Performance Evaluation of a Local Approach for VPC Capacity Management

Sven-Olof LARSSON[†] *and* Åke ARVIDSSON[†], *Nonmembers*

SUMMARY   By reserving transmission capacity on a series of links from one node to another, making a virtual path connection (VPC) between these nodes, several benefits are obtained. VPCs will simplify routing at transit nodes, connection admission control, and QoS management by traffic segregation. As telecommunications traffics experience variations in the number of calls per time unit due to office hours, inaccurate forecasting, quick changes in traffic loads, and changes in the types of traffic (as in introduction of new services), there is a need to cope for this by adaptive capacity reallocation between different VPCs. We have developed a type of local VPC capacity management policy that uses an allocation function to determine the needed capacity for the coming updating interval, based on the current number of active connections. We suggest an allocation function that is independent of the actual traffic, and determine its optimal parameters and the optimal updating interval for different overhead costs. The local approach is shown to be able to combine benefits from both VP and VC routing by fast capacity reallocations. The method of signaling is easy to implement and evaluations indicate that the method is robust.
key words: ATM, virtual paths, VP capacity management

1. Introduction

To accept a new call a check must be made to ensure that there is enough capacity left to establish the call through a series of links between the end nodes. When a route is found, the required amount of capacity is reserved for the call. (By capacity we mean equivalent bandwidth [1], [2] or transmission capacity needed for a certain traffic.) The established call uses this logical connection, which is called a virtual channel connection (VCC). A virtual path connection (VPC) groups VCCs together to be handled as an entity. A VPC can be seen as reserved bulk capacity between two nodes. By using VPCs the acceptance of a new call is simplified because the routing and reservation of capacity has already been done.

A VPC network constitutes a higher layer which is logically independent of an underlying physical network. Having several VPC networks each supporting one type of traffic simplifies statistical multiplexing and quality-of-service (QoS) management.

There are always variations in telecommunications traffics. Traditional telephone networks have been di-

mensioned for the so called busy hour to cope with the maximum traffics. This means that much of the capacity will stay unused for most of the time. By using VPCs, the capacity allocation can be altered dynamically. This allows us to meet traffic variations by reshaping the VPCs in order to match the current demands. This means savings on the amount of capacity required in a network, if we can utilize non-coincidental busy hours to reallocate the capacity. The concept of VPCs and VCCs is supported in the asynchronous transfer mode (ATM) and in the synchronous digital hierarchy (SDH/SONET).

We have grouped VPC capacity management approaches into groups which are fundamentally different. This definition is based on the amount of information used in the calculation of the VPC capacity reallocation. We have found that the calculation can be centralized e.g. [3]–[5], distributed e.g. [6], or local e.g. [7]. (A lot of papers describing distributed approaches falls into our definition of a local approach.) The central approach has the ability to make VPC capacity reallocation based on global information. The idea of the local and distributed approaches is to increase the robustness and improve performance compared to a central approach, which is depending on a central computer. By assigning costs for rejected calls and overhead such as control messages, the performance of the various approaches can be evaluated and compared to each other.

Section 2 describes our proposal for a local approach for VPC capacity management. In Sect. 3 we describe the evaluation of the approach and give the results in Sect. 4. Finally, we conclude the paper in Sect. 5 and discuss further work in Sect. 6.

2. The Local Approach

Each node periodically makes a decision about whether to seize or release capacity on the VPCs originating from that node. The decision is based on the actual number of occupied connections on each VPC. In our study we have used two VPCs between each node pair (though the method supports an arbitrary number of VPCs). The VPC with the smallest number of hops is preferred and is labeled **PVPC**, while the other one is referred to as an optional VPC and labeled **OVPC**.

The VPC capacity management is done with help of signaling. The following four control messages are

used:
- Path finding   (PTH)   + Answer
- VPC Establishment   (EST)
- Increment request   (INC)   + Answer
- Decrement   (DEC)

2.1 Control Messages

PTH is used for path identification by broadcasting it from all nodes to all other nodes. The broadcasting can be done from time to time or at command to recover from faulty links[8],[9]. In our evaluation we have only used it once to initiate the management system. When a PTH arrives at the destination, an answer message will be sent back to the originating node, the same way the actual PTH has traveled. This message contains a route to the destination.

