

Chapter 2

Related Work

There has been a wide spectrum of work during the last decade on Quality of Service (QoS) control and management in packet switched networks. This chapter only presents a survey of related work that serves as background to our research. In particular, we will survey recent development in: QoS control and router mechanisms in Section 2.1, resource provisioning techniques and bandwidth brokering architecture in Section 2.2, admission control based on passive measurements and active probings in Section 2.3, and traffic policing in Section 2.4. We will also discuss how our proposed architecture and mechanisms are distinct from or influenced by these prior efforts.

2.1 QoS Control in Packet Networks: An Overview

The emergence of IP telephony, video conferencing and other applications with very different throughput, loss and delay requirements are calling for substantial changes in the Internet infrastructure that was originally designed to offer a single, best-effort level of service. Providing different levels of service in the network requires new QoS control and management capabilities, which can be classified along two major axes: *data path* and *control path*. Data path mechanisms are responsible for classifying and mapping user packets to

their intended service class and enforcing the treatment (e.g., amount of resources consumed or delay experienced) received by each service class. Control path mechanisms allow the users and the network to agree on service definitions. They are also needed to determine which users to grant service to, and appropriately allocate resources to each service class. The QoS mechanisms discussed in this section have largely been proposed for the IP layer (Layer 3) and developed to be application independent.

2.1.1 Schedulers and Buffer Management

Data path mechanisms are basic building blocks in a QoS-aware infrastructure. They control how packets access network resources, such as buffers and bandwidth, to provide service differentiation. The two corresponding mechanisms that have been long-standing research topics are (a) scheduling algorithms and (b) buffer management schemes, respectively. Scheduling mechanisms control which packets are selected for transmission on the link, while buffer management schemes decide which packets can be stored or dropped as they wait for transmission.

G. Apostolopoulos et al. reviews the different scheduling and buffer management schemes in [8] and discusses their associated trade-offs in terms of fairness, isolation, efficiency, performance and complexity. For example, Weighted Fair Queuing (WFQ) [26] and its many variants provide rate and delay guarantees to individual flows, while class based scheduling mechanisms, e.g., CBQ [27] provide aggregate service guarantees to the set of flows mapped into the same class. The finer granularity of per-flow information required for the former case comes at the cost of greater complexity. Examples of buffer management schemes include: First Come First Serve (FCFS), Early Packet Discard (EPD) [28] and Random Early Drop (RED) [29].

2.1.2 QoS Control Architecture

There are two major approaches currently under development in the Internet Engineering Task Force (IETF)¹: Integrated-Services (Int-Serv) [12, 13] and Differentiated-Services (Diff-Serv) [14, 15].

Integrated Service

The philosophy behind Int-Serv is that routers must be able to reserve resources for individual flows to provide QoS guarantees to end users. Int-Serv QoS control framework supports two additional classes of service besides "best effort": (a) Guaranteed service [30] and (b) Controlled-load service [31]. Guaranteed service provides quantitative and hard (deterministic) guarantees, e.g., lossless transmission and upper-bound on end-to-end delay. This is useful for hard real-time applications that are intolerant of any datagram arriving after their play-back time [32]. Controlled-load service is intended to support a broad class of applications that are highly sensitive to overloaded conditions. It promises performance as good as in an "unloaded" datagram network, and provides no quantitative assurance. Both services must ensure that adequate bandwidth and packet processing resources are available to satisfy the level of service requested. This must be accomplished through active admission control. The following are the various Int-Serv components that are needed to provide end-to-end QoS, and many research contributions have been made to define their functionality and study their implementation issues:

- A signaling protocol to set up and tear down reservations, e.g., Resource ReSerVation Protocol (RSVP) [16].
- An application-level interface (API) for applications to communicate their QoS needs, e.g., Unix RSVP API (RAPI).²

¹<http://www.ietf.org/>

²The technical standard of Resource ReSerVation Protocol API (RAPI), developed by the Open Group,

- Per-flow scheduling in the network (e.g., WFQ [26]).

Unfortunately, Int-Serv faces other challenges that make immediate deployment infeasible. The increase in per flow state maintenance at routers is proportionally to the number of flows. This incurs huge storage and processing overhead at routers, and therefore does not scale well in the Internet core backbone. In addition, RSVP/Int-Serv Model needs to work over different data-links such as Ethernet, and ATM. Therefore, mechanisms to map integrated services onto specific shared media are needed.

