A Scalable Model for Interbandwidth Broker Resource Reservation and Provisioning

Haci A. Mantar, *Member, IEEE*, Junseok Hwang, *Member, IEEE*, Ibrahim T. Okumus, *Member, IEEE*, and Steve J. Chapin

***Abstract*—As the Internet evolves into global communication and commercial infrastructure, the need for quality-of-services (QoSs) in the Internet becomes more and more important. With a bandwidth broker (BB) support in each administrative domain, differentiated services (Diffserv) is seen as a key technology for achieving QoS guarantees in a scalable, efficient, and deployable manner in the Internet.**

In this paper, we present a scalable model for inter-BB resource reservation and provisioning. Our BB uses centralized network state maintenance and pipe-based intradomain resource management schemes that significantly reduce admission control time and minimize scalability problems present in prior research. For inter-BB communication, we design and implement a BB resource reservation and provisioning protocol (BBRP). BBRP performs destination-based aggregated resource reservation based on bilateral service level agreements (SLAs) between peer-BBs. BBRP significantly reduces the BB and border routers state scalability problem by maintaining reservation state based only on destination region. It minimizes inter-BB signaling scalability by using aggregated type resource reservation and provisioning. Both analytical and experimental results verify the BBRP achievements.

***Index Terms*—Bandwidth broker (BB), BB signaling, domain, differentiated services (Diffserv), interdomain resource management, quality-of-service (QoS), scalability.**

I. Introduction

WITH THE rapid growth of the Internet into a global communication and commercial infrastructure, it has become evident that Internet service providers (ISPs) need to implement quality-of-service (QoS) to support diverse applications' requirements (e.g., packet delay, packet loss ratio) with their limited network resources.

Integrated services (Intserv) with resource reservation protocol (RSVP) signaling provides per-flow end-to-end QoS guarantees by reserving adequate resources in all the nodes along the path. While this architecture provides excellent QoS guarantees, it has significant scalability problems in the network core because of per-flow state maintenance and per-flow operation in routers. Because of scalability problem with Intserv/RSVP, the Internet Engineering Task Force (IETF) has proposed differentiated services (Diffserv) [12] as an alternative QoS architecture for network core data forwarding plane, and bandwidth broker (BB) [2] for control plane.

A. Differentiated Services (Diffserv)

Diffserv requires no per-flow admission control or signaling and, consequently, routers do not maintain any per-flow state or operation. Instead, routers merely implement a small number of classes named per hop behavior (PHB), each of which has particular scheduling and buffering mechanisms. A packet's PHB is identified with the Diffserv field (DSCP) assigned by the ingress router (IR).

To this end, Diffserv is relatively scalable with large network size because the number of states in core routers are independent of the network size. Thus, it is considered as the de facto standard for the next generation of the Internet. However, unlike the Intserv/RSVP, Diffserv only addresses forwarding/data plane functionality, whereas control plane functions still remain an open issue. Hence, Diffserv alone cannot provide end-to-end QoS guarantees. In fact, providing end-to-end QoS is not one of the goals of Diffserv architecture [13]. In particular, these limitations and open issues are the following.

1) As its name indicates, a PHB defines the forwarding behavior in a single node. Unlike Intserv/RSVP model, there is no QoS commitment for the traffic traversing multiple nodes or domains.
2) With the exception of expedited forwarding (EF) [14], all the PHBs that currently have been defined provide *qualitative* QoS guarantees. Hence, the requirements of real-time applications, which need *quantitative* bounds on specific QoS metrics, cannot be guaranteed even in a single node.
3) The lack of admission control: There is no admission control mechanism to ensure that the total incoming traffic to a node or domain does not exceed the resources for the corresponding PHBs.
4) Knowing that more than 90% of the traffic today traverses multiple domains [22], [28], there is a need for interdomain SLA negotiation for border-crossing traffic.

From the above issues, it is envisioned that Diffserv needs a control path mechanism to achieve end-to-end QoS guarantees. BB [2] is one of the strongest candidates for this.

B. Bandwidth Broker (BB)

The BB [2] is a central logical entity responsible for both intradomain and interdomain resource management in a

Manuscript received September 30, 2003; revised March 15, 2004. This work is supported in part by the National Science Foundation (NSF) Award NMI ANI-0123939.

H. A. Mantar is with the College of Engineering, Harran University, Urfa 63200, Turkey (e-mail: hamantar@harran.edu.tr).

J. Hwang is with Syracuse University, Syracuse, NY 13244 USA and also with Seoul National University, Seoul, Korea (e-mail: jshwang@syr.edu).

I. T. Okumus is with Mugla University, Mugla, Turkey (e-mail: okumus@mu.edu.tr).

S. J. Chapin is with Syracuse University, Syracuse, NY 13244 USA (e-mail: chapin@ecs.syr.edu).

Digital Object Identifier 10.1109/JSAC.2004.836010

Diffserv domain.[1] The goal of the BB is to provide *Intserv-type* end-to-end QoS guarantees in Diffserv-enabled networks. With such a centralized scheme, control functionality such as policy control, admission control, and resource reservation are decoupled from routers into the BB and, thus, a BB makes policy access and admission control decisions on behalf of its entire domain. The BB is also responsible for setting up and maintaining reservations with its neighboring BBs to assure QoS handling of its border-crossing traffic. The BB has several appealing aspects.

- By decoupling control path functions (e.g., signaling, link QoS states, admission control) from routers, a BB increases network core scalability.
- Because of the minimal changes required in network infrastructure, it increases the likelihood of QoS deployment.
- Simplifies accounting and billing associated with QoS.
- Minimizes the inconsistent QoS states faced by distributed approaches in which edge routers make admission control decision independent from each other.
- Interdomain level resource reservation and provisioning can be automated with the BB. It can perform sophisticated QoS provisioning, reservation and admission control algorithm to optimize network utilization in a *network-wide* fashion.

However, the BB model is still in its initial stage, and no substantial study has been done. Many scalability-related issues, which are the fundamental problems of any QoS model, remain unclear and, therefore, it is questioned by many researchers if this model will ever be widely deployed. Among many others, the following problems are related to the subject of this paper: 1) how to get dynamic network states; 2) how to assure quantitative QoS guarantees with no reservation in core routers; 3) how to obtain QoS and cost information about networks beyond its domain (e.g., which provider it should choose for border-crossing traffic); 4) how to manage domain resources in an efficient and scalable manner; and 5) how to communicate and reserve resources with a neighboring BB for border-crossing traffic.

C. Organization of This Paper

The rest of the paper is organized as follows. Section II presents the background and previous work. In Section III, we briefly describe the simple inter-BB signaling (SIBBS) that is used to evaluate the pipe model. In Section IV, we introduce our proposed architecture for inter-BB resource reservation and provisioning. Section V presents the analytical evaluation of the proposed model compared to the pipe model. In Section VI, we present the implementation and simulation results that validate our achievements. A summary of this paper and motivation of future work are given in Section VII.

II. Background and Problem Definition

Several studies have addressed *scalability* problems in providing QoS across single or multiple domains [1], [8], [19], [28], [29]. The common approach in these studies is the *pipe model*. In the pipe model, for each ingress-egress pair a pipe is established and all the traffic that share the same ingress and egress points is aggregated into the same pipe. By using utilization-based admission control at the ingress point of the pipe, the required QoS guarantees can be achieved between ingress and egress points. An interesting work called border gateway reservation protocol (BGRP) was proposed by Pan *et al.* [9]. By relying on BGP-4's aggregation scheme, BGRP significantly improves network scalability compared to RSVP. However, since BGRP does not relies on the BB, it cannot take the BB advantages described above. Furthermore, BGRP does not address the network resource utilization and business aspect of the Internet.

Khalil *et al.* [19] have used the BB for providing virtual private network (VPN) across multiple Diffserv domains. In [6], we have used the BB to establish label switching path (LSP) [10], which is another example of the pipe model, across multiple Diffserv domains. TEQUILA [29] and GlobalCenter [30] have used the BB for pipe-based QoS provisioning across a domain.

The SIBBS [1] protocol, which we developed as the Qbone Signaling Team, is another example of the pipe model. SIBBS is used for interdomain pipe setup and inter-BB communication in BB-supported Diffserv networks.

One of the common issues with these pipe schemes is that there is neither experimental nor analytical evaluation. We use our SIBBS implementation to evaluate the pipe models. Note that although these schemes are different from each other in some details, they have the same behavior in terms signaling and state scalability and resource utilization. Thus, the experimental and analytical results obtained for SIBBS throughout this paper will be very similar for all the above schemes. Because SIBBS uses the common pipe paradigm in terms of signaling and state scalability and resource utilization.

By aggregating individual reservations into an existing pipe, the pipe model can improve network scalability in terms of the signaling and state load and admission control time (compared with intserv/RSVP model). However, the application of the pipe model is limited to small-scale networks such as VPNs across a single domain [28]. It has the following problems, when it is applied to large-scale networks (e.g., the entire Internet).

- **State scalability**: The number of pipes in core transit domains scale with $O(n^2)$, where n is the number of domains in the Internet. Currently, there are approximately 13 500 domains and 130 000 networks [22]. This makes more than 10^8 domain-to-domain and more than 10^{10} network-to-network pipes, which are much higher numbers than a router can handle [9].
- **Signaling scalability**: Since the pipes are isolated from each other in transit domain, meaning that there is no aggregation among the pipes destined for the same domain, each pipe is provisioned separately. Thus, the number of inter-BB signaling messages is proportional to the number of pipes.
- **Statistical multiplexing gain**: Since the aggregation is only performed at source domains, the transit domains cannot take advantage of statistical gain across the pipes.

[1]Although the BB was originally proposed for Diffserv networks [2], it can also be applied to non-Diffserv networks. Because the BB is independent of the forwarding plane schemes.

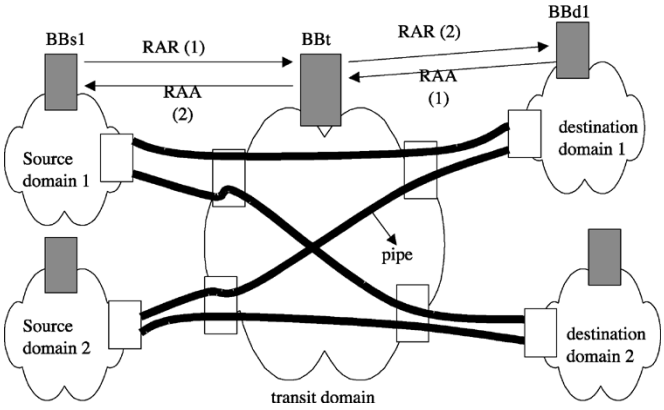


Fig. 1. SIBBS pipe setup steps.

