Comparisons of Different Approaches for Capacity Management in ATM Networks

S. O. Larsson

Department of Telecommunications and Signal Processing,
University of Karlskrona/Ronneby, S–371 79 Karlskrona, Sweden

Abstract

The fact that there is a lack of comparisons between methods for VPC management makes it important to evaluate fundamentally different methods. We have tried to find representatives of different ways of calculating the capacity distribution. A central approach uses global information about demands and resources, while decentralized approaches can be categorized into iterative and local. The local approaches that we have evaluated use the number of ongoing connections to decide how much capacity is needed during the next updating period. An iterative approach uses a distributed way of calculating the capacity distribution. By evaluating these, we have then been able to do a survey over the performances.

1. Introduction

There are a lot of different methods for virtual path connection (VPC) management [6] to improve the performance. They mainly differ in the objective function(s) to be optimized, the number of different paths between node-pairs, and constraints like having a target maximum call blocking probability. This makes conclusive comparisons very difficult. However, some general aspects can be mentioned. The problem with a comparison is to give costs for control messages and processing which enables calculation of comparable numbers. One can consider to have an upper limit on the number of messages that can be handled within a certain time interval in a switch. If there exists a signaling network with a fixed capacity (out--hand signaling), a certain amount of the link capacity can be regarded as belonging to the signaling network. This capacity is not included in the dynamic capacity allocation. If for instance RM--cells are used (in--band signaling), a cost can be given on messages, which is related to the loss in revenue due to less handled customer traffic. In [3] a cost model based on transmission, switching and setup costs are investigated. This shows that the optimal policy is having both VPCs and virtual channel connections (VCCs). In [2] a cost--benefit study is done. The cost of traffic carrying capacity is related to the control costs. When capacity is cheap relative to control, it becomes economic to use VPCs and to periodically update the reservations. In [1] the total network cost is considered including the architecture and VPC management. In this evaluation we consider in--band signaling and we do not consider different network architectures and costs related to hardware. All of the capacity in the network is free to use and the evaluations is done with a call--by--call simulation.

2. Different approaches

We have evaluated seven approaches. A central, two iterative, two local, a fixed, and one which do not use reservations in advance. In the central approach (denoted as CENT) all the nodes monitor current traffic demands and periodically report their results to a network management center (NMC). At the end of each such cycle, the NMC collects all reports on user traffic demands from which a VPC network is designed, i.e. a capacity distribution for all demands. Capacity is distributed one capacity unit (c.u.) at a time. This is first done in an attempt to equalize the VPC connection setup blocking probability (CBP) to be less than the target maximum VPC CBP (MAXCBP). In other words, capacity is distributed as long as any VPC has an estimated CBP larger than the target MAXCBP and it is possible to allocate capacity to such a VPC. The remaining capacity is distributed to maximize the total handled traffic. For this we use the concept of marginal utilization (MU). The MU is the extra number of connections that the VPC is expected to carry if allocated an extra c.u. For each VPC, the MU is also divided with the number of traversed links to maximize the total number of handled connections in the network. When the new capacity allocation has been calculated, the information concerning each node is gathered together and sent at the same time (i.e. one control message is considered enough).

An approach, which is a compromise in the amount of information needed for the VPC management, is a distributed--iterative one. There are very few presentations of this kind of approach in the literature. We have developed one denoted as ITER [7]. (A central approach
Profitability = \( \frac{(Call_{hi}(1-C_{vc}) - (AtrR+LinkR)\ C_{vc}}{Sigs \cdot C_{s} - \ VP_{s} \cdot C_{c}) / Calls_{o}} \) (2)

The number of handled calls (connections) is denoted as \( Call_{hi} \), the number of offered calls as \( Call_{o} \) and the number of control messages as \( Sigs \). \( AtrR \) is the number of alternate routed calls, and \( LinkR \) is the number of link routed calls. \( C_{vc} \) is the VC link setup cost (0.05 according to [2]), including control and selection of VPI and VCI numbers (and an extra cost for switching). The number of changes of reserved capacity in the nodes is denoted as \( VP_{s} \) and \( C_{c} \) is the cost for updating reserved capacity in a node (0.01). \( C_{s} \) is the cost for a control message (10^{-4}). The cost of 10^{-4} is the result when comparing an RM-cell to an average telephone call.

The measure of the profitability can be discussed from different points of views, such as in-band or out-band signaling. One can discuss if the capacity of the signaling network should be taken into account and if the costs are correct. We have seen that if the VC link setup cost is decreased from 0.05 to 0.01, then the mutual order of profitability (best to worst) between all approaches stays the same, except between the FIXED and the NOVPC approach. In this case, NOVPC gets better than FIXED when having high network load. However, a comparison can not rely on one measure only. In the following, complementary measures are discussed.

The number of control messages that can be handled by a switch is limited. Measurements on commercial ATM switches available today shows a mean value of maximum handled messages of 10–50 per second [11]. New technology will, however, increase the speed of the switches. In the evaluations we set the signaling CPU service time to 20 ms and simply measure the maximum number of messages per node and time unit (t.u.).

