, these factors could be modelled more accurately and followed more closely.

3 NETWORK ELEMENTS

Dimensioning a network element, similar input data as for the network dimensioning processes can be identified:

- Load described by arrival process for each class, usage of resources and corresponding service times.
- Characteristics of components used to implement the network element. Both hardware and software architectures must be given. In addition, mapping of software blocks onto the hardware units have to be described. This may also describe candidate policies for handling the load.
- Requirements to be fulfilled, e.g. given by thresholds for delays and blocking probabilities.

For each type of network element, a suitable algorithm taking these input data and finding the following output data should be described:

- Number of units for each type of hardware component. In case these are grouped, differing in functionality or accessibility, the corresponding grouping must also be given.
- Resulting performance for each class.
- Resulting element cost for the network element.
- Service demand that is handled for each class.

A number of scopes for dimensioning a network element could be relevant. For instance, estimating resulting performance for an element when the load is given could be one task. At the other end, optimising the design of the element, e.g. minimising its cost for a mixture of loads could also be carried out. For a network operator, these scopes may differ for the different elements as some of them are tightly integrated with other functions, like an SSF, while others could even be implemented based on specifications from the operator.

The performance models relevant for the different network elements must capture the role played by the element. For instance, related to an STP, signalling load including processing of signalling messages has to be considered. For an SSP, both signalling load and load related to circuit switched connections have to be included. This is also influenced by whether or not the IN-based service handling is integrated or not with other services. In the former case, additional load resulting from IN-related traffic could be considered in the performance studies. However, when the situation for IN-based services is examined, any other services may be modelled as an additional class.

Which hardware and software architectures that are used for an element have to be reflected in the model. This may strongly influence which component that may become the bottleneck. Such a bottleneck, however, may depend on the mixture of classes that is assumed.

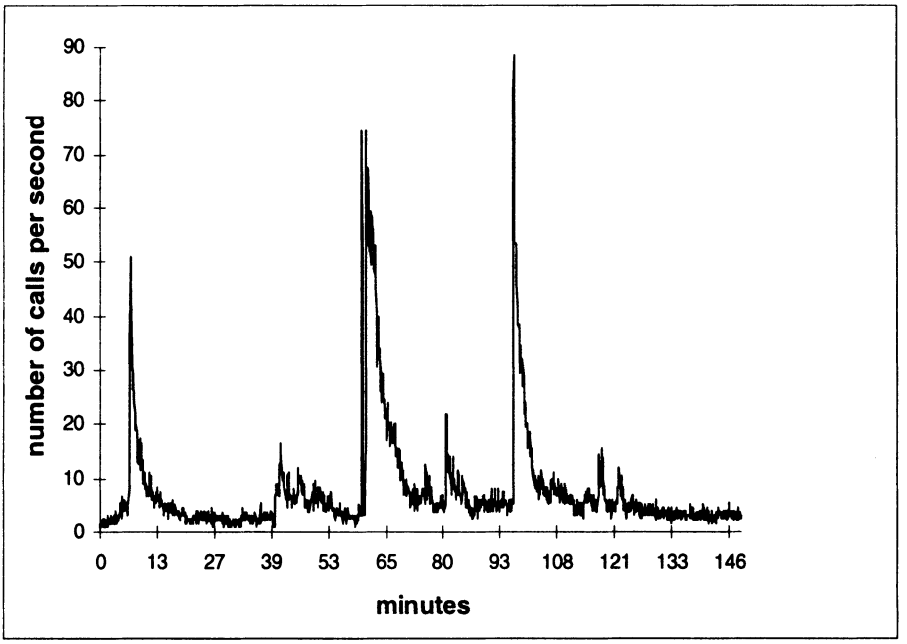

Figure 4 Mass calling situation

Related to IPs and SNs (or SRFs in general), specific concerns on how the load is distributed on the device groups should be taken. In particular, if specialised devices are present, their capacities should be tuned according to the mixture of loads foreseen.

In addition, aspects of the management system could be of concern for performance studies. In case users may access their own profiles through this system, examinations of the load and corresponding performance variables should be undertaken.

