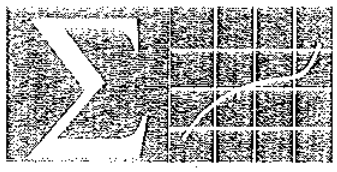


Capacity Management and Routing Policies for Voice over IP Traffic

Partho P. Mishra, Gigabit Wireless
Huzur Saran, Indian Institute of Technology Delhi

Abstract

This article addresses the problem of designing capacity management and routing mechanisms to support telephony over an IP network. For this service, we propose two distinct architectural models. The first relies on enhancements to the basic IP infrastructure to support integrated service transport and QoS routing. The second assumes that the IP network can support an overlay virtual private network with dedicated capacity for the VoIP service, thereby allowing standard capacity management and routing mechanisms from circuit-switched networks to be reused. We evaluate the performance of these two architectural models and their associated policies via simulations using configuration and usage data derived from operational networks.

There has been substantial interest in the recent past in migrating telephony service away from circuit-switched networks onto an IP-based packet-switched network infrastructure. One critical issue to be resolved in achieving this migration is how to support the quality of service (QoS) requirements of a high-quality telephony service over an IP network. A telephony service requires stringent bounds on end-to-end packet delay, jitter, and loss. Ensuring that these requirements are met requires the use of resource management mechanisms, such as scheduling and admission control, in the network. A telephony service also requires that the probability of blocking offered calls be fairly small (≤1 percent). This requires the use of capacity planning and provisioning mechanisms to ensure that the network has adequate capacity to handle the expected traffic volume, and routing mechanisms to ensure that the offered traffic is routed over the network in a manner that makes the most efficient use of available network capacity.

The design of capacity planning, routing, and resource management mechanisms that are suited to the requirements of telephony traffic has been very well addressed in the realm of the circuit-switched networks that have traditionally carried telephone traffic [1]. Sophisticated capacity planning, routing mechanisms, and resource management mechanisms have also been developed in IP networks. However, since the services that run on IP networks have different QoS requirements and traffic characteristics than telephony, these mechanisms cannot easily be modified to support voice-over-IP (VoIP) services.

This work was done while Partho P. Mishra was with AT&T Labs-Research, and Huzur Saran was visiting AT&T Labs-Research.

In this article we examine how the capacity management and routing mechanisms used in IP networks can be augmented to support an IP telephony service. We focus on a particular style of IP telephony service in which the IP network is used for "trunking" voice traffic between telephony switches, such as private branch exchanges (PBXs), local exchange carrier switches, or long distance toll switches. For this service, we evaluate the performance of two distinct architectural models for capacity management.

In the first model, we assume that the backbone IP network is enhanced to support an IP integrated services framework [2]: IP endpoints express the QoS requirements and traffic characteristics for a voice call using Resource Reservation Protocol (RSVP) [3], and all routers on the end-to-end path use the information in the RSVP message to perform admission control. The routing of the RSVP signaling messages follows the path determined by the IP routing policy. We evaluate three routing policies: Shortest Path First (SPF), Shortest Available Path First (SAPF), and Widest Available Path First (WAPF). In SPF, a voice call is routed along the shortest path between the IP endpoints. If capacity is not available on this path, the call is blocked. In the SAPF and WAPF policies, the network attempts to route the call along alternate paths when capacity is not available on the shortest path.

In the second model, we assume that the voice service is provided with its own virtual private network(VPN), consisting of a set of virtual leased lines (VLLs) [4–6] interconnecting the telephony switches over the IP network. The routing of VLLs over the underlying IP network is along the shortest path, as measured by the hop count, between the two switches. Each VLL has a specific capacity associated with it; as long as the volume of traffic generated by the customer does not

0890-8044/00/$10.00 © 2000 IEEE

exceed this capacity, the service provider guarantees bounded packet loss and delay. Since VLLs are functionally similar to dedicated links, capacity management and routing can be done in a manner analogous to circuit-switched networks. We evaluate four routing policies: Direct Path Only (DPO), Success to the Top (STT), State-Dependent Routing (SDR), and Approximate State-Dependent Routing (ASDR). In DPO a call is admitted only when there is adequate capacity on the direct path of the call (i.e., the VLL between the ingress and egress telephony switches). In the other policies, a call is admitted when there is adequate capacity on either the direct path or an alternate path consisting of two VLLs that interconnect the ingress and egress switches.

We evaluate the performance of the two capacity management models and the various associated routing policies using simulations. To ensure that our simulation results are realistic, we use configuration and usage data that are derived from operational carrier networks. The simulated IP network is based on an ISP network topology, and the IP telephony traffic is simulated using traces derived from actual calling patterns on the carrier's long distance network. The specific results reported in this article are based on simulating a 16-hr interval using the set of calls placed on the long distance network between midnight and 4 p.m. on Monday, May 17, 1999.

The rest of this article is organized as follows. We briefly review the routing algorithms currently used in voice (circuit-switched) and data (IP) networks. Next, we describe two architectural models for supporting telephony services over an IP network, and the associated capacity management and routing options. The article then presents simulation results evaluating the performance of each of these options. We then conclude the article.

Routing Policies Used in Existing Circuit-Switched and IP Networks

The design of the routing policies used in existing circuit-switched and IP networks has been strongly influenced by the structure of the network topology. Most circuit-switched networks used to carry voice traffic have evolved to a topology consisting of a core network with full mesh connectivity between a set of *toll switches* and an access network with a hierarchical tree-like structure. Simple static routing techniques (with provision for failover routes) are used in the access network. More complex routing techniques are used in the core, which is much more richly connected, thereby providing a richer choice of routing options.