Differentiated Service

Diff-Serv, on the other hand, aggregate multiple flows with similar traffic characteristics and performance requirements into a few classes. This approach requires either end-user applications, first hop routers or *Ingress* routers (interface where packets enter an administrative domain) to mark the individual packets to indicate different service class, e.g., low delay, high throughput, etc. Currently this QoS information is carried in band within the packet in the Type of Service (TOS) field in IPv4 header or Differentiated Service (DS) field in IPv6 [33]. The backbone routers provide per-hop differential treatments to different service classes as defined by the Per Hop Behaviors (PHBs) [34]. Two service models have been proposed: assured service [17] and premium service [18]. Assured service is intended for customers that need reliable services from service providers. The customers themselves are responsible for deciding how their applications share the amount of bandwidth allocated. Premium Service provides low-delay and low-jitter service, and is suitable for Internet telephony, video-conferencing and for creating virtual lease lines for Virtual Private Networks (VPNs).

Diff-Serv approach has several advantages over Int-Serv:

- Diff-Serv is simpler than Int-Serv and does not require end-to-end signaling.

is available at <http://www.opengroup.org/onlinepubs/9619099/toc.htm>.

- Diff-Serv is efficient for core routers since classification and PHBs are based on a few bits rather than per-flow information. Since there are only a limited number of service classes indicated by the TOS field, the amount of state information is proportional to the number of classes rather than number of flows, and therefore Diff-Serv approach is more scalable than Int-Serv.
- Diff-Serv requires minimum change to the current network infrastructure. End hosts, routers or firewall can mark packets while intermediate routers/switches can employ active queue management to provide service differentiation based on bits in the packet headers.

Although flow aggregation improves scalability in Diff-Serv, it becomes unclear what level of statistical guarantees Diff-Serv can provide to individual flows, and if there exist such "guarantees" at all. In [35], Bolot analyzes the performance of two Diff-Serv service models: assured service and premium service. Several studies [36, 37] examine the loss and/or delay behaviors of Diff-Serv architecture using a variety of traffic models.

2.1.3 The Big Picture

Figure 2.1 shows how the Internet infrastructure is evolving. While smaller-scaled local networks might be able to support RSVP signaling and per-flow queuing as in Int-Serv, we envision that the backbone will be more likely to provide coarse-grained service differentiation as in Diff-Serv approach. In this model, the heterogeneous devices and access networks are connected to the backbone through boundary routers, or edge routers (ER), which will mark the TOS or DS field value of the packets accordingly.

We adopted the Diff-Serv approach that separates data path mechanisms from admission control and resource allocation mechanisms that belong to the management plane. The network is viewed as a collection of various domains based on administrative boundaries,

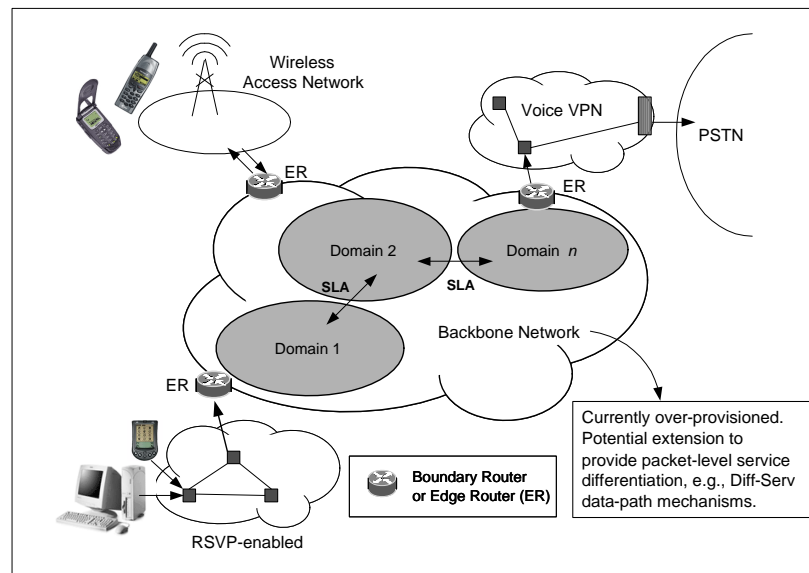


Figure 2.1: Distributed administration of Internet infrastructure with multiple ISP (Internet Service Provider) domains and heterogeneous access networks.

