

Chapter 4

The Clearing House: A Distributed QoS Control Architecture

The lack of a well-studied policy architecture to regulate network resource provisioning in a scalable manner has motivated our design of a Clearing House (CH) as an alternative solution. Examples of *network resources* include link capacity, buffer space, processing cycles at intermediate routers and storage space. In this dissertation, the word *resource* refers specifically to link capacity (bandwidth). The Clearing House is a distributed architecture that coordinates resource reservations within and across multiple domains based on statistical estimates of aggregate traffic demand. This chapter focuses on the CH architectural design, including its logical structure and the functionalities of its different components. In Section 4.1, we describe the design goals of the CH, and the assumptions we make about the network. In Section 4.2, we introduce background knowledge about the Internet network topology and traffic characteristics, and discuss how this knowledge affects our key design decisions. Section 4.3 and 4.4 provides an overview of the hierarchical CH-tree formation and the various resource control mechanisms. Section 4.5 summarizes the key features of the CH architecture, and discusses an example application where the

CH is deployed to manage virtual private networks (VPNs).

4.1 Introduction

The concept of a clearing house has long been existent in the banking industry as an establishment where financial institutions adjust claims for checks and bills, and settle mutual accounts with each other. Even in the context of the Internet, the concept of the Clearing House is not entirely new. In 1995, a consortium of leading California Internet Service Providers formed the Packet Clearing House (PCH) [92] to coordinate the efficient exchange of data traffic from one network to another. The PCH member agreement includes cost of membership, peering connections and routing policy. For example, PCH members may exchange traffic between networks without any settlement fees. However, the PCH agreement does not provide any performance assurance or reflect any monetary compensations based on relative amount of traffic exchanged between members. Many architectural design issues involved in such an Internet Clearing House remain unexplored. On the other hand, increasing number of Internet companies are now offering on-line network resource brokerage by gathering guaranteed demand from the prospective customers and matching it with the sellers' capabilities. Examples include RateXChange's Real-Time Bandwidth Exchange (RTBX) [93], Arbinet Global Clearing Network's trading floor for minutes [94] and Priceline.com's future plan to offer time-block brokerage for domestic and international long-distance calls [95]. Such business models involve Clearing House mechanisms, which have not been studied carefully for the Internet scenario where bandwidth efficiency and QoS assurances are important.

4.1.1 Design Goals and Assumptions

Even today, most ISP backbone networks are managed manually. For example, ISPs rely on human operators to monitor their operational networks and reconfigure the

routers when necessary, e.g., changing the preferred route between two endpoints to an alternate shortest path if congestions or link failures occur on the original path. In addition, the link weights for intra-domain routing protocols (OSPF [42] or IS-IS [44]) are chosen manually to perform load balancing. Several measurement studies [96, 97, 98] reveal that the distribution of Internet traffic over an ISP network can be highly unbalanced, e.g., link utilization on some critical links can be as high as 65-90% while other links are only 10% utilized. Obviously, there is a strong need for a more efficient way of allocating intra-domain resources (link capacity) and automate the required control process, especially for large ISPs that span over 500 nodes. Similar techniques are essential for managing resource allocation across multiple ISP domains to improve end-to-end network performance. Inter-domain resource control encounters an additional set of challenges that are not existent in the intra-domain case, including trust issues between different competing ISPs. For example, an ISP may not be willing to share information such as internal topology or backbone measurements with its neighboring peers. The lack of such knowledge can impact the effectiveness of inter-domain traffic engineering and resource management schemes.

In this chapter, we address the above issues by proposing a distributed control architecture, which we call Clearing House (CH), to coordinate intra- and inter-domain bandwidth allocation. An ISP can deploy a local CH to manage its backbone network, while relying on a third party (global) CH-node to coordinate inter-domain SLA negotiation and resource management. One of the basic design requirements of the CH architecture is to extend rather than replace the existing network devices, protocols, and implementations of inter-domain policies to minimize the development cost. The CH enhances the services and performance of the network by adding some functionality to the network access routers (or edge routers) and leveraging information from traffic monitoring devices.

The main goals that drive our design of the CH architecture are:

- **QoS Provisioning:** The CH attempts to provide an end-to-end coarse-grained QoS

assurance by performing aggregate resource reservation along the path from source to destination host networks. This approach tradeoffs the fine-grained QoS assurance in order to preserve scalability and reduce signaling complexity. Per-flow admission control is performed only at the edge routers, but it should take into consideration the reservation status and traffic fluctuations within the domain.

