INTERNET-DRAFT

A Two-Tier Resource Management Model for Differentiated Services Networks

Francis Reichmeyer, Nortel Networks
Lyndon Ong, Nortel Networks
Andreas Terzis & Lixia Zhang, UCLA
Raj Yavatkar, Intel

November, 1998

Abstract

This draft proposes a two-tier resource management model for differentiated services networks. Following the approach taken by the Internet routing architecture, we propose that bilateral service agreements are made for aggregate border-crossing traffic between neighboring administrative domains. We also propose that administrative domains individually make their own decision on strategies and protocols to use for internal resource management and QoS support, both to meet internal client needs and to fulfill external commitments.

We sketch out one specific realization of this two-tier model by having a Bandwidth-Broker (BB) as the resource manager for each domain and a BB-to-BB protocol, equivalent to BGP in routing, for inter-domain resource management. We believe that this two-tier resource management model matches the direction of, and complement the work by, the diff serv effort. We also expect this two-tier model to scale well in the global, heterogeneous Internet .

Note: This draft contains pictures that could not be included in the text version. A postscript version of the draft (including the pictures) can be found at http://irl.cs.ucla.edu/publications.f.html

1 Introduction: a High-Level Model of QoS Control

The ultimate goal of network QoS support is to provide users and applications with high quality data delivery services. From a router's view point, however, QoS support is made of three basic parts: defining packet treatment classes, specifying the amount of resources for each class, and sorting all incoming packets into their corresponding classes. Over a year-long effort the IETF Differentiated Services Working Group is reaching agreement on initial definitions on "per-hop behaviors" (PHB), a set of differentiated packet treatments. At each router IP a packet is treated in a specific way based on the TOS field value (called the "codepoint") carried in the IP header of the packet. Diffserv effort addresses both the first and third issues above: it specifies traffic classes as well as provides a simple packet classification mechanism – routers easily sort packets into their corresponding treatment classes by the TOS value, without having to know which flows or what types of applications the packets belong to.

As work on diff serv progresses in the IETF, there has been a continued discussion on the second issue, that is whether differentiated services would need any signaling protocols for dynamic resource management. A commonly perceived notion is that manually configured resource allocations at network boundaries should be able to provide us a jump start in differentiated services deployment, offering preferential treatment to some packets relative to others. However, many people also expressed concerns on how to achieve high quality delivery services from end to end using the differentiated services model.

We believe that end-to-end performance can be met through the concatenation of PHB's along packet delivery paths. We also believe that certain automatic protocol mechanisms will be needed in near future to assure that adequate amounts of resources for each PHB class, in order to meet the ultimate goal of satisfying users and applications performance requirements effectively and efficiently. In the remaining of this document we propose a hierarchical approach to scalable bandwidth allocation support for the global Internet.

1.1 A Picture of the Internet Today

The Internet today is made of the interconnection of multiple autonomous networks called autonomous systems, or administrative domains, each under a separate administrative control. This is illustrated in Figure 1, where the differently shaded regions represent different administrative domains. Each domain contracts its neighboring domain(s) for data delivery service; the neighbor domain, in turn, may pass the

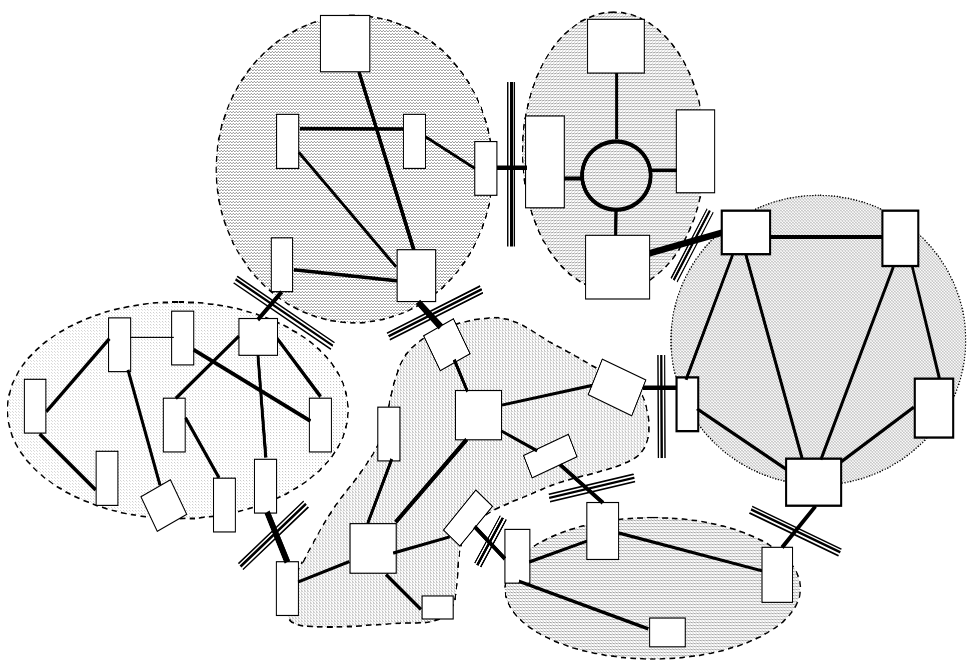

Figure 1: The Internet Today

traffic to next neighbors, so on and so forth until packets are delivered to final destinations. For example, a campus contracts one ISP (or a few for redundancy) to deliver its traffic; the ISP delivers the campus'
traffic either directly if the destinations are connected to the same ISP, or otherwise passes the packets to other ISPs for further forwarding.

Following the administrative-domain based network topology, today's Internet routing architecture is a two-level hierarchical design. Each of the administrative domains, or

Autonomous Systems (AS), is free to choose whatever routing protocol it deems proper to run. To assure global connectivity, neighbor domains speak BGP (Border Gateway Protocol) with each other to exchange network reachability information. Reachability information can be aggregated, for example, if nearby networks share common prefixes. Their reachability reports are merged so that a remote site will keep only one entry in its forwarding table showing the common prefix. The separation of the Internal Gateway Protocols (IGPs) and the Border Gateway Protocol (BGP), coupled with the ability to aggregate reachability information, provides the global routing with proven flexibility and scaling characteristics.

