On State Aggregation for Scalable QoS Routing*

LIANG GUO IBRAHIM MATTA†
College of Computer Science
Northeastern University
Boston, MA 02115
`{guol, matta}@ccs.neu.edu`

Technical Report NU-CCS-98-02
February 1998

1 Introduction

We consider the problem of routing diverse traffic streams with diverse Quality of Service (QoS) requirements. This requires the regular exchange and maintenance of network state information so as to compute paths (or routes) between end-users which can satisfy varying QoS. This process is often referred to as *QoS routing.* QoS routing has been largely left out of architectures proposed for integrated services. Only recently, routing has been recognized as an important component of QoS architectures and the IETF (Internet Engineering Task Force) has chartered a working group with the mission of writing a framework document [9]. We particularly consider large networks where it becomes necessary to organize the network hierarchically in order to scale in terms of processing, memory and communication requirements. This allows for the aggregation and hence reduction of state information [13], although this *may* have a negative effect on the performance of the network.

In this paper, we study hierarchical QoS routing as proposed in the *Private Network-Network Interface (PNNI)* specification of ATM networks. Here source routing is used to select a path for an incoming connection. In other words, a forwarding path that is likely to satisfy the requested QoS is first computed at the source and then listed in the header of each packet. All packets are then forwarded along that same path. The aggregation of network state is done by dividing nodes and links into groups (or areas). In this case, the source node has less detailed knowledge about the state of links and nodes outside its area. Thus the route that the source node selects outside its area is only a *logical path.* The actual selection of a physical path outside the source area has to be made by each entry (border) node of areas along the logical path. Clearly, routing is then determined by the logical view the source node has about remote areas, which in turn is determined by the scheme used to aggregate state information of each area.

We compare three different aggregation schemes: two extreme alternatives, namely the *Simple-Node* approach, which collapses a group with multiple nodes into one virtual node, and the *Full-Mesh* approach, which uses a complete graph connecting the externally visible border nodes of an area. Obviously, *Simple-Node* offers the greatest reduction of advertised state information, while *Full-Mesh* maintains the accuracy of all transit path metrics within an area. The third scheme, namely the *Star* approach, attempts to achieve a

---

*Supported by NSF CAREER grant ANIR-9701988 and NEU grant RSDF-377090.

†Corresponding author. Address: College of Computer Science, Northeastern University, 119 Cullinane Hall, Boston, MA 02115. Tel: (617) 373-3694, fax: (617) 373-5121, e-mail: matta@ccs.neu.edu.

compromise between these two extreme approaches by representing a group by a star-like topology connecting all the border nodes via a central virtual node. The *Star* scheme has been adopted by the ATM Forum [19]. For path selection, we consider several path cost functions (routing metrics) based on hop-count (number of links), current available bandwidth and feasibility to satisfy the requested delay QoS. Also, it has been shown in a number of non-hierarchical (flat) QoS routing studies (e.g. [15]) that efficient resource utilization can be achieved if routing is restricted to short paths (even though longer paths might satisfy the QoS requirements). In this paper, we investigate the effect of such restriction on candidate paths in hierarchical QoS networks.

Related work: Compared to other QoS routing studies, our model considers *transient* (or instantaneous) performance measures (instantaneous blocking probability, revenue, etc.), other QoS controls such as weighted round-robin scheduling, both uniform and skewed workload distribution of source-destination pairs, and multiple routing metrics (namely hop-count and utilization). In [10], Guerin and Orda suggest several heuristics to path selection accounting for inaccuracy due to state aggregation and distribution. In [21, 6], Baras *et al*. propose a hierarchical reduced load approximation to evaluate PNNI routing. However, no simulation results are presented. Some graph theoretic studies examined compact graph representations [7, 2] with emphasis on the maximum cost distortion between any pair of nodes in the aggregated graph. Lee presents a spanning tree method for link state aggregation in [14], and shows the effect of state aggregation theoretically. Awerbuch and Shavitt [4] present a distortion-bounded solution for aggregating the state of a subnetwork, and show how to apply it to PNNI. The effect of different aggregation schemes on network performance is compared by simulation in [3]. However, the model only considers a single source-destination pair.