In the core network, each pair of toll switches has a set of links providing direct connectivity. The simplest possible routing policy, DPO, is to accept a call only when there is adequate capacity on the direct path between the originating and terminating toll switches. However, since the link capacities are typically provisioned based on estimated call arrival patterns, it is always possible that these capacities could be exceeded due to a transient overload, resulting in call blocking. To reduce the call blocking probability, most routing algorithms also consider alternate routes before blocking a call. An alternate route typically consists of a two-hop route, where the call from switch i to switch j is routed through an intermediate switch, k. Since a call routed over two hops occupies double the network capacity, two-hop paths are used only when the direct path is unable to admit the call. While, it is possible to consider even longer alternate paths, the additional performance gains are marginal. In a core network consisting of n switches, there are $n - 2$ two-hop paths between any pair of switches. For each (i, j) switch pair, a routing policy has to determine the sequence in which it will try to route a call over each of these alternate paths and how many alternate paths it will try before blocking the call. Ideally, a routing policy should quickly discover an alternate path that has adequate capacity to minimize call setup latency. It is also desirable to spread out the offered load evenly among all the available links to make the most efficient use of network capacity.

A simple but naive alternate routing policy would be to search through the set of alternate routes in a predetermined order. For example, assume that the set of switches is numbered 1 through n, and a call is being set up between switches i and j. The routing policy might start the search for alternate routes by first trying to route the call through switch 1, then switch 2, and so on. For such a policy, the links incident on the lower-numbered switches are likely to be heavily loaded, while those for higher-numbered switches will rarely be required to carry alternate routed traffic. Such an approach would also require evaluating a larger number of alternate routes, on the average, before a call can be accepted. This increases the call setup latency. Clearly, this policy is not very efficient. A simple variation of this naive policy, which addresses some of these problems, is to randomize the starting point of the search. This makes it more likely that the offered load due to alternate routed calls will be distributed evenly across all the links. Another variant is to precompute a sequence in which to try the intermediate points for each (i, j) pair based on historical load information.

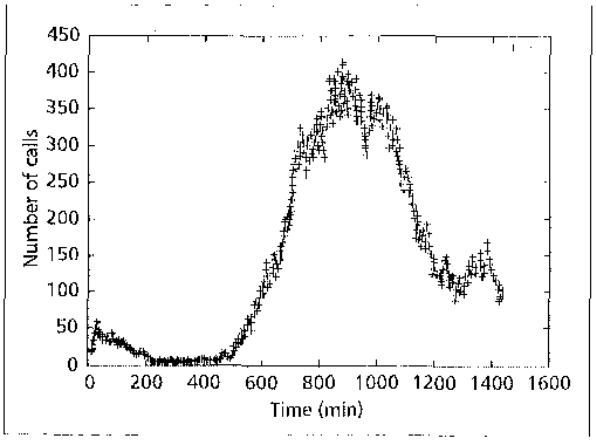
A simple and effective routing policy that has been used in operational networks is STT. For each (i, j) pair, STT keeps track of the identity of the node k through which an alternate call was last successfully routed. Whenever a new call from i to j needs to be alternate routed, STT first tries to route via k. If there is no capacity on this alternate path, STT uses the randomized search procedure described earlier to pick a new alternate path. The intuition behind STT is that if a particular alternate path has been used successfully in the recent past, it is still likely to have available capacity.

All of the policies discussed so far do not take into consideration the existing state of a link. For instance, say that a call from switch i to switch j needs to be alternate-path routed, and that the capacity available on links (i, k), (k, j), and (l, j) is adequate for 10 calls, while the capacity available in (i, l) is adequate for only one call. A routing policy that chooses to try the path $<i, l, k>$ first will cause link (i, l) to become saturated. This may in turn cause a call arriving on i for l to be alternate routed or blocked. To address this problem, state-dependent routing policies have been defined that make use of current link loads in selecting an alternate path. In SDR the originating switch i determines the available capacity on each of its links. It also queries the destination switch j to determine the available capacities on all links incident on j. The call is then routed over an intermediate switch l such that $\min(\text{Avail}(i, l), \text{Avail}(l, j))$ is maximum over all nodes. Thus, SDR tries to minimize the likelihood of forcing a future call to be alternate routed/blocked. SDR, as described above, requires instantaneous load information. An approximate version of SDR, ASDR, studied in this article, makes the routing decision based on predicted link loads derived using periodic link state updates.

Alternate path routing algorithms such as STT, SDR, and ASDR can demonstrate instability under heavy loads. Suppose, for instance, that each link has offered load close to its provisioned load. If a small transient overload were to cause a call to be alternate routed, this alternate routed call may take up one unit of capacity each on two links which are themselves running at close to provisioned capacity. This in turn may cause two calls which could have been direct rout-

ed on those links to be alternate routed. This can cascade quickly into a situation where the network is carrying half the original traffic volume and is still saturated since every call is alternate routed. To address this potential instability, circuit-switched networks use alternate-path routing in combination with *trunk reservation* techniques [1]. In trunk reservation, a certain amount of capacity on each trunk (referred to as the *trunk reservation parameter*) is reserved for the use of direct routed calls. When the available capacity on a link falls below its trunk reservation parameter, the link will refuse to admit any alternate routed traffic, although it will still admit direct routed traffic. It has been observed that a trunk reservation of even 10 calls can extend the effective operating range of alternate routing algorithms to a wide range of offered loads.

In comparison to the routing strategies for circuit-switched (voice) networks, routing in IP networks has traditionally employed only a single forwarding path at a time between a pair of IP routers. The distance vector or link state routing protocols, such as Border Gateway Protocol (BGP) or Open Shortest Path First (OSPF), used in IP networks assign a weight (or cost metric) to each link and then route traffic along the least weight path (shortest path) between every source–destination pair. The consequence is that there is only one path chosen at any particular point in time between a source–destination pair. In case the available capacity on that path is lower than the amount of traffic between the specified source–destination pair, packets will be dropped even if an alternate path exists to route the overflow traffic. By basing the weights/cost metric on the link capacities along the path, it is possible to bias the routing procedure to select higher-capacity paths, which may alleviate this problem, but there may exist situations where no individual path has sufficient bandwidth to carry all the traffic between a source–destination pair. Certain extensions to OSPF offer a limited form of multipath routing capability. These extensions allow a router to split the traffic for a source–destination pair between all the minimum weight paths for that source–destination pair. By a careful choice of weights, it is possible to configure more than one shortest paths between a specified source-destination pair and thereby enable traffic splitting. More recently, multiprotocol label switching (MPLS) offers another mechanism through which multiple logical paths or links may be configured between specific source–destination pairs [4]. These links can potentially be used to override the default OSPF routing mechanisms.