4 SERVICE CHARACTERISTICS

Characterising services includes the task of estimating the arrival processes and the sets of network resources requested for every class of service. A suitable division into classes has to be made, balancing the trade-off between having a intractable number of classes and sufficient level to assess the specific characteristics. Based on the assumed usage of services, the arrival processes might be given as Poissonian or not. The former could be used when there is no correlation between the different users. However, for some services, a common event could trigger the service invocation. One example of such a phenomenon is the mass calling service as illustrated in Figure 4. From these measurements, instants when the directory number was announced can clearly be identified.

Capturing parameter values for the arrival processes, a reference period has often been chosen. As the demand patterns for the different services change during the day, a number of reference periods could be used. A simple case could be to select one period for the working hours and another period for the evening period. The former may be related to mostly business users while the latter to mostly private users. Finding penetration and usage of the different services for the different market segments could be undertaken in order to better describe the market.

Values for traffic interests describing which directory number series that have been used describe the flow of traffic loads in the network. In particular, distribution of the load on the legs after SSPs could be derived, although this may also be influenced by the service implementation, like time dependent routing.

As basis for network dimensioning traffic matrices are often elaborated. These matrices should take into account the network elements into where the functional entities have been mapped. Matrices both for circuit switched portion and signalling portion could be used. Relationships between values in these matrices could be identified, e.g. derived from a description of the service. One example is that number of messages exchanged between an SSP and an SCP could be found from examining the path through the service script. In case the script path may vary depending on the user's behaviour, adequate information must be added.

Requirements for the service quality are either stated as blocking or as delays (E.724). Delays can be estimated by applying queueing network models. For certain cases, additional models could be requested. Estimates for delay and blocking are required for networks and individual elements. In addition, as mass calling services can be based on IN, transient studies of blocking due to abrupt changes of the service demand are requested, e.g., see (Jormakka 1995).

Currently, most requirements are given as seen from the users' point of view. When IN-based services are considered, these requirements could be given for the additional delay/blocking implied by using the IN concept. For several services, however, current implementations may not exist or several options could exist. Different approaches for defining values of service quality parameters could be examined. One way is to decompose the chain of elements which interactions need for handling the corresponding call. The resulting values could then be estimated by proper additions and/or multiplications of the components involved. Corresponding estimates for each of the components must also be found which is in line with describing a reference connection and allocating the service quality degradation to the sections involved. Requirements for each network element must be stated by identifying relevant contributions to the resulting service quality.

Load control both for individual network elements and for networks should also be included. Proper mechanisms, ways of deciding values or parameters and effect of the load control mechanisms must be treated. In principle, the traffic load can be limited by introducing mechanisms in several places in the network,

like in the circuit switched part, in SSPs and in SCPs. A specific objective to be achieved by applying load control is to ensure fairness between services as well as to reserve capacity for certain services (e.g., emergency services). In addition, load control applied for mass calling services should be included.

Measurement schemes for IN-based services must be described. As far as possible, existing measurement methods should be applied, like measurements of circuit switched traffic, number/length of signalling messages and processing loads. The introduction of additional functional entities, like SCF (Service Control Function) and SRF (Specialised Resource Function), requires the definition of additional measurements both of service usage and network performance.

The service demands as appearing from the networks' point of view, are derived from the users' interest although influenced by a range of circumstantial factors. The service quality and charges are also two factors having impact on these transformations. Network cost could be used as basis for deciding the charges (cost-based charges), although other principles could also be applied. The cost of the network is found after dimensioning, taking the assumed service demands into account. The service quality variables can also be estimated. Other inputs are needed for those calculations, as described in Section 2. The service quality and cost could be fed back in order to estimate service demands. In case no changes of service demands result, the task is finished, while changes invites for recalculating the network solution. In principle this could be regarded as an outer iteration loop. The iterations are continued until the convergence criteria are met. Naturally, this could be a tedious activity depending on the dimension of the problem and the rate of convergence. In addition, convergence may not be guaranteed. However, a major challenge is to describe the relationships between service demands and uses' interests and to capture values for the other effects. To a certain extent historical observations could be applied, but new services and new environments may limit the validity of those observations. Then, simple relationships could be used in order to gain insight into the dynamics and expressiveness of the model. The more important issue is also to identify factors on which the outcome is more sensible to. That is, factors strongly influencing the results. Then, these factors could be further detailed in order to increase the accuracy and the understanding.