By taking the above issues into account, this work presents a novel BB model to achieve *quantitative* QoS guarantees in multidomain Diffserv networks. Although the BB was originally proposed for Diffserv networks [2], it can be used with the other underlying technologies (non-Diffserv) such as asynchronous transfer mode (ATM) and RSVP. Because in the BB model each domain is free to choose its own intradomain resource management scheme and data forwarding plane scheme as long as its SLAs with neighboring domains are met.

III. Simple Interdomain Bandwidth Broker Signaling (SIBBS)

The SIBBS [1] was proposed as an interdomain QoS resource reservation protocol for the BB-supported Diffserv model. A BB uses SIBBS to establish pipe with other BBs for its border-crossing traffic. The source domain's BB preestablishes pipes to every other possible destination domain and then multiplexes all the reservation requests (initiated by end hosts) that have the same destination domain and QoS class into the same pipe.

A. Interdomain Pipe Setup

SIBBS is a simple query-response protocol. Common SIBBS messages are resource allocation requests (RARs), resource allocation answers (RAAs), cancel (CANCEL), and cancel acknowledgment (CANCEL ACK). The communication between BBs is handled via a long transmission control protocol (TCP) session.

Assuming that all the policy issues (such as SLA and SLS) are satisfied, we briefly describe pipe setup steps shown in Fig. 1. (See [1] for details.) Suppose the BB of source domain 1 wants to establish a pipe to destination domain 1 for a particular class, the procedure is as follows.

- The BB of source domain 1 (BBs1) builds an RAR message with appropriate parameters such as service ID, BW amount, duration, and the destination domain IP prefix, and then sends it to the transit (downstream) domain BB (BBt).
- Upon receiving the RAR message, BBt checks its intradomain resource availability and ingress-egress links' capacity by querying intradomain pipe database and border/edge links resource database, respectively. If both of these checks succeed, BBt builds an RAR message and sends it to BBd1.
- BBd1 checks its egress router (ER) link capacity. If sufficient capacity is available, it reserves bandwidth in its link, builds an RAA, and sends it to BBt.
- When BBt gets RAA, it reserves resources, builds RAA, and then sends to BBs1.
- When BBs1 receives the RAA, the tunnel establishment procedure ends.

Note that both pipes and reservations are unidirectional. As a typical pipe model approach, a pipe resources can only be used by the source domain. The intermediate domains cannot use it, meaning that the aggregation is done only at source domains. An important point is that the traffic conditioning in border routers is pipe-based. When a BB accepts the pipe setup request, it configures the corresponding border routers with the traffic parameters associated with the pipe for traffic conditioning. The conditioning is performed based on <*destination IP, source IP,* and *DSCP*>.

B. Dynamic Pipe Size Update

We extend SIBBS by adding pipe update scheme. Since a pipe is established between the ER of the source domain and the IR of the destination domain, and carries only the traffic of source domains, only the source domain's BB initiates the pipe resizing process.

The source BB dynamically estimates the traffic rate of each pipe. If there is a significant change in traffic rate compared with the pipe size, it signals the downstream BB to resize the pipe. Depending on the QoS class, rate estimation can be either parameter-based or measurement-based described in the next section.

C. Admission Control and Aggregation

At this point, it is assumed that pipes are preestablished and dynamically resized in advance. When a stub domain's BB receives a reservation request from an end host within its domain to another end host in a different domain, it simply checks the resource availability of the pipe that corresponds the request's *destID* and QoS class. If this test succeeds, the BB grants the reservation, otherwise it rejects the request. As we can see, although the destination is in another domain, the BB does not go beyond its domain, making admission control depend only on local knowledge.

Note that end-to-end QoS guarantees depends on the resource availability in end hosts (such as availability of a multimedia server) as well. Although, in this paper, we focus only on network resources, both SIBBS and our proposed protocol BBRP can work with any scheme that provides tools to identify available resources in end hosts. For example, in [32], we demonstrated how SIBBS can be used with Globus toolkit in a complementary fashion to provide end-to-end QoS. After identifying available servers, Globus uses SIBBS to check the network resources along the paths to the particular servers and to reserve the required network resources.

For simplicity, in this paper, we assume that the end hosts have sufficient resources to handle requests and the BB makes

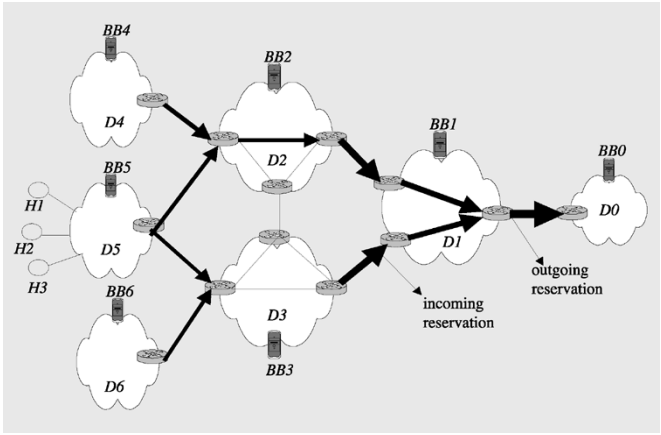


Fig. 2. Network example that consists of multiple Diffserv domains.

its decision based only on network resources. In reality, however, both network and end hosts resources need to be available in order to grant an end-to-end QoS guarantees.

IV. Scalable Inter-BB Resource Reservation and Provisioning Model

A. Network, Service, and Reservation Model

Network Model: Fig. 2 illustrates a network example that consists of multiple Diffserv domains. Following the current Internet structure, each domain manages its own network resources and establishes service agreements with its neighbors for its border-crossing traffic. A domain can have multiple adjacent domains numbering from just a few up to hundreds, each of which can be a potential customer and provider.

Knowing that *multilateral* SLAs, with which a domain need to have SLAs with all the domains along the path from the source to the destination, are too complex to be managed [1], [2], our model relies purely on *bilateral* SLAs, with which domains only need to establish relationships of limited trust with their adjacent peers. Each domain has only one BB and the resource negotiation between domains is handled solely by peer BBs. The end-to-end QoS connectivity is provided by concatenation of piece-to-piece bilateral commitments.

QoS Service Model: We define a limited set of network services supported by each domain. It is assumed that these services are globally well-known (GWK) and that each of them associates with a unique service ID, *servID*. Each service requires *quantitative* bounds on the packets' loss ratio and queueing delay in each node in turn across each domain.

We assume that each service is assigned to a certain share of link capacity and that each service can use only its share. The surplus capacity of a service can only be used by best-effort traffic. To have quantitative QoS guarantees in a scalable manner, we associate an upper delay bound d, and upper loss ratio bound l to each service (e.g., $d < 3$ ms, $l < 10^{-2}$ in a link). A service delay and loss ratio bounds across a domain are simply calculated in terms of the number of links along the path. It is also assumed that d and l are predetermined at network dimensioning (configuration) stage [1], [13], [29], which is done over long time intervals (e.g., days, weeks).

Quantitative QoS guarantees, of course, require explicit admission control to make sure that the traffic rate in a link does not exceed its capacity share. In our model, knowing the domain topology and state information and QoS constraints in each node, a BB performs *utilization-based* admission control to make sure that the utilization of a link never exceeds its capacity.

Interdomain Reservation Model: The resource negotiation between BBs is made, based on a particular *servID* and destination region. Depending on the granularity (scalability concern), the destination region can refer to a domain, a region that consists of multiple networks, or a single network or an end host. A destination region address (*destID*) is in the format of classless interdomain routing (CIDR) [31] (i.e., for $IPv4/X$, where $X < 32$). Reservations are classified as incoming or outgoing (Fig. 2).

An incoming reservation represents a commitment that a BB provisions to a particular customer (upstream domain or end host) for the traffic coming from that customer. For example, in Fig. 2, BB1 has two incoming reservations provisioned to BB2 and BB3 for the traffic destined for $D0$.

An outgoing reservation represents the resources (commitments) that a BB is provided by its provider (the downstream domain's BB) for its aggregated outgoing traffic. For example, in Fig. 2, BB1 is provided resources by BB0 for the traffic destined for $D0$. Multiple incoming reservations are multiplexed into the same outgoing reservation if their *destID* and *servID* are the same.

B. Architecture Overview

The core idea of this model is very similar to the notion of the *wholesale-retailer* paradigm in the sense that a BB reserves (buys) resources from its provider(s) for a particular destination region and QoS class and then grants (sells) these resources to its customers upon a request. For example, BB1 buys resources from BB0 for destination $D0$, and then sells to its customer BB2 and BB3. Unlike hierarchical BB schemes [11], [29], where each BB has a single provider for a destination, in our model, a BB can act as both customer and provider for the same destination. For example, BB2 can act both as customer and provider to BB3 for destination $D0$ (Fig. 2). The end-to-end QoS guarantees are achieved by concatenation of the piece-by-piece bilateral commitments between customer and provider.

The proposed model, in particular, consists of four key components: Inter-BB resource reservation and provisioning (BBRP) protocol, the dynamic provisioning algorithm (DPA), a BB routing information base (BB-RIB), and the routing setting controller (RSC). We assume that bilateral SLAs exist between neighboring BBs, and that initial resources are reserved during the startup time. The initial resources do not have to reflect the future traffic demand, because a BB dynamically adapts the reservation rate to the actual traffic demand during the time. In general, a BB has two operation phases: *resource reservation* and *resource provisioning*.

In the resource reservation phase, a BB modifies the outgoing reservation rate. The DPA dynamically monitors the actual traffic rate for each (<*destID, servID*>) and compares with the associated outgoing reservation rate. When there is a substantial change in the actual traffic rate (exceeds predefined

threshold points), the DPA triggers BBRP to make appropriate changes in the corresponding outgoing reservation. Based on the parameters (the required *BW* amount and *servID, destID*) received from DPA, the BBRP first queries BB-RIB to select an appropriate provider BB. As shown in Fig. 2, there might be multiple providers for a particular destination. The BBRP applies its own criteria (e.g., the least costly) to select the appropriate one. It then sends a reservation request to the selected BB. When the request has a positive outcome, it updates the outgoing database and invokes the routing setting controller (RSC) to configure the corresponding ER with the new reservation parameters.

In the resource provisioning phase, the BB performs admission control for incoming customer requests. When a customer request arrives, the BB can reject or accept it, based on resource availability in the corresponding outgoing reservation, without signaling the downstream domain's BB. Note that since the intradomain resource reservation is made in advance, the BB's task for intradomain admission control is limited to direct access to the intradomain database. An important point here is that all the incoming customer requests are multiplexed into the same outgoing reservation if they have the same <*destID, servID*>. This is one of the key features that make this scheme different from the pipe-based models.