■ Figure 1. *The number of active calls from a New York area code to a Los Angeles area code on May 17, 1999.*

Capacity Management and Routing Policies for Voice over IP Traffic

Our focus in this article is on capacity management and routing policies for a VoIP service in which the IP network is used for "linking" voice traffic between telephony switches, such as PBXs, local exchange carrier switches, or long distance toll switches. In this configuration, a public switched telephone network (PSTN) gateway acts as the interface between the circuit-switched and IP network segments, performing the analog–digital and circuit–packet conversions. These PSTN gateways are connected to an edge router at an Internet service provider (ISP) point of presence. Edge routers are connected to backbone routers through an access network. A backbone network segment interconnects the backbone routers.

Supporting high-quality telephony services over an IP network requires the network to meet fairly stringent packet loss and delay requirements; typically, the end-to-end round-trip delay has to be less than 300 ms, and packet loss rates need to be 1 percent or lower. Meeting these QoS requirements requires the use of bandwidth management and admission control mechanisms in the IP network to ensure that adequate capacity exists before a new voice call is admitted. There are two distinct architectural models for how this can be accomplished.

The Integrated Services Model for Capacity Management

The first model for supporting the QoS required by VoIP traffic is to extend the backbone IP network to support an IP integrated services framework [2]. In the IP integrated services framework, IP endpoints express the QoS requirements and traffic characteristics for an IP flow using the RSVP signaling protocol [3]. Routers on the end-to-end path of the IP flow use the information in the RSVP message to perform admission control and to set up per-flow state in the data path in the router to ensure that the packets for this flow get the appropriate forwarding treatment. For the IP voice service, the RSVP message would need to be initiated by the PSTN gateways; a call would be admitted only if the RSVP message successfully reserved resources at every intermediate access and backbone IP router between the two gateways.

The routing of the RSVP signaling messages follows the path determined by the IP layer routing protocol (e.g., OSPF). If capacity is not available on this path, the call is blocked, even if available capacity exists on an alternate path between the two endpoints (i.e., the PSTN gateways). To address this potential inefficiency, there have been several proposals to use more sophisticated QoS-sensitive routing policies in which the routing policy explores alternate paths after learning that the direct path does not have capacity. These policies may be viewed as analogous to STT/SDR-like routing policies in the IP world. However, there is an important difference between these two classes of policies due to the different topologies used in the circuit-switched and IP networks.

Since the circuit-switched network topology is a clique, there are a very large number of alternate paths between source and destination that are two hops long. Therefore, STT/SDR-like policies only explore the set of alternate paths that are two hops long. The order in which the set of alternate paths is explored when trying to set up a call is determined based on the maximum (currently) available capacity on each of these paths, or some other similar criterion that evaluates the likelihood of making a successful reservation on each of these alternate paths. However, IP networks typically have a more sparsely connected and irregular topology. Therefore, it is not adequate to only compare among the set of available

alternate paths that are one hop longer than the shortest path (in fact, there many not be even a single alternate path that is one hop longer than the shortest path). When the set of alternate paths being evaluated have different hop counts, the order in which they are tried can depend on both the hop counts for these paths and the likelihood of making a successful reservation. In this article we explore the performance of two policies that represent the two extremes in this space of options:

- Shortest Available Path First (SAPF): The set of alternate paths is explored in the order of hop count.
- Widest Available Path First (WAPF): The set of alternate paths is explored in the order of maximum (currently) available capacity.

The integrated services model allows very efficient use of network bandwidth since capacity on a physical link is tied up only when it is being used by an active voice call. However, the IP integrated services framework has not been deployed very widely. This is because of scalability problems resulting from the need to implement per-flow signaling and retain per-flow state at every router on the end-to-end path of a flow. For example, it may not be feasible to implement per-flow queuing and scheduling for hundreds of thousands of active flows at a backbone router. More scalable capacity management models have been proposed in which per-flow signaling and scheduling is used in the access network, but backbone routers manage capacity on a more aggregated basis [6, 7]. An example of such a model is one in which the voice service is provided with its own VPN over the IP backbone network.

The VPN Model for Capacity Management

IP-based VPNs typically attempt to emulate the service model provided by private line or frame relay service: a customer buys VLLs (over the ISP's network) to interconnect customer routers. Each VLL has a specific capacity associated with it; as long as the volume of traffic generated by the customer does not exceed this capacity, the service provider guarantees bounded packet loss and delay. This is ensured by provisioning adequate bandwidth along the set of physical links along which a VLL is routed.[1] Several mechanisms have been proposed to allow this style of provisioning, for example, through the use of bandwidth brokers or hop-by-hop signaling [5]. When a VPN is implemented over a traditional IP network, the topology of the VPN is a clique, since there exists a forwarding path and hence a logical link between every pair of customer edge routers.

When the VoIP service has its own dedicated VPN, admission control and routing can be done using fairly straightforward extensions to the techniques described in earlier. For example, if DPO routing is being used, a call is admitted only when there is adequate capacity on the direct path (i.e., the VLL between the ingress and egress PSTN gateways). Similarly, when using one of the alternate path routing algorithms, such as STT or SDR, a call is admitted when there is adequate capacity on either the direct path or any of the alternate two-hop paths, consisting of two VLLs interconnecting the ingress and egress gateways.