- **Scalability:** The CH has a hierarchical tree structure that can incrementally scale to support a large user base (i.e., large geographic regions and large volume of simultaneous calls). We strive to minimize the number of states maintained in each node of the CH and the backbone routers.
- **Efficient Network Utilization:** The CH attempts to optimize the overall throughput while preserving the QoS of admitted calls by performing admission control based on information of the entire network stored in the CH database, e.g., reservation status and available bandwidth of inter-domain links. The accuracy of this information depends on the time granularity by which the database is updated.
- **Robustness Against Malicious Flows:** To preserve the overall QoS assurance and minimize the impact of malicious activity on the performance of innocent flows, the CH should provide mechanisms to police admitted flows, detect and isolate the malicious flows as fast as possible.
- **Secure Real-time Billing:** The CH is a distributed database that can store the billing prices, quality and latency provided by various ISPs. It can inform ISPs and customers about the available bandwidth, bandwidth demand, and reservation costs. This aspect of CH has been explored in [99] and is not a focus of this dissertation.
- **Support for Multicast Operations and Mobility:** The CH infrastructure can be easily extended to support multicast operations by coordinating resource reserva-

tions and cost-sharing between the group members at different level of the multicast tree. The CH can also keep track of the dynamic path changes and modify resource reservations accordingly to support mobility. This is part of future work, and is out of the scope of this dissertation.

We address mainly the first four design goals in our thesis work. Specifically, we design control mechanisms within CH to establish and negotiate aggregate resource reservations both within an ISP and between neighboring domains in a hierarchical manner. For example, the CH ensures that there is sufficient link capacity on intra- and inter-domain links to carry the VoIP traffic so that the maximum loss rate and delay are below the acceptable thresholds for perceived voice quality. We will not discuss how the reservation requests are translated to a specific traffic control agreement (TCA) that can be understood by the edge devices, or how these TCAs are delivered to the edge routers.

In designing the CH architecture, we make the following assumptions:

- The networks are capable of providing different service levels through a combination of packet marking, scheduling and queue management mechanisms. We assume network edge routers can verify whether the QoS assurance agreement is met by measuring the packet loss, average queuing delay, delay variance, etc.
- Every routing domain has the capability to monitor and collect statistics of the incoming and outgoing traffic. We assume this information is trustable, and will be used by CH to negotiate resource reservations with neighboring domains.
- Control paths (e.g., reservation requests) and data paths are separated. We decouple call setup and resource reservation procedures to reduce the overall response time and increase the system throughput.
- Only latency sensitive applications (LSA) such as voice and video need resource reservations and are classified as “high-priority”. For the rest of our discussions, *flow*

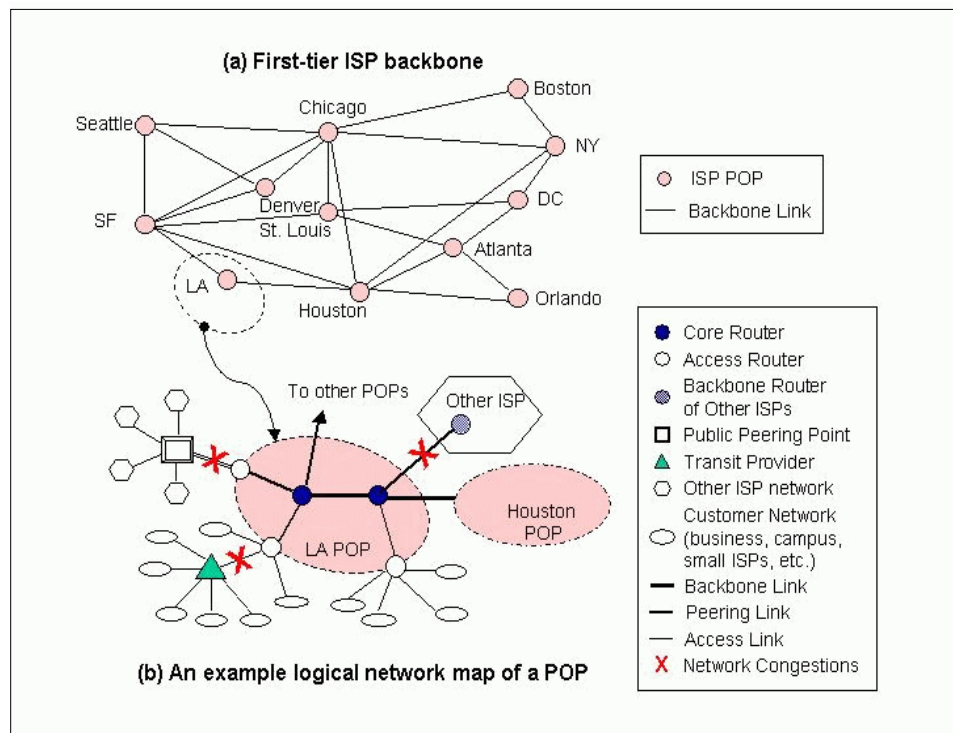


Figure 4.1: An example first-tier ISP backbone and its logical network map.

or *traffic* refer to high-priority packet streams that are affected by our session-level control mechanisms.

4.2 Design Rationale

In this section, we describe some basic properties of current Internet network topology, traffic characteristics and economic models based on our discussion with two major ISPs, and discuss how these considerations affect our design rationale.