The node of origin selects some routes received as answers to PTH and puts them in a table. We have selected node disjoint paths (shortest and link disjoint paths give almost the same performance[10]). The paths are ordered by the number of links and the total physical distances. The VPCs are finally established by sending **EST**, which is a source routed message along the path enabling the intermediate nodes to set up their routing tables.

When the VPCs have been established the periodic management starts in each node. When more capacity is needed an **INC** message is sent on the VPC to find out the capacity allowed for the whole path. This means that a VPC get the minimum allowed capacity on the series of links. The amount of available capacity is stored in the INC on successive links. When it reaches the end node, indicating the available capacity, an answer message is sent back to the originating node.

An increment request message is first sent on the PVPC. If the request cannot be satisfied, the OVPC is tried. When trying to get capacity on the links a temporary reservation must be made. This makes interference from other requests impossible, but can result in deadlock. To avoid deadlocks the following procedure is applied. If a request message reaches a node where the next link is already reserved, a message will be sent back to the node of origin releasing its current reservations. The node of origin tries again after a random delay (sufficiently long).

When an originating node determines that capacity should be released, a **DEC** is sent. The capacity reservation is decreased on each traversed link. No answer message has been used for this message. Capacity on the OVPC is released first.

2.2 Calculation of Needed Capacity

The developed approach for local VPC capacity management is inspired by the one developed by Mocci et al.[7]. This method allocates just enough VPC capac-

ity to meet a predefined limit of call blocking (target blocking)[11].

The idea of this approach is to handle traffic variations in a short time scale, i.e. larger than the mean interarrival time but smaller than the average call holding time. At regular intervals (with length T_u) the needed VPC capacity for the coming interval is determined. This is done by calculating the expected blocking probability over the interval for various capacities given the offered traffic and present occupation state. The capacity that meets the target blocking is the needed one.

The calculation is quite complex, hence it is suggested to use precalculated table to achieve real time applicability. Another approach is to apply simplifications along the line of Virtamo and Aalto[12]. In [13] an allocation function is developed which does not depend on the actual offered traffic. It is based on the formula

$$N(n) = \lceil n + K\sqrt{n} \rceil \qquad (1)$$

where N is the required capacity, n is the number of currently active calls, and K depends on the offered traffic, target blocking probability, and updating interval. The idea behind this function is that for a specified traffic intensity the mean occupation state is equal to the intensity (if the blocking probability is low). K can be seen as a safety factor which adds extra capacity in units of the standard deviation of the occupation state. It is possible to rewrite (1) in a way that makes K only dependent on n[13].

However, this approach does not take into account the interaction between several VPCs on a physical link of fixed capacity. For example, although Ks computed for 1% target blocking will result in this value (as long as the physical link permits), the result will be under utilization of the link if not all capacity is seized. If, however, a larger K is used, the blocking will actually decrease and the link will be fully utilized. In this case, a higher K will thus better exploit the traffic fluctuations, i.e. when some VPCs temporary increase their number of allocations, others decrease theirs.

Another complication is K's dependence on the updating interval. The optimal choice of updating interval is determined by trade off between increased traffic (which partly depends on K) and overhead associated with updating. When doing frequent updating, it is better not to allocate too much capacity, because it should only be allocated when needed (less marginal needed for a short interval). However, having a longer T_u makes it necessary to allocate more. To optimize total performance, an optimal allocation strategy in terms of K and T_u must thus be found.

The local approach multiplexes VPCs in a special way. Considering a deterministic multiplexing of the VPCs, i.e. each VPC gets a fair portion of the capacity for a long time, the call blocking probability can be calculated using the Erlang's B-formula for each VPC.

When having full statistical multiplexing of all VPCs, the blocking probability can be calculated from the same formula by adding all traffics together. Since the local approach reallocates the capacity on a rather frequent basis, the blocking probability gets lower than for the deterministic multiplexing, but higher than for the full multiplexing. (When using a sufficiently short T_u the blocking reaches the same level as for full multiplexing over one link.)

2.3 The Allocation Function

To get an optimal allocation function the effect of K when having many VPCs on the same link must be studied. The problem is very complex since in a real network the number of VPCs and their traffics vary on each link traversed by a VPC.