Our results: We obtain our results using our recently developed time-dependent evaluation method [16]. We consider both skewed and uniform workloads. In a *skewed workload*, some nodes are selected as the destination of the majority of the connections (or calls) and are thus considered "hot-spots." In a *uniform workload*, source-detination node pairs are uniformly distributed over the network. Under skewed workload, we found that *Simple-Node* performs better or as well as *Full-Mesh* and *Star*. This is contrary to the common belief that more performance loss is inevitable as more aggregation would result in greater loss of QoS/routing information. However, under uniform workload, both *Full-Mesh* and *Star* outperform *Simple-Node*. The reason is that with more aggregation, the network view maintained at the source is less detailed and hence the source becomes less restrictive in its route selection (considering more candidate paths) and its admittance of calls. A larger set of candidate paths allows for better distribution of the load. This improves the overall network performance under a (realistic) skewed workload, but worsens it under a uniform workload. Under uniform workload, it becomes more beneficial to have a more detailed view and hence to be able to reduce the set of candidate paths. This avoids using alternative longer paths, which might have been used by many short length connections. Note that with the unpredictable traffic patterns of today's networks, it is the case that workload is largely skewed and that aggressive state aggregation can then considerably reduce memory, processing and communication requirements without significant loss in performance.

Furthermore, we found *Star* to perform slightly worse than *Full-Mesh* as it results in less detailed view about available bandwidth than *Full-Mesh*. Although *Star* provides more candidate paths than *Full-Mesh* and so results in better traffic distribution under skewed workload, this does not compensate for the loss of information about available bandwidth.

Organization: The rest of the paper is organized as follows. Section 2 presents our hierarchical network model. This model is solved using our recently developed time-dependent evaluation method. The method solves dynamic flow models of multi-service networks such as ATM—see [16]. Section 3 discusses the computation and aggregation of routing metrics. Section 4 presents numerical results for a sample network, and concludes with future work.

2 Evaluation Model

Network and routing models: We consider a connection-oriented network. To support real-time QoS, necessary resources are reserved along the connection's route from the source to the destination. The network employs hierarchical routing as in PNNI routing [19]. Figure 1 shows an example two-level hierarchical network.

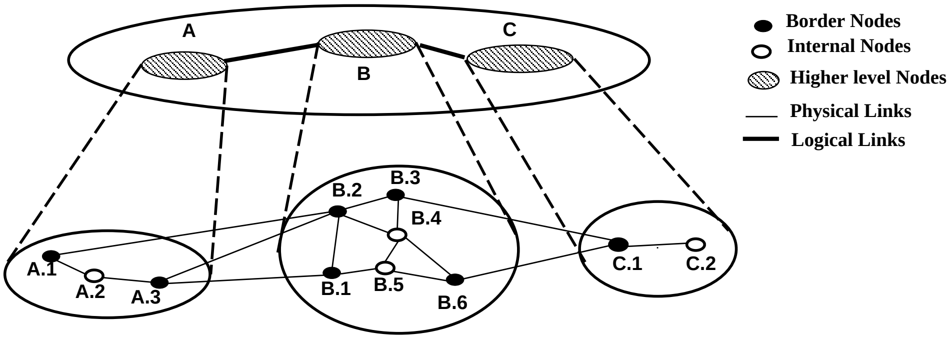


Figure 1: A two-level hierarchical network.

A group of nodes may be abstractly represented by a single point known as a logical node. One of the nodes associated with a logical group is elected to advertize its *aggregated state* based on which route selection is done at nodes outside the group. Figure 2 illustrates the local view for group A nodes.

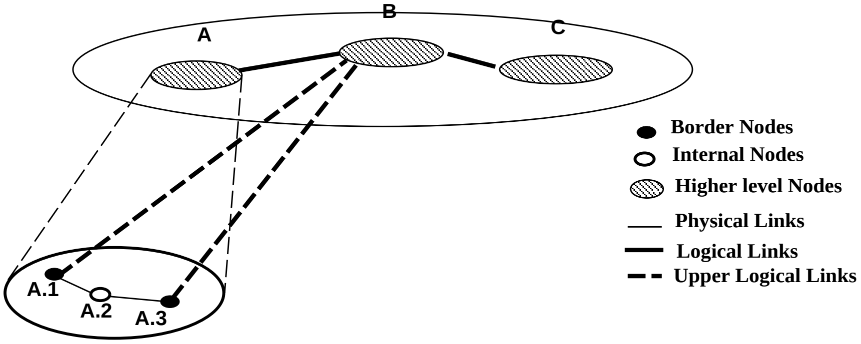


Figure 2: View of the network at nodes A.1, A.2 and A.3.