4.2.1 Background: The Internet Structure

Internet Service Providers (ISPs) are segmented into tiers depending on the size of their network and number of subscribers. According to the xSP Forums¹, there are three

¹<http://www.xspsite.net/isp/tiers.htm>

ISP Tier-levels:

Tier-1: Backbones These are big ISPs either have their own nationwide backbone or over 1 million subscribers. There are about ten Tier-1 ISPs in the United States.

Tier-2: Regional Tier-2 ISPs usually own local regional backbone and/or support over 50,000 subscribers. These ISPs can also offer state or nationwide access services by peering or subscribing to Tier-1 ISPs.

Tier-3: Little, Local & Lots Local service providers that make up the majority of the ISPs belong to this category. Tier-3 providers typically support less than 50,000 users and offer local services only.

A typical first-tier ISP backbone² in the United States generally has 15-25 Points-of-Presence (POPs) located in major cities throughout the country, as shown in Figure 4.1. The fan-out structure of POPs varies from city to city. For example, the number of edge routers (ERs) connected to the core routers (CRs) inside a POP, usually in the range of 10-20 [101], depends on the number of customers in that region of the country. A POP generally consists of high-speed backbone links (0.6-10 Gbps) connected to the core backbone and low-bandwidth edge links (45-155 Mbps) connected to Local Access Providers (LAP) or customer networks through ERs. 30-50 of such edge links can be terminated at the same ER. Customers that have direct connections to a POP include corporate networks, university campuses, web-hosting co-location sites, and modem pools. Individual users gain Internet access through LAPs. As mentioned in [102], the ISP backbone may also connect to neighboring private peers, public exchange points or transit providers via separate peering links. Past studies have indicated that it is at these interconnection points, where large amounts of traffic converge and the backbone pipes meet the narrow access links, that con-

²For proprietary reasons, we do not have access to the exact backbone topology. Example ISP network maps are available from <http://www.cybergeography.org/atlas>.

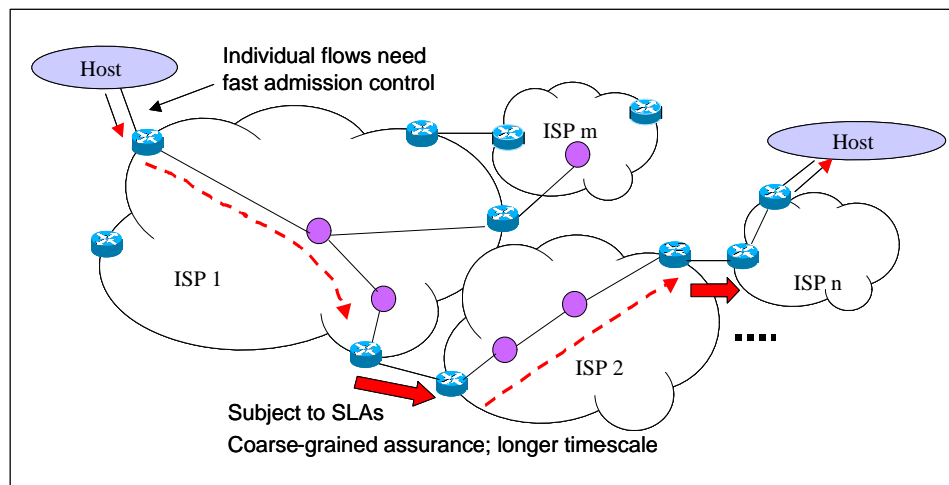


Figure 4.2: Multiple-ISP scenario.

gestion occurs³. This often results in packet loss and unreliable QoS. The CH architecture is designed to improve end-to-end performance by rationing the number of high priority flows admitted by the edge router and managing network resources on congested links based on continuous network monitoring.

The task of resource provisioning is not confined to be just within an ISP domain, since end-to-end connections often span multiple domains, as shown in Figure 4.2. We treat the traffic demand coming from a private peer or transit provider differently from the high priority flows generated by end-hosts, because the resource allocation for both cases happen in vastly different time-scales and granularity. In the former, the traffic is usually subjected to peering agreements or Service-Level-Agreements (SLAs) [10] that reflect aggregate traffic performance (e.g., maximum round-trip delay). SLA renegotiation and the corresponding resource allocation decisions take place over longer time-scale, e.g., weeks or months. In the latter, the reservation requests from individual flows usually need fast admission control decisions, e.g., within ms, and the aggregate traffic demand fluctuate in a smaller time-scale, e.g., hours.