The VLL model requires the customer to request adequate capacity to accommodate the worst-case traffic requirements for each VLL. While this is similar to the manner in which capacity is provisioned in circuit-switched networks, it can be wasteful if the traffic on every VLL is bursty or time-varying.

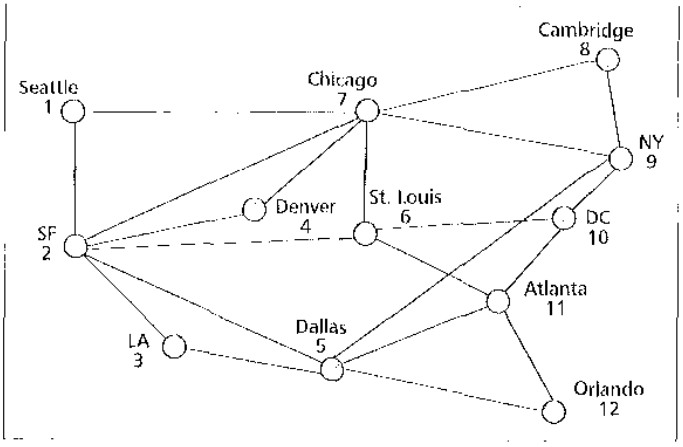


Figure 2. *The physical topology of the simulated IP backbone.*

Figure 1, which plots the count of active calls between two area codes in the New York and Los Angeles metropolitan areas on a minute-by-minute basis, illustrates that telephony traffic is both bursty and time-varying. In fact, the use of alternate path routing techniques such as STT or SDR can be viewed as a way to compensate for traffic burstiness and improve the efficiency with which network capacity is utilized by allowing VLLs to share capacity on very small timescales.

Another way of dealing with the burstiness or unpredictability of the traffic on a VLL is to do a greater amount of traffic aggregation before the voice traffic enters the IP backbone. In other words, the number of VLLs is reduced so that each VLL carries a greater volume of traffic, thereby reducing the burstiness of the traffic on each VLL due to a law-of-large-numbers effect. Below, we show that even small reductions in the number of PSTN gateways (e.g., by a factor of 4) can result in significant improvements in performance.

Results

In this section we present experimental results that compare the performance of the various capacity management and routing policies described above. These results are obtained using call-level simulations, in which we model the arrival and departure of voice calls (flows) onto a simulated network consisting of a set of PSTN gateways interconnected via a backbone IP network. Each voice call is assumed to generate data at a constant rate and therefore require a fixed amount of bandwidth for the duration of the call. To ensure that our results are realistic, we use configuration and usage data derived from operational networks:

- The arrival and departure processes for the voice calls are simulated using traces derived from the *call detail record* database. This database has a record for every telephone call ever placed on the carrier's long distance network. Each record includes the originating, dialed, and terminating telephone number, as well as the origination time and duration of the call.[2]
- The simulated backbone IP network topology is derived from an ISP network topology. This network comprises 12 core routers and 42 links spanning the continental United States (Fig. 2). For the purpose of shortest-path computations, it is assumed that each of these links has the same weight. The simulated access network topology comprises a set of PSTN gateways that are directly connected to one of these 12 back-

[1] *The path along which the packets belonging to a VLL are routed is determined by the IP-layer routing algorithm (e.g., OSPF).*

[2] *The origination time and duration are captured at the granularity of seconds.*

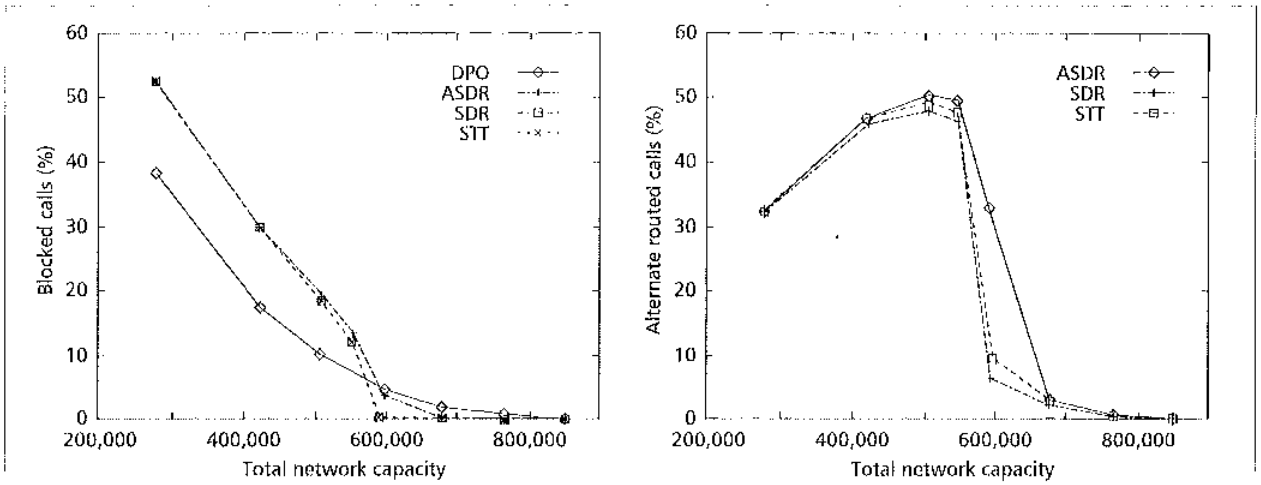


Figure 3. *Comparison of direct and alternate path routing policies with no trunk reservation.*

bone nodes. In most of the simulations, we assume that there are a total of 168 PSTN gateways, one for each of the 168 area codes in the United States and Canada.

The specific results reported in this article simulate a 16-hr interval using the set of all calls (approximately 30 million) placed on the carrier's long distance network between midnight and 4 p.m. on Monday, May 17, 1999. For each of these calls, we generate a call origination event and a call departure event in the simulation, at the time instants corresponding to the call origination and call departure time recorded in the CDR.