For a connection setup request, the source node (or switch in PNNI terminology) selects a path to the destination based on its local view. Such a path is not a fully detailed source route because it does not contain the details of the path outside the source's group. Instead, those portions of the path are abstracted as a sequence of logical group nodes to be transited. When the call setup arrives at the entry (border) node of a group, that entry border node is responsible for selecting a lower level source route describing the transit path across that group, and this transit path must reach the "next hop" destination specified by the higher level (source) path. A node along the path may reject a setup request due to lack of resources.

We assume state information is periodically broadcasted. A node uses its new state information to compute new routes to be used for incoming connections until the next broadcast. We assume a source node or an entry border node uses probabilistic routing, a type of routing proposed in many studies (e.g., [5, 8]). Here, a probability is assigned to every candidate path and arriving connections are routed independently according to these path probabilities.

Unless otherwise stated, the set of candidate paths that a source node or an entry border node chooses from is restricted to the set of minimum-hop and minimum-hop+1 paths. This is desirable because using a longer path for a connection ties up resources at more intermediate nodes, thereby decreasing network throughput. Furthermore, it also ties up more resources at each intermediate node because satisfying the end-to-end QoS requirement would require more stringent local (per link) QoS requirements. We discuss routing in more detail later in Section 3.

Traffic model: We consider multiple services, where a service represents connections with the same source-destination node pair and the same traffic and QoS parameters. The parameters of a service include: time-dependent arrival rate of requests for a connection setup; average lifetime of a connection; QoS requirement of a connection (namely, end-to-end statistical delay bound); and packet (or cell) generation characteristics of a connection (namely, peak and average transmission rates and duration of the busy period).

For a connection setup request on a multi-link route, the requested end-to-end QoS is divided equally among the links. This is the so-called "equal allocation" policy [17]. Note that because of state aggregation, the source does not have an accurate measure of the (physical) path length, which affects the (local) delay requirement on each individual link.

Scheduling and admission control model: We assume each link in the network uses a "per-connection" weighted round-robin type scheduling [18]. A connection is allocated (and guaranteed) a certain amount of bandwidth on a link along its path such that its local QoS requirement is satisfied. This required bandwidth is often referred to as effective or equivalent capacity [11]. It depends of course on the local QoS and the packet generation characteristics of the connection. It is feasible for a link to accept a connection if the required bandwidth does not exceed the current available (idle) capacity of the link. We assume each link has adequate buffer space.

3 Routing Schemes

We consider two main routing metrics: utilization (or equivalently available bandwidth) and hop-count. According to [20], utilization is a concave routing metric, whereas hop-count is an additive metric. It has been shown that QoS routing with such multiple metrics is NP-complete [20]. We use a heuristic method that first reduces the candidate path set to include only those short length paths, and then use a mixed metric of utilization and hop-count. Specifically, from the set of candidate paths, the source node or an entry border node selects a path p probabilistically using path weights W_p where

$$W_p \propto \frac{F_p\ (1 - U_p)}{H_p}$$

The path metrics H_p, U_p and F_p are defined as follows:

- **Hop-count H_p:** The hop-count defines the length of a candidate path (this gives preference to shortest paths). For a low level (physical) path, it is simply the number of links along the path. For a high level (logical) path, it is the sum of the length of all logical links and the diameter of all logical nodes along the path, where the values of length and diameter depend on the underlying aggregation scheme.
- **Utilization U_p:** The utilization of a low level link (physical link) is defined as the fraction of link capacity reserved averaged over the last routing update period. For a low level (physical) path, it is simply the maximal utilization of links along this path. For a high level (logical) path, it is the maximal utilization of logical links and nodes along the path, where the utilization values depend on the underlying aggregation scheme.
- **Feasibility F_p:** F_p is either 1 or 0 depending on whether the path p is feasible or not. A path p is said to be feasible if the source "expects" a successful setup on p [1]. To test the feasibility of the path, the source considers the effective capacity requested by the new connection in addition to the actual available bandwidth of physical links in the source group and aggregated available bandwidth of logical links and nodes. Although this source's routing view contains aggregated information and is periodically updated and thus might be outdated, the source assumes that the view is accurate. In other words, F_p is the result of the source node evaluating the admission control function for the remote (physical and logical) nodes on the path.