Interface identity

performance variable/meassure	traffic condition	performance threshold	reaction pattern	measurement scheme

Figure 5 Potential content of interconnect agreement related to an interface with corresponding content

5 MULTI ACTOR CONSIDERATIONS

Operating telecommunication networks have long traditions for interworking. That is, it should be possible to establish connections between users in networks managed by different organisations. Interconnect agreements must be described correspondingly. Although generic agreements could be proposed, the interfaces in relation to INs have different characteristics which should be reflected in the corresponding contract. Typically, such an agreement covers several issues, like legal, financial and technical. Some of the technical ones will be treated in the following.

An agreement should include aspects like the conditions for operation, how to assess these conditions and actions to be taken in case agreed levels on any of the conditions are exceeded. Naturally, the particularities on interfaces have to be considered when describing these aspects. For instance, delays may be more essential on signalling relationships while blocking probabilities are used for circuit switched connections. In the general case, both variables describing delays and variables describing blocking are relevant.

Interfaces could be identified horizontally and vertically. Horizontal means interconnecting network domains at the same functional level, e.g., between two exchanges. Vertical means connecting different levels in a functional view or a network architecture view. An example is connections between SCPs and SDPs. Functional relationships would exist between the different network elements. In addition, interfaces could be present without separate physical elements, like when service logic/data are executed on a platform provided by another actor. That is, an interface may be more diffuse than a physical link between network elements. Naturally, when interfaces are incorporated in an element, assessing the conditions may become more involved unless external equipment can be used.

Deciding thresholds for performance variables, the load conditions for when these are valid must be described, see Figure 5.

In addition, an actor would also like to know the characteristics of the loads at the ingress points. For instance, loads with higher variability or having sever correlation may influence the network such that the performance is significantly worse compared to situations where such effects are not present. Therefore, descriptions of traffic characteristics should also be given. Alternatively, counteracts could be taken if the offered traffic load deviates from the

characteristics. It might be tempting to either avoid such situations or making the relevant traffic flows confirm with the better characteristics (e.g., by applying shaping). Naturally, if the situation is such that no significant reduction of performance is foreseen, the traffic flow could be treated as it is. This may, in one way, seem similar to a "best effort" manner of treating the traffic load and could be stated as such in a contract. In that way the thresholds in the agreement may be regarded as minimum values which commonly are met and where higher values can be found when the network states allow for it.

Naturally, the actor will usually be the only instance having a complete view of the network state. In order to limit any disturbances following a network condition, it is a sound principle to choke the relevant traffic flows at the edges of the network. Considering interconnects, this means that suitable mechanisms should be present in the network elements associated with the ingress points.

Often, having described the conditions to appear on an interface, belonging measurements schemes are identified. Values for variables of the arrival processes, service mixture and the resulting performance are to be captured. As for every measurement, decision of when, where and what to measure must be made. That is, topics like time and duration, interface/location and events have to be specified. These are measurements which may be carried out by both parties of an interface. It must also be decided whether or not continuous measurements are to be performed. As measuring could be regarded as sampling, it is to be agreed upon if terms in a contract can be questioned based on a single measurement period or if a number of measurement periods have to be done of which several indicate that the terms can be questioned before the contract is renegotiated or other means are applied. This is also seen as a trade-off between the time for reactions (responsiveness of a scheme) and the effort needed for preparing for and carrying out the reactions.

In addition, measures treating sudden changes in the arrival processes must be present. Load control is an example of such quick response measures. Which measures to apply for an interface should also be stated in the interconnect contract. A number of measures, operating on a range of times scales, may be thought of.