We make a few observations from the above picture. First and foremost, to get its data delivered, each domain makes a bilateral agreement with each of its directly connected neighbor domains, rather than multi-lateral agreement with each of all ISPs along the paths to all possible destinations. That is, the campus contracts one or a few ISPs for its data delivery services to *all* destinations. The local ISP in turn contracts its neighboring ISPs for delivery to those destinations that it does not directly connect to. Such concatenation of hop-by-hop forwarding through transit ISPs results in global IP delivery service.

Secondly, each individual domain makes simple delivery commitments externally, while it retains freedom in choosing its own routing approach internally. One may choose a preferred IGP from multiple candidates, such as OSPF, RIP or manual router table configuration. One's choice of IGP does not impact routing function between domains. By keeping inter-domain and intra-domain routing independent, the system allows routing to scale, and still to be easily administered and to provide flexible granularity of control within each administrative domain.

Thirdly, forwarding entries to *all* destinations are *pre-computed*, based on routing protocol message processing, rather than being computed in real time upon packet arrival. In addition, the pre-computed routing database is also dynamically adjusted to account for changes in topology or policy. The separation of routing computation and packet forwarding allows a network being up and operating while its routing protocol continues to evolve, and allows routing adjustments to be made on time scales independent from individual flow duration, providing system stability.

1.2 A Framework for Scalable QoS Support

Following the development of the global routing architecture, we suggest that individual administrative domains be the basic control unit for resource management. Bilateral service level agreements (SLA), expressed in diff serv terms, are made between neighboring administrative domains regarding the aggregate border-crossing traffic. Meanwhile, each administrative domain individually makes its own decision on strategies and protocols to use for internal QoS support to meet client needs and to fulfill external commitments.

With this two-tier hierarchical approach, end-to-end QoS support can be achieved through a concatenation of inter- and intra-domain resource allocations, as indicated in Figure 2, as long as those allocations match the level of the aggregated demand.

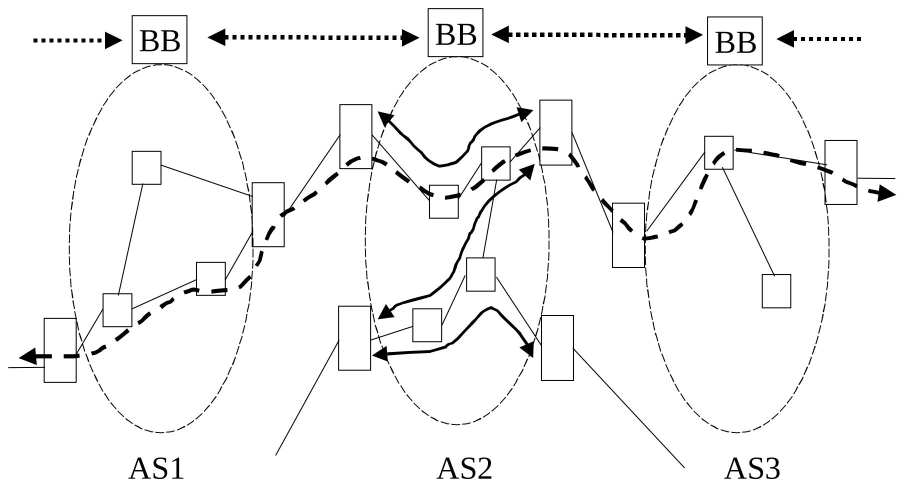

Figure2: End-to-end QoS through concatenated QoS Resource Management

We assume that a resource manager, named the Bandwidth Broker (BB) by Van Jacobson of LBNL, exists in each administrative domain. A BB will be in charge of both the internal affairs and external relations regarding resource management and traffic control. Internally, a BB may keep track of QoS requests from individual users and applications, as necessary, and allocate internal resources according to the domain's specific resource usage policies. Those policies specify which users may use how much resource or resource shares (and perhaps also under what specific conditions). The internal resource allocation can be done in a number of ways. For bandwidth-rich domains, for example, perhaps little needs to be done other than closely monitoring the network utilization level and re-provisioning accordingly. On the other hand, for bandwidth-poor domains, or those domains with either high variation in link capacities or high variation in traffic load, the BB may need to use some internal signaling protocol, such as RSVP, to reserve bandwidth for individual applications [e2e].

Externally, a BB will be responsible for setting up and maintaining bilateral service agreements with the BBs of neighbor domains to assure QoS handling of its border-crossing data traffic. The dotted arrows in Figure 2 show this relation. These agreements can be achieved, and in fact are currently achieved, through human communication between network managers of neighbor domains. The SLAs between domains will be in terms of differentiated traffic classes. A BB collects from internal users/applications the requests for external resources, and make its SLA arrangement based on these aggregate requests; it may also readjust the SLA according to the changing demand and conditions.

The BB for a transit domain (i.e. a provider network) must also keep those external service commitments to be within its internal resources capacity. The solid arrows within Autonomous System AS2, in Figure 2, represent intra-domain signaling; here signaling is used to allocate resources between ingress and egress points of the domain. Individual BBs instruct their own border routers how much traffic each border router should export and import for each PHB class.

The two levels of resource management must be coordinated in order for the network to provide the appropriate end-to-end QoS to quantitative applications. For example, regardless of how resource management is done within an individual domain in Figure 2, to concatenate the intra-domain resource commitment in each domain at borders, the amount of resources committed for each PHB class between neighboring domains must be consistent in order to provide quantitative end-to-end performance for host applications.

There remain a number of challenges in realizing this proposed two-level resource management. For example, the BB-to-BB communications must be secure, robust, and scalable. To scale well it is desirable that BB-to-BB resource requests be destination-independent, that is one domain tells its neighbor domain how much bandwidth should be reserved for premium traffic, without having to enlisting all possible destination domains. We are also yet to understand what is the best way to implement BB. The BB is a logical entity; actual implementations may take either a centralized or a distributed approach, or a combination of both.