We label the K (for a specific traffic intensity, T_u, and traffic load situation) that minimizes the total number of blocked calls (for a specific link) as K_{opt} (for that link). To see if the number of VPCs have an impact on K_{opt}, we have evaluated a link that is dimensioned so that each VPC can expect 1% loss if deterministic multiplexing is used. Results are shown in Figs. 1 and 2 for $T_u = 0.01$. The lines are interpolated between evaluated Ks (circles). The dashed lines show the 95% confidence intervals. Each VPC gets the same blocking probability. The curve in Fig. 1 shows the mean blocking when having 5 VPCs each carrying 152 Erlangs (This value was chosen in accordance with our test networks, see below). When more VPCs are interacting, the blocking probability decrease as seen in Fig. 2 where 10 VPCs are used. (With full multiplexing the blocking probabilities are $4.55 \cdot 10^{-5}$ and $1.07 \cdot 10^{-7}$ respectively.) The somewhat lower blocking noted for more VPCs is consistent with the well known concept "the bigger the better."

To simplify the problem for an inhomogenous network we consider the situation on an average link in the network under study. On this link we put a mean traffic consisting a couple of "background-VPCs."

In our test networks the mean number of PVPCs on a link is about 4 and each have on the average a capacity of 171. (The mean number of PVPCs per link vary for each test network, but most of them are in the range from 2 to 6.) The number of OVPCs is about 6 and each carries less than 10 Erlangs. With this in mind, we have formed an average link with 4 VPCs as background traffics each carrying 152 Erlangs together with 6 VPCs each carrying 8 Erlangs.

The dependency of K_{opt} on the traffic indicates that we need two K_{opt}s, one for each background traffic. If the K_{opt} is changed for one of these traffics it will affect the K_{opt} for the other one. The procedure to find these two K_{opt}s is as follows.

We first search for the K_{opt} for the PVPCs. This is done by having 5 PVPCs on a link where one of them can be seen as an aggregate of the 6 OVPCs. The found
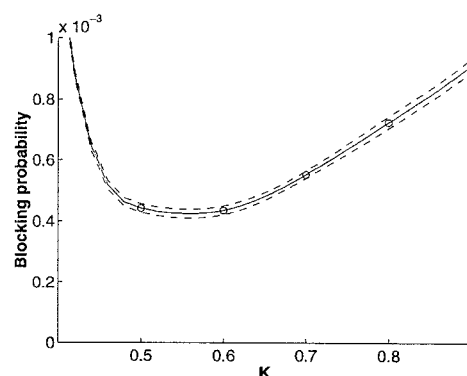

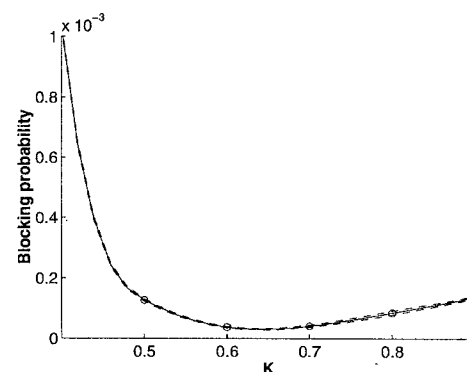**Fig. 1** Multiplexing gain for 5 VPCs on a link.


Fig. 2 Multiplexing gain for 10 VPCs on a link.

Table 1 K_{opt} for the background VPCs.

T_u	0.05	0.07	0.09	0.11	0.15	0.20
PVPC K_{opt}	0.87	0.97	1.03	1.12	1.22	1.28
PVPC $Bl.$ [%]	0.33	0.44	0.54	0.59	0.70	0.83
OVPC K_{opt}	0.79	0.88	0.90	0.90	0.91	0.92
OVPC $Bl.$ [%]	1.5	1.9	2.2	3.1	4.6	6.0

K_{opt} is used by the 4 PVPCs when we in the second step search for the K_{opt} for the 6 OVPCs. When this is found the OVPCs use this value when we once again check the K_{opt} for the PVPCs. These steps are repeated until no changes in either of the two K_{opt}s are detected. Table 1 and Fig. 3 show the K_{opt}s found. Figure 3 show crosses for PVPCs and circles for the OVPCs. (The exact values are not critical because the optima are flat as indicated in Figs. 1 and 2.) In the table are given the blocking probabilities for each background traffic. Since we do not try to equalize the blocking probabilities, these get different when choosing Ks that minimize the total number of blocked calls.