When a call origination event occurs, we first identify the originating and terminating PSTN gateway for the call using the first three digits (area code) of the originating and terminating numbers in the CDR. Next, an admission control check is executed to verify whether the voice call can be admitted. The mechanism used to performed admission control depends on the capacity management and routing policy being simulated. For example, with the (VPN capacity management, direct-path routing) policy, the originating PSTN gateway checks whether it has adequate capacity on the VLL connecting it to the terminating PSTN gateway for the new call. If the admission control fails, the voice call is blocked. When a call departure event occurs, the capacity assigned to the call is freed on the VLL or the set of physical links over which the call is routed.

In our simulations we first compute the minimum edge capacities needed to ensure that there is no blocked call and all calls are routed using the direct path. To obtain this we run a simulation, using the VPN capacity management model and the direct-path routing policy, for the topology shown in Fig. 2. In this simulation run, the link capacities are set to be infinitely large so that no calls are blocked. For this run, we measure

$$V_i = \max_{t=0}^{t=57,600} v_i^t \qquad (1)$$

where v_i^t is the number of active calls on link i at time t (the tick interval in our simulation is 1 s). Therefore, V_i represents the maximum call volume on link i for this particular simulation. Clearly, if we were to set the link capacity, C_i, equal to V_i, no call would ever be blocked, for this particular combination of call arrival process, capacity management model, and routing policy.

To gain an understanding of the benefits of alternate path routing, aggregation, and so on, we examine the performance in terms of blocked calls when we reduce the capacity by 10–50 percent below these baseline capacities. Although one can study situations where the capacity reduction is nonuniform, in our work we focus on the case when we have a uniform reduction on the capacity of each edge (or VLL), by setting the individual link capacities to be $\alpha * C_i$, where $0 \le \alpha \le 1$.

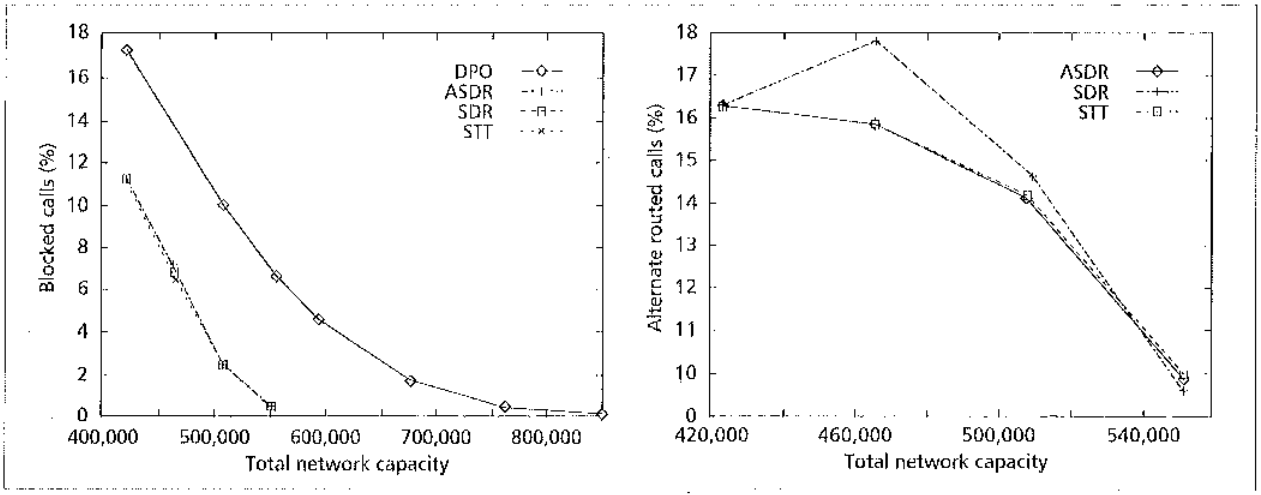


Figure 4. *Comparison of direct and alternate path routing policies with trunk reservation.*

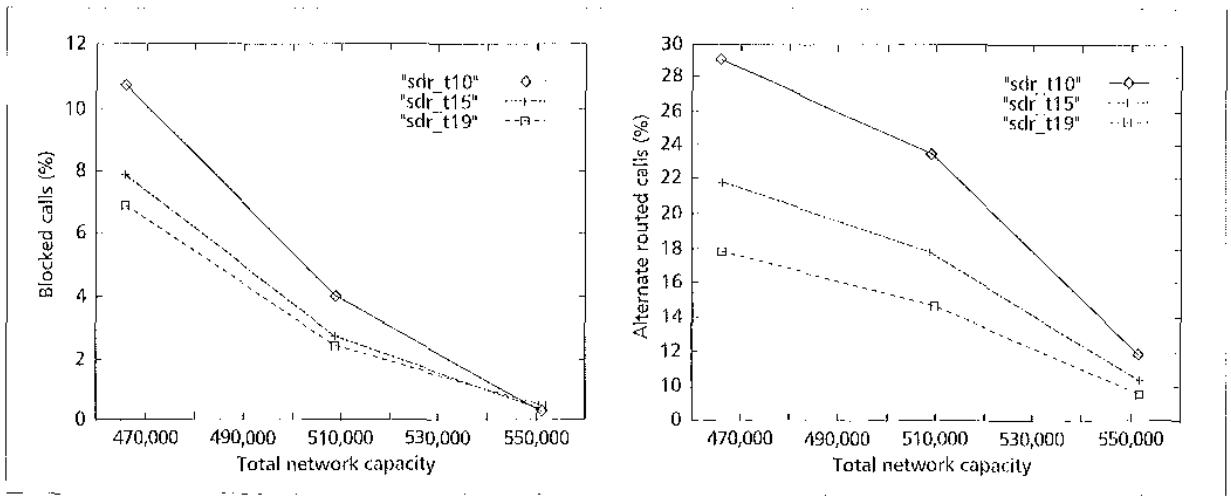


Figure 5. *The effect of trunk reservation thresholds on the performance of SDR.*

Evaluation of Routing Policies with the VPN Model

Our first set of experiments evaluates the effect of the routing policy when using the VPN model of capacity management. In these simulations there are 168 PSTN gateways, so there are a total of 168*168 VLLs. Each VLL is routed over a set of intermediate links corresponding to the shortest path between each gateway pair. Figure 3 shows the performance of DPO, SDR, STT, and ASDR routing policies as measured by the call blocking probability as the aggregate network capacity is varied between 847,816 and 279,779 calls. In this experiment, the SDR, STT, and ASDR policies did not make use of trunk reservation.