³<http://www.internap.com>

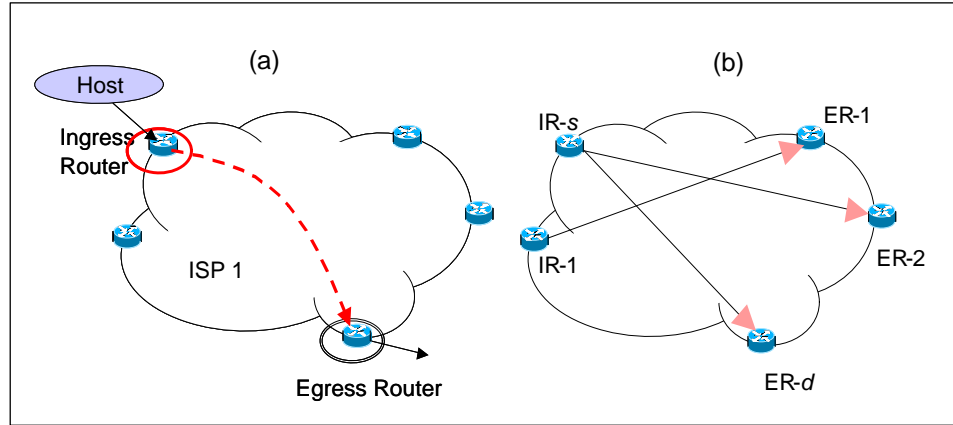


Figure 4.3: New service model: IE-Pipes(s, d) between specific ingress router IR- s and egress router ER- d .

4.2.2 Key Design Decisions

The above observations about the Internet have led to three major design decisions that contribute to the scalability and robustness of our architecture.

Decision 1. New Service Model: IE-Pipes

Our first design decision is to treat ingress and egress routers within an ISP domain as endpoints instead of individual hosts. An ingress router (IR) is the point at which a flow enters into a domain, while an egress router (ER) is the exit point from the domain (Figure 4.3a). We define a new service model called IE-Pipe(s, d) that provides performance assurance for aggregate traffic between a specific IR- s , and a specific ER- d (Figure 4.3b).

Decision 2. Hierarchical Logical Structure

The Clearing House has a hybrid of flat and hierarchical logical structures. We introduce a local hierarchy within large ISP domains to distribute network management tasks to various CH-nodes. This helps reduce the amount of state information maintained at each CH-node, and there is no single point of failure. In addition, the distributed and hierarchical nature of the CH allows us to build in redundancy and system fault tolerance.

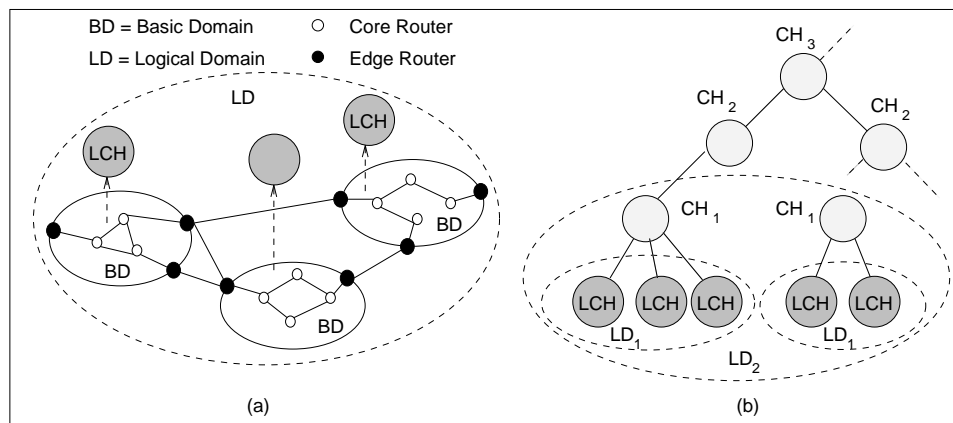


Figure 4.4: (a) Local Clearing House (LCHs) associated with their basic domains that lie within a single logical domain. (b) An hierarchical CH-tree with multiple levels of logical domains.

In our model, various basic domains (based on administrative or geographic boundaries) are aggregated to form *logical domains (LD)*, as shown in Fig. 4.4. These logical domains are then aggregated to form larger logical domains and so forth. This introduces a hierarchical tree of the LDs and a distributed CH architecture is associated with each LD. Individual CH-nodes can be thought of as agents that maintain aggregate reservations for all the links within the same domain at a particular hierarchical level. The reservations between neighboring domains are monitored by the parent CH-node. This hierarchical tree of CH-nodes form a “virtual overlay network” on top of existing wide-area network topology.

In the multiple-ISP scenario where the ISPs do not share mutual trust, the top-level CH-node for each ISP maintain peer-to-peer relationships with one another to regulate resource allocation for aggregate traffic exchanged between two domains. In Section 4.3.1, we illustrate how the logical CH-tree is mapped to the physical network, and how CH can be deployed by various independent ISPs.

Decision 3. Decoupling Reservation and Admission Control

For scalability reasons, the CH does not maintain per-flow reservation at any routers. Instead, aggregate reservation is set up for a group of flows that share the same link and request for the same class of service (high priority). We assume that the path for a specific pair of ingress and egress routers (or IE-Pipes) can be determined from the underlying routing protocol, and it remains stable relative to the individual flow lifetime. The core routers only need a simple two-level priority scheduler to provide Expedited Forwarding (EF) service to high-priority packets. As a result, the edge routers only maintain aggregate state information, while the core routers remain stateless.