e.g., an organization's Intra-net or an ISP makes a domain. At domain boundaries, *service-level agreements* (SLAs) are made regarding to the amount of resources allocated to traffic that cross domains. A Service Level Agreement (SLA) also refers to the contract between a service provider and a customer. In this case, the SLA specifies a fixed peak bit rate, which the customer is responsible for not exceeding. All excess traffic is dropped. For better link utilization, dynamic SLAs should be supported so customers can request bandwidth on demand. However, the proper configuration of Diff-Serv mechanisms and traffic handling within each domain cannot work in isolation to achieve predictable end-to-end service model. They must be coordinated in a scalable manner across many devices in multiple domains to provide useful end-to-end services.

Figure 2.2 summarizes the various QoS control components of IP-networks in both data and control planes. As mentioned earlier, packet forwarding mechanisms (e.g., schedulers and buffer management) have been the subject of various studies while it is within the last five years that control architectures such as Diff-Serv and Int-Serve are being developed. We have described some of these earlier works in Section 2.1.1 and 2.1.2. The rest of this

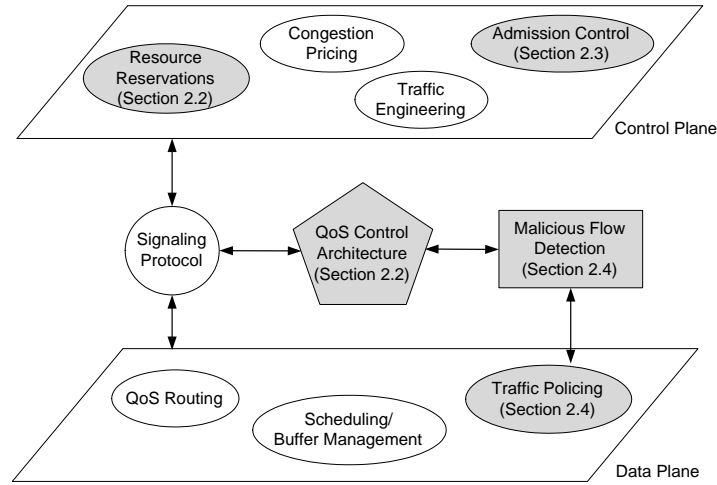


Figure 2.2: QoS control mechanisms in both control and data planes.

chapter will survey the related work in the areas of resource provisioning, admission control, and traffic policing (the shaded areas in Figure 2.2), which are the focus of our dissertation.

2.2 Network Resource Provisioning

As we described in the previous section, the management plane of the Diff-Serv architecture is a design space that remains to be explored. In late 1997, the concept of "Bandwidth Brokers" was introduced by K. Nichols et al. in [38] as an entity in charge of resource management in an administrative domain. The Internet2 QoS working group [39] has then made an attempt to harmonize the different ideas and proposals to define a model of Bandwidth Broker (BB) to be deployed in an inter-domain Diff-Serv test-bed called Qbone [40].

The issues of resource allocation and management are not unique to Diff-Serv architecture. In fact, they are long-standing research topics on their own. Besides discussing the prior work on Diff-Serv bandwidth broker, this section also reviews other resource management techniques and architecture relevant to this dissertation, including: dynamic

allocation for virtual private networks, capacity planning in telephone networks, advance reservation techniques, and pricing-based approach.

2.2.1 Diff-Serv Bandwidth Broker Architecture

Several bandwidth broker (BB) implementations have been proposed and analyzed in [19, 20, 21] as a scalable QoS provisioning mechanism over the Diff-Serv architecture. A BB is an agent that performs a subset of policy management functionality, including:

- admission control to limit the number of connection requests based on available resources in the network,
- intra-domain resource allocation to support the QoS services offered to users, and configuring routers with correct forwarding behavior, and
- automate inter-domain SLA negotiations.

In [21], the authors presented the broker signaling trade-offs in the context of the Swiss National Science Foundation project CATI [41], but they do not optimize end-to-end path selections. The Internet2 QoS working group have been investigating the inter-broker signaling to automate the adaptive reservation within and across domains. However, the BBs are currently configured manually, and many design decisions remain open.