Note that path selection is performed at the source node as well as at each entry border node. Thus the path weights used at the source node are derived from the actual metrics inside the source group and the aggregated metrics outside the source group. While the path weights used at each entry border node are only based on the actual metrics inside its group. In Section 3.1 we present the various aggregation methods.

We consider various route selection policies, which differ in how they assign the weights W_p:

- **UTIL** defines path weight W_p as
$$W_p = (1 - U_p)$$
- **UTIL-HOP** defines path weight W_p as
$$W_p = (1 - U_p)/H_p$$
- **UTIL-FES** defines path weight W_p as
$$W_p = \begin{cases} 0 & \text{if path is not feasible} \\ 1 - U_p & \text{otherwise} \end{cases}$$
- **UTIL-HOP-FES** defines path weight W_p as
$$W_p = \begin{cases} 0 & \text{if path is not feasible} \\ (1 - U_p)/H_p & \text{otherwise} \end{cases}$$

3.1 Metrics Aggregation

We now present the aggregation methods. We first define the aggregated (logical) links between two adjacent groups as *inter-group links*, e.g., the logical links between group A and group B in Figure 1. We define the logical links between every two border nodes of the same group as *intra-group links*, e.g., the logical link between nodes B.2 and B.6. The process of aggregation includes the aggregation of nodal and link information. For our routing purposes, we need to aggregate both hop-count and utilization.

Full-Mesh aggregation: In *Full-Mesh* scheme, all inter-group and intra-group links should be advertised. Note that a logical intra-group link may correspond to several low level (physical) paths connecting the same source and destination border nodes. The utilization of an intra-group link is defined as the minimum utilization of all the corresponding candidate low level paths. Recall that the candidate path set includes paths of length less than or equal to minimum-hop+1. For the hop-count metric of an intra-group link, we define it as the maximum length of the corresponding candidate low level paths. The source node uses this maximum length to conservatively estimate per link QoS requirements.

In this and the *Star* scheme, any inter-group link only corresponds to one physical link, therefore, no aggregation is needed in advertising inter-group link metrics. Furthermore, no nodal information is advertized.

Star aggregation: Because *Star* scheme advertises messages in the order of the number of border nodes, to reflect more asymmetry, we use the following method to aggregate the intra-group link utilization information:

1. Construct the same full-mesh topology as in the *Full-Mesh* scheme.
2. For each border node, compute and advertize two metrics: incoming utilization, which indicates the maximal utilization of all logical intra-group links that end at this node; and outgoing utilization, which is the maximal utilization of all logical intra-group links that start from this node.
3. A node outside this group reconstructs a logical intra-group link within that group by assigning it a utilization value as the maximum of the outgoing utilization of the source border node and the incoming utilization of the destination border node.

For simplicity, we define the length of all intra-group links to be the diameter of the group, i.e., the maximum hop-count of all intra-group links within this group.

Simple-Node aggregation: In *Simple-Node* scheme, the inter-group link may correspond to several physical links. We define the link utilization of such inter-group link as the minimum utilization among all of the corresponding physical links. No information about border nodes is advertised in this scheme, so we only broadcast a network diameter as in *Star*, and a logical group utilization which is defined as the maximum utilization of all logical intra-group links within this group.

4 Numerical Results

We report simulation results for two types of workload. The first type is a *skewed* workload where network services are concentrated around some hot-spot nodes. Hence load is also concentrated over a number of "hot" links. The second type of workload is *uniform* where services are uniformly distributed among node pairs. Our goal is to study the effect of service distribution on the performance of aggregation schemes.