The results show that when the network capacity is around 847,816 calls, all of the routing policies result in zero call blocking. As the network capacity is reduced from 847,816 to 593,471, the call blocking probability increases steadily with the DPO policy. However, the STT and SDR policies are able to ensure zero blocking by routing an increasing fraction of the calls over an alternate path. For example, when the network capacity is 593,471 calls, SDR routes 6.24 percent of the calls over a two-hop path, while STT routes 9.43 percent of its calls over a two-hop path.

As the network capacity is reduced further from 593,471 calls, the call blocking probability with the DPO policy continues to increase steadily. However, the call blocking probability obtained with the STT and SDR policies increases very abruptly from 0 to 12.15 percent as the network capacity is reduced from 593,471 to 551,080 calls. The reason for this abrupt falloff is that by the time the network capacity is reduced to 551,080, almost 50 percent of the admitted calls are routed on two-hop paths. In this regime, every call traversing a two-hop path is potentially preempting two calls that could have been supported over each of these two hops. This is a well-known behavior in the circuit-switched world [1], and is typically dealt with using trunk reservations as described earlier.

The Effect of Trunk Reservations — To validate that our hypothesis is correct, we reran the same simulations using trunk reservation. The trunk reservation threshold is set to 19 percent, that is, at least 19 percent of a VLL is always held in reserve for direct traffic. The results are shown in Fig. 4. We only show the behavior in the capacity region that is of interest, that is, where the blocking probability is not zero but not very high either.

The results show that the call blocking probability drops from 12.15 to 0.45 percent for a network capacity of 551,080 calls when using SDR. A similar effect is observed with both the STT and ASDR policies. This happens because the trunk reservation policy drastically reduces the fraction of alternate routed calls. Even at a capacity of 508,689, the call blocking is

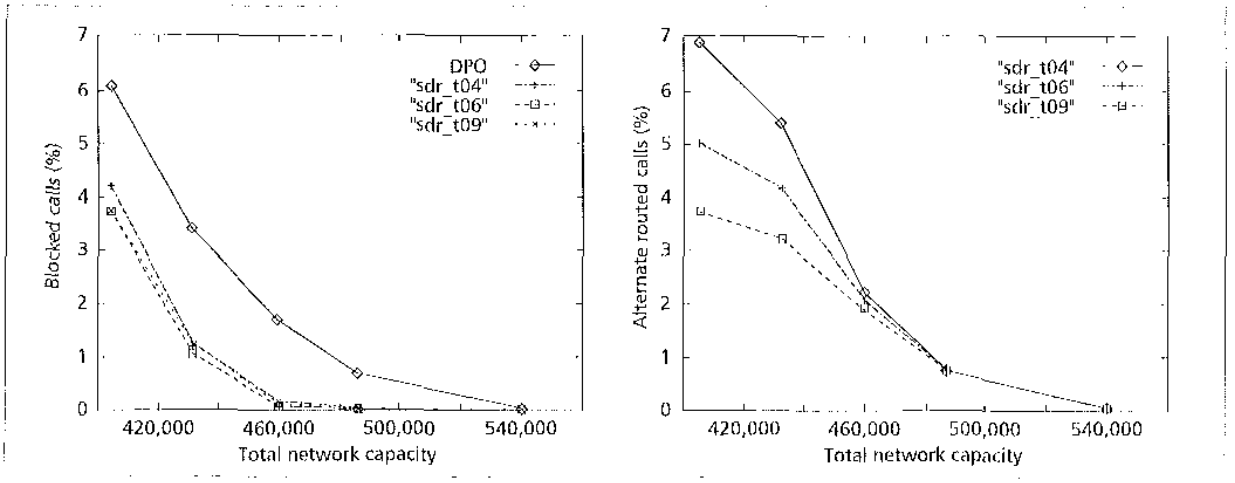


Figure 6. *The effect of alternate path routing at higher levels of aggregation: 48 access ports on the VPN.*

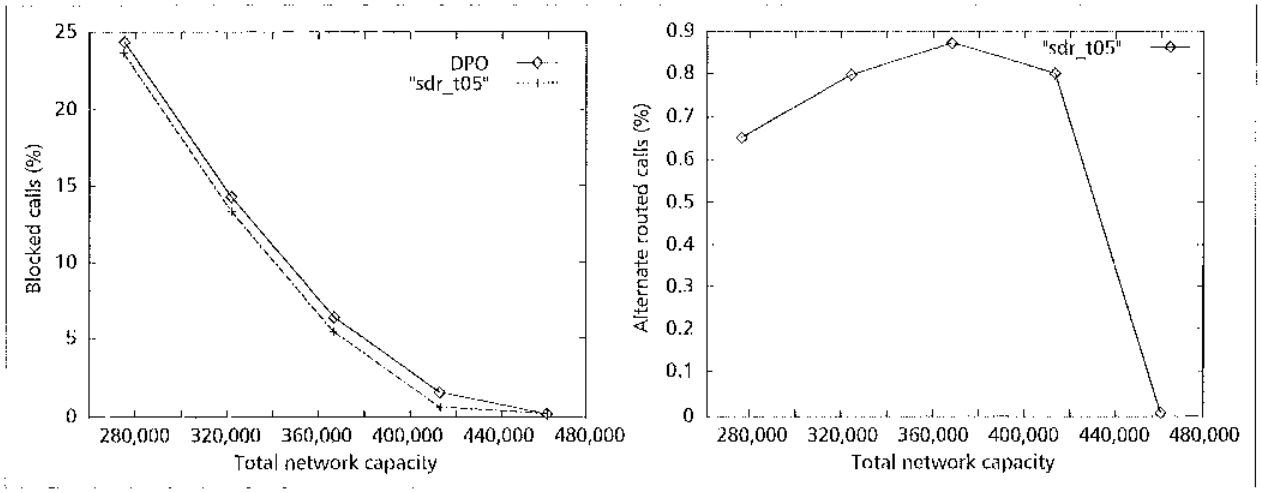


Figure 7. *The effect of alternate path routing at higher levels of aggregation: 12 access ports on the VPN.*