Terzis et al. proposed a *Two-Tier* model [20] where resource allocation control is separated into two level hierarchy: inter-domain allocation and intra-domain allocation (Figure 2.3). In this case, the Bandwidth Broker has a dual role:

- manages internal resources within each administrative domain, which can be fined grained (per flow), and
- maintains bilateral SLA agreements with its neighboring BBs to allocate resources to aggregate traffic crossing domain borders.

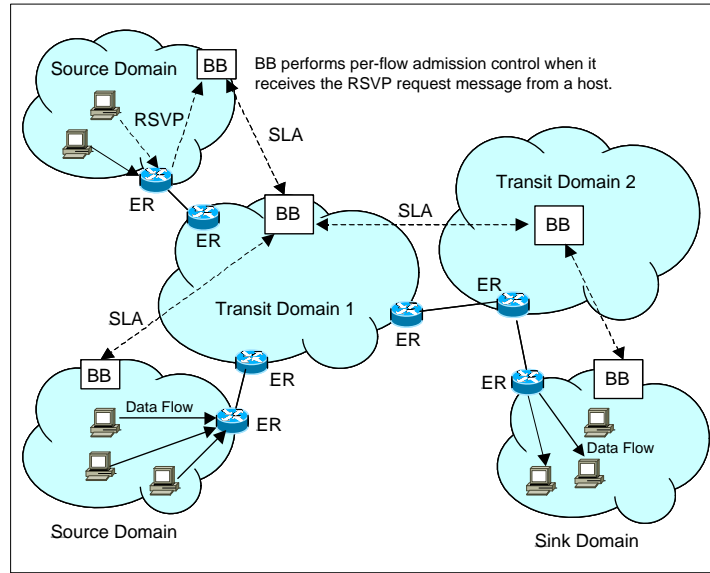


Figure 2.3: Two-tier model for Diff-Serv Bandwidth Broker architecture.

This design reflects today's Internet two-level routing architecture that allows each Autonomous Systems to freely choose its own routing protocol internally (OSPF [42], RIP [43], or IS-IS[44]). To achieve cross-domain connectivity, neighboring domains use the Inter-domain routing protocol BGP [45] to exchange network reachability information.

A simplified RSVP is used in [20] as an intra-domain resource allocation protocol for traffic between edge routers, and admission control is performed on a hop-by-hop basis. The inter-domain allocation is adjusted dynamically based on a *additive increase multiplicative decrease* approach. The reservations are performed locally between two neighboring domains without reflecting the traffic and network variation in other domains that lie in the end-to-end path between source and destination networks. End-to-end QoS support is achieved through the concatenation of bilateral SLAs and adequate intra-domain resource allocations.

Discussion

Although we adopt the fundamental principle of the *two-tier* model for intra- and inter-domain resource allocation, our mechanisms and emphasis are significantly different from [20] or other BB proposals [40, 41].

1. We introduce a hierarchical structure, called the Clearing House (CH), within each administrative domain to manage resource allocation instead of employing a single Bandwidth Broker. The intra- and inter-domain resource control tasks are partitioned and distributed to various CH-nodes, so that the amount of state maintenance and processing per node is reduced.
2. Per-flow admission control is only performed by a local resource manager at the edge (leaf nodes) without propagating the reservation request hop-by-hop across the entire administrative domain. Our approach considers network-wide traffic distributions in choosing the admission threshold, rather than measuring the impact of admitting a new flow on a single node (ingress point) as in many previous solutions.
3. Inter-domain resource reservations are established in advance based on real-time measurements of aggregate traffic statistics, rather than the congestion control approach detailed in [20].
4. We propose a scalable approach to perform traffic policing and detect misbehaving flows within each domain, which are not addressed in the previous BB literature.

2.2.2 Virtual Private Networks

The paradigm of allocating resources for traffic aggregates has also been applied to managing Virtual Private Networks (VPNs). Reference [46] proposed a service model, called *hose*, which specifies the capacity required for aggregate traffic from one endpoint to the set

of other endpoints in the VPN customer sites. Each *hose* is associated with a performance guarantee. The share of resources allocated to a *hose* is resized based on a set of traffic predictors, and the performance of this adaptive scheme is compared to static provisioning through trace-driven simulations. The authors consider traces of telephone calls over the AT&T national long distance network as well as data traffic on a large corporate private network. Results show that dynamic resizing method achieves a factor of 2 to 3 in capacity savings on access links over statically provisioned customer-pipes.