It is important for a dynamic capacity allocation scheme to be robust and able to handle unexpected traffic changes. Evaluations of the performances should be done for different traffic loads and distributions in the networks. The MAXCBP is important since the impacts of large CBP on redialing as well as fairness among the customers must be taken into account.

Having a target MAXCBP might limit effects like redialing and customers getting so disappointed that they rather move to other service-providers. In case of having a large MAXCBP, we think one could see the situation as having an underdimensioned network.

4. Assumptions

In this evaluation we implement predefined routes and prereserved resources by means of VPCs upon which individual VCCs are established and terminated for each connection. The nodes have both VP and VC routing capabilities. Ten fully connected non-hierarchical networks with ten nodes each are used (which can be seen as core networks). A smaller evaluation of a 50-node network is also done. As a test of robustness, ten different traffic patterns were generated for each network by randomly selecting a busy center. Nodes inside the center increase their traffic above the average and those outside the center decrease theirs. We have denoted this traffic as high traffic demand imbalance (HI). We also consider a general overload traffic demand (GO) where the traffic has been changed proportionally. The total network traffic load is related to the basic traffic (for which the network was dimensioned) and denoted as the network load. For example, a network load of 1.2 with GO traffic, means 20% increase of the basic traffic. For HI traffic the resulting greatest increase is +90% and greatest decrease is −55%. The total traffic offered to the network at any time is typically about 6500 Erlangs for a network load of 1.0. Our test networks have the capacity to handle the mean traffic with 1% CBP. We neglect the effect of redialing.

For the sake of simplicity we limit the numerical examples of this study to the case of a single, uniform service class. However, the results are readily extended to multi-service networks. Multiplexing in the burst-scale (e.g. for VBR services) is hidden in the use of equivalent bandwidth [5] hence extensions to bursty traffics is straightforward. Requests for connections arrive independent negative exponentially distributed intervals for all node pairs which means that the statistical multiplexing gain in the call-scale can be determined by the Erlang B-formula. The offered traffics are estimated by arrival counting. The connection holding time is assumed to be negative exponentially distributed with unit mean.

We will only use the shortest paths in this comparison since the performance only increase slightly when using several paths for our test networks and traffics. A VCC between two nodes is normally routed over the corresponding, direct VPC. The one-link VPCs are able to use all unreserved capacity on the link they use. All approaches use one c.u./connection (i.e. the finest granularity). Reallocations are done periodically and they use the same networks and traffic patterns. The updating interval (\( T_{u} \)) has been set to 3 t.u.s (optimal for HI traffic) for CIENT, ITER, and NONCOOP, and for the local methods \( T_{u} \) is set to 0.1.

As a compliment to VPC rearrangements, we have applied dynamic alternative routing by means of DAR [13]. Two control messages are used to determine the status of the transit nodes. A central approach can use a special algorithm for global optimization of trunk reservation, TR. The local and iterative ones can not use this algorithm, since they do not have access to global information. For normal traffic loads the optimal TR has
takes the link loads into account gives a better performance. The far most important mechanism to use for increasing the profitability, is the ability for one-hop VPCs to use all unreserved capacity on the link. Without this extension, the performance can be worse than for FIXED and using link routing will further increase the performance for high loads.

5.2. Maximum CBP

As can be seen in figure 4 the mean MAXCBP gets high for the decentralized approaches for high traffic loads. This depends on the different ways of allocating capacity link-by-link. Only CEINT can control the MAXCBP. On the other hand, reserving a small amount of capacity for many-hop VPCs will decrease the CBP for these. It is possible to reserve different amount depending on the number of hops each VPC have. In a capacity request message, the number of hops can be included. The node will have a table to get the amount of VPC hop count trunk reservation. The longest VPC will be able to allocate all capacity that is left. An evaluation has been done with having two connections reserved for each “class” of VPC hop length. The mean MAXCBP will, for ALCRA, be very similar to NOVPC with only a slight decrease in profitability in high load situations.

5.3. Capacity reallocations

For all decentralized approaches, the amount of unallocated capacity is large for low network load but decreases when the load is increased (figures 6 and 7). ALCRA can utilize the low load indication when the allocation factor gets large enough.

The decentralized approaches decrease the number of reallocations for high loads (seen in figure 8), because fewer changes can be made, i.e. VPCs more often needs additional capacity due to more unsuccessful allocations.
5.5. Maximum CBP and call routing

The effect of alternative routing has been studied for CENT in a general overload situation. Figures 12 and 13 show the profitability and mean MAXCBP for the possible combinations of routing schemes. The target MAXCBP is set to 20%. For comparison reasons, we also give the results for the fixed approach (straight line without circles). The extra profit of also using link routing (LINKR) when having DAR is minimal. For DAR to work a TR has to be used. It is also clear that dynamic capacity reallocation is the most important mechanism to increase profitability, when the traffic load is higher than anticipated. At an overload of 10% the effect of DAR is equal to the dynamic capacity reallocation. The fixed approach can not utilize the fact that the statistical multiplexing gain increases non-linearly. It should be noted that the mean MAXCBP also reflects (to some extent) the mean MAXCBP when not having DAR.