The actual resource allocation is decoupled from the admission control process, which is necessary to ensure that there is sufficient network resources to deliver the end-to-end QoS assurance. In our architecture, admission control for individual flows are performed only at the edge routers but it leverages the knowledge of the global *traffic matrix* and topology within an ISP domain. A traffic matrix captures the distribution of aggregate traffic demand between different pairs of ingress and egress routers. The CH infers this traffic matrix through passive monitoring of the traffic arrivals.

4.3 Clearing House Architecture

The CH is a distributed control architecture that performs four major resource management tasks:

- **Monitoring and Measurements**

The CH collects aggregate traffic statistics through passive monitoring and measures the network performance such as packet loss rate and delay. It also estimates traffic matrix within each ISP domain by inferring the traffic demand distributions between every pair of ingress and egress routers.

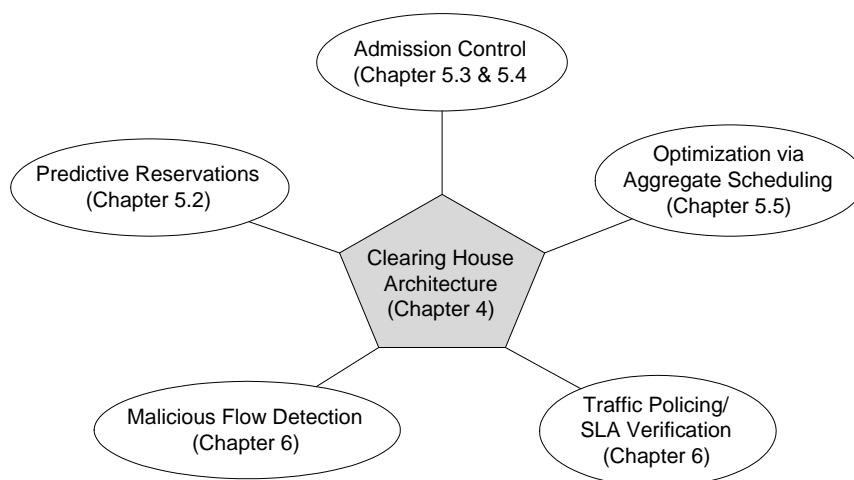


Figure 4.5: Thesis roadmap: The CH architecture and its various resource control mechanisms.

- **Intra- and Inter- Domain Aggregate Reservations**

Since LSA traffic has more stringent delay bounds, best-effort traffic must not preempt LSA traffic, and hence the latter should be served in a higher priority class. The CH architecture provides aggregate reservations for high-priority traffic within and across domains.

- **Admission Control**

Since the LSA traffic must co-exist with current best-effort traffic, we need to limit bandwidth allocated to the LSA traffic to prevent starvation of best-effort class. In addition, the admission of high-priority flows that compete for network resources should be controlled to ensure satisfactory end-to-end performance. The CH leverages the knowledge of traffic matrix in performing admission control at the edge.

- **Traffic Policing for Malicious Flow Detection**

Traffic policing is useful for monitoring admitted flows and detecting malicious behavior. The words *malicious* and *misbehaving* are used interchangeably to describe flows that violate their allocated share of bandwidth.

Figure 4.5 shows our thesis roadmap and the various CH mechanisms that we will address in Chapter 5 and 6. In the rest of this section, we illustrate how the hierarchical CH-trees are formed and discuss how the various CH control blocks interact with one another (the shaded area in Figure 4.5).

4.3.1 Hierarchical CH-Tree

First, we define several terms that we use in our discussions:

- A *basic domain* refers to a basic routing domain in the network. For example, a basic domain can be a small subset of backbone networks owned by a specific Internet Service Provider (ISP) which serves multiple host networks. We assume that the Internet can be divided into non-intersecting basic domains.
- A *logical domain (LD)* is a collection of adjacent basic domains that are clustered to form a larger domain, which may refer to geographic boundaries (e.g., states, or small countries) or for administrative reasons (e.g., campus, company etc). On the other hand, a big ISP backbone network can span across multiple domains.

The various logical domains can be clustered to form a larger logical domain. We can repeat the same process until we are left with one logical domain that represent the whole network. Together, these domains form a hierarchical tree, which we call a *CH-tree*. A distributed CH architecture is associated with every LD represented by a node in this tree. A CH-node at a particular level of the CH-tree maintains the reservation states of the LD, which is the union of all the sub-LDs whose states are maintained by its children CH-nodes. The actual number of CH-nodes in the distributed architecture will vary as a function of the size of the LD, and the level of the LD in the hierarchy. Mirror sites can be added to every CH-node to support fault tolerance and higher availability.

A CH in the hierarchy aggregates all inter-LD call requests to a particular domain and sends this aggregated request to the parent CH. In other words, all call requests between two LDs would be aggregated as a single request at a parent CH. Therefore, a CH of a LD that is a collection of K sub-LDs would contain $O(K^2)$ aggregate reservation requests. Only the CH at the local operators (at the leaf nodes of the CH-tree) maintain per-flow state information.