only around 2.5 percent. In contrast, DPO needs much more capacity for the same level of blocking; for example, SDR+TR yields a blocking probability of 6.89 percent with an aggregate network capacity of 466,298 calls, while for DPO the equivalent numbers are 6.95 percent and 551,080 calls, respectively.

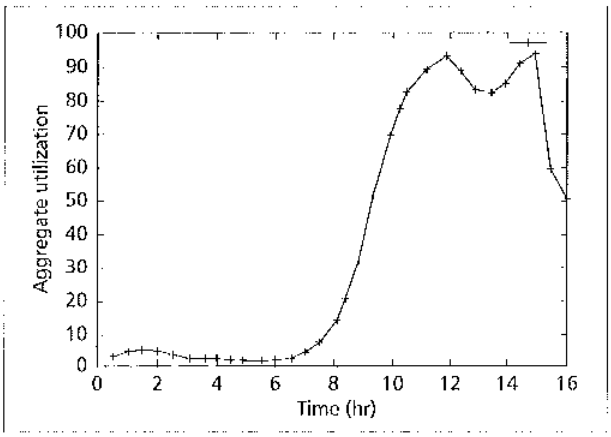


Figure 8. *Aggregate utilization.*

For smaller target blocking probabilities, the advantage of the alternate path routing policies is even more pronounced.

Since trunk reservation has such a dramatic impact, an important question is understanding whether adjusting the threshold significantly affects the performance. Figure 5 shows the effect of varying the trunk reservation threshold from 10 to 19 percent. The results suggest that a threshold of 10 percent seems to result in the best performance in the network capacity region that corresponds to the "sweet spot" of the SDR policy. At higher and lower loads, setting the threshold to 19 percent results in the best performance. In ongoing work we are studying how to set the "optimal" value of the trunk reservation threshold.

The Effect of Greater Traffic Aggregation — The next set of experiments evaluates whether SDR-like alternate routing policies are useful when there is a much greater level of traffic aggregation. We aggregate the traffic from multiple area codes onto the same VLI, so there are 48 x 48 VLIs instead of 168 x 168 VLIs.

For this level of aggregation, we study the performance of DPO and SDR with three different trunk reservation parameters. Figure 6 shows that at this greater level of aggregation, the network capacity required for near zero blocking is dramatically reduced. For example, even with DPO, the call