However, this work only considers a single ISP scenario. It is important to study how this technique can be extended to address inter-domain resource allocation in the case of multiple domains (ISPs). The ultimate challenge is how to provide end-to-end performance assurance through intelligent coordination of intra- and inter-domain resource control mechanisms without requiring a complex signaling protocol. Another desirable property of the solution is scalability with respect to both the number of users and the number of domains traversed.

2.2.3 Capacity Planning in Telecommunication Networks

The concept of hierarchical databases has long been used in telephone network switching, and for user mobility management in the PCS network. In both cases, the sessions are circuit switched or connection oriented, and each session generates a constant bit rate (CBR) traffic. The hierarchy of increasingly aggregated flows is common in the telephone network, but it is based on a fixed bit-interleaved digital multiplexing, as defined in the PDH standard [47], e.g., 24 telephone channels are carried at the T1 level (1.544 Mb/s). Each session is assigned a fixed time-slice of the resources.

The hierarchical structure of our proposed Clearing House architecture that allocate resources at different levels of aggregation is very similar to the telephone network. However, this dissertation explores a different problem space where all the sessions are

connection-less, and individual flows can generate variable bit-rate traffic (due to compression), which allows statistical multiplexing at the packet level. The CH-architecture aggregates call requests and perform admission control decision in real-time based on the available bandwidth and network performance. As a result, we see constantly varying statistical multiplexing gains at different links, as opposed to a fixed multiplexing ratio achieved in the telephone trunks. In addition, routing (setting up a dedicated circuit) and resource allocation (allocating time-slice) are tied together in the telephone network when a connection is established. In our case, the two components are separate and we do not provide establish per-flow end-to-end reservations.

2.2.4 Advance Reservations

The need for advance resource reservation (ARR) has been recognized for applications such as video conferencing where network resources are required at a specific start time and for a given duration. Many studies [48]-[51] on this subject focus on performance modeling of ARR on a single link. The co-existence of immediate and advance reservations is addressed in [52, 53] where the authors show that network resources can be shared between the two without being pre-partitioned. Immediate and advance admission control are performed by agents [52] so that reservations can be provided without requiring any state maintenance at the routers. An important parameter is the *lookahead time*, the point at which the agents start making resources available for approaching advance reservations by rejecting immediate requests. It is assumed that individual users specify the bandwidth requirement at the time of requests, and for advance reservations, the duration is also specified.

We, on the other hand, consider advance reservation for the intra- and inter-domain traffic aggregate, instead of for individual sessions. The advance reservations are established based on aggregate traffic measurements without relying on how well individual flows keep

to their bandwidth specifications.

2.2.5 Dynamic Packet State

I. Stoica et al., have proposed a new architecture called Scalable Core (SCORE) [54, 55] in which only edge routers perform per flow management, while core routers do not. The authors have shown that a SCORE network can achieve fair bandwidth allocation in [54], and provide end-to-end per flow delay and bandwidth guarantees (like Int-Serv) in [55]. The key technique behind SCORE is the Dynamic Packet State (DPS) approach that carries additional state information in each packet header. The packet header state is initialized by the ingress routers. The core routers process each packet based on the state carried in its header, update the state in the packet's header and forward it to the next hop. With DPS, the actions of edge and core routers along the path of a flow can be coordinated to implement distributed algorithms that deliver QoS assurance without maintaining per flow state at the core routers.

2.2.6 Pricing-based Approach

The Internet access has been predominantly sold based on flat monthly rates depending only on the size of access links, not on usage. However, there is a strong motivation to adopt “usage-sensitive pricing”, as advocated by Professor Pravin Varaiya in the INFO-COM'99 keynote lecture:

Flat-rate pricing encourages waste and requires 20 percent of users who account for 80 percent of the traffic to be subsidized by other users and other forms of revenue. Furthermore, flat-rate pricing is incompatible with quality-differentiated services.