We consider the NSFNET-backbone topology shown in Figure 3. We consider 52 services using this network. We arbitrarily divide the 14 nodes into 5 groups, represented by dotted circles in Figure 3. We assume links have different capacities of 6.4 Mbps or 45 Mbps. Two types of services representing video and audio were considered. A video type connection has an average lifetime of 20 mins, a peak rate of 2.1Mbps, average rate of 0.7Mbps, a busy period of 0.3 sec, and a statistical delay requirement Prob[end-to-end packet delay > 50 msec]$< 10^{-4}$. An audio type connection has an average lifetime of 5 mins, a peak rate of 30Kbps, average rate of 10Kbps, a busy period of 0.3 sec, and a statistical delay requirement Prob[end-to-end packet delay > 50 msec]$< 10^{-4}$.

To simulate skewed workload, we deliberately set the destination node of 26 services to be node 1, and the destination node of another 26 services to be node 13. The parameters of these services are shown in

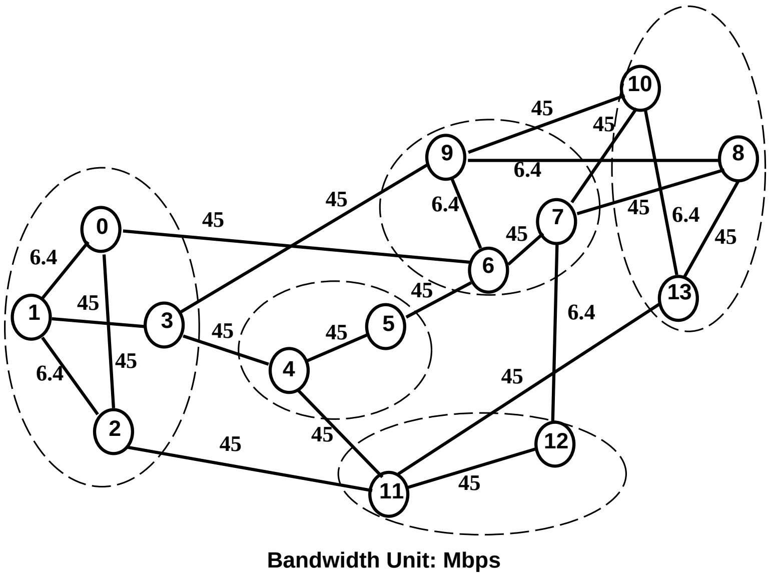


Figure 3: Sample network: 14 nodes, 22 bidirectional links, average degree 3, divided into 5 groups.

Table 1: Parameters of services using network in Figure 3.

$(SRC_s, DEST_s)$	$(M_s,\ m_s,\ b_s,\ D_s,\ \varepsilon_s)$	$(\lambda_s,\ 1/\mu_s)$
(0, 13),(1, 13),(2, 13),(3, 13),(4, 13),(5, 13),(6, 13)	(30, 10, 0.3, 0.05, 10^{-4})	(2, 5)
(7, 13),(8, 13),(9, 13),(10, 13),(11, 13),(12, 13)	(30, 10, 0.3, 0.05, 10^{-4})	(2, 5)
(0, 13),(1, 13),(2, 13),(3, 13),(4, 13),(5, 13),(6, 13)	(2100, 700, 0.3, 0.05, 10^{-4})	(0.4, 20)
(7, 13),(8, 13),(9, 13),(10, 13),(11, 13),(12, 13)	(2100, 700, 0.3, 0.05, 10^{-4})	(0.4, 20)
(0, 1),(2, 1),(3, 1),(4, 1),(5, 1),(6, 1),(7, 1)	(30, 10, 0.3, 0.05, 10^{-4})	(2, 5)
(8, 1),(9, 1),(10, 1),(11, 1),(12, 1),(13,1)	(30, 10, 0.3, 0.05, 10^{-4})	(2, 5)
(0, 1),(2, 1),(3, 1),(4, 1),(5, 1),(6, 1),(7, 1)	(2100, 700, 0.3, 0.05, 10^{-4})	(0.4, 20)
(8, 1),(9, 1),(10, 1),(11, 1),(12, 1),(13,1)	(2100, 700, 0.3, 0.05, 10^{-4})	(0.4, 20)

Table 1. Services with the same traffic and end-to-end QoS parameters, but with different source/destination pairs, are grouped in the same row. The routing update period equals 1 min.