2.4 Optimizing the Allocation Function

When the K_{opt}s for the background traffic have been found another VPC is added and K_{opt}s are found for different traffic intensities on this VPC. (Additional capacity on the link is added, based on the resulting total
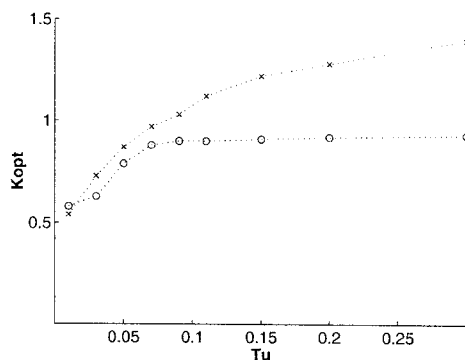
Fig. 3   The found K_{opt}s.
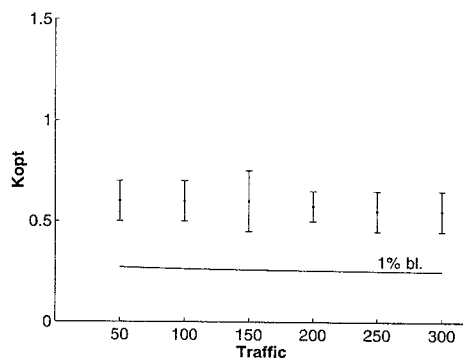
Fig. 4 K_{opt}s and fixed-target K (for $T_u = 0.01$).
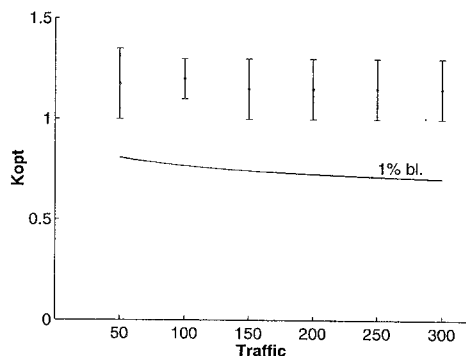
Fig. 5 K_{opt}s and fixed-target K (for $T_u = 0.10$).

traffic.) To find the optimal T_u, the previous method of finding the optimal K must be done for different T_us.

The result is displayed in Figs. 4 and 5 where each K_{opt} are given as ranges for a number of different traffics because there is no particular K for which there is a distinct minimum of the total number of lost calls. As a comparison, the K values calculated for only one VPC and a fixed target of 1% call blocking, as described in [13], are also shown (solid lines).

By studying the resulting intervals, it is clear that it is possible to construct an allocation function which does not depend on the actual traffic.

In Fig. 3 the K_{opt} for the OVPC is generally smaller

than for the PVPC. It is therfore tempting to use a K that increases with traffic. However, if one tries to find Ks for the two background traffics that make the blocking equal among them, the OPVC gets a larger K than the PVPC. For example, for $T_u = 0.05$, we get $K = 0.8$ for the PVPC and 1.1 for the OVPC and the resulting blocking probability 0.5%. Thus there is a trade-off between fairness (smaller K for large traffics) and maximal utilization (larger K for large traffics). With this and Fig. 3 in mind, we conclude that a fixed K_{opt} can be used for the different traffics.

3.   The Evaluation

For the sake of simplicity we limit ourselves in the numerical examples of this study to the case of a single, uniform service class. Multiplexing in the burst scale (e.g. for VBR services) is hidden in the use of equivalent bandwidth [1], [2] hence extensions to bursty traffics is straight forward. Requests for connections arrive according to independent Poisson processes for each node pair. The connection holding time is assumed to be negative exponentially distributed with unit mean.

We have used ten non-hierarchical networks with ten nodes each (which can be seen as core ATM networks). The nodes have both VP and VC routing capabilities. The VCCs use the existing VPCs but we have also used dynamic alternative routing (DAR) [14] on the call scale, i.e. if the direct VPC does not have room for an arriving call, rerouting with two VPCs in tandem over a selected transit node is tried. If this does not succeed, the call is rejected and a new transit node is chosen (at random) for the next time a call needs to be rerouted. Two control messages are used to determine the status of the transit nodes (question + answer). To ensure stability, we have applied a trunk reservation parameter of ten connections for direct traffic on all VPCs.

Our test networks have the capacity to handle the mean traffics with 1% call blocking probability. To simulate actual traffic, ten different traffic patterns were generated for each network by randomly selecting a busy center. Nodes inside the center increase their traffics above the average and those outside the center decrease theirs. Further details are described in the appendix.