2 Inter-Domain Bandwidth Management

Inter-domain resource management is concerned with provisioning and allocating resources at network boundaries between two domains. Typically, the two domains are separately owned and administered, for example two neighboring ISP networks or an ISP and an enterprise network. In the case of an ISP-ISP boundary, the two providers are usually customers of each other, each providing to each other, packet forwarding services over its transit network. In the case of an ISP-enterprise boundary, the ISP provides transit network services to the enterprise customer. A bilateral service-level agreement (SLA) specifying the amount and types of traffic each side agrees to send and/or receive must be established on the boundary between two domains.

For best-effort service, an SLA might specify the amount of traffic a network can reasonably handle from a customer, usually based on the capacity of the connecting link, and possibly some “guarantee”. The network is then provisioned in order to accommodate the aggregate traffic expected from its customers. When customers are added, or existing customers re-negotiate for more traffic, more bandwidth is added to the network.

When differentiated service is provided, the SLA specifies a profile for the traffic that is to receive a particular service and the ingress and egress border routers provision resources for the PHB(s) employed to provide the service(s). The SLAs may be communicated between the participating networks in a number of ways, for example via a phone call, e-mail exchange, or “automatically” via an inter-domain resource management protocol. In this section we discuss inter-domain resource management related to differentiated services and describe some likely properties of such a protocol. Besides being needed to allocate resources on the ingress and egress boundary devices, the information contained in SLAs may also be used to allocate resources within a domain. Intra-domain resource management is discussed in Section 3 of this document.

For initial diff serv deployment, SLAs negotiation is expected to occur relatively infrequently and network resources may be statically provisioned based on expected SLAs. For example, an ISP network might be provisioned such that it can support 10% of the bandwidth on its border links with an enterprise network for diff serv “Premium” traffic. An SLA is then established with the enterprise customer, to use some or all of this Premium capacity and the border routers are configured with traffic conditioners to police, shape, and mark the data packets as appropriate, based on the bilateral agreement. This provisioning might be sufficient for several months as the enterprise customer grows or deploys more applications requiring the differentiated service.

When the customer does need more Premium resources, the network is re-provisioned to support the additional traffic and the SLA is re-negotiated but, again, for initial diff serv deployment, this is not expected to occur frequently.

As diff serv is more widely deployed it can be envisioned that bilateral agreements between domains will be dynamically negotiated, for example to request certain services which are more conducive to a "pay-per-usage" model. An example of such a service is IP-telephony where the diff serv provider may allow customers to signal for the resources at the time they are needed as opposed to "statically" allocating (and paying for) the resources as described above. Thus, inter-domain resource management must account for varying temporal granularity with which SLAs are re-negotiated, and how this affects the requirements on network provisioning.

In addition to temporal granularity, diff serv providers might also wish to support SLAs for traffic at different flow-level granularities. For example they may specify aggregate flows based on the various service classes offered and classified by DS-byte marking, or they may specify microflows based on individual users or applications that originate the traffic [dsarch]. Support for qualitative QoS applications can be provided with SLAs for aggregate flows, while quantitative applications that require tighter “guarantees” from the diff serv network will require SLAs for finer flow-level granularity [e2e]. The latter may be provided across enterprise-ISP boundaries but, typically, will not be supported between two diff serv ISP networks. Also, either may require dynamic inter-domain signaling and admission control from the diff serv network, i.e. dynamic SLA negotiation as described above.

Dynamic inter-domain communication can be achieved Bandwidth Brokers. The idea of a Bandwidth Broker (BB) was introduced as part of the Differentiated Services architecture [twobit]. The BB plays several roles in administering a diff serv resource management, one of which is management of inter-domain provisioning to support the enforcement of bilateral agreements, or SLAs. Signaling messages are sent between BBs of adjacent domains to request from the adjacent BB the necessary resources in the adjacent network, and to communicate the information about the resources required on the links connecting the domains. That is, inter-domain signaling between adjacent BBs is employed to achieve dynamic SLA negotiation between the domains.

Figure 3 illustrates the inter-domain signaling between different networks, including a stub network, shown running the RSVP resource reservation signaling protocol, and two transit networks, AS1 and AS2. The BB in the stub network communicates with the BB in AS1 requesting resources for traffic originated by hosts on the stub network. How the stub network BB may determines the QoS needs over the link connecting the stub network and AS1 is a matter of local concern. For example, the edge routers may process individual RSVP messages and forward the appropriate information to the local BB, shown in the figure, and the BB may in turn aggregate the flows from the stub network. The BB in AS1 may, in turn, talk to the BB in AS2 in to manage the resources for the aggregate flow(s) passing between domains, etc. Examples of such inter-domain communication (BB-to-BB) are given in [twobit].

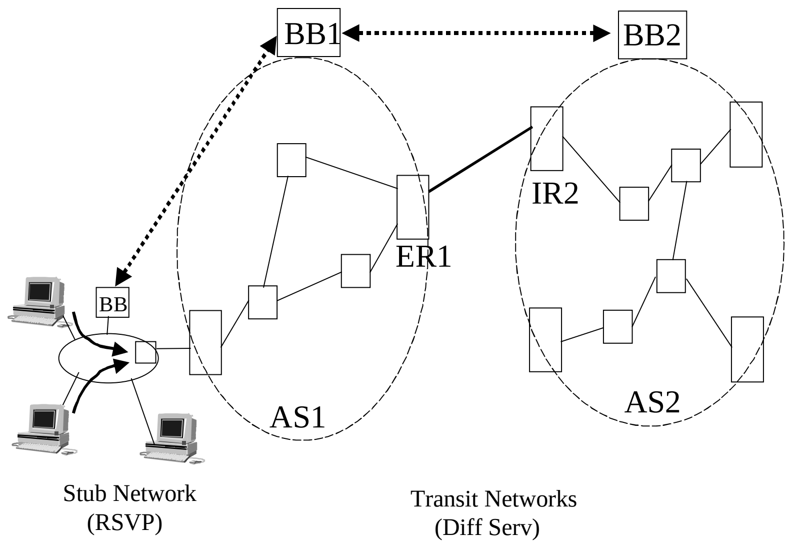

Figure 3: Inter-domain Resource Management

2.1 Static vs. Dynamic Management

One issue with signaling for inter-domain resources is the temporal granularity of the protocol. That is, how often the BBs exchange messages to update/renegotiate their bilateral agreement. SLAs can be either static and dynamic. In the case of static SLAs, an inter-domain protocol may not actually be required, except maybe for the purpose of automating the resource management function. We describe how the BB may manage these types of inter-domain agreements below, by way of examples.