For uniform workload, we randomly chose source-destination pairs for each service so as to get a uniform distribution. We do not change the parameters of services. Thus, the overall external load remains the same.

We obtain several instantaneous performance measures, including the total number of connections established at time instant t, the average connection blocking probability at time instant t, and revenue defined as the total amount of bandwidth in use by connections (reserved) at time instant t. We only show here revenue versus time for selected parameters and routing policies [12].

Figure 4(a) shows revenue for the three aggregation schemes under skewed workload when **UTIL-HOP-FES** routing policy is employed. Initially, *Full-Mesh* and *Star* perform better than *Simple Node*. However, as time increases and the network becomes more loaded, *Simple Node* performs better or as well as both *Full-Mesh* and *Star*. *Star* performs slightly worse than *Full-Mesh*.

Figure 4(b) compares the three aggregation schemes under uniform workload when **UTIL** routing policy is employed. In this case, we see that *Full-Mesh* and *Star* outperform *Simple-Node*. Again, *Full-Mesh* performs slightly better than *Star*.

Figure 5 shows the effect of various policies for assigning path selection weights on network performance. A hop-count metric H_p in the weight function slightly enhances routing efficiency. We found this effect to be most significant with *Star*.

Figures 6(a) and 6(b) indicate that when we relax the *minhop+1* restriction to *minhop+2*, i.e., the candidate path set is enlarged to include those paths that have a length less than or equal to minimum-hop+2, revenue generally decreases. This decrease is more pronounced in the uniform workload case (cf. Figure 6(b)). *Full-Mesh* is the most affected (about 11%), followed by *Star* (about 9%), and finally *Simple-Node* being the least affected (about 5%).

5 Conclusions

We conclude by summarizing our observations.

- The workload distribution can strongly affect the performance of different aggregation schemes. When network resources are scarce due to heavy load, the performance loss caused by less accurate routing/QoS information may not be as significant as that caused by poor load distribution. Moreover, the principle of selecting candidate paths of short length and giving preference to shortest paths minimizes the amount of resources allocated to a connection, however, it may not provide a good balance of the load over the network. Consequently, when the network is subjected to skewed workload, some links may become saturated quite early so that they can not be allocated for further connections. Therefore, more accurate routing information, resulting in a more restricted (smaller) set of candidate paths, may provide worse overall network performance.

- With state aggregation, the source node has an inaccurate view of remote nodes and links. This inaccuracy affects the computation of QoS parameters such as per link delay requirements. Obviously, per link delay requirement with *Simple Node* and *Star* is more stringent than with *Full-Mesh* under the same conditions since we assume *Simple Node* and *Star* advertize the maximum intra-group link hop-count. Consequently it would result in a more conservative way to allocate resources to connections. For this reason, our results show that *Simple Node* and *Star* yield slightly higher connection blocking probabilities compared to *Full-Mesh*.

- The improvement in performance obtained by giving preference to shortest paths when assigning path selection weights is not significant since long paths are already filtered out of the set of candidate paths.

- Restricting the size of the candidate path set can effectively enhance performance. This is because long paths would always result in increased utilization and hence reduced throughput. But candidate path restriction can also have a negative effect as the load may not be well balanced over the network. Therefore, a good routing policy must achieve a good compromise between link usage efficiency and load balancing. This is affected by the distribution of the workload as well as the sensitivity of different aggregation schemes to the candidate path restriction. Obviously, *Full-Mesh* is the most sensitive because it advertises the most detailed routing information.

Our future research includes studying other aggregation schemes and route selection policies.

References

[1] H. Ahmadi, J. Chen, and R. Guerin. Dynamic Routing and Call Control in High-Speed Integrated Networks. In Proc. *Workshop on Systems Engineering and Traffic Engineering, ITC'13*, pages 19–26, Copenhagen, Denmark, June 1991.

[2] I. Althofer, G. Das, D. Dopkin, D. Joseph, and J. Soares. On Sparse Spanners of Weighted Graphs. *Discrete and Computational Geometry*, pages 981-100, 1993.

[3] B. Awerbuch, Y. Du, B. Khan, Y. Shavitt. Routing Through Teranode Networks with Topology Aggregation. http://www.cnds.jhu.edu/publications/.