With the optimal allocation functions for different T_us, we can evaluate the overall performance in a test network to get the optimal T_u.

4.   Results

Our evaluation is based on the reached profitability (2). The profitability is a normalized measure where 100% profitability means that all calls are handled without any overhead costs. (100% is infeasible for high traffic load situations.)

$$\text{Profit.} = \frac{\text{Calls}_{\text{Handled}} - (\text{Messages} \cdot \text{Cost})}{\text{Calls}_{\text{Offered}}} \quad (2)$$

The profit of handling one call is set to one unit. However, to be able to handle calls, several control messages (by means of RM-cells using some of the bandwidth) have to be used and these affect the total profit. The messages included are INC, DEC, and status messages for DAR. If a control message is seen as an RM cell, the cost can be related to an average phone call. Suppose that a phone call uses 167 cells/second, then the RM cell could be given a cost of $1/(167 \cdot$ seconds per mean holding time) which is $\sim 10^{-4}$. This cost might be too optimistic because there are also costs other than the ones related to bandwidth, e.g. processing. These are difficult to estimate. By using a higher message cost, the messages can be seen as having an overhead. The profitability is used to enable a reasonable evaluation of the overall performance by combining gains and costs.

An interesting aspect is the occasionally occurring capacity violations. These are caused by excess calls on links which have been granted less capacity and that are not disconnected in time before new calls arrive on VPCs which have been granted more capacity as a result of the reallocation. There are three ways to deal with this. One way is to move ongoing surplus connections to a path that can accommodate them. Another way is to wait for the connections to finish until capacity is released. The third way is to use "guard bands" which will not be allocated to any particular VPC. By this one hopes that there will be enough capacity to deal with over allocations. The amount of capacity violations depends not only on the guard band but also on the actual network and traffics. In general, the impact of a capacity violation depends on the degree of violation and the time during which it persists. In this evaluation a guard band is not needed to cope with link violations. Instead, a decrease of capacity for a particular VPC is simply not allowed if some of this capacity is in use.

In Fig. 6 the optimal updating interval can be found for different message costs. Clearly, the higher the cost, the lower the profitability. The upper line is the result when having a message cost of 0.01 and the lower line when having a message cost of 0.09. The step between the message costs is 0.02. The dotted lines are interpolated between evaluated Ks (circles). It can be seen that the optimum is further right the higher the updating cost. This is in perfect agreement with the assumption that more frequent reallocations are preferable when the message cost is small.

Since we have used a very simple model to calculate the K_{opt}s, a check can be done to see the real optimal K value for a particular T_u. Figure 7 shows the result for different values of K when $T_u = 0.1$ and a message cost of 0.01. The mean profitability of ten networks is given. The optimal K lies between 1.3 and 1.4. The K_{opt} for the PVPCs in our simple model is in this case between 1.0 and 1.3 (see Fig. 5).

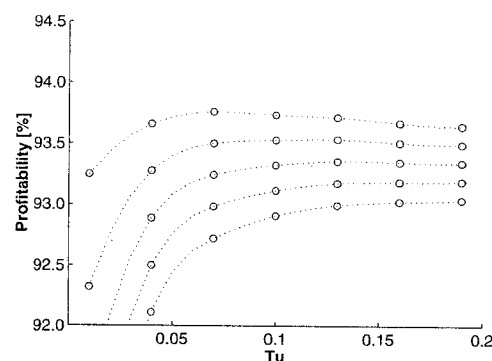Comparing to a fixed allocation of the capacities (capacities distributed according to the basic traffics and

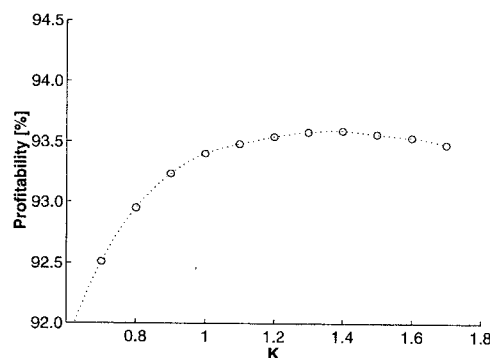**Fig. 6** Profitability for different T_us and message costs.