The key point here is: a BB makes one single reservation with its provider for a particular <*destID, servID*> and this reservation is shared by all of its customers. Unlike traditional pipe models, an individual customer reservation is not visible beyond its provider domain (e.g., BB0 is not aware of the reservations that BB2 and BB3 made with BB1, rather, it simply knows what BB1 reserves). The reservations from BB4, BB5, and BB6 are aggregated toward the destination as they merge. A BB can handle its customer requests without further communication with its downstream domain. Assume that BB2 requests to increase its reservation rate with BB1 due to the high reservation requests that it receives from its customers (BB4 and BB5), while at the same time BB3 requests to decrease its reservation rate due to lack of requests. In this case, BB1 can simply grants BB2's request by assigning the resources released by BB3. BB1 does not have to negotiate with BB0 for individual requests that it receives from its customers (BB2 and BB3) as long as the total aggregated demand is between certain thresholds.

The key advantages of this model are the following.

- **Inter-BB signaling scalability**: It damps interdomain signaling frequency. Since a BB makes one reservation that is shared by all of its customers, the frequency of signaling messages exchanged with downstream BB depends on the change of aggregated demand rather than on individual demand.
- **BB-state scalability**: The number of reservation states that a BB needs to maintain is substantially reduced compared to traditional pipe-based models. The n^2 problem is reduced to nm, where m is the number of adjacent peers and n is the number of domains or networks in the Internet ($m \ll n$) [21], [22].
- **Data forwarding path scalability**: The number of state messages that a border router maintains for traffic conditioning and forwarding (i.e., classification and scheduling state) purposes is proportional to the number of destination regions. The typical n^2 problem with pipe-based approaches is reduced to n.
- **Multiplexing gain**: Due to aggregation, the outgoing reservation rate can be less than the sum of incoming reservation rates.
- **Efficient resource utilization**: There is no worst case provisioning. By using the inter-BB signaling protocol, a BB dynamically adjusts the reservation rate with respect to actual traffic rate.
- **Interdomain traffic engineering support**: The BB has the flexibility to select the appropriate downstream domain (BB) based on its outgoing link utilization or the resource availability that downstream BB provides. It can even send traffic with the same destination through multiple providers. For example, in Fig. 2, BB5 can choose either BB2 or BB3, or both of them as provider for its traffic destined for $D0$. Furthermore, the BB can select the next provider based on the price of the service.

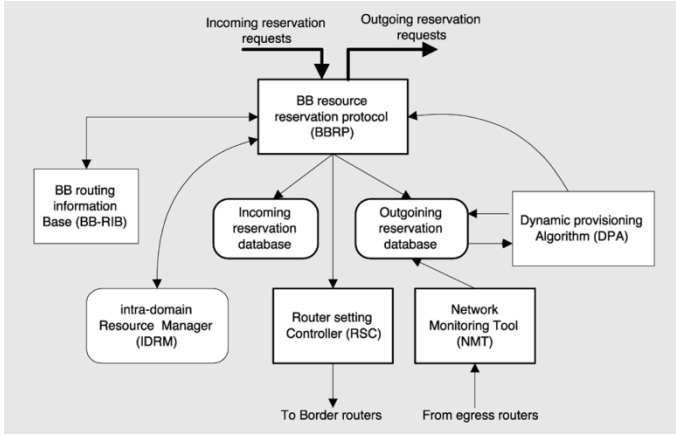


Fig. 3. Components of a BB for reservation and provisioning.

C. Inter-BB Resource Reservation and Provisioning Protocol (BBRP)

The inter-BB resource reservation and provisioning protocol (BBRP) originates from SIBBS [1] described in the previous section. The BBRP uses typical SIBBS messages named RAR, RAA, CANCEL, and CANCEL ACK. The BBRP has two operational phases: resource reservation performed upon receiving a message from DPA and resource provisioning performed upon receiving a reservation message from a customer (Figs. 3 and 4). These two events can be handled independently of each other at different times.

1) Resource Reservation Phase: In this phase, a BB acts as a customer requesting resources from the downstream domain for its outgoing traffic. This can be done either by making a new reservation or updating the existing one. Unlike pipe-based models, the outgoing reservation is not customer-specific, but rather reflects the aggregation of all the incoming customer reservations.

Upon receiving a request from the dynamic provisioning algorithm (DPA), the BBRP requests resources from its

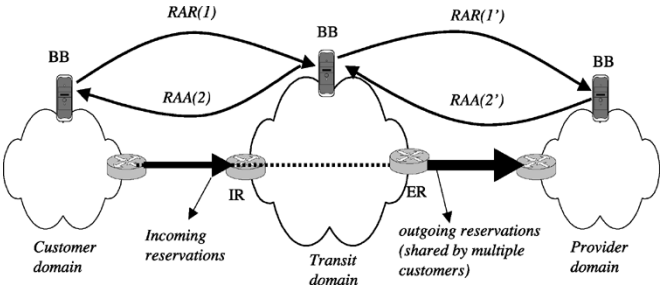


Fig. 4. Operation steps of BBRP (transit domain BB is chosen as reference); RAR (1) and RAA (2) represent messages received from and sent to customer, respectively; RAR (1') and RAA (2') represent messages sent to and received from provider, respectively.

provider(s). This is performed when a substantial change occurs in the instantaneous traffic rate compared to the outgoing reservation rate. The DPA dynamically monitors the outgoing reservation database, which is updated in online either by parameter-based or measurement-based rate estimation methods. When a modification is needed, it predicts the new reservation rate and invokes the BBRP to reserve the required resources. (This will be made clear in the next sections.) Upon receiving a modification message from the DPA, the BBRP performs the following steps.

Step 1) Builds an RAR based on the parameters (*servID, destID*, BW amount) received from DPA.
Step 2) Queries BB-RIB for appropriate next BB (for the given <*destID, servID*>).
Step 3) Sends RAR to the selected BB and waits for RAA.
Step 4) If RAA has positive outcome, updates outgoing reservation rate, and configures the corresponding border router's traffic conditioning parameters.
Step 5) If RAA has negative outcome, it tries the next possible BB if there is any.

To reduce the frequency of inter-BB signaling messages and to minimize interdomain admission control time, the outgoing reservation rate is usually chosen higher than the actual traffic rate with some thresholds. As shown in Fig. 2, a BB may have multiple providers for the same destination. In this case, the BB sends the RAR message to the first BB, which is associated with <*servID, destID*>, in the BB-RIB (Step 2). If the selected BB does not have enough resources, it queries the BB-RIB again for another possible candidate (Step 5). This continues until a positive RAA is received. An important point here is that the new reservation does not have to use the same provider as the existing one. Thus, there might be multiple outgoing reservations for the same <*destID, servID*>, each handled via different providers.

2) Resource Provisioning Phase: In this phase, the BB acts as provider and determines if the requested resources can be granted. It checks the outgoing reservation rate to find out if there are sufficient available resources to handle the new request. Unlike pipe-based models such as original SIBBS, it does not have to signal the downstream BB for individual customer request.

Upon receiving an RAR from any of its customers, the BBRP:

1) Determines IR and ER by querying BB-RIB; the IR associates with the sender BB, while the ER associates with <*servID, destID*>.
2) Checks ingress link resource availability from interdomain link database.
3) Checks intradomain resource availability by querying intradomain resource manager (IDRM).
4) Assuming that (2) and (3) have positive outcomes, checks if the corresponding outgoing reservation has sufficient resources, by accessing the outgoing reservation database, *Query(servID, destID, BW)*; depending on the outcome, the result may be one of the following.

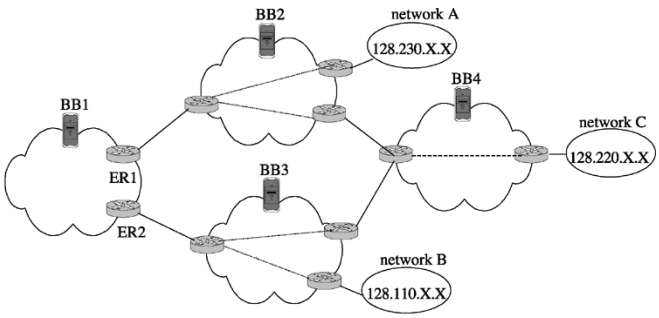
Case 1) Outcome is positive:
Build an RAA message and send it to the customer, then update the corresponding ingress link resources, outgoing reservation rate, and traffic conditioning parameters in IR.
Case 2) (4) Outcome is negative:
Send message (including *destID, servID* and BW amount) to DPA, which predicts the corresponding outgoing reservation rate and returns it to BBRP. BBRP negotiates with provider for new resources by sending RAR; if it gets a positive RAR message, the same task is performed as in Case 1); if negative RAA is received, it is sent to the negative RAA customer.

Since an outgoing reservation task is performed in advance, when an RAR is received the BB can handle it without further communication with downstream BB. Although, in some special cases, the resource availability (thresholds) in outgoing reservation may not be sufficient to accommodate incoming RAR, resulting in further communication with the downstream domain [Case 2) in Step 4)], this is expected to happen rarely.

The protocol operation presented above is basically for RARs that indicate either a new reservation or an increase in the existing reservation rate. To decrease or cancel an existing reservation, the operation is much simpler. The BB just modifies the outgoing and incoming reservation databases and the traffic conditioning parameters in the border routers.

D. BB Routing

A BB can use BGP-4 [21] to determine the next BB (provider) with which to communicate for its border-crossing traffic [1], [2], [4], [7]. However, BGP-4 does not provide any QoS information. While the path provided by BGP-4 may not support the required QoS or may not have sufficient resources, some alternative paths may exist. Also, one of the main assumption in previous routing schemes is that whenever a better route is found, the existing reservations are shifted to the new path. Knowing that a reservation has a certain time duration, this may not be a realistic assumption. This is especially true in interdomain, where domains (BBs) buy and sell QoS resources to each other for a particular time duration [7], [11], [22]. Furthermore, this assumption is the one that causes route flapping, which is one of the main concern of any QoS routing. Thus, our main assumption for the BB routing is that a BB uses QoS routing information only when it is needed to make a new reservation.



(a)

destID	servID	egress router	neighbor BB	cost
128.230/16	service 1	ER1	BB1	X
128.220/16	service 1	ER1	BB1	X
128.220/16	service 1	ER2	BB2	X
128.210/16	service 1	ER2	BB2	X

(b)

Fig. 5. (a) Simple multidomain Diffserv network. (b) BB1's BB-RIB.

We use a simple and scalable BB routing scheme. Each BB has a neighboring BBs' routing information base (BB-RIB) that gives not only the set of destination regions that can be reached through those domains (BBs), but also the supported QoS classes and their associated costs, as well as SLS rules for using these services.

Consider Fig. 5(a) where each domain and each network are represented by a unique IP address prefix and there is only one QoS class. Fig. 5(b) shows the BB1's BB-RIB. When BB1 needs to make an interdomain reservation, it consults BB1-RIB to determine the next (provider) BB and the associated ER. As shown in the figures, BB1 has two possible providers (BB2 and BB3) for the requests destined for network C (128.220.X.X). In this case, it can choose the provider based on its cost and resource availability.