Although it is easy to extend the depth of the CH-tree to represent the whole network, our preliminary analysis considers the case of a two-level tree with one parent CH-node associated with an ISP domain and multiple children nodes (associated with basic domains). We quantify the performance of Clearing House and reservation strategies in this simple case and the simulation results are presented in Chapter 5.

4.3.2 An Illustration

A Single ISP Case

First, we consider a single ISP scenario. Since all the routers and CH-nodes within an ISP belong to the same administration (with the same AS number), we assume they share the same trust entity and therefore can exchange sensitive information such as traffic statistics and preferred routes among themselves. For ease of discussion, we denote the leaf nodes of the CH-tree as *Local Clearing House (LCH)* and the CH-node on top of the hierarchy for an ISP is called a *Parent Clearing House (PCH)*. Within an ISP, a basic domain (BD) typically represents a local Point of Presence (POP) that is connected to downstream customer (host) networks. In general, a BD can be a subtree of a tier-1 or tier-2 ISP, or even the entire domain for a tier-3 ISP (as discussed in Section 4.2.1) that forms a regional network (e.g., a city or a county). There can be 10-25 nodes within a BD, depending on its geographical scope (city vs. county) and user population.

Figure 4.6a shows how a simple two-level CH-tree can be constructed within an

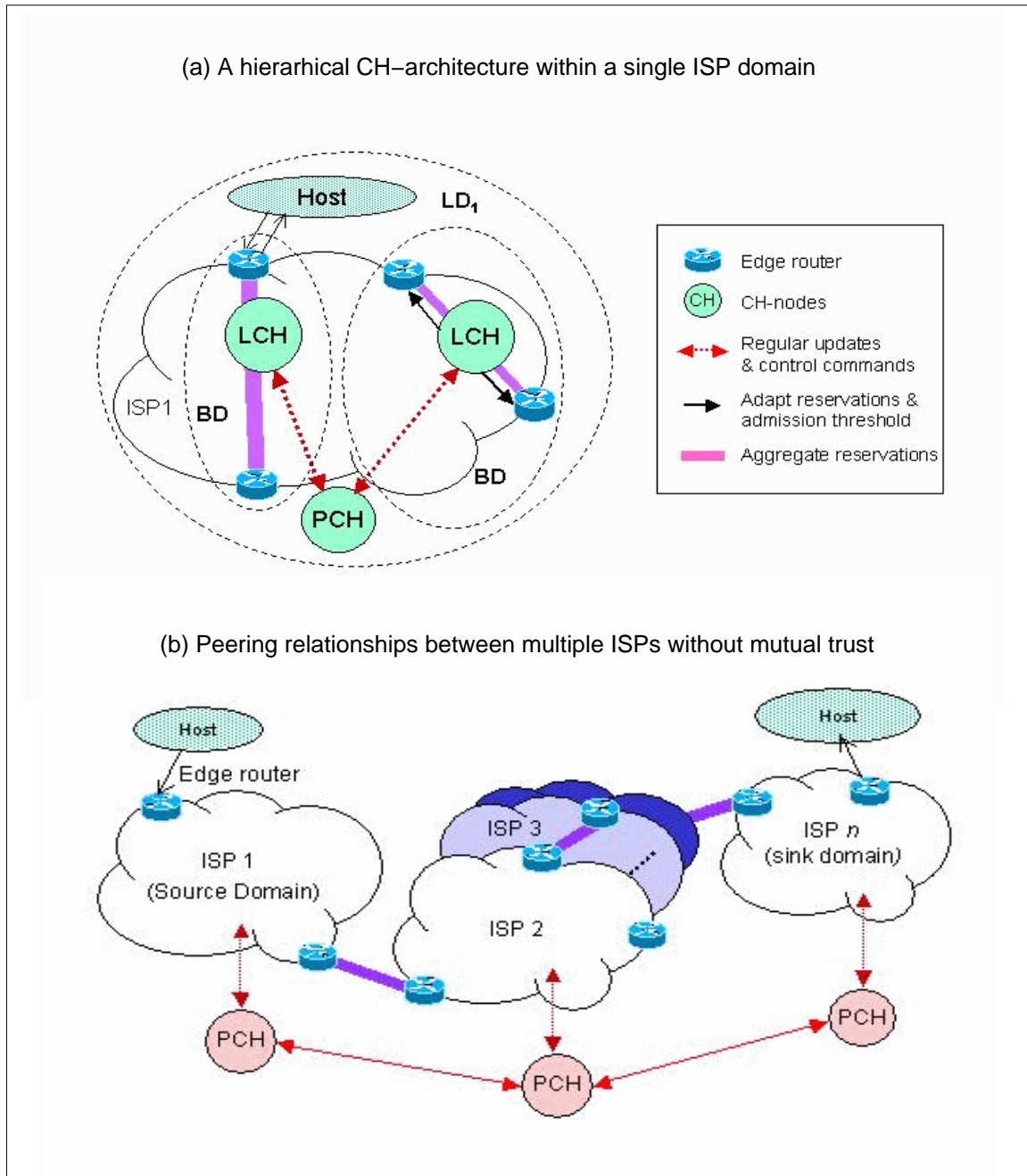


Figure 4.6: An illustration: A hybrid of flat and hierarchical CH-architecture within and across ISP domains.