[4] B. Awerbuch, Y. Shavitt. Topology Aggregation for Directed Graph. http://www.cnds.jhu.edu/publications/.

[5] S. Bahk and M. Zarki. Dynamic Multi-Path Routing and How it Compares with Other Dynamic Routing Algorithms for High Speed Wide Area Networks. In *SIGCOMM '92*, pages 53–64, Baltimore, Maryland, August 1992.

[6] J. Baras, M. Corson *et. al.* Tactical and Strategic Communication Network Simulation and Performance Analysis. *In the proceeding of the Advanced Telecommunications/Information Distribution Research Program (ATIRP) 1st Annual Conference*, January 1997.

[7] Y. Bartal. Probablistic Approximation of Metric Space and its Algorithmic Applications. *In 37th Annual IEEE Symposium on Foundations of Computer Science*, October 1996.

[8] L. Breslau, D. Estrin, and L. Zhang. A Simulation Study of Adaptive Source Routing in Integrated Services networks. ftp.catarina.usc.edu:pub/breslau.

[9] E. Crawley, R. Nair, B. Rajagopalan, and H. Sandick. A Framework for QoS-based Routing in the Internet. Internet draft draft-nair-qos-based-routing-01.txt, March 1997.

[10] R. Guerin and A. Orda. QoS-Based Routing in Networks with Inaccurate Information: Theory and Algorithms. *IBM Research Report* RC 20515, July 1996.

[11] R. Guerin, H. Ahmadi, and M. Naghshineh. Equivalent Capacity and its Application to Bandwidth Allocation in High-Speed Networks. *IEEE J. Select. Areas Commun.*, SAC-9(7):968–981, September 1991.

[12] L. Guo and I. Matta. Technical Report, extended version. Under preparation.

[13] W. Lee. Topology Aggregation for Hierarchical Routing in ATM Networks. *Computer Communication Review: Special Issue on ATM*, pages 82-92, April 1995.

[14] W. Lee. Spanning Tree Method for Link State Aggregation in Large Communication Networks. *IEEE INFOCOM '95*, pages 297-302, April 1995.

[15] Q. Ma and P. Steenkiste. On Path Selection for Traffic with Bandwidth Guarantees. *In the Proceedings of IEEE International Conference on Network Protocols*, Atlanta, Georgia, October 1997.

[16] I. Matta and A. Udaya Shankar. Fast Time-Dependent Evaluation of Multi-Service Networks. To appear in Computer Networks and ISDN Systems - Special Issue on Modeling of Wired and Wireless ATM. http://www.ccs.neu.edu/home/matta/publications.html.

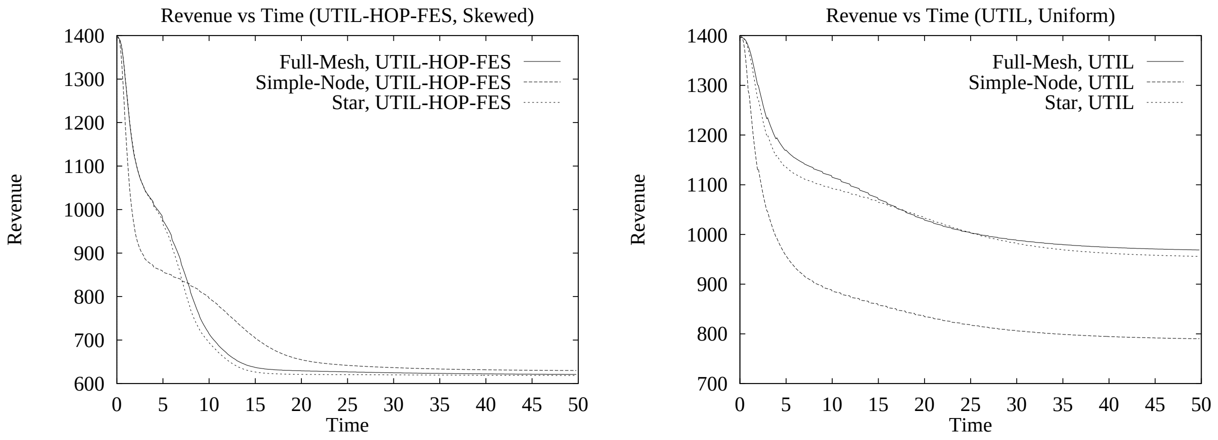
[17] R. Nagarajan, J. Kurose, and D. Towsley. Local Allocation of End-to-End Quality-of-Service in High-Speed Networks. In Proc. *IFIP TC6 Workshop on Modeling and Performance Evaluation of ATM Technology*, page 2.2, January 1993.