Referring to Figure 3 above, the BB in AS1 (BB1) knows what the current allocation is at the border with AS2 and what traffic is currently traversing the link across the border. Originally BB1 sets up a service agreement with BB2 to send, say, 10 Mb Premium traffic across a link from ER1 to IR2 (egress router to ingress router). Currently there is, on average, only 1 MB Premium going on that link now. Now, at some point (weeks or months) later, BB1 recognizes that now there is, on average, 8 or 9 Mb Premium traversing that link (for example, because AS1 has more customers subscribing to its Premium service). BB1 may notify the administrator of AS1 who in response notifies the administrator of AS2 and increases the level of Premium resources on the link, i.e. adjusts the service agreement. Another possibility is that BB1 could be configured to signal this adjustment without human intervention ("when avg. Premium traffic reaches 85% of agreement, send a signal to adjacent BB to request 2x the current level").

If the frequency of such updates changes from weeks or months to hours or days, so that billing more closely reflects the actual resources used for example, the resource management moves from static to dynamic. Also, reply/ack messages of the inter-domain protocol between BB's may be propagated back to end users to facilitate admission control, necessary to provide quantitative QoS service.

2.2 Aggregate Flow Management

It may be undesirable for individual flow information to be communicated across diff serv network boundaries, for scalability reasons. Only aggregate flow information should be contained in inter-domain resource management signaling messages. However, within a domain, resource management may be performed at a fine-grain level, for example using RSVP, if the network size is such that scaling is not an issue.

Referring again to Figure 3, a bilateral agreement is established between the stub network (running RSVP) and the transit network AS1. The terms of the agreement might state that AS1 will provide Premium service to packets received from the stub network but to avoid paying for unused resources, the BB in the stub network dynamically notifies BB1 of the requested resources. The BB in the stub network still aggregates individual RSVP requests and sends the aggregate flow requirements to BB1. However, the inter-domain requests are now updated *for each* individual RSVP request that effects the aggregate flow into AS1, and resources for that aggregate flow are dynamically allocated on the edge devices. That is, individual session requests within a stub network may influence how often inter-domain SLAs are updated, but the details of the individual flows are hidden from the BBs involved.

2.3 Bandwidth Broker Message Processing

One of the responsibilities of the bandwidth broker, as discussed previously, is inter-domain resource management, including inter-domain message processing. In the case of dynamic resource management, messages are sent BB-to-BB via some inter-domain signaling protocol. For the static case, messages may be received via some network management interface, issued by a network administrator. The BB's inter-domain functionality, in general, may include the following.

Upon receiving an inter-domain resource management message, a BB may:

- determine whether resources on the ingress router are appropriately allocated for the (aggregate) traffic flow from the sending domain/stub network
- calculate the egress point(s) based on destination AS(s) – *if* provided in the message from the adjacent BB – if no destination AS information is available, some other means of determining egress points or estimating paths will be necessary (there is lots of room for work in this area)
- determine whether resources on the egress router are appropriately allocated for the (aggregate) traffic flow specified in the message
- if necessary or desired, perform intra-domain resource management (see next section)
- if resource allocation changes are necessary for the new aggregate flow at the egress, it may be necessary to generate an inter-domain message to the next-hop domain BB (to request resources from that domain for the new aggregate flow or simply inform the next-hop BB of changing traffic conditions across the boundary), depending on the SLA on that boundary. The next-hop BB repeats this process.

2.4 Inter-Domain Signaling Protocol Issues

The purpose of this document is not to specify, or recommend, an inter-domain resource management signaling protocol. Several protocols that could possibly be enhanced for this use, have been suggested, such as RSVP [rsvp], COPS [cops], and DIAMETER [diam]. The choice of protocol is still an open research issue. However, having discussed general inter-domain message processing, we now look at some possible message content for inter-domain messages. A requirement of an inter-domain resource management protocol is the need to be able to express resource requirements for all types of applications and/or SLAs. Some of the issues to consider are discussed below.

Aggregate Flow Information

QoS information, in inter-domain messages, is with respect to the aggregate traffic crossing the boundary between two adjacent domains. In general, aggregation may be with respect to any data characteristics and may be negotiated by the BBs as part of the bilateral agreement. For the purpose of differentiated services, however, flows are aggregated based on the diff serv per-hop-behavior (PHB) to be received by the packets belonging to the flow. Thus, an inter-domain traffic "flow", for the purposes of diff serv, is made up of *all packets going from the egress router of the sending domain to the ingress router of the receiving domain, receiving a particular PHB.*

Along with the aggregate flow information, a next-hop AS and a profile for the portion of the aggregate flow (may not be entire flow) going to that destination AS may also be included in inter-domain messages. This information may allow each BB along the way to properly aggregate the flows when sending request(s) to the next domain(s).

For example, an inter-domain message might contain the following aggregate flow information. Note this is just an example of some possible data items that might be part of an inter-domain resource management (BB-to-BB) protocol specification:

- ingress_address; interface where the aggregate flow is entering the domain
- ingress_profile; e.g. rate, peak rate, burst size and PHB of flow coming into ingress_address
- dest_AS; implies an egress point for some portion of the aggregate flow, as specified in the egress_profile information
- egress_profile; e.g. rate, peak rate and burst size (some percentage of ingress_profile) destined to the egress router implied by dest_AS.
- Num_of_dest_ASes.

The first two items may be used by the BB to perform admission control or resource monitoring based on the bilateral agreement in place between the sending network and the receiving network. The ingress_address informs the BB where the aggregate flow is to enter the network so the bilateral agreement at that interface can be checked. The ingress_profile tells the BB what resources are required at ingress_address to service the aggregate flow.