Professor Pravin Varaiya, University of California, Berkeley

In fact, the role of prices as essential resource allocation control signals has long been established. J. Sairamesh et al. proposed a new QoS provisioning methodology based on mathematical economic models in [56]. They compute the equilibrium prices based

on the user demands, and from this determine the optimal allocation of buffer and link resources to each of the traffic classes. Results in [56] are based on a single-node model that has multiple output links with an output buffer. In another independent work, N. Semret et al. [57] introduce the Progressive Second Price (PSP) auction as a bandwidth pricing mechanism, and show that it achieves economic objectives (efficiency and incentive compatibility), while requiring small signaling and computational load. Further studies are needed to investigate the applicability of these results [56, 57] to large networks, and develop market based mechanisms to admit and route sessions over multiple domains. References [58] and [59] examine some of these issues. The former [58] considers a game theoretic model of capacity provisioning in a Diff-Serv Internet to maintain stable and consistent SLAs across multiple networks. The latter [59] introduced a hierarchical economy consisting of two types of markets (retail and wholesale) and three types of entities (service provider, domain broker, and users). The authors in [59] use retail market estimation to determine the optimal buying/selling strategies that maximizes profit while maintaining low blocking probability.

There is also a huge literature on Internet charging and billing mechanisms, mostly in the context of how users value services and react to price changes. Recently, a large-scale experiment called INDEX [60] was deployed to test users' willingness to pay for various Internet access options. The INDEX investigators conclude from the experimental data that differentiated services and usage-sensitive pricing would be better for both ISPs and users. However, the data also shows that metered billing has dramatically decreased usage. A. Odlyzko [61] attributed this decrease to very strong consumer preferences for simplicity, especially flat-rate pricing. The author's argument is based on an extensive and detailed analysis of the history of communication technologies reported in [61], including ordinary mail, telegraph, wired voice phone, cell phone, residential Internet access and private lines. For example, when AOL switched from usage-based pricing to flat-rate pricing in October 1996, the usage per person tripled in a year because the users found flat-rate pricing easier to

understand. As a result of the increased usage, the total profit was greater when AOL used flat-rate pricing as opposed to usage-based pricing. This result indicates that the main weakness of usage-based pricing is its relative complexity compared to flat-rate pricing, which makes it less attractive to end users.

A detailed analysis of the effect of the various Internet pricing on user behavior and network efficiency is out of scope of this dissertation. We briefly mention recent work on pricing here because it provides an orthogonal degree of freedom to achieve Internet service differentiation. This thesis mainly focuses on network resource management and addresses the tradeoffs between efficiency and end-to-end performance.

2.3 Admission Control

Admission control is an essential component of any control architectures providing service differentiation. It determines whether flows requesting services are accepted (or rejected) depending on the available network resources to ensure that acceptable QoS levels are delivered to the admitted traffic. There are typically two classes of approach: *parameter-based* or *measurement-based* admission control. Parameter-based admission control algorithms are based on worst case bounds derived from the parameters describing the flow, and are typically more appropriate for providing hard-real time services. Their effectiveness depends on the ability to predict the traffic behavior based on client-specified parameters, and hinges on the ability of the flows to provide the best guesses of what these parameters are, and their lack of incentives to lie. These algorithms may result in low network utilization if the traffic is bursty. On the other hand, measurement-based admission control (MBACs) algorithms base their decisions on measurements of existing traffic rather than on worst-case bounds. Therefore, MBACs are best suited for providing soft real-time service, i.e. an enhanced QoS without hard guarantees.

In our architecture, we chose measurement-based over parameter-based admission control for two reasons. First, MBACs yield higher network utilization, and secondly it is difficult to describe Internet traffic with such diversity and unpredictability with a reasonably small set of parameters.

2.3.1 Measurement Based Admission Control

Many algorithms and principles outlined in the MBAC literature apply in our work, and we definitely benefit from results in [62, 64, 65] to name a few. L. Breslau et al. evaluated six different MBACs in [62] and results showed that all these algorithms achieved nearly identical performance in terms of their ability of balance the tradeoff of losses (QoS seen by individual users) and load (network utilization). Their study also revealed the following insights:

- measurement estimation and admission decision processes can be decoupled for many algorithms
- Differences in performance caused by flow heterogeneity should be addressed by policy, and rather than by algorithmic differences.
- MBACs appeared to cope well with long range dependence. In some of the simulation scenarios, they perform better than parameter-based algorithms.
- None of the MBACs evaluated are able to provide reliable performance tuning knobs that allow network operators to set a target performance level and actually match it.

In [65], the authors implemented and evaluated a new MBAC algorithm that exploits measured peak rate envelopes of the aggregate traffic. The “maximal rate envelope” is a function of a chosen interval length and captures the temporal autocorrelation structure of the aggregate flow. The MBAC uses this envelope to bound future packet arrivals, and

ensures that the admission of new flow will not cause any buffer overflow. Packet losses and delays may occur due to the uncertainty of the prediction. The authors presented new theory to quantify the confidence level of a schedule-ability condition and predict loss probability when the condition is violated.