Fig. 7 Profitability for different Ks, $T_u = 0.10$.

use of DAR) the results are better. When having a message cost of 0.01, the profitability for the fixed allocation is 92.3%, while the local approach reaches 93.8%. (We have earlier studied a central approach [15] which in this case gives more profitability, but this approach tends to give a lot of link violations.)

We have seen in [15] that for high traffic loads the situation is reversed and the local approach reaches better profitability than the central one.

5. Conclusions

We have proposed a type of local VPC capacity management policy that uses regular updates and a simple allocation function to determine the needed capacity for the coming updating interval. With our proposed procedure for setting its unknown parameter K, the number of parameters is limited to one, viz. the current number of active connections. We have also shown how to determine the optimal interval for different message costs.

The optimal K values for the allocation function are not critical but optimal over a wider range as shown in Figs. 4, 5, 6, and 7. (For the used message costs, the optimal updating interval can be selected from Fig. 6.) Moreover, using a simplified model with an average background traffic for the calculation of the parameter K works although the actual distribution of PVPCs

and OVPCs covers a wide range. This further emphasize the robustness.

An improvement is achieved compared to a fixed capacity allocation and the approach takes advantage of the benefits from both VP and VC routing, i.e. enabling fast CAC and using multiplexing of VPCs. We also notice that the method of signaling is easy to implement and that link violations can be avoided.

6. Further Work

Sensitivity analysis of resulting blocking probabilities depending on the number of VPCs and traffic load variations might be helpful to show the robustness of the parameter K. It is expected that further performance improvements can be obtained by deploying VPC-dependent Ks rather than a global one. Evaluating the gains and relating them to the additional complexity of such estimations remains for further study. Another issue refers to the grouping of VPCs. Using the local approach there might be no need for group VPCs[16] and it could simplify the trade-off between VP and VC routing[17]. Finally, the aspect of deploying the OVPCs as backup-VPCs leads into the issue of self-healing networks[8],[9],[18]. The issues that arise when integrating fault management into the capacity management constitute an interesting area to explore.

References

[1] A.I. Elwalid and D. Mitra, "Effective bandwidth of general Markovian traffic sources and admission control of high-speed networks," IEEE/ACM Trans. Networking, vol.1, no.3, pp.329–343, 1991.

[2] R. Guérin and H. Ahmadi, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," IEEE J. Sel. Areas Commun., vol.9, no.7, pp.968–981, 1991.

[3] Å. Arvidsson, "High level B-ISDN/ATM traffic management in real time," Performance Modelling and Evaluation of ATM Networks, vol.1, pp.177–207, ISBN 0-412-71140-0, Chapman & Hall, 1995.

[4] Å. Arvidsson, "Real time management of virtual paths," Proc. IEEE Globecom '94, pp.1399–1403, 1994.

[5] S. Shioda, "Evaluating the performance of virtual path bandwidth control in ATM networks," IEICE Trans. Commun., vol.E77-B, no.10, pp.1175–1187, 1994.

[6] S.-O. Larsson and Å. Arvidsson, "A study of a distributed approach for VPC network management," Proc. IEEE Globecom '97, paper no.S50.3, 1997.

[7] U. Mocci, P. Pannunzi, and C. Scoglio, "Adaptive capacity management of virtual path networks," Proc. IEEE Globecom '96, paper no.19.2, 1996.

[8] N.D. Lin, A. Zolfaghari, and B. Lusignan, "ATM virtual path self-healing based on a new path restoration protocol," Proc. IEEE Globecom '94, pp.794–798, 1994.

[9] R. Kawamura, H. Hadama, and I. Tokizawa, "Implementation of self-healing function in ATM networks based on virtual path concept," Proc. IEEE Infocom '95, pp.303–311, 1995.

[10] S.-O. Larsson and Å. Arvidsson, "Performance evaluation of different local and distributed approaches for VP

network management," 13th NTS, pp.98–107, ISBN 82-993980-0-2, NTNU, 1996.

[11] C. Bruni, P. D'Andrea, U. Mocci, and C. Scoglio, "Optimal capacity management of virtual paths in ATM networks," Proc. IEEE Globecom '94, pp.207–211, 1994.

[12] J. Virtamo and S. Aalto, "Blocking probabilities in a transient system," COST TD257(97)14, Netherlands, Jan. 1997.

[13] J. Virtamo and S. Aalto, "Remarks on the effectiveness of dynamic VP bandwidth management," COST TD257(97)15, Netherlands, Jan. 1997.