To maintain BB-RIB, we use a lightweight scheme, which is similar to the idea of distance vector protocol used in BGP-4 [21]. Each BB floods its routing information to its neighboring BBs. Routing information includes <*servID, destID, service cost*>. Upon receiving an update message, the neighboring BBs update their BB-RIB and flood it to their neighbors. This may continue from destination domains upto source domains. An important issue here is the frequency of the update messages, which is a common problem in QoS routing schemes. Due to the nature of interdomain SLA and SLS, this is expected to be done within a medium time scale such as minutes, hours, and days. Furthermore, an update in one BB may not affect all the upstream domains.

The format of BB-RIB is: <*destID, servID, nextBB,cost*>. If there is more than one candidate BB for a <*destID, servID*>, the candidates are sorted in the order of increasing cost. It is important to note that BB-RIB updates do not affect *existing* reservations, but only *future* reservation requests. This is one of the key point that damps the frequency of route flapping.

E. Class-Based Traffic Rate Estimation and Admission Control

In order to perform admission control and dynamic provisioning and reservation properly, it is essential to have the accurate instantaneous traffic rate of outgoing reservations. That is, the traffic rate for each <*destID, servID*> tuple needs to be estimated.

As described in [16] and [24], for the services that require deterministic QoS guarantees, a peak-rate-based approach (parameter-based) is used. In peak rate-based approach, a reservation rate is updated with the request's peak rate, independent of whether the source transmits continuously at peak rate or not. The peak rate-based rate estimation is very simple because only the knowledge of peak rate is required. However, this method is not feasible for statistical services, which can tolerate limited delay and loss, due to poor resource utilization.

For services that require statistical QoS guarantees, either effective BW (parameter-based) methods [5], [24] or measurement-based methods [16] can be used. Our experimental results have shown that the choice of method for statistical services not only depends on service-type but also the location of traffic (whether it is in a stub domain or transit domain) [32].

1) Stub Domains: In stub domains, most of the reservation requests are made by individual end hosts. The numbers of end hosts are relatively large (e.g., thousands), and each end host may have different applications and the traffic characteristics may vary with application-type (e.g., peak rate, mean, and variance may be different for each application's traffic). In such a heterogeneous environment, effective BW approaches may not be feasible because of the complexity in estimating traffic rate [17] and the lack of statistical multiplexing gain. Hence, in stub domains, we use measurement-based approach.

In measurement-based models, traffic samples are collected at regular small time intervals called sampling period S during the duration of a measurement window. Previous measurement-based schemes [16], [24] do not take QoS constraints (i.e., delay and loss bound) into account in traffic rate estimation. Hence, they may not be feasible for Diffserv networks, because the rate estimation of the same traffic samples varies with *class-type* (because of different the delay and loss constraints associated with each class-type).

The objective of our class-based rate estimation is: Given a class QoS constraints (i.e., delay and loss bound), estimate the traffic rate of each outgoing reservation based on real-time measurement statistics. By assuming that a large number of reservations are being aggregated into the same queue, we use a Gaussian (normal) approximation model under the conditions of the central limit theorem (CLT) [15]. The CLT states that the aggregation rate approaches a Gaussian distribution when the number of aggregated reservations is large [11], [15]. For a moderate number of aggregated reservations, one may employ other candidate distributions such as [33].

There are several advantages of using Gaussian model. First, from the given definition, it is seen that the Gaussian distribution becomes a more realistic model for the Diffserv network to estimate the traffic rate of a class because of the coarse granularity of the aggregation. The individual reservations' traffic rate fluctuations are smoothed due to aggregation. Second, the

traffic can simply be characterized by mean and cumulative variance alone. Thus, the Gaussian model is computationally simple compared with other traffic models. Third, unlike previous measurement-based schemes [16], [24], the rate can be estimated based on QoS metrics.

Let us donate m_i as the mean traffic rate of sampling period (S) i, N as the number of samples in a window W, m as the mean traffic rate of a W, $(m = (1/N)\sum_{i=1}^{N} m_i)$, and σ^2 as the average variances of N samples $(\sigma^2 = (1/(N-1))\sum_{i=1}^{N}(m_i - m)^2)$. Let R and $R(t)$ represent the estimated and instantaneous traffic rate, respectively. To meet a class loss ratio constraint l, the following probability condition must be held:

$$\Pr\left(R(t) > R\right) \leq l \tag{1}$$

(1) can be solved with the well-known Gaussian approximation [15]

$$Q\left(\frac{R-m}{\sigma}\right) \leq l \tag{2}$$

where $Q(x) = (1/\sqrt{2\pi})\int_x^{\infty} \exp^{-y^2/2}\, dy$. Taking inverse transform, (2) can be rewritten as

$$R \geq m + Q^{-1}(l)\sigma. \tag{3}$$

Equation (3) computes the link traffic rate with respect to l. As seen, by changing the l, the estimated rate can vary. The multiplier $Q^{-1}(l)$ controls the estimation rate to meet the constraint l. (The values of $Q^{-1}(l)$ are obtained from well-known Gaussian probability table.)

Another important factor is the buffer effect, in turn the length of S. This controls the sensitivity of the rate measurement according to the delay constraint. While a small S can be more sensitive to bursts, it results in poor resource utilization due to overestimation. On the other hand, a large S makes traffic smoother and, therefore, allows more aggressive resource utilization, but it degrades QoS performance because of a longer packet delay under burst conditions. Since, in our model, delay is one of the constraints, the value of S is set based on the class's delay bound d. That is, $S = d - \triangle d$; $\triangle$ is a cushion to prevent delay violation. By setting S based on d, the traffic rate is estimated with respect to d.

The ER dynamically measures and estimates the traffic for a particular outgoing reservation as (<*servID, destID*>), and then sends it to the BB when a substantial change occurs in the traffic rate. The BB performs admission control and resource provisioning and reservation based on the results received from ER. With admission of a new reservation request, the QoS metrics' bound (i.e., l and d) violation will not occur if

$$m + Q^{-1}(l)\sigma + r^{\text{new}} \leq C \tag{4}$$

r^{new} represents the reservation rate of new request. Equation (4) ensures that the packet loss ratio over a certain time interval is less than l as long as the BB performs the admission control based on this condition.

2) Transit Domain: The traffic characteristics in transit domains might be significantly different from those in stub domains. In transit domains, most requests are made by BB of neighboring domains, rather than by individual hosts [1], [4]. As described before, a customer BB makes its outgoing reservation rate more than its current incoming demand in order to accommodate near-future requests and to allow short-lived traffic rate fluctuations. To some extent, this implies an *advance* reservation. From this perspective, the measurement-based scheme may not be feasible for providing reliable QoS commitments for guaranteed services, because the measurement results are obtained based solely on the instantaneous traffic rate.

Two typical problems of a parameter-based approach are the lack of statistical multiplexing and the complexity of characterizing heterogeneous requests into a single model. After examining these problems more closely, it is evident that in transit domains the effect of these problems are relatively very small compared to those in stub domains [3], [24] for the following reasons. First, the requests are mostly from peer BBs that use aggregated-type reservations, which implies that the requester has already aggregated multiple requests and exploited the advantages of statistical multiplexing. As the number of hosts in an aggregation increase, the required BW tends to the mean rate [5]. Second, the heterogeneity of the original traffic sources is minimized in the aggregation. Thus, the BB can simply estimate the traffic rate for a given class based on the requests' BW requirements.

F. Dynamic Provisioning Algorithm (DPA)

The DPA determines when a BB should modify the outgoing reservation rate and how to modify it (i.e., how much to reduce, how much to increase, and when). Two essential issues that need to be considered are inter-BB signaling scalability and efficient resource utilization.

The DPA modifies the reservation rate according to a simple threshold-based scheme. The following are the parameters used throughout this section.

- R_{out} Outgoing reservation rate.
- R_{cur} Instantaneous outgoing traffic rate.
- HT High threshold $(HT = R_{\text{out}} * h)$. This is the utilization level where the BB is triggered to increase the outgoing reservation rate.
- LT Low threshold, $(LT = R_{\text{out}} * l,\ 0 < l < h < 1)$. This is the utilization level where the BB is triggered to reduce the outgoing reservation rate.
- OR Operation region $(OR = HT - LT)$.

Fig. 6 illustrates the operation of the DPA. It is assumed that the instantaneous traffic rate for each outgoing reservation (<*servID, destID*>) is estimated and the BB outgoing reservation database is updated accordingly (described in previous section). The algorithm dynamically checks if the current traffic rate is within the OR. As long as the traffic rate fluctuates within the OR, no negotiation takes place. Once the traffic rate crosses the OR boundaries (HT, LT), the algorithm predicts the OR width, and then triggers the BBRP to negotiate resources with the provider BB.

Here, the width of OR is critical in terms of the tradeoff between resource utilization and the frequency of BBRP invocation. To maintain the balance, the OR width is chosen by taking previous, current, and future traffic demand into account. For simplicity, similar to the mechanism used by Jacobson for estimating TCP round-trip time [27], the DPA uses the first order of

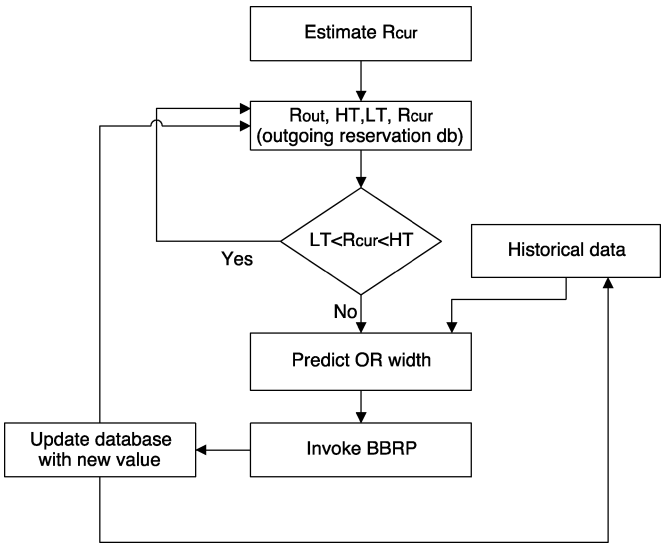


Fig. 6. Functional operation of DPA.

autoregressive integrated moving average (ARIMA). The idea here is to make OR width adaptive to traffic characteristics.