It can be tempting to compare different approaches that do not use alternate routing. Since the differences are small in the profitability, alternate routing will have a noticeable effect. The difference in the time scale for reallocations, also increase the differences of the effects of using alternate routing.

The target MAXCBP is set to a value that is supposed to be an acceptable decrease in service at occasional high traffic load situations. Figures 14 and 15 show the profitability and mean MAXCBP when the target MAXCBP is set to 10%, 20%, and 30%. A more constant CBP is noted for the 20% line for network loads between 1.2 and 1.3. This illustrates the effect of the target MAXCBP. The reason why this does not appear for the 10% line is the large step between the evaluated traffic loads. These figures illustrate another aspect than figures 12 and 13. When using DAR, an increased target MAXCBP will increase the possibilities to increase the
total mean length of an overallocation period is ranging from approximately 0.003 to 0.017 t.u.s among the networks. The mean number of overallocated connections per overallocation period and test network is ranging from approximately 2 to 17. This is an indication that the cell loss probability will in fact be affected if reallocations are not handled carefully. When the load is light, abrupt reallocations can be used in smaller networks too, but in this situation the benefit from using dynamic reallocations is small. In this case, the benefit comes from rerouting in case of network failures. The load will then, on some parts of the network, not be light any longer. For the GO traffic the profitability can be improved if a longer measure interval is used for estimating the traffic demands.

For the local VPC capacity management policies we have used a simplified model with an average background traffic for the estimation the relation between \( K \) free capacity, and traffic load. This works well although the actual distributions of the number of VPCs per link cover a wide range. Considering this and adjusting \( K \) according to the estimated experienced VPC congestion state (given by the amount of free capacity), we are able to fine-tune the setting of it. Although the profitability is not increased very much, this suggests that our method is robust with respect to traffic load situations, accuracy of traffic measurements or forecasts and the numerous combinations of VPCs. The method takes advantage of the benefits from both VP and VC routing, i.e. enabling fast connection admission control and using multiplexing of VPCs. In other words, a local approach combines the benefits from both hop-by-hop allocation (dynamic routing) and dynamic capacity allocation (transport routing). This is a result of the ability to quickly adapt to changes in the capacity demands. We also notice that the method of signaling is easy to implement and that overallocations are totally avoided. If link routing is omitted the profitability will only decrease slightly. A local approach can be used when capacity is paid for, i.e. not more than "enough" capacity is to be reserved. On the other hand, a network provider can allocate all capacity and an NMC is useful to control the network. In a sparse logical topology, such as a ring network, the profitability will be better due to the statistical multiplexing effect among the VPCs that can be accomplished by frequent reallocations on high capacity links. For the 50-node network having a general overload, ALCRA is not as good as CENT and FIXED, mainly because of the larger update interval. The update interval was chosen to decrease the maximum number of messages per t.u. to any node. It could possibly be decreased since the maximal length of the control message queue is 50 and the mean number is 25. We have also seen that the local approach can be used for a 100 node network by setting \( T_w = 0.5 \) without decreasing the profitability notably.

7. Conclusions

The central approach presented (CENT), calculates a nearly optimal trunk reservation for each link, which in fact reserves more capacity for direct traffic than other approaches in the high load situations. It can also take into account the loads on different links. For a 50-node network, it seems as if it is difficult for CENT to guarantee that capacity is not overallocated. One can state that a controlled reallocation performs best for small networks and abrupt (global) reallocation performs best in big networks.

We have developed a distributed link-iterative VPC management policy. The method uses iteration cycles to reallocate the capacity. The computations needed are simple but the complexity is instead moved to the management of control messages (i.e. time-outs, delays).

We have also evaluated a distributed-iterative approach based on the method called Gauss-Seidel. It seems as if the scheme is a bit tricky to use. Much capacity will not be reserved, which makes link routing very important to increase the amount of handled traffic. Otherwise, the performance can get much worse than for the fixed approach that do not alter the capacity distribution. DAR does not help in this case. Having a large network, results in longer locking times of the links while computing the capacity allocation. The number of iterations needed at each update instant will increase. This will, together with link routing, decrease the scalability of the approach since long VPCs will starve in a big network, i.e. never get a chance to reserve capacity.

We have proposed local VPC capacity management approaches (isolated and adaptive) that use regular updates and a simple allocation function to determine the needed capacity for the coming update interval. With our proposed, simple procedure based on averages for setting the unknown parameter \( K \), the number of parameters is limited to one, which is the current number of active connections.

The fixed approach performs well for low and moderate traffic imbalance and low loads. For high traffic load and high traffic imbalance (when some links are heavily overloaded) the fixed allocation is not as good as the other approaches. The fixed approach only uses messages for DAR.

For the approach that does not use VPCs, connection establishment is done for each VCC along the shortest path. This results in a large number of control messages, which decreases the profitability.

The local approaches are interesting alternatives to the other so frequently studied central approaches. They may also complement each other. The central approach has the ability to find VPCs and order them (i.e. finding which one is to be used as a preferred path) and a local one can fine-tune the capacity allocation.