2.3.2 End-point Admission Control

Admission control in the traditional Int-Serv approach requires a signaling mechanism such as RSVP [16] to carry per-flow request to all the routers along the path. The routers must perform local admission control and keep per-flow state to ensure delivery of desired QoS. The significant burden placed on the routers limit the scalability of this approach. An alternative solution is *end-point admission control* where end-hosts probe the network to check for resource availability before establishing any connections. This is combined with the course-grained Diff-Serv router mechanisms and proper provisioning in the network to achieve QoS.

Recent proposals on end-point admission control [66]-[72] share similar architectures but differ significantly in the control algorithms. Prior to call establishment, the end host send probe packets at the data rate it would like to reserve. In [66] and [67], all data and probe packets are indistinguishable, and there is no differentiation of best-effort vs. real-time traffic. The packets are marked upon congestion (ECN congestion marks) [68], and flows must pay for the marked packets. In this case, admission control is an implicit service provided through price discrimination. The schemes described in [69] and [70] use packet drops instead of congestion marks to indicate congestion, and probe packets are sent in a separate (lower) priority class than data. The *endpoint* in [71, 72] refers to the edge router and not the host. In this setting, edge routers passively monitor paths to derive better estimates of the current network load. L. Breslau et al. provided a careful study of the architectural and performance issues inherent in endpoint admission control in [73].

Our proposed framework shares similar features as end-point admission control, that is the per-flow admission control is only performed at the edge. However, the details of our scheme differ significantly. We do not rely on per-hop signaling protocol or end-host probing to determine whether sufficient resources are available. Instead we leverage the knowledge of aggregate traffic distribution in the ISP domain between different ingress and egress routers to make admission control decisions. The details are discussed later in Chapter 5.

2.4 Traffic Policing

Traffic policing in the Diff-Serv literature usually refers to parameter-based packet filter mechanisms, which are useful in tracking and shaping per-flow usage. In this dissertation, policing refers to monitoring admitted traffic and identifying malicious flows. We use the words “malicious” and “misbehaving” interchangeably to describe admitted flows that violate their allocated share of bandwidth.

2.4.1 Stochastic Fair Blue

The most relevant work with respect to our traffic policing mechanism is Stochastic Fair Blue (SFB) proposed by W. Feng, et al. in [74]. SFB provides a scalable way to identify and rate-limit non-responsive flows using two independent algorithms BLUE [75] and a Bloom filter. BLUE is an active queue management algorithm that uses packet loss and link utilization history to manage congestion. It marks packets in the queue based on a probability that is incremented when a buffer overflow occurs. The rate at which it sends back congestion notification also increases with the marking probability. On the contrary, if the queue becomes empty or the link is idle, BLUE decreases this marking probability. Bloom filters are designed to uniquely classify objects through the use of multiple, independent hash functions. They are commonly used in word processing

software applications as an efficient means to do spell checking or web caches to efficiently determine the existence of an object. Using bloom filters, SFB is able to classify flows with an extremely small amount of state and a small amount of buffer space.

The goal of SFB is to manage congestion and enforce fairness among a large number of flows. The basic algorithm is as follow:

- SFB maintains $N \times L$ accounting bins. The bins are organized in L levels with N bins in each level. There are L independent hash functions, and each is associated with one level of the accounting bins.
- Each hash function maps a flow into one of the N bins in that level. When a packet arrives at the queue, it is hashed into one of the N bins in each of the L levels.
- The accounting bins keep track of a marking/dropping probability, p_m as in BLUE, which is incremented when the bin goes above a threshold.
- The decision to mark a packet is based on p_{\min} , the minimum of p_m of all bins to which the flow is mapped into. If p_{\min} is 1, the packet is identified as belonging to a non-responsive flow, and is rate-limited.

In short, SFB can effectively identify a single non-responsive flow in n^L flow aggregate using $O(L \cdot n)$ amount of state.

The idea of classifying good versus bad (non-responsive in SFB or misbehaving in our case) flows is similar, but the associated algorithm is different. Instead of Bloom filters, we classify packets into different groups for policing based on the Flow-Identifiers (*Fids*) carried in the packet header. We employ a set of token bucket filters (TBF) to police the traffic, and packets are dropped when the TBFs overflow. Active queue management such as BLUE is not considered in our scheme. The details of the detection scheme are outlined in Chapter 6.