Let us define $t_n, t_{n-1}, \ldots, t_1$ as the time of negotiations, and T as the expected negotiation time interval. As mentioned earlier, a change can only be made when the traffic rate crosses any of the OR boundaries. Thus, the negotiation might be performed before or after T. Assume that at $t = t_n$ a change is needed, and T_{cur} and T_{prev} are defined as

$$T_{\text{cur}} = t_n - t_{n-1}, \quad T_{\text{prev}} = t_{n-1} - t_{n-2}.$$

T_{next} is predicted as

$$T_{\text{next}} = \lambda T_{\text{cur}} + (1-\lambda) T_{\text{prev}} \tag{5}$$

where $0 < \lambda < 1$ (e.g., $\lambda = 0.2$). OR can be adjusted as follows:

$$OR = \frac{T}{T_{\text{next}}} OR \tag{6}$$

$$R_{\text{out}} = R_{\text{cur}} + \left(\frac{OR}{2}\right) R_{\text{cur}}. \tag{7}$$

If the last two values of period (T_{cur}, T_{prev}) are equal to T, there will be no change on the OR width. If T_{cur} is shorter than T_{prev}, in order to avoid scalability caused by the high frequency of inter-BB signaling the OR is increased. If T_{cur} is longer than T_{prev}, meaning that the traffic rate is changing slowly, the OR is decreased to increase resource utilization.

The blocking, or rejecting of reservations during the renegotiation time between BBs, is prevented by introducing a cushion $(R_{\text{out}} - HT)$. Since the inter-BB renegotiation process is relatively long [1], [5], [7], once the traffic rate reaches HT, the BBRP attempts to increase the reservation rate. By the time new resources are reserved, the incoming reservation requests can be accepted, because there will still be some available resources.

V. Scalability Analysis

In this section, we analyze our model in terms of control and data/forwarding path scalability. For simplicity of analysis, Fig. 7 is chosen as a reference network, where $BB_1, BB2_2 \ldots BB_n$ represent source stub domains, $BB_{n+1}, BB_{n+2} \ldots B_{2n}$ represent destination stub domains, and BB_a, BB_b, BB_c represent transit-only domains. Since the network core has the highest scalability concern (for most QoS schemes), the analysis results are obtained based on the BB_b. Although others are possible, BBRP enhancements are compared with the pipe-based models, including SIBBS [1], extended-RSVP [8], and VPN [19], which span multiple domains. Even though these models are quite different in details, they have similar manners in terms of control and data forwarding path scalability. Thus, we simply use the term "pipe model" for all of them. It is assumed that both pipes and reservations (in BBRP) are made at domain level granularity. It is also assumed that the network has only one QoS class in addition to best-effort class.

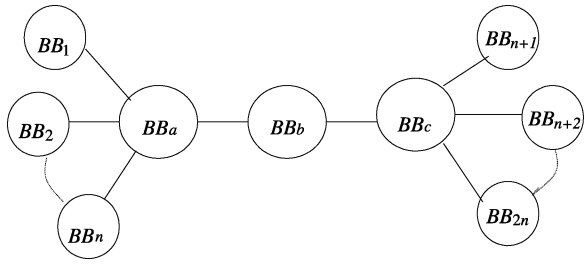


Fig. 7. Reference network for scalability analysis.

A. BB State Scalability

Typically, a BB maintains states for each reservation in its database. In the pipe model, a pipe is identified by source and destination region address, and a BB keeps state for a pipe based on the <*srcID, destID*> tuple. In BBRP, a BB has two different databases, an incoming reservation database and an outgoing reservation database. Both of these databases keep state per *destID*.

Let us assume that the rate of a pipe setup request for a destination domain is Poisson distributed with λ, and the pipe duration is exponentially distributed with a mean of $1/\mu$. Based on these assumptions, the average number of pipe states in BB_b for a particular destination domain is λ/μ (according to the Little theorem [18]). There are n destination stub domains and, therefore, the total number of states is $n\lambda/\mu$. For BBRP, there is only one state for a particular destination, regardless of the number of incoming reservations. That is, if there is at least one request for a destination, there will be a state for that. The probability of there being at least one reservation request for a domain is $1 - \exp^{-\lambda/\mu}$. Since in the reference network BB_b has only one upstream domain, the numbers of incoming and outgoing reservation states will be the same. The average number of states is $2n(1 - \exp^{-\lambda/\mu})$, which includes both incoming and outgoing reservations.

Comparing two schemes, the gain is $(\lambda/\mu)/2(1 - \exp^{-\lambda/\mu})$. When each domain establishes a pipe to every other domain in the network (e.g., SIBBS core tunneling), the λ/μ approaches n. That is, the number of states in the pipe model becomes n^2, while in BBRP it is $2n(1 - \exp^{-n})$. When n gets larger, the gain approaches n.

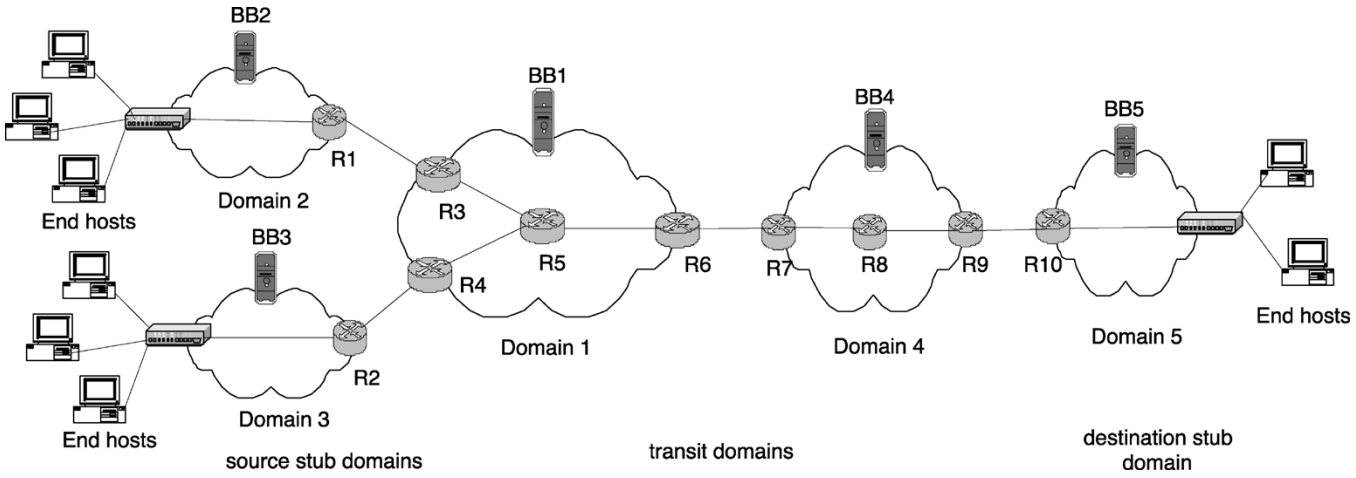


Fig. 8. Simple experimental topology.

In the above analysis, the BB_b has only one upstream domain (one customer). In reality, however, there might be multiple customer domains. So, the number of incoming reservation states will be higher (in BBRP). Assume that BB_b has m upstream domains and all of them make reservations to each destination (worst case condition). Thus, there will be mn incoming reservation states. By taking this into account, the above results will be $O(n(m+1))$ for BBRP and $O(n^2)$ for the pipe model. In the Internet today, while m can be from just a few to hundreds, n is more than 10 000. Furthermore, when pipes are made between networks, n is more than 100 000 [22], while m is still the same.

B. Inter-BB Signaling Scalability

The inter-BB signaling messages include a new reservation setup and updating and canceling of an existing reservation.

By multiplexing all the customer requests destined for the same destination into the same outgoing reservation and performing reservation setups and updates based on the aggregated traffic rate, the BBRP significantly reduces the number of signaling messages. Let us consider the worst case situation in Fig. 7, meaning that every source domain has a reservation to every destination domain. As explained above, there will be n^2 pipes in the pipe model and n outgoing reservations in BBRP. Since in the pipe model each pipe is isolated from the others and, therefore, a signaling message for a pipe needs to be processed by all the BB along the path from source to destination, the number of signaling messages in BB_b will be proportional to the number of pipes. Thus, the scalability problem for the pipe model is $n^2O(1)$, while it is $nO(1)$ for BBRP.

If we assume that the traffic in each pipe is statistically identical, and independently distributed, due to multiplexing gain in BBRP the scalability will be further reduced to $nO(1/\sqrt{n})$ (obtained based on normal distribution under the assumption of the central limit theorem [15]).

C. Forwarding Path Scalability

Forwarding path analysis includes the number of states that ingress (border/edge) routers keep for traffic conditioning. Since the traffic conditioning is performed per-packet-based, the number of reservation state is very critical in terms of routers' performance. This is because in the data forwarding path the packets need to be processed at line speed.

In BBRP, the traffic conditioning is performed per-reservation-based. Since a reservation is identified solely by *destID*, the scalability is $O(n)$, where n is the number of possible destination regions. In the pipe model, because a pipe is identified by <*srcID, destID*> tuple, each pipe needs to be conditioned separately. This makes the scalability of pipe model $O(n^2)$. (These results are obtained in the same way with those in BB state scalability.)

D. Resource Utilization

Assume that the traffic in each pipe is statistically identical, and independently distributed with average rate m and variance σ^2. As described in Section IV, under the assumption of Gaussian model, the reservation rate for a pipe will be $R^{pipe} = m + Q^{-1}(l)\sigma$.

In BBRP, where the pipes are aggregated, the aggregate mean rate m_a and variance r_a are Nm and $N\sigma^2$, respectively. The aggregated reservation rate will be $R_a = Nm + Q^{-1}(l)\sqrt{N}\sigma$. The equivalent reservation rate for a pipe is $R^{bbrp} = m + Q^{-1}(l)(\sigma/\sqrt{N})$. Compared to the pipe model, the BBRP gain is $\sqrt{N}$.

As shown, the equivalent reservation rate of a pipe approaches to the mean rate as the aggregation granularity increases.

VI. Evaluation

In this section, we evaluated inter-BB signaling scalability, resource utilization, and QoS assurance in our prototype BB implementation test bed. To evaluate BB and border/edge router state scalability, a large-scale network is needed. Unfortunately, it is very difficult to build such a network in a lab environment. Thus, we evaluated these features with a simulation scenario.

The model was evaluated in a simple topology as depicted in Fig. 8. We configured domains 2 and 3 as source stub domains, domains 1 and 4 as a transit-only domains, and domain 4 as a destination stub domain. There were 30 source end hosts and two destination end hosts. There were ten routers. Each router was a Pentium III with 997-MHz PCs running Linux-2.4.7 as the operating system. These Linux PCs were configured

to act as routers with Diffserv functionalities, which involved installing *iproute2* [25], and reconfiguring kernel to enable QoS and networking options. Traffic conditioning specifications such as class-based-queueing (CBQ) parameters, token bucket parameters, and performance parameters such as drop probability, delay, and throughput were configured based on the service requirements.

End hosts were Pentium PCs running Windows 2000 and Linux-2.4.7. Stub networks were implemented on separate VLANs on an ALCATEL 5010 and 6024 switches with ten base-T connections. Each end host had a traffic generator (TG) tool [23] that was used to generate UDP traffic. TG had a provision to generate traffic with DSCP and with different traffic distribution patterns.