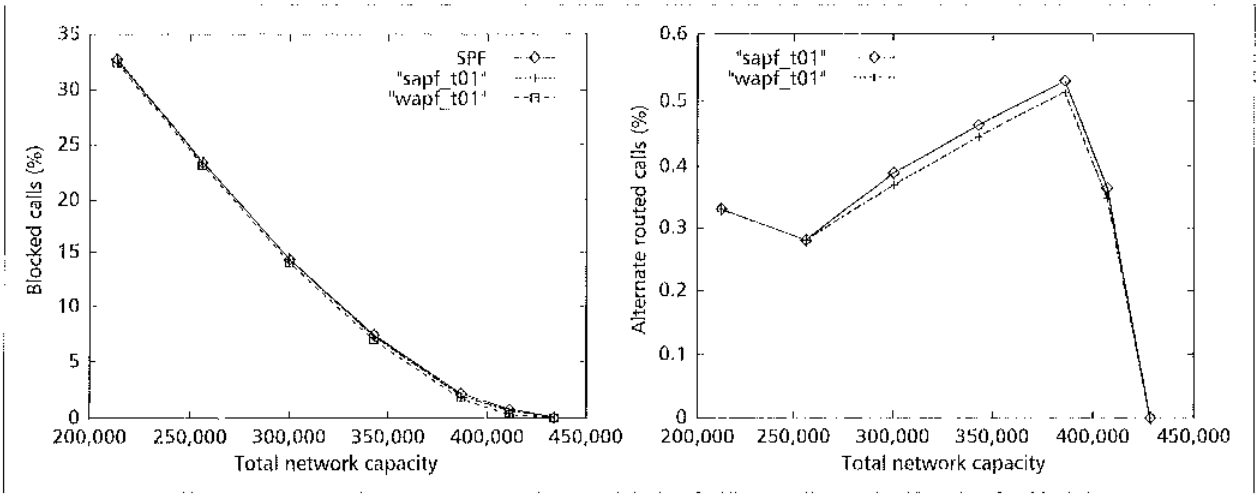


Figure 9. *The effect of alternate path routing in the int-serv model.*

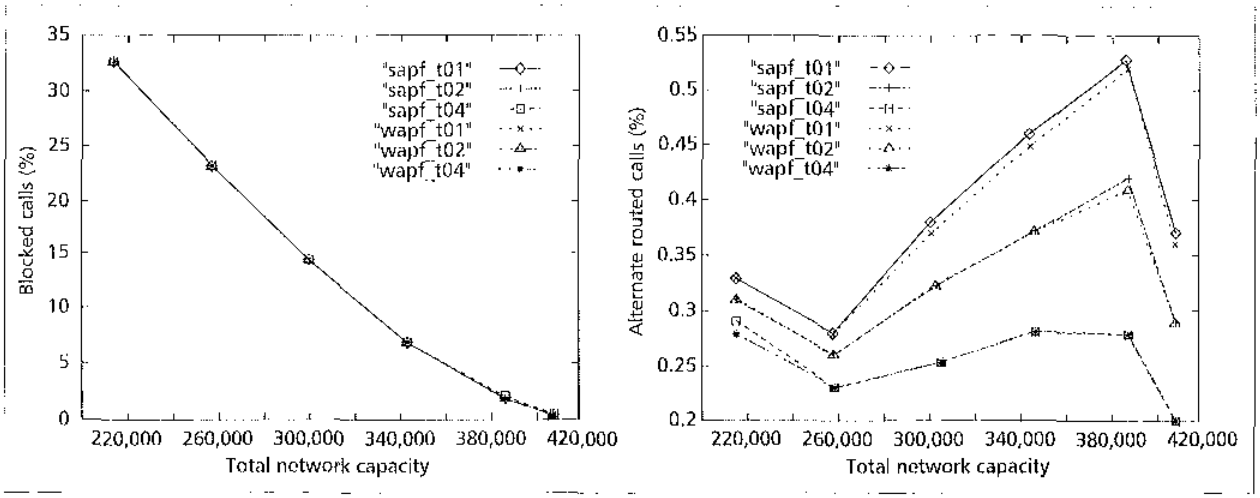


Figure 10. *The effect of the trunk reservation threshold on SAPF/WAPF performance.*

blocking probability is only 0.63 percent with a network capacity of 486,010 calls. The table also illustrates that at these greater levels of aggregation, the relative gain provided by SDR over DPO is much reduced. Figure 7 shows that these trends get more pronounced as the degree of aggregation is increased further with little difference in the performance of the DPO and SDR policies.

How Much Better Could the Best Routing Policy Do? — Since STT and SDR are both based on heuristics, one interesting question is how much margin for improvement there is: would it be possible, with an optimal routing policy, to reduce the aggregate network capacity significantly, while providing the same call blocking probability as SDR or STT? A simple way to estimate this is by looking at the aggregate network utilization. Figure 8 plots the aggregate network utilization averaged every half hour, for the case when we use DPO and the aggregate capacity in the network is 414,612. (Aggregate utilization = total hop-seconds of traffic carried over each half hour period divided by 1800*414,612) The results show that the network is utilized up to 94 percent at 3.00 p.m. Therefore, there is little margin for improvement beyond what STT and SDR are providing.

Evaluation of Routing Policies with the Integrated Services Model

The next set of simulations is for the integrated services model. We study three routing policies: direct routing on the shortest path between ingress and egress PSTN gateways (SPF), as well as alternate path routing based on SAPF or WAPF. In these simulations, SAPF and WAPF are always used in conjunction with trunk reservation. Figure 9 illustrates the performance of these three routing policies.

The most important observation is that all three policies require much less network capacity for the same level of call blocking probability than do the routing policies used with the VPN model. For example, even SPF results in only 0.68 percent blocking at a network capacity of 408,004 calls. This happens because the int-serv model allows much better sharing of the network capacity, resulting in much higher statistical multiplexing gain.

The second observation is that the advantage provided by alternate path routing over SPF is relatively small. Less than 0.5 percent of the traffic is alternate routed. We speculate that the reason for this is that due to the large amount of traffic aggregation on each of the links, all of the links are operating at close to 100 percent utilization during the busy hours. Finally, for this set of experiments, both the WAPF and SAPF policies perform almost identically.

Figure 10 shows that the trunk reservation threshold does not have a significant impact on the performance of the WAPF and SAPF policies.

Conclusions

There has been substantial interest in the recent past in migrating telephony service away from circuit-switched networks onto an IP-based packet-switched network infrastructure. One critical issue to be resolved in achieving this migration is how to support the QoS requirements of a high-quality telephony service over an IP network. In this article we have proposed two models for capacity management and routing in an IP network that allows these requirements to be met while making efficient use of network capacity. We have used simulations to evaluate the performance of these capacity management models.

References

[1] G. Ash, *Dynamic Routing in Telecommunications Networks*, McGraw-Hill, 1998.
[2] Wroclawski *et al.*, "The Use of RSVP with Integrated Services," RFC 2210, NIC, 1997.
[3] Braden *et al.*, "Resource Reservation Protocol (RSVP) - - Version 1 Functional Specification," RFC 2205, NIC, 1997.
[4] Callon *et al.*, "A Framework for Multiprotocol Label Switching," Internet draft, NIC, 1998.
[5] Nichols *et al.*, "Definition of the Differentiated Services Field (ds Field) in the IPV4 and IPV6 Headers," RFC 2474, NIC, 1998.
[6] P. Goyal *et al.*, "Integration of Call Signaling and Resource Management for IP Telephony," *IEEE Network*, May 1999.
[7] R. Guerin *et al.*, "Aggregating RSVP-Based QoS Requests," Internet draft, NIC, 1999.

Biographies

PARTHO P. MISHRA (pmishra@gigabitwireless.com) received a B.Tech. degree from the Indian Institute of Technology, Kharagpur, in 1988, and M.S. and Ph.D. degrees from the University of Maryland in 1991 and 1993, all in computer science. He is currently at Gigabit Wireless, Mountain View, California. He was previously at AT&T Labs-Research from 1996 to 1999, and with AT&T Bell Labs-Research from 1993 to 1996. His research interests include traffic management, wireless networking, and packet telephony and video services.

HUZUR SARAN (saran@cse.iitd.ernet.in) received a B.Tech degree in electrical engineering from the Indian Institute of Technology, Delhi, in 1983, and a Ph.D. degree in computer science from the University of California Berkeley in 1989. He is on the faculty of the Department of Computer Science and Engineering at the Indian Institute of Technology, Delhi. He his research interests include algorithms, wireless networking, scheduling, and traffic management.