2.5 Summary

Our survey indicates that QoS control mechanisms in the data path have been well studied. Some of the solutions such as Weighted Fair Queuing (WFQ) and Random Early Drop (RED) have been implemented in existing routers while other proposals are under development in the IETF (Section 2.1.1). On the other hand, we have relatively limited understanding of the control plane and many open issues remain to be resolved. One of the challenges is how to coordinate resource allocation within and across multiple domains in a scalable manner to provide end-to-end performance guarantees. Towards this end, Int-Serv and Diff-Serv have been developed as QoS-aware control architectures but each of these two solutions has its own limitations that hinder its wide-spread deployment (Section 2.1.2). Int-Serv, which requires per-flow signaling and state maintenance, does not scale well as the user population grows. Although Diff-Serv approach is scalable, it only manages to provide coarse-grained performance assurance. In either case, end-to-end QoS is impossible without inter-domain resource control mechanisms.

The earliest work to address inter-domain resource provisioning issues and attempt to bridge the gap between Int-Serv and Diff-Serv is the Bandwidth Broker (BB) Architecture (Section 2.2.1). However, the reservation and admission control mechanisms within the BB proposal only consider local measurements at a single node (ingress router) or between a single pair of neighboring domains (for inter-domain reservations) and fail to reflect the traffic fluctuations and congestion levels in other parts of the network.

In this dissertation, we propose a new architecture called Clearing House (CH) to provision the intra- and inter-domain link capacity to provide statistical QoS such as maximum packet loss rate and latency. The two key design principles that make CH scalable are: *hierarchical approach* and *aggregation*, which we will explain in detail in the next chapter. In short, an ISP can be partitioned to several smaller domains, each associated

Table 2.1: Comparisons between the Clearing House approach and previously proposed architectures.

Proposed Architectures	Properties	Scalability	QoS Guarantees
Int-Serv	Flat structure. Uses RSVP protocol & soft state approach.	Limited. Per-flow signaling & state maintenance at all routers.	Strong per-flow, end-to-end QoS.
Diff-Serv	Flat structure. Uses DHCP in IP-headers to indicate traffic requirements.	Scales well. Only edge routers keep per-flow states.	Coarse-grained, per-hop performance assurance for traffic aggregates.
Bandwidth Broker	Two-tier model. One BB per domain to manage resources.	Scales well, except for large domains. BB & edge routers keep per-flow states.	Coarse-grained end-to-end performance via concatenating pair-wise SLAs.
Clearing House	Hierarchical structure.	Scales well. Edge routers keep aggregate states.	Statistical end-to-end QoS, e.g., maximum loss rate and delay. Btw Int-Serv and Diff-Serv.

with a CH-node. The resource control tasks are then distributed to the various CH-nodes that form a hierarchical tree. The CH exploits the predictability of aggregate traffic to establish and adapt intra- and inter-domain reservations while requiring only aggregate state maintenance. In our model, per-flow admission control is only performed at the ingress routers, but our algorithm considers network-wide traffic distribution in making admission control decisions. We also provide a mechanism to police admitted flows and detect malicious flows to ensure that the end-to-end performance of legitimate flows is protected.

Table 2.1 and 2.2 compare how our approach is different from the previous work. The details of the CH architecture and its various resource control mechanisms are presented in Chapter 4, 5, and 6.

Table 2.2: Comparisons between our resource control schemes and related work.

Proposed Solutions	Reservations (resv)	Admission Control (adc)	Traffic Policing (tp)
Int-Serv	Per-flow, end-to-end resv via RSVP based on user-specified parameters.	Per-flow and per-hop on end-to-end path using worst-case bounds.	Per-flow policing at all routers.
Diff-Serv	Per-traffic class, via configuring schedulers like WFQ.	Per-flow, only at ingress routers using single-node measurements.	Per-flow policing only at ingress routers.
Bandwidth Broker	Per-traffic class like Diff-Serv, and through pair-wise SLAs.	Per-flow, only at ingress routers using single-node measurements.	Not addressed.
Clearing House	Aggregate reservations for intra- & inter-domain traffic aggregates based on real-time traffic measurements.	Per-flow, only at ingress routers using estimated network-wide traffic distributions.	Aggregate policing with ability to detect individual malicious flows.