Load control implying rejecting or delaying calls is considered as a feature utilised during operation. One of the purposes may be to avoid that a single group of services/call types seizes too large fraction of the capacity leading to degradation of the quality for other services/call types. This may be particularly relevant when mass calling services are introduced.

However, mass calling services could also mean that specific means should be taken by the neighbouring operators. One potential solution is that the dialled number is recognised as a mass calling type in those network domains as well, and the operators co-operate in order to collect the results, e.g., in case of televoting. Another potential solution is that the neighbouring operators recognise these calls and may be allowed to throttle a certain fraction (stated in the

contract). It may, however, happen that directory numbers not belonging to the predefined mass calling series are announced leading to mass calling situations towards these numbers. To cater for such circumstances, appropriate load control schemes have to be applied on the basis of directory number series. In case information about application of directory numbers is not exchanged between actors, such means could be needed.
Another example where load control between operators may be requested is when an operator chooses to reroute calls to other domains through a network not prepared for that situation. This may happen when the more direct connection is not available (e.g. the circuit group is disconnected because of failure). Although accounting rules may treat this situation by introducing financial compensations, using measures for not degrading the service quality for the remaining services could be more fruitful in order to keep ones reputation.
Ensuring availability and successful calls for calls originated in a network domain and destined for other domains, is also an issue. In particular, as the users may require explanations and possible compensations by the operator/provider dealing with the originating side. Therefore, an interconnect agreement has to incorporate such cases as well. That is, on the call level both outgoing and incoming situations must be considered and the view of both parties must be taken on.

6 CONCLUSIONS

Effective utilisation of the involved equipment related to IN implies that methods for service quality calculations and network planning are needed. In particular, additional elements and potential service demand patterns may request that methods currently applied for telecommunication networks should be revisited. Most operators seem to base their future service portfolios on solutions similar to INs. Having appropriate methods covering the issues raised in this paper will therefore be essential.
Handling the IN-based services implies more signalling and processing. In order to carry out the performance evaluations, service demands have to be characterised, meaning that the resource usage of a call and the users' requests for calls are described. Commonly, an IN is integrated with networks also dealing with non-IN-based services. Specific phenomena associated with IN-based services in the circuit switched network, the signalling network and the network elements should be examined.
As more customers are basing their businesses on available telecommunication services, utilising proper mechanisms for achieving dependable solutions which at the same time are cost effective will also be an issue of specific interest. In particular, as more actors can result in higher competition, providing services with appropriate level of service quality will be essential. More actors, may also

mean that more interconnect arrangements are needed. These interconnections should have agreements associated which also cover service quality issues.

REFERENCES

Bafutto, M.; Kühn, P.J. and Willmann, G. (1994) Capacity and Performance Analysis of Signaling Networks in Multivendor Environments. *IEEE JSAC.* Vol. 12, no. 3, 490-500.

Demounem, L. and Arai, H. (1992) A Performance Evaluation of an Integrated Control and OAM Information Transport Network with Distributed Database Architectures. *IEICE Trans. Commun.* E75-B, no. 12, 1315-1326.

E.724; ITU-T recommendation E.724: GOS parameters and target GOS objectives for IN-based services.

E. 734; ITU-T recommendation E.734: Methods for allocating and dimensioning Intelligent Network (IN) resources.

Jormakka, J. (1995) Calculation of blocking probability in televoting. *12th Nordic Teletraffic Seminar.* Helsinki, 97-107.

Kwiatkowski, M. (1995) Performance modelling of UPT networks. *ICUPC'95.* Tokyo, 543-547.

Pandya, R. (1994) Emerging Standards for PCS Traffic Performance. *ICUPC'94.* San Diego, CA, 581-585.

Q.12xx; ITU-T recommendation series Q.1200: Intelligent Network recommendation.

Ramaswami, V. (1995) The essential role of traffic performance analysis in Intelligent Networks. *Globecom'95.* Singapore, 1254-1259.

Saito, H. and Asaka T. (1994) Traffic aspect of personal telecommunications in intelligent networks. *Computer Networks and ISDN Systems*, 26, 1089-1099.