All the links in the network had 10-Mb/s capacity and unidirectional (from the source domains to the destination domain). The routers in stub domains (R1, R2, and R10) acted as both ingress and egress. In R1 and R2, per-flow traffic conditioning was applied. In R3, R4, R7, and R10, per-destination-based (aggregated) traffic conditioning was applied.

The performance evaluation of signaling scalability and resource utilization depends heavily on traffic behavior. Unfortunately, it is difficult to represent the current Internet traffic behavior with any of the existing traffic models [20]. Therefore, it was important to use the traffic traces collected from real networks for our experiments. We chose the traffic traces provided by CAIDA [26], which has advanced tools to collect and analyze data based on source and destination address (domain level) and traffic-type for short time intervals (every 5 min). Trace data was gathered for 150 min on February 21, 2003. From more than 100 different destination ASes, we chose the top one (AS2641), which has the highest traffic rate, to represent our destination domain (domain 5). Similarly, from more than 100 source ASes, we chose the top five domains (AS7377, AS1909, AS1668, AS1227, and AS33), which sent traffic to the selected destination domain, to represent the source domains in our test bed.

To generate traffic according to the traced data, we normalized the traced data rate and mapped it to our test bed, meaning that the aggregated traffic behavior of each source domains changed according to the traced data characteristics during the experiment time. All the experiments were run for 20 min. The duration of a reservation was exponentially distributed with a mean of 1 min (which corresponds to 5 min in traced data). The reservation rates varied over the duration of the experiment, based on the rate of profile generated from the traced data.

Note that we compared our model with the pipe model. Since the SIBBS implementation is the only implemented interdomain pipe model, the comparison was done via SIBBS. However, to some extent, the SIBBS results obtained in this work represent the pipe model. Because the SIBBS features described in this paper are the same with the basic pipe paradigm. Thus, we did not need to implement any other pipe scheme as far as this paper concerns.

A. Signaling Scalability

In this section, we show how BBRP can reduce the number of inter-BB signaling messages. The OR boundaries LT and HT were set to 99% and 80%, respectively. The traffic rate estimation for each outgoing reservation was performed with the measurement-based method. The measurement window size W and the factor $Q^{-1}(l)$ were set to 10 and 2 s, respectively.

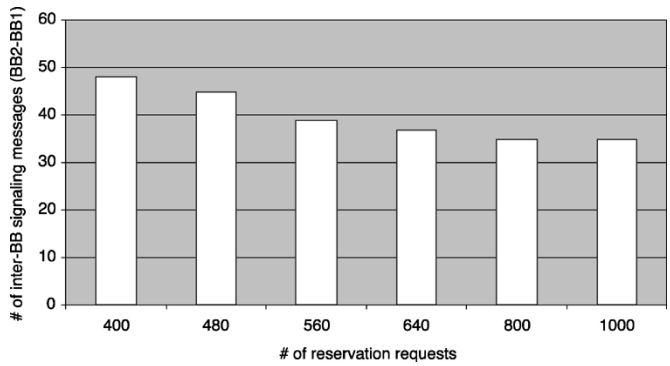


Fig. 9. Number of signaling messages between BB2 and BB1.

Fig. 9 shows the cumulated number of signaling messages between BB2 and BB1 with respect to the number of reservation requests that BB2 received during the experiment time. As expected, the number of signaling messages (initiated by BB2 to adjust outgoing reservation rate) between BB1 and BB2 is relatively small compared with the number of reservation requests that BB2 receives. The figure also shows that inter-BB signaling scalability can get even better as the number of requests increases. If we consider the steady-state, where the average number of reservations in the network remain the same, a BB may not initiate any signaling message while handling a large number of end host requests.

In Fig. 9, we have shown the behavior of a source stub domain BB that handles only the requests from its domain. Although aggregating reservation in stub domains can significantly improve BB signaling scalability compared to per-request-based reservation schemes such as RSVP, performing aggregation only at stub domains (source domains), a typical pipe-based approach, may still not be sufficient to keep the number of signaling messages in transit domains to a scalable amount. Thus, in the next step, we examined the signaling scalability in transit domain BBs, which is one of the main concerns of BBRP.

In SIBBS, the transit domain needs to process each pipe message separately. In other words, all the individual pipes setup or modification messages need to be processed by all the BBs along the path, from source to destination BB. Thus, the number of signaling messages in a transit domain BB is proportional to the number of pipes that use that domain. In our test bed (Fig. 8), BB1 has two pipes for the traffic destined for domain 5, one for BB2 (domain 2), and one for BB3 (domain 3). For the BBRP case, BB1 aggregates the reservation of BB2 and BB3 for destination domain 5 and then makes only one reservation with BB4. Thus, it needs to signal BB4 for only one aggregated reservation.

Fig. 10 shows the signaling messages that BB1 processes for SIBBS and BBRP. In the case of SIBBS, BB1 forward every signaling message received from BB2 and BB3 to BB4. Thus, the number of signaling messages between BB1 and BB4 is the sum of the messages received from BB2 and BB3, plus the number of messages that BB1 forward to BB4. In the case of BBRP, BB1 receives same amount of messages as in SIBBS, but it does not have to forward each message to BB4. Instead,

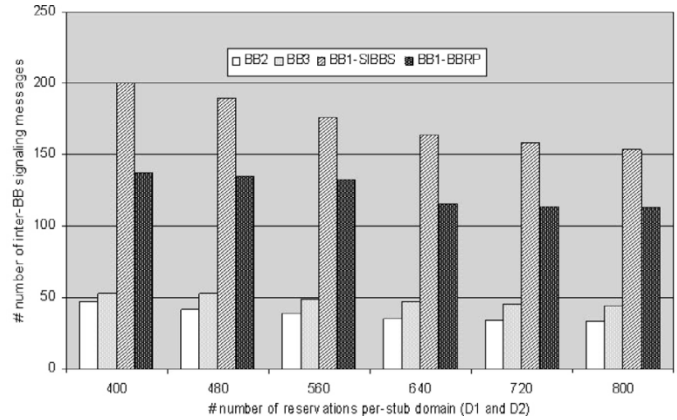


Fig. 10. Effect of BBRP in transit domain (BB1).

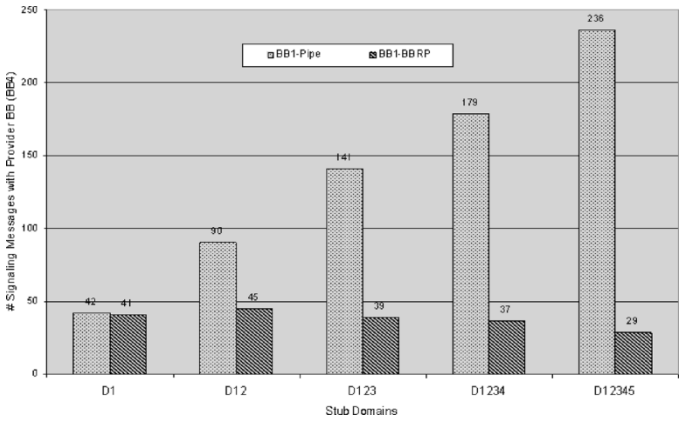


Fig. 11. BB5 signaling load for BBRP and SIBBS.

it forward one aggregate message. Thus, as can be seen even in this simple topology, BBRP significantly reduces the signaling load in transit domain.

The BBRP signaling gain increases as the length of the AS path increases. When a path spans only two ASes, the gains of BBRP and SIBBS are the same. However, when it spans more than two AS hops, the BBRP significantly outperforms SIBBS. Knowing that the average AS level path length in the Internet today is around 4.9 [22], [26], the usage of BBRP aggregation scheme becomes more important. To show these enhancements, we added three more source stub domains (AS1668, AS1227, and AS33) to domain 1 in our basic test bed (Fig. 8), and evaluated the BBRP gain compared with SIBBS.

Fig. 11 depicts the number of signaling messages that BB4 receives for BBRP and SIBBS. As expected, in the case of SIBBS, the signaling load increases as the number of stub domains increase. This is because BB5 gets signaling messages for each pipe, which are owned by a particular stub domain. On the other hand, the signaling load changes very slowly (relatively) when BBRP is used. A comparison of Figs. 10 and 11 show that the BBRP gain increases as the path length (the number of AS hops) increases. It is important to note that due to heterogeneous traffic characteristics, the signaling load for BBRP may not always decrease as the aggregation level increase. For example, in Fig. 11, the signaling load of the aggregated two domains is more than the load for a single domain (for this particular experiment). On

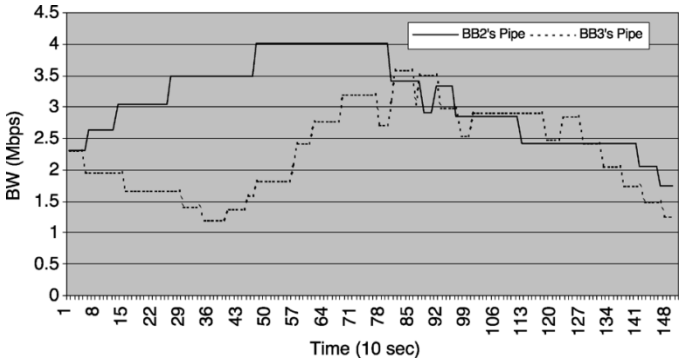


Fig. 12. BB1 resource reservation with BB4, using SIBBS.

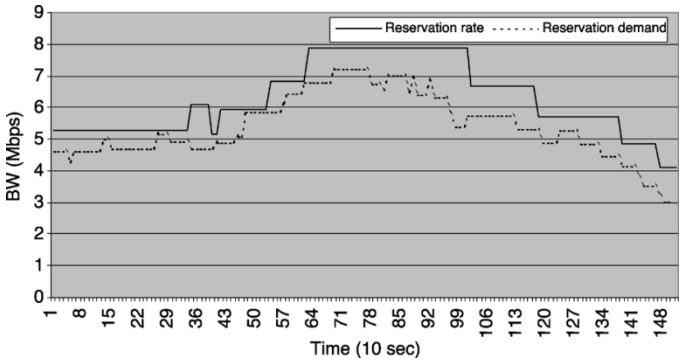


Fig. 13. BB1 resource reservation using BBRP.

the other hand, the load decreases as more than two stub domains are in the network. Depending upon variable traffic characteristics (the traffic rate fluctuation in long time-scale) in each domain, the BBRP gain may vary. However, as the figure illustrates, this variation is practically very small compared to SIBBS. At least, the load is not increased proportionally to the number of stub domains.

B. Resource Utilization Versus Signaling Scalability

In this section, we investigate the tradeoff between resource utilization and signaling scalability.

Figs. 12 and 13 show BB1 resource reservation for its customers' (BB2 and BB3) traffic destined for domain 5. The experiment was performed for a deterministic service (EF) by using the parameter-based admission control scheme. The OR boundaries LT and HT were set to 80% and 99%, respectively.