The other two items may be used by the BB to monitor agreements on egress interfaces and, if necessary, formulate inter-domain messages to BBs in adjacent networks. The dest_AS provides information that may be obtained from routing information at the sending BB. For example, if the sending BB is that in the stub domain in Figure 2, the IP destination of a flow is known from RSVP messages sent by the end systems. BGP routing information can be used to match the IP dest with a particular AS in the network that includes the address. That AS information can then be sent as the dest_AS in an inter-domain message to the next-hop BB and the next-hop BB can match the AS to an egress interface to be used to reach that AS. The egress_profile tells the BB what portion of the aggregate flow will go to the dest_AS. Since the sending BB is signaling for the aggregate flow made up of all flows being forwarded to the receiving BB's network, not all flows in the aggregate will necessarily be destined for the same AS. Therefore, a list of <dest_AS, egress_profile> pairs may be sent an inter-domain message. The egress_profile describes an aggregate flow which is a subset of the aggregate flow in ingress_profile. The Num_of_dest_Ases field provides the number of egress profiles for this specific ingress_profile

If the entire traffic flow entering the domain at ingress_address is exiting the domain at a single egress, then ingress_profile and egress_profile will be identical. If a flow (flow1) exists already, say entering at ingress1 and exiting at egress1 (for AS1), and a new flow (flow2) is added across the link, say exiting at egress 2 (for AS2), the ingress_profile and egress_profile will be different. Ingress_profile will contain the aggregate of flow1 and flow2, but egress_profile will contain only the profile for flow2. The receiving BB would then aggregate egress_profile with any other flow in place exiting egress2, and pass on a message to the BB of the next-hop domain (AS2).

State sharing between BB's

The interaction between two BB's can be further distinguished based on whether or not the two BBs are located within the same domain (AS). The level of trust and kind of information exchanged between BB's may vary based on this relationship. An important question related to this relationship is that of "state (and fate) sharing". The bandwidth broker architecture may allow for a bandwidth broker function be provided by a primary bandwidth broker with secondary brokers acting as backups. In addition, depending on the granularity of resource allocation and time-scale for negotiation, the amount of state information shared between two BB's may vary. For the sake of robust, fault tolerant operation, any sharing of state between BB's must be based on the "soft state" model, similar to that described in RFC2205 [rsvp], so that necessary state can be re-established and recovered quickly when a BB recovers from a crash or a BB is replaced by another one as part of fault recovery. Therefore, we stipulate that any interaction among BB's that requires establishment of shared state must involve periodic timeout and refresh of shared state for robust operation.

Security Requirements BB-BB Interaction

The communication between BB's requires establishment of trust and use of a secure communication channel for protectinng the privacy and integrity of the bi-directional communication. The IPSEC infrastructure [IPSECarch] should meet the requirements for this purpose.

Multicast Support

The SLA between two neighboring BB's concerns resource allocation for the aggregate border-crossing traffic. On the other hand, multicast groups are set up for specific application instances, and are likely to extend over multiple administrative domains. Generally speaking, the BB interactions have a larger time scale but a smaller topological coverage than the lifetime and coverage of multicast groups. This does not mean to say, however, that the SLA between two BB's cannot be adjusted dynamically in order to accommodate for the resource needed for multicasting a major event (e.g. a White House address event). Instead, we emphasize that the SLA's between BB's must be able to manage resource allocation at coarser granularity than per-application, and with longer time scale.

As we described earlier, although a BB communicates with directly connected neighbor AS BB's only, unicast end-to-end QOS support can be achieved by concatenating these pair-wise SLA's along the path from source to destination domains. The same can be said about multicast traffic support. At the inter-domain level, a multicast tree may be made of many border-crossing links. Multicast traffic can use reserved resources at each "link" if:

- Packets are carrying the correct DS field value, and
- Adequate resources have been allocated.

There is one common question that is often raised in the context of multicast: given multicast data flow is receiver driven (that is data only goes to places where the receivers have expressed interest), how can the receiving ends cause adequate resources to be allocated to achieve good reception quality? The answer to this question has two parts. One, we assume a SLA covers agreements on traffic volume going both directions. Secondly, the only function needed to allow resource allocation be adjusted from the receiving end is being able to forward the adjustment request up the multicast tree towards the source. While BGMP provides us with the information how to reach the root domain from leaf domains containing receivers, it does not provide us with the information of how to reach from the root domain the domains containing the senders. For now, we don't have any solution to this problem.

3 Intra-Domain Resource Management

As discussed above, resource management techniques used in any single domain should be left to the discretion of that domain's administrator. In this section we discuss some general approaches to performing intra-domain resource management for the stub network and transit network.

3.1 Stub Networks

The stub network is the sender or receiver's local network, consisting of hosts and QoS-capable routers or switches. Individual information flows are created or terminated by end systems connected to the stub network. In the paragraph that follows we give a brief description of the scheme proposed in [e2e] where Intserv/RSVP is used for resource management in stub networks. However, stub networks may also utilize differentiated services mechanisms such as a Bandwidth Broker (BB) internally for providing QoS to the end user. In any case, BBs are still suggested in any network where a neighboring network is accessed and some bilateral agreement is negotiated between the networks.

If RSVP/Intserv QoS is used in a stub network, resources are reserved on a per-flow basis, hop-by-hop, at each RSVP-enabled router in the stub network and data packets are classified and serviced at each router according to the contents of the IP packet header (source, destination, ports, protocol). When a flow exits the stub network, and enters an adjacent transit network, the resources on the egress interface must be managed in accordance with the bilateral agreement in place between the stub and transit networks. For this reason, the stub network may still employ a bandwidth broker, to manage the resources on the links connecting the stub network to its neighboring transit networks and to aggregate the individual RSVP entering the transit (diff serv) nework. In this case the BB may also provide information such as how to appropriately set the DS Field of packets before forwarding them into a particular diff serv enabled transit network.

In very simple networks, it may be possible for the BB to do resource management by applying methods as described in [dspres], which do not require knowledge of detailed network topology. In one example in [dspres] (Figure 4), the stub network consists of LANs supporting at least 10 Mbps connected by higher bandwidth core links. Each end system has at least 10 Mbps connectivity to the core network. A simple network resource model that assumes a total of 10 Mbps capacity within the network for Premium traffic to prevent oversubscription can still support 300 simultaneous voice/video sessions using the available 10 Mbps pool.