ISP. This specific example shows two BDs (or POPs), each associated with a LCH, that are clustered to form a LD that represents the entire ISP. The LCHs send regular updates such as estimated traffic demand distributions and reservation status to upper-level CH-nodes (in this case, the PCH). The PCH collects traffic reports from all LCHs and construct the traffic matrix for the entire ISP.

Multiple-ISP Scenario

The local CH-hierarchy within ISP domains are hidden in Figure 4.6b. The independent service providers may not trust each other to share information about internal routing topology for traffic distributions. Therefore, we assume a flat structure at the top level where the PCH associated with each ISP form a peer-to-peer network to coordinate inter-ISP resource allocations. It is sufficient to reveal only aggregate traffic statistics, e.g., total bandwidth requirement to exchange high-priority traffic between any given pair of ISPs.

4.3.3 Local, Intermediate, and Parent Clearing House

We assume that the basic domains are non-overlapping to ensure that a user at a particular location has a unique LCH to contact for resource reservation or billing purposes.

The LCH is responsible for the following set of operations:

- An LCH keeps track of the amount of existing reservations and the available bandwidth on all the links between edge routers within its own BD. Based on the statistics of the intra-domain traffic, an LCH performs advance resource reservations on the intra-domain links. It also makes local admission control decisions when a new intra-domain reservation request arrives.
- An LCH also monitors the aggregated incoming and outgoing traffic exchanged with other neighboring BDs and uses these statistics to estimate the future bandwidth

usage. The predicted bandwidth requirement for high-priority traffic that traverse between its own BD to every other BD within the ISP is reported to the CH-node higher up in the hierarchy (PCH in this example). If there are K BDs within the ISP domain, the traffic report would be $K - 1$ vector.

- Based on the available network resources on the end-to-end path, the PCH will adjust the inter-domain reservations accordingly and sends updates to the LCH. Upon receiving acknowledgments from the PCH, the LCH will adapt resource allocation on its edge routers as well as the admission threshold for different IE-Pipes.
- If the existing inter-domain load is less than the allocated bandwidth, new requests will be admitted. Otherwise, the LCH aggregates inter-domain reservation requests as a single request and forwards it to the PCH. If there are sufficient network resources on the end-to-end path, the LCH will receive acknowledgments from the PCH to enhance the aggregate reservations, and the new requests will be admitted. Otherwise, the requests will be rejected.

In general, an intermediate CH-node acts as the coordinator among the various BDs and handles resource allocation for all inter-domain calls:

- A CH-node keeps track of the links that run between children sub-domains and their corresponding reservation status and network performance such as latency, average queuing delay, and packet loss rate.
- Based on the traffic statistics collected from all the children-LCHs, a CH estimates bandwidth usage on a particular inter-domain link and performs advance reservation accordingly (see Chapter 5).
- A CH-node aggregates reservation requests received from its children LCHs, and performs advance reservations for the inter-domain links that lie within its LD. If the

reservation request involves links that connect to neighboring LDs at the same level, the reservation request will be forwarded to the upper-level CH. A CH-node services reservation requests for aggregated traffic instead of individual calls.

The parent CH-node (PCH) sits on top of the hierarchy for a particular ISP and adapts trunk reservations between different domains. In addition to the general tasks for CH-nodes described above, the PCH can advertise the costs of reserving bandwidth on their internal links to other neighboring PCHs. For example, the service providers can offer various prices based on the domain of the final destination (e.g., call Canada 7/9 cents/min) and the traffic load [103]. The PCH can then choose the optimal route that satisfied the performance constraints while minimizing the total costs.

4.4 CH Control Flows

Our architecture requires explicit REQUEST and TEARDOWN messages for each high-priority flow for admission control and accounting purposes. These control messages are generated by either a customer router or a proxy and are sent as UDP packets at the same priority level (high) as the data packets.

Figure 4.7 shows the essential control blocks within the LCH and how it interacts with the edge routers. It also shows the REQUEST, ACCEPT/REJECT and CONFIGURE control messages between the LCH, the edge routers and the host proxy. When a new REQUEST message arrives, the LCH performs admission control based on its current knowledge of intra- and inter-domain reservations and existing load. It will then respond with an ACCEPT or REJECT message. If the flow is accepted, the LCH needs to update the traffic-policing unit in the associated edge routers through CONFIGURE message. If necessary, it uses the same message to adapt the resource allocation at the routers.