[14] R.J. Gibbens, F.P. Kelly, and P.B. Key, "Dynamic alternative routing—Modelling and behaviour," Proc. ITC 12, Torino, paper no.3.4A.3, June 1988.

[15] S.-O. Larsson and Å. Arvidsson, "A comparison between different approaches for VPC bandwidth management," Proc. 5'th IFIP Workshop on ATM Networks, UK, pp.52/1–10, 1997.

[16] M. Omotami and T. Takahashi, "Network design of B-ISDN using the group virtual path scheme," Electronics and Commun. in Japan, Part 1, vol.79, no.7, pp.10–22, 1996.

[17] D. Hughes and K. Wajda, "Comparison of virtual path bandwidth assignment and routing methods," Annals of Telecom., vol.49, no.1–2, pp.80–89, 1994.

[18] P.A. Veitch, D.G. Smith, and I. Hawker, "Alternative routing strategies for virtual path restoration," IFIP Workshop TC6, IFIP working groups 6.3 and 6.4 participants proc., p.860, 15/1–10.

Appendix: Test Networks

The test networks have been made with a program that generates networks with N nodes. Call holding times are assumed to be negative exponentially distributed with a mean holding time of 1 time unit. User demands are fully characterized by a sequence of known end-to-end traffic demand matrices $A(k)$ (of size $N \cdot N$), where $a_{o,d}(k)$ denotes the traffic from o to d at time k, $k : (k = 1, \ldots, K)$ (see Fig. A·1). The time index k indicates intervals such as hour, day of week, or day of year. For each origin-destination pair an offered traffic was assigned to give 1% expected loss for a given transmission capacity. This basic traffic was modified to yield different situations by the use of "busy center" (Fig. A·2). Traffics between busy center nodes were increased randomly between 20 and 60%, traffics between
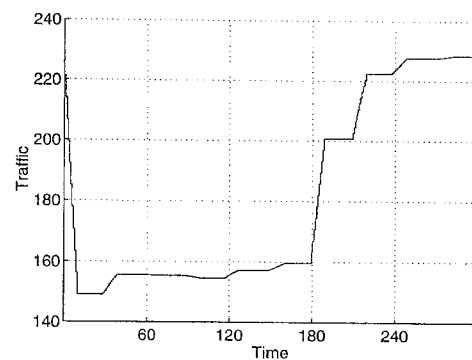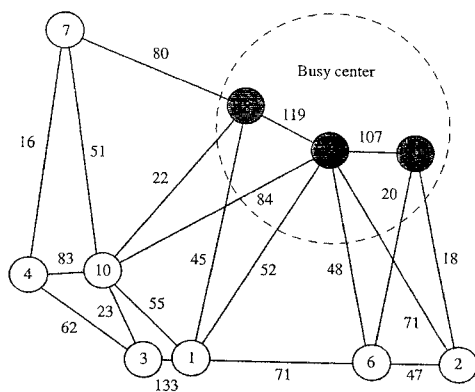

Fig. A·1   Sample traffic.

Fig. A·2 Sample network layout.

nodes outside the busy region were decreased randomly between 20 and 60%, and the traffic between a busy center node and a node outside the center was modified randomly between -20% and $+20\%$. After the modification the traffics have been normalized to give the same total amount of offered traffic as before. The resulting greatest increase is 97% and greatest decrease 60%. In Fig. A·2 the link capacities are given in capacity units which can accommodate ten connections. With $N = 10$, the total traffic offered to the network at any time is typically about 6800 Erlangs.

Sven-Olof Larsson received M.Sc. in computer science from the University of Lund in 1988. He has worked at Ellemtel and Ericsson for three years and 1992 he joined Karlskrona Shipyard, where he worked with embedded real-time systems and human computer interaction. Since 1995 he is a Ph.D. student at the University of Karlskrona/Ronneby. His current research involves distributed and local VPC capacity management functions for ATM networks.

Åke Arvidsson received his Ph.D. in 1990 from the Lund Institute of Technology in Lund, Sweden, where held a position as associate professor until 1994. He has held positions as visiting professor at the Bond University, Qld., Australia, and the University of Adelaide, S.A., Australia, and is currently acting professor of teletraffic systems at the University of Karlskrona/Ronneby. His main research interests include performance evaluation and optimization of traffic management and control in telecommunication systems, in particular ATM-, IP-, and signaling networks.