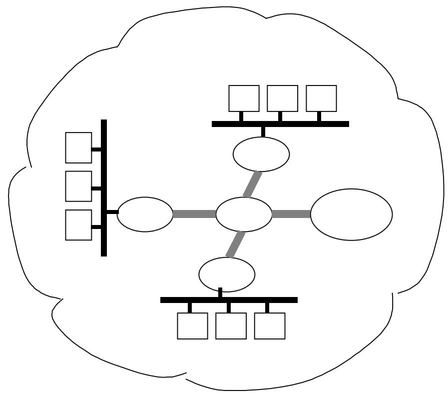

Figure 4: Campus with 10 Mbps Minimum Access

A second example from [dspres] (Figure 5) is where two campuses are connected by a lower speed WAN link, so that 10 Mbps can be supported within each campus, but not between campuses (Figure 5). In this case, Brokers can be implemented in each campus to limit the intra-campus resource allocation to 10 Mbps maximum for Premium traffic, and allocate bandwidth out of the T1-size bandwidth pool available between the campuses, when the ingress and egress points of the information flow are in different campuses.

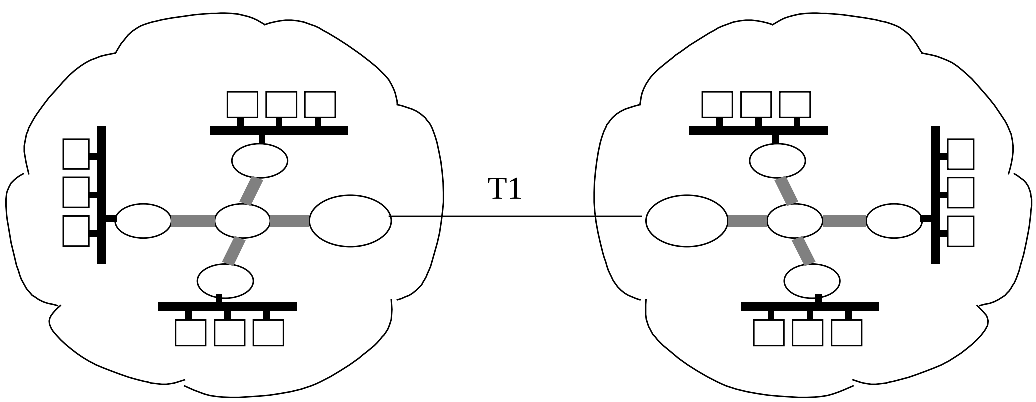

Figure 5: Campus Connected by WAN Link

In most stub networks, however, there may be a variety of link rates and access methods, ranging from switched 100 Mbps Fast Ethernet access to 56 Kbps modem or frame relay connections to remote users. In this case a much more sophisticated BB is needed to evaluate resources available for a new information flow.

Instead, RSVP and diffserv methods can be combined to take advantage of RSVP signaling and diffserv aggregation. RSVP can be used to carry per-flow reservation requests hop-by-hop as a means of ensuring the necessary resources are available within the boundary of the stub network. However, if the RSVP messages now include a DS Field value set by the ingress router based on mapping to the diff serv PHBs, this will indicate the desired PHB to the intermediate nodes. The DS Field of the data packets in the flow will be marked at the ingress router and packets will be processed at intermediate nodes based on the DS Field alone, as in the diff serv QoS model. Thus, per-flow RSVP state is used for resource management while the DS Field in the data packets is used for classification.

The advantages of such a method are that the current RSVP model would not need to change to accommodate aggregate flows (see next section), while at the same time, BB functionality is reduced to basically mapping Tspec values to PHB/DS Field values. Alternatively, each router in the stub network can apply the appropriate PHB based on the RSVP message contents, rather than having to interpret the DS-byte marking to determine the PHB to apply to incoming packets. This may simplify some cases where the network administrator must deal with a heterogeneous network of new and embedded devices.

3.1.1 Using RSVP in the stub networks

The authors of [e2e] have proposed a scheme whereby RSVP is used in the stub networks for reserving resources for individual traffic streams that have their source or destination(s) in these networks. For completeness reasons, we give here a brief summary of how the scheme works.

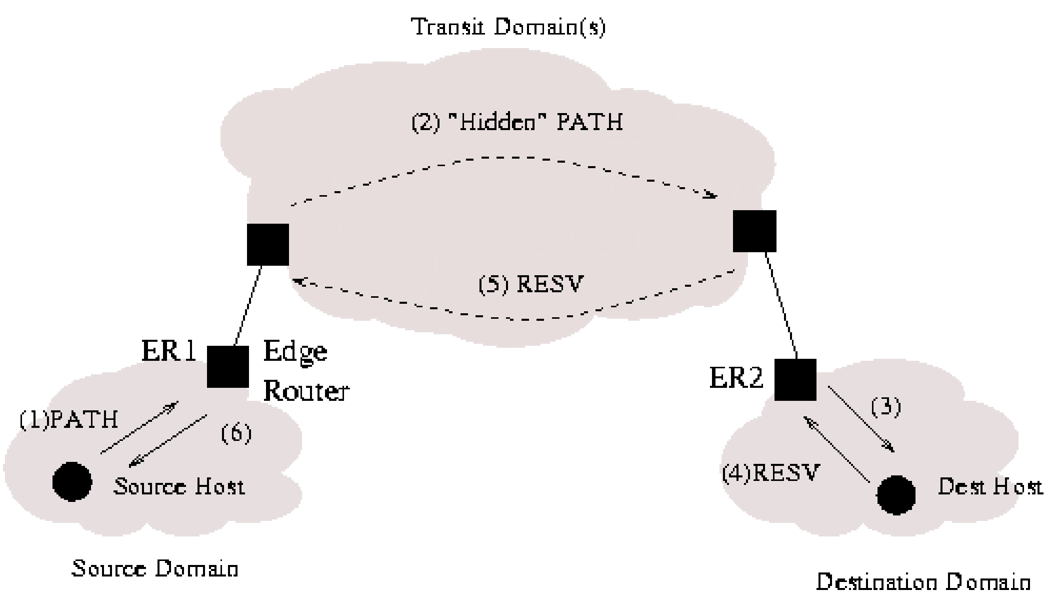

Figure 6. Support for Integrated Services

As we can see from Figure 6, the sender initiates the exchange by sending a RSVP PATH message towards the receiver. Standard RSVP processing is applied within the sender's domain. Once the PATH message reaches the domains edge router, it is ''transparently" tunneled through the transit diffserv domains until it reaches the egress router(s) at the destination domain(s). The reason that RSVP PATH messages should be tunneled through transit domains is to avoid the scalability problems associated with processing of end-to-end RSVP messages by all core routers.