[18] A. Parekh. A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks. Technical Report LIDS-TR-2089, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, 1992.

[19] The ATM Forum technical committee. Private Network-Network Interface Specification Version 1.0 (PNNI). *ATM Forum af-pnni-0055.000*, March 1996.

[20] Z. Wang and J. Crowcraft. Quality of Service Routing for Supporting Multimedia Applications. IEEE J. Select. Areas Communication, 14(7):1228-1234, 1996.

[21] H. Xie and J. Baras. Performance Analysis of PNNI Routing in ATM Networks: Hierarchical Reduced Load Approximation. *Institute of System Research Technical Report*, TR 97-65, Univ. of Maryland. 1997.



(a) (b)

Figure 4: Network revenue versus time: (a) Skewed workload (b) Uniform workload.

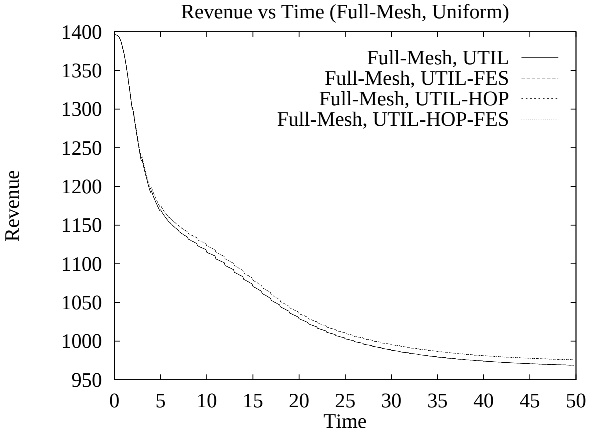
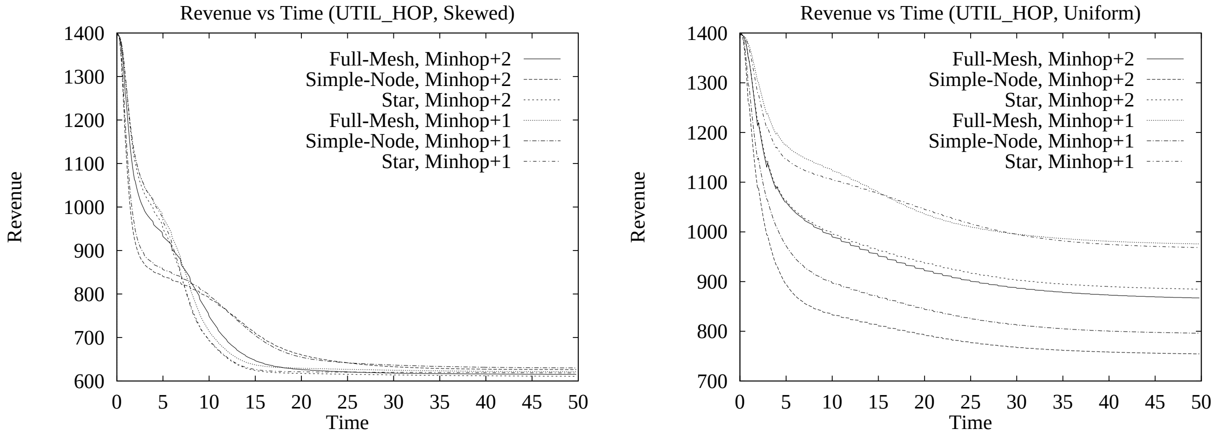


Figure 5: Network revenue versus time. Full-Mesh scheme. Uniform workload.



(a) (b)

Figure 6: Revenue versus time, different candidate path restrictions: (a) Skewed workload (b) Uniform workload.