Fig. 12 depicts BB1 the resource negotiation with BB4 for the pipes of its customers (BB2 and BB3) using SIBBS. As shown, each pipe is resized independently, meaning that the resources reserved for one pipe cannot be used by the other. Another important point here is that BB1 resizes the pipes based only on the requests received from customer domains. In this sense, there is no over provisioning in transit domain and, therefore, BB1 can achieve maximum resource utilization. However, as mentioned above, on demand-based resource reservation results in a high inter-BB signaling load.

Fig. 13 shows BB1 resource reservation using BBRP. The BB1 makes its reservation rate always more than the actual reservation demand (the aggregated demand from BB2 and BB3) in order to damp the future short-lived traffic fluctuations without signaling BB4 for each individual request. As shown,

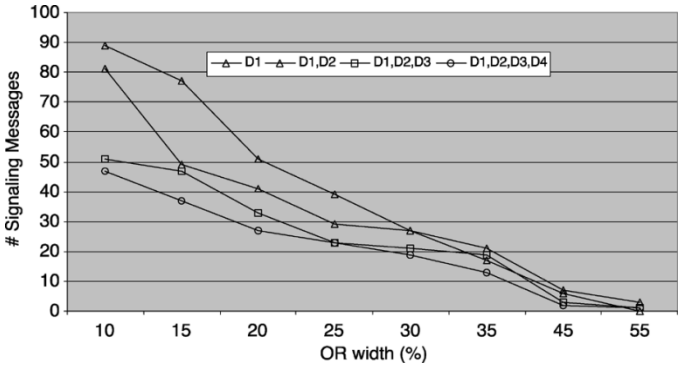


Fig. 14. Effect of OR width and traffic load on signaling.

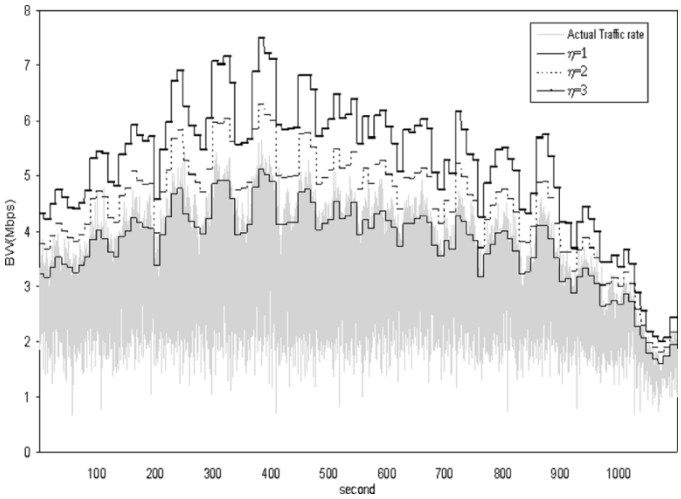


Fig. 15. Class-based rate estimation ($\eta = Q^{-1}(l)$).

the frequency of reservation rate adjustment is very small compared to SIBBS (Figs. 12 and 13). However, this gain comes with an overprovisioning tradeoff. While BBRP reduces the signaling load by 52%, it wasted resources by 12.5%.

Fig. 14 shows the signaling load can be changed with the traffic rate. The experiment was performed separately for 1, 2, 3, and 4 stub domains connected to domain 1. By adding more stub domains, we increased the reservation demand on BB1, which in turn increased the traffic rate. As the figure clearly shows, the signaling load reduced significantly as the traffic rate increased because the aggregated traffic rate becomes smoother as the aggregation granularity increases (due to the statistical multiplexing gain). By taking this feature into account, the DPA adjusts OR width based on the traffic characteristics in order to increase resource utilization. It increases the OR width when the traffic is bursty and decreases when the traffic is smooth.

C. Traffic Rate Estimation and Admission Control

One of the main concerns of all measurement-based schemes is QoS assurance. By using normal approximation, our measurement scheme measures the traffic rate based on the given service-specific statistical QoS constraints such as loss ratio so that the given constraint will not be violated. Fig. 15 shows how the estimated traffic rate of the same traffic samples varied with class loss ratio (l) constraint ($\eta = Q^{-1}(l)$). There are three classes, each of which has different l, class 1 ($\eta = 1$), class 2

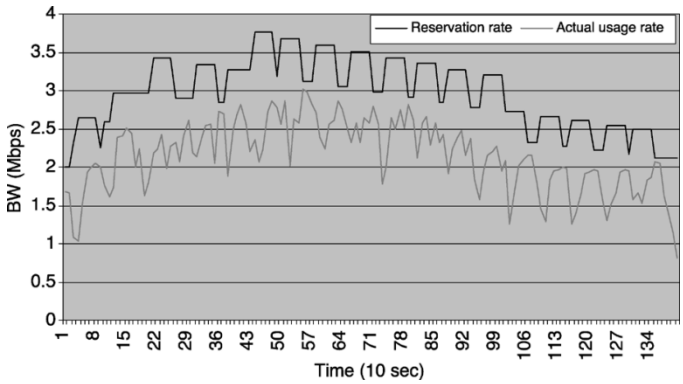


Fig. 16. BB2 measurement-based resource reservation.

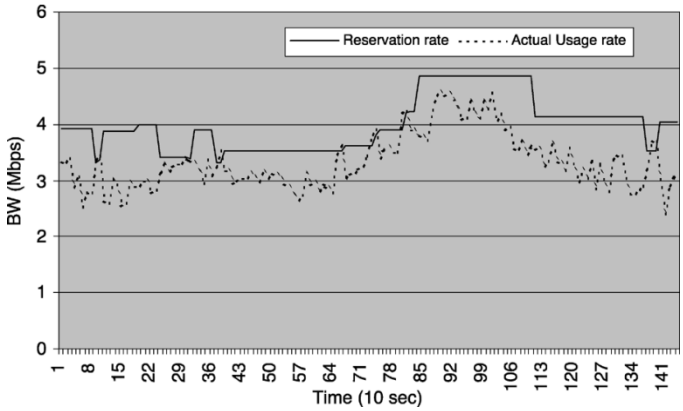


Fig. 17. BBI measurement-based resource reservation.

($\eta = 2$), and class 3 ($\eta = 3$). While the lower η (higher tolerable loss ratio) achieves higher resource utilization, it substantially degrades the QoS by dropping some packets. On the other hand, the higher η increases the QoS, but results in overestimation. This highlights the importance of the traffic rate estimation scheme. Since the traffic rate is estimated based on l, the rate is neither too conservative nor too aggressive.

Since the measurement results are obtained based on instantaneous traffic rate, the QoS constraint might be violated in the following cases (in transit domain): 1) a customer may not send the traffic immediately after its request is granted and 2) a customer may over-reserve by taking near future requests into account.

Figs. 16 and 17 present the measurement-based admission control results for source stub domain (domain 2) and transit domain (domain 3). As shown in Fig. 16, the actual usage rate (obtained by measurement) is always less than the reservation rate, meaning that the QoS commitments given to the customer are not violated. This is because all the end hosts make immediate reservations, and no overprovisioning occurs from the end host's point of view. In reality, since most stub domain customers are end hosts, and the number of end hosts is relatively large, using a measurement-based approach may not have a significant negative effect on QoS assurance.

One of the disadvantages of measurement-based schemes is that the actual reservation rate of the customer is not taken into account. Although this may not be a problem in stub domains where most customers make immediate reservations, it may result in a serious QoS violation in transit domains. As depicted in Fig. 17, the actual traffic rate exceeded the reservation rate

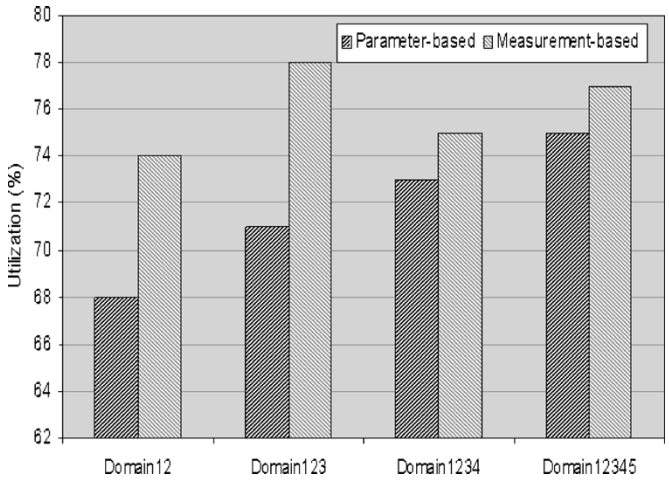


Fig. 18. Measurement based versus parameter-based in transit domain.

several times, meaning that the QoS commitments given to BB2 and BB3 were violated. This happened because both BB1 and BB3 can have 19% unused reserved resources (overprovisioning) and use these resources whenever needed without signaling BB1. This is one of the main reasons that we do not use the measurement-based scheme in transit domains.

One of the main motivations for using a measurement-based scheme is to take advantage of statistical multiplexing. To analyze the effect of statistical multiplexing gain, we applied measurement-based scheme in BB2 and BB3, and the parameter-based scheme in BB1. BB2, and BB3 made no overprovisioning. The results in Fig. 18 shows that the statistical multiplexing in transit domain is very small compared with the one in stub domains because the traffic entering the transit domain is already aggregated. The traffic rate gets closer to the mean rate as the level of aggregation increases.

D. BB and Border Router State Scalability

The evaluation of state scalability in the BB and in border routers requires a large-scale reference network. Due to the limitations of our test bed, we simulated state scalability in a dumbbell reference network (Fig. 7). It was assumed that a pipe is established when at least one end host makes a reservation and is canceled when there is no reservation request from any host.

The reservation arrival request from an individual host (to its BB) for a particular destination domain was Poisson distributed with an arrival rate of 0.000 01/s and the reservation duration was exponentially distributed with a mean of 100 s. The number of clients located in each stub domain varied from 5 to 200 during the simulation period.

Fig. 19 shows the number of states in BB_b for SIBBS and BBRP with respect to the number of hosts in each source stub domain. As shown in the figure, BBRP significantly reduces the number of states. When every source domain has traffic being sent to every destination domain, the number of states approaches 10 000, and 100 for SIBBS and BBRP, respectively. The fact of $n^2 \rightarrow n$ can be easily seen from the figure.

For traffic conditioning purposes, we changed our reference network by adding nine more border routers to transit domain b. The incoming traffic is equally distributed among these routers,

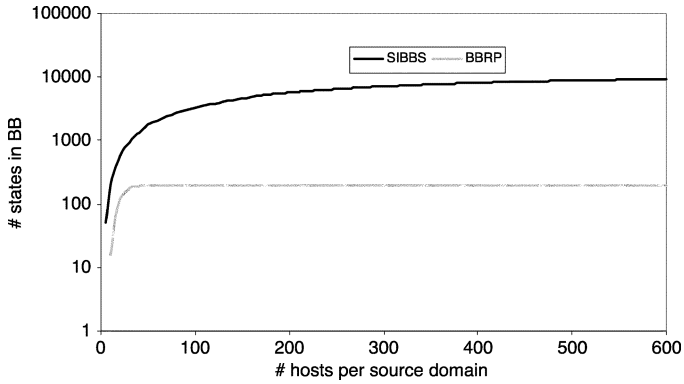


Fig. 19. BB state scalability (BB_b).