Once the PATH message arrives at the egress router of destination leaf domain it is processed as usual and it is further forwarded inside the leaf domain towards the receiver host. At that time, the receiving host creates a RESV message indicating interest in the offered traffic at a certain Intserv level. The RESV message is carried back towards the sending host. Once the RESV message reaches ER2, it will be transparently transported over the transit networks, arriving at ER1. At this point, ER1 has to do two things: (1) transform the IntServ request to it's Diffserv equivalent and (2) apply some form of admission control for this extra flow. ER1 can take this decision either by looking up some configured mapping or by consulting the domain's bandwidth broker.

Once the mapping has been done, the egress router has to decide if the total amount of traffic crossing the domain, including this new flow is less than the contracted amount. If this is the case, the request can pass. If the total amount is larger that the contracted one then, depending on the type of agreement between the leaft domain and the service provider there are two possible cases. If the agreement is static then an error message has to be sent back to the originator of the RESV message. If the contract allows renegotiations, then possibly the reservation can go through. Assuming that enough resources are available the RESV message is admitted and allowed to travel upstream towards the sending host. If not rejected on the way, the RESV message arrives at the sending host. The receipt of a RESV message is an indication that the specified traffic has to be admitted for the specified Intserv service type (in the Intserv-enabled parts of the path) and for the corresponding diffserv service level (in the diffserv-enabled parts of the path). The host then begins to set the DS-field in the headers of transmitted packets, to the value which maps to the Intserv service type specified in the admitted RESV message.

The scheme presented here assumes that all the leave domains involved use RSVP for resource management. We feel that, while this scheme provides an end-to-end solution, the ultimate goal is to de-couple the resource management schemes used in the peering leaf domains. We are currently working towards this goal.

3.2 Transit Networks

For transit networks, resource management is primarily required to support information flows across the domain from an ingress point to an egress point, as shown in Figure 7. Here the scale of transport requirements may make it impossible to use RSVP on a per-flow basis (e.g., attempting to reserve 40 Kbps flows filling an OC-12 link).

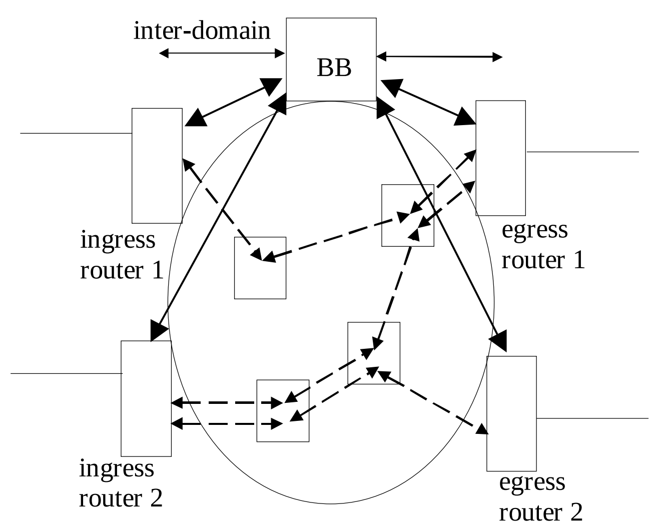

Figure 7: Intra-Domain Resource Management

When a BB receives an inter-domain resource management message, the message contains: information about an aggregate flow entering the domain at a particular ingress; the PHB requested for the flow (see previous section). Ideally, it should also identify the destination AS for the egress_profile portion of the aggregate flow (i.e. an egress). Then the existing aggregate flow, if there is one, between ingress and egress can be updated with the egress_profile flow information. However, this information may not always be available, or may not be required in the agreement between domains.

The mechanism for performing the intra-domain resource management is entirely up to the individual network administrator. In transit networks, employing RSVP/Intserv QoS is typically not a viable option, due to the size of the network and the scalability problems imposed by the per-flow processing. Differentiated Services can be used in transit networks to provide users with end-to-end quality of service [e2e] with greater scalability.

The differentiated services framework [dsarch] suggests that a Bandwidth Broker (BB) is used to manage the allocation of resources of an administrative domain to support the diff serv traffic traversing a transit network. The BB keeps track of the DS traffic that enters and leaves the domain across its boundaries, making sure that the bilateral agreements with adjacent domains are adhered to. The BB communicates with the ingress and egress border routers to configure traffic conditioners within the routers, according to the bilateral agreements. The COPS protocol is suggested for BB to border router communication [copsds].

3.2.1 RSVP as the intra-domain management protocol

We present in this paragraph an example realization of the intra-domain protocol for transit networks using RSVP for internal resource management. We use here the assumption that upstream neighbors who contract the domain for delivering their traffic do not specify the set of destinations. The agreement only specifies an aggregate amount of diff-serv traffic that enters the domain through a particular interface of an ingress router.

This assumption is realistic since customers may not always know in advance all possible destinations of their traffic. Furthermore it makes the SLA easier to create, maintain and understand therefore making the service more attractive to customers. The downside is that it makes it more difficult for the transit network to allocate local resources to satisfy the requirements of the transit traffic.

Destination	Outgoing	Egress		PHBs	
	Interface	Router	PHB1	...	PHBn
18.0.0.0/4	1	D	120		240
132.123.0.0/16	1	F	230		450
207.30.64.0/24	2	E	80		70
158.3.34.0/24	2	E	340		10

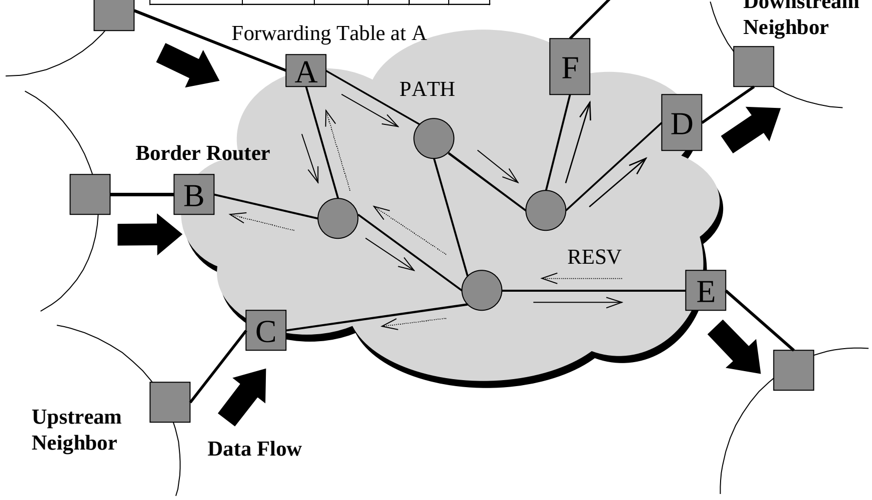

Figure 8: RSVP as intra-domain protocol

Since upstream neighbors do not specify the set of their destinations it is the task of the local domain to estimate this set along with the aggregate amount of traffic destined to each of the downstream neighbors. Once the aggregate amount of traffic destined to each of the downstream neighbors is known then usual RSVP signaling can be used for local resource management.

Each border router has an enhanced forwarding table, where it keeps a counter per PHB of packets destined to each of it’s known destination prefixes. Counters are used to measure to the amount of traffic destined to each of the domain’s downstream neighbors. Fig. 8 shows the forwarding table at ingress router A. There are four destinations and two outgoing interfaces. For each of the known destinations a counter per PHB is maintained.

Each time that packet arrives at an ingress router, the router looks up it’s destination address and consults the forwarding table to properly forward the packet towards it’s destination. In addition to that, the ingress router increases the counter corresponding to packets of the same class as the one the packet belongs to (as this is specified by the DS field). The counter can count packets or bytes depending on the PHB definition and the SLA between the two domains.

We assume that each of the border routers in the transit network participates in the BGP routing exchange and therefore has knowledge about the AS topology and the egress router towards each destination. Each of the ingress routers periodically consults its forwarding table to figure out the amount of traffic flowing towards each of the egress routers in the domain. The following procedure is repeated at each of the domain’s border routers periodically:

```
for (k=0;k<num_of_BRs;k++) {
      egress_router = BR[k];
      for (i=0;i<num_of_destinations;i++) {
            if(egress(dest[i]) == egress_router) {
```

```
                for(j=0;j<num_of_PHBs;j++) {
                    counter[i,j] = dest[i].PHB_counter[j];
                }
            }
        }
}
```

`num_of_destinations` is the number of destinations in the router's forwarding table, `num_of_BRs` is the number of the domain's border routers while the table `BR[…]` holds the border routers of the domain. The function `egress()` gives the border router towards a destination (by looking at the BGP routing table). The table `counter[i,j]` holds the counter values for egress router `i` and PHB `j`.

Once the amount of data traffic to each of the other border routers is known, an edge router starts sending PATH messages to each of the other border routers. The Tspec in those PATH messages reflects the amount of data traffic indicated by the counters and possibly the PHB. Moreover the Tspec in the PATH message should be somewhat higher than the currently observed traffic load so that adjustments in the Tspec sent should relatively infrequent. A threshold mechanism should be used, so that when the amount of actual traffic surpasses a given threshold below the volume described by the Tspec, then an updated Tspec should be sent.

Once an egress router receives PATH messages from the other border routers of the domain it sends a RESV message with a flowspec equivalent to the "sum" of all the Tspecs it has received. The style of the reservation is Shared Explicit and the set of filterspecs contains all the border routers that sent PATH messages.

As an alternative to RSVP, resource management in transit networks can be performed using MPLS [MPLS]. In this case, based on the inter-domain information, the network administrator or some automated entity such as the BB can configure LSPs (Label Switched Paths) for aggregate flows entering the domain at a specific ingress router to the egress router(s) identified by the egress_profile. Using mechanisms such as those defined in the PASTE draft [PASTE] the appropriate level of resources can be reserved inside the transit network for these LSPs.

4 References

[e2e] Y. Bernet, R. Yavatkar, P. Ford, F. Baker, L. Zhang, K. Nichols, and M. Speer. *"A Framework for End-to-End QoS Combining RSVP/Intserv and Differentiated Services"*, IETF <draft-ietf-diffserv-rsvp-00.txt>, June, 1998.

[dsopdef] K. Nichols, S. Blake, *"Differentiated Services Operational Model and Definitions"*, IETF <draft-nichols-dsopdef-00.txt>, February 1998.

[cops] J. Boyle, R. Cohen, D. Durham, S. Herzog, R. Rajan, A. Sastry, *"The COPS (Common Open Policy Service) Protocol"*, IETF <draft-ietf-rap-cops-01.txt>, March, 1998.

[dsarch] K. Nichols, L. Zhang, *"A Two-bit Differentiated Services Architecture for the Internet"*, IETF <draft-nichols-diff-svc-arch-00.txt>, December 1997.

[rsvp] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, *"Resource Reservation Protocol (RSVP) Version 1 Functional Specification"*, IETF RFC 2205, Proposed Standard, September 1997.

[diam] P. Calhoun, *“DIAMETER Resource Management Extensions”*, IETF <draft-calhoun-diameter-res-mgmt-00.txt>, March, 1998.

[IPSECarch] S.Kent, R.Atkinson, “*Security Architecture for the Internet Protocol*”, IETF < draft-ietf-ipsec-arch-sec-07.txt>, July 1998.

[dspres] V. Jacobson, *“Differentiated Services for the Internet”*, presentation at the Internet2 QoS Workshop, May 21, 1998,
http://www.internet2.edu/qos/may98Workshop/html/presentations.html

[copsds] F. Reichmeyer, K. Chan, D. Durham, S. Gai, K. McClourghie, *"COPS usage For Differentiated Services"*, IETF draft, August, 1998.

[MPLS] E. C. Rosen, A. Viswanathan, R. Callon, “*Multiprotocol Label Switching Architecture*”, IETF draft, July 1998.

[PASTE] Y. Rekhter, T. Li, “*Provider Architecture for Differentiated Services and Traffic Engineering (PASTE)*”, IETF draft, January 1998.