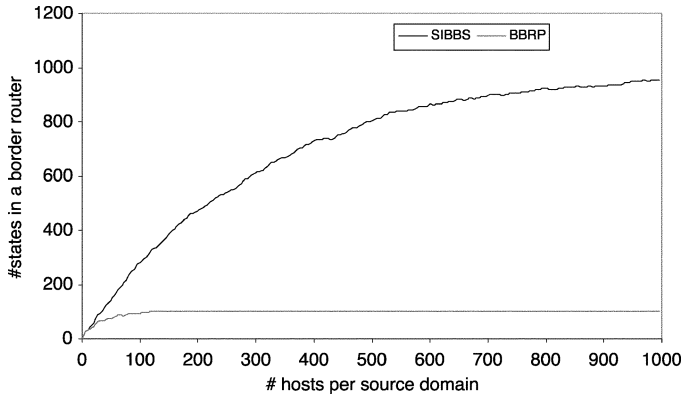


Fig. 20. Border router state scalability in a core transit domain (in the IR of BB_b).

with each router receiving traffic from ten different source domains). Fig. 20 depicts state scalability in a border router. As can be seen, the BBRP improves border router state scalability. From Figs. 19 and 20, we can see that while the number of states increase exponentially with the number of source domains for SIBBS, it stays almost constant for BBRP.

VII. Conclusion and Future Work

In this paper, we presented the design, implementation, and analysis of a scalable BB resource management scheme. We developed and implemented a scalable inter-BB resource reservation and provisioning protocol (BBRP). By maintaining reservation state based only on destination region, the BBRP significantly reduces the BB and border routers state scalability problems. It also minimizes inter-BB signaling scalability by using an aggregated-type resource reservation and provisioning. Another key novelty of BBRP is that it uses a lightweight BB routing scheme for selecting next BB (provider) instead of BGP-4. Furthermore, it supports a multiprovider scenario, where multiple providers can be used for the same destination region. This, to some extent, implies interdomain traffic engineering.

Our design, implementation, and experimental results show the important role that bandwidth brokers can play for service providers. The experimental results show that an ISP can substantially improve its resource utilization, thereby increasing its

revenue, while requiring minimal changes in underlying infrastructure. The scalability and utilization results give the basic guidelines that an ISP may consider in defining services and SLAs.

In the future, we will investigate a dynamic interdomain SLS negotiation scheme that is used for exchanging service usage rule and pricing. In this paper, we relied on GWK services, however, it is clear that there is a potential need for other services, so we will enhance our model to support additional services. An important point with centralized models is reliability. We are currently implementing active replication scheme to increase the BB reliability.

REFERENCES

[1] QBone Signaling Design Team. Simple interdomain bandwidth broker signaling (SIBBS). Work in Progress. [Online]. Available: http://qbone.internet2.edu/bb/

[2] K. Nichols, V. Jacobson, and L. Zhang, "A two-bit differentiated services architecture for the Internet," RFC 2638, July 1999.

[3] H. A. Mantar, "Scalable resource management framework for QoS-enabled multidomain networks," Ph.D. dissertation, Syracuse Univ., Syracuse, NY, Aug. 2003.

[4] H. Mantar, J. Hwang, I. Okumus, and S. Chapin, "Interdomain reservation via third agent," in *Proc. SCI 2001/ISAS 2001*, Orlando, FL, July 2001, pp. 561–567.

[5] H. Mantar, J. Hwang, S. Chapin, and I. Okumus, "A scalable and efficient interdomain QoS routing architecture for DiffServ networks," in *Proc. IEEE IM2003*, Mar. 2003, pp. 463–467.

[6] I. T. Okumus, J. Hwang, H. A. Mantar, and S. J. Chapin, "Interdomain LSP setup using bandwidth management points," in *Proc. GLOBECOMM*, San Antonio, TX, 2001, pp. 7–11.

[7] A. Terzis, L. Wang, J. Ogawa, and L. Zhang, "A two-tier resource management model for the Internet," in *Proc. Global Internet 99*, Rio de Janeireo, Brazil, Dec. 1999, pp. 1779–1791.

[8] F. Baker, C. Iturralde, F. Faucheur, and B. Davie, "Aggregation of RSVP for IPv4 and IPv6 reservations," RFC 3175.

[9] P. Pan, E. Hahne, and H. Schulzrinne, "BGRP: A tree-based aggregation protocol for interdomain reservations," *J. Commun. Networks*, vol. 2, no. 2, pp. 157–167, June 2000.

[10] E. Rosen, A. Viswanathan, and R. Callon, "Multiprotocol label switching architecture," IETF, RFC 3031, 2001.

[11] C.-N. Chuah, "A scalable framework for IP-network resource provisioning through aggregation and hierarchical control," Ph.D. dissertation, Univ. California at Berkeley, Berkeley, CA, 2001.

[12] S. Black *et al.*, An architecture for differentiated services, RFC 2475, Dec. 1998.

[13] K. Nichols and B. Carpenter, "Definition of differentiated services per domain behaviors and rules for their specification," RFC 3086.

[14] V. Jacobson, K. Nichols, and K. Poduri, "An expedited forwarding PHB," RFC 2598, June 1999.

[15] A. Papoulis, *Probability, Random Variables, and Stochastic Process*, 3rd ed. New York: McGraw-Hill, 1991.

[16] S. Jamin, P. B. Danzig, S. J. Shenker, and L. Zhang, "A measurement-based admission control algorithm for integrated services packet networks," *ACM/IEEE Trans. Networking*, vol. 5, pp. 56–70, Feb. 1997.

[17] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high speed networks," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 1091–1100, 1995.

[18] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1992.

[19] I. Khalil and T. Braun, "Edge provisioning and fairness in VPN-DiffServ networks," *J. Network Syst. Manage.*, vol. 10, no. 1, pp. 357–373, Mar. 2002.

[20] V. Paxson and S. Floyd, "Wide-area traffic: The failure of Poisson modeling," *IEEE/ACM Trans. Networking*, vol. 3, pp. 226–244, June 1995.

[21] Y. Rekhter and T. Li, "A border gateway protocol 4 (BGP-4)," RFC 1771, Mar. 1995.

[22] G. Huston, "Analyzing the Internet's BGP routing table," *Internet Protocol J.*, vol. 4, no. 1, pp. 2–15, 2001.

[23] Traffic Generator Software. [Online]. Available: http://www.postel.org/tg/

[24] G. Knightly and J. Qiu, "Measurement-based admission control with aggregate traffic envelopes," in *Proc. IEEE IWDC'98*, Italy, Sept. 1998, pp. 199–210.

[25] [Online]. Available: http://snafu.freedom.org/linux2.2/

[26] [Online]. Available: http://www.caida.org/dynamic/analysis/workload/sdnap/

[27] V. Jacobson, "Avoidance and control," in *Proc. ACM SIGCOMM*, 1998, pp. 314–329.

[28] T. Li and Y. Rekhter, "A provider architecture for differentiated services and traffic engineering," RFC2490.

[29] P. Trimintzios *et al.*, "An architectural framework for providing QoS in IP differentiated services networks," in *Proc. IEEE IM*, Seattle, WA, 2001, pp. 17–34.

[30] X. Xiao, A. Hannan, B. Bailey, S. Carter, and L. M. Ni, "Traffic engineering with MPLS in the Internet," *IEEE Network Mag.*, vol. 14, pp. 28–33, Mar. 2000.

[31] V. Fuller, T. Li, J. Yu, and K. Varadhan, "Classless interdomain routing (CIDR)," RFC 1519.

[32] J. Hwang, S. Chapin, H. Mantar, and I. Okumus, "An implementation study of a dynamic interdomain bandwidth management platform in DiffServ networks," in *Proc. IEEE IM*, 2004, pp. 321–334.

[33] F. Alagoz, "Approximations on the aggregated MPEG traffic and their impact on admission control," *J. Elec. Eng.*, vol. 10, no. 1, pp. 73–84, Mar. 2002.

Haci A. Mantar (S'99–M'02) received the B.S. degree in electronics and telecommunication engineering from Istanbul Technical University, Istanbul, Turkey, in 1993, and the M.S. and Ph.D. degrees in electrical engineering and computer science from Syracuse University, Syracuse, NY, in 1998 and 2003, respectively.

From 1997 to 2000, he worked as an Instructor at Syracuse University. From 2000 and 2003, he was supported by a Graduate Students Research Fellowship from National Science Foundation. He is currently with the College of Engineering, Harran University, Urfa, Turkey, as an Assistant Professor. His research interests include wide area network management, QoS, grid computing, routing, IP telephony, and network security.

Junseok Hwang (S'96–M'00) received the B.S. degree in mathematics from Yonsei University, Seoul, Korea, in 1989, the M.S. degree from the Interdisciplinary Telecommunications Program (ITP), University of Colorado, Boulder, in 1996, and the Ph.D. degree in information science and telecommunications from the University of Pittsburgh, Pittsburgh, PA, in 2000.

From 1989 to 1994, he worked as a Computer Systems Researcher, Electronic R&D Center of Hyosung Group, Seoul, Korea. In 2001, he was with Hewlett Packard Laboratory, Bristol, U.K. as an Invited Researcher of Internet bandwidth management. In 2000, he was with the School of Information Studies, Syracuse University, Syracuse, NY, as an Assistant Professor. In 2003, he joined Techno-Economics and Policy Program, Seoul National University, Seoul, Korea, where he is currently an Assistant Professor in the College of Engineering. His research interests include international telecommunication policy, network economics, electronic commerce, design, performance, and economic analysis of real-time (voice) integrated network services, Internet telephony, QoS and bandwidth management, grid computing and networks, and resource management issues of next-generation Internet.

Ibrahim T. Okumus (S'98–M'03) received the B.S. degree in electronics and telecommunication engineering from Istanbul Technical University, Istanbul, Turkey, and the Ph.D. degree in electrical engineering and computer science from Syracuse University, Syracuse, NY, in 1995 and 2003, respectively.

He is currently with Mugla University, Mugla, Turkey. His research interests include wide area network management, QoS routing, and MPLS.

Steve J. Chapin received a dual B.S. in computer science and mathematics from Heidelberg College, Tiffin, OH, and the M.S. and Ph.D. degrees in computer science from Purdue University, West Lafayette, IN.

He is an Associate Professor in the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY, where he has served as Director of the Systems Assurance Institute. Prior to joining Syracuse, he served on the faculties of the University of Virginia, Charlottesville, and Kent State University, Kent, OH. His research interests are operating and distributed systems and computer security.