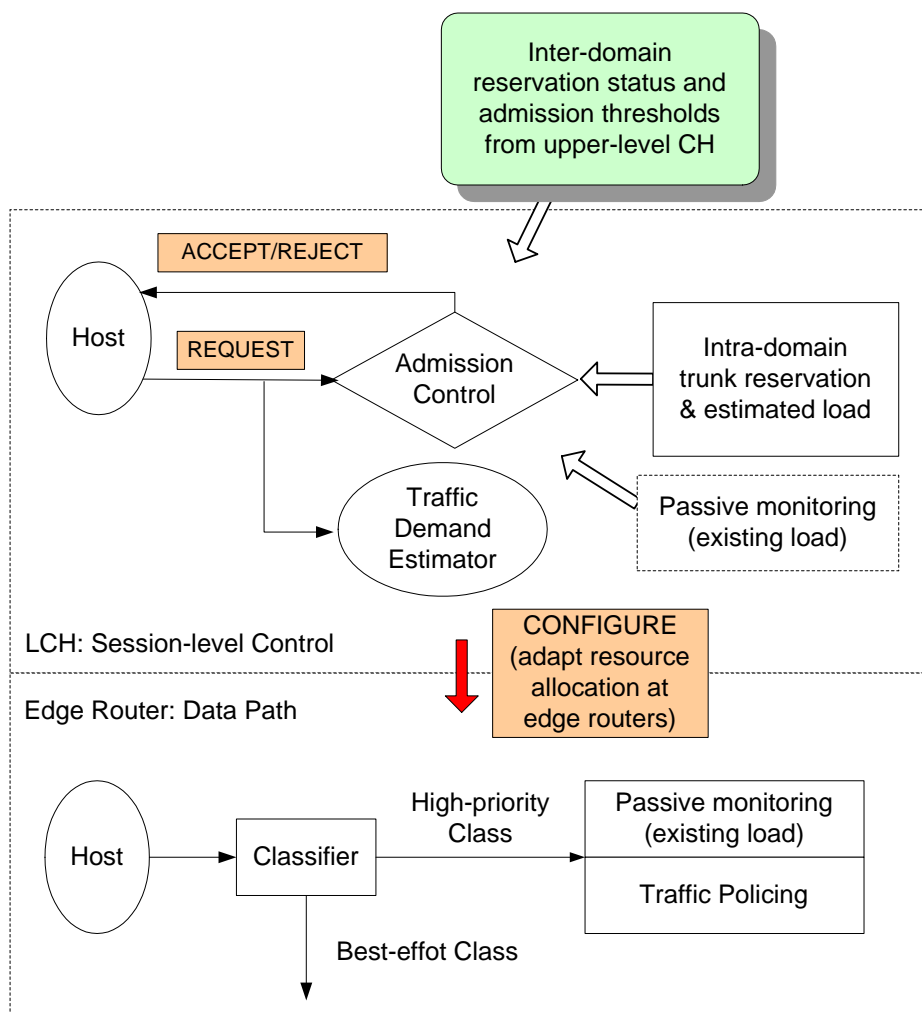


Figure 4.7: A logical view of the Local Clearing House (LCH) and its interaction with edge routers.

4.4.1 Resource Reservations

The CH architecture can support two types of reservations: advanced and immediate reservations. An advanced reservation (AR) is time-limited and resources are allocated in advance based on statistical estimates of aggregate traffic over a particular link. We use advance reservations to reduce the call setup time, and the potential violation of QoS assurance if the traffic arrives before the resources are properly reserved. Such approach has been used for resource management in Virtual Private Networks (VPNs) as reported in [46]. Traffic statistics can be easily obtained by leveraging the existing traffic monitoring and measurement systems, through either third party organizations, e.g., MIDS Internet Weather Report (IWR) [104], Internet Traffic Report [105], or the ISPs themselves, e.g., Cable & Wireless USA [106] and AT&T IP Services [107]. We can also gather information from end nodes using software toolkit such as SPAND [108], which enables the networked applications to report the performance they perceive as they communicate with distant Internet hosts. Advance reservations only track the aggregate traffic pattern at a large time-scale (e.g., different hour of the day) and do not reflect the rapid fluctuations of local traffic volumes produced by end-users. Immediate reservations (IR), on the other hand, can be made on demand when the existing reservations become insufficient to accept the new admission requests. The local CH-nodes performs admission control to ensure that QoS assurance to the existing connections are not violated.

For our initial analysis, we consider the case with a two-level CH-tree within each ISP. We evaluate, with simulations, the performance, robustness and overheads of the CH architecture. The simulation framework and results are discussed in Chapter 5.

4.4.2 Caching and Aggregate Scheduling

We can employ two enhancements to improve the performance of the Clearing House, namely caching and RxW scheduling [109]. An LCH or GCH can cache intra-domain

and inter-domain computed paths for previous reservation requests. This can reduce the service time of a reservation request at a CH. Since the number of logical domains maintained by a CH is small (10-50), a local cache can typically store all inter-domain paths. A local cache in a LCH can also store the price listings of various service providers to different destinations. RxW scheduling [109] is a very good algorithm for increasing the throughput of the CH. It schedules the aggregated call requests with the maximum value of $R \times W$, where R is the number of requests aggregated and W is the maximum waiting time of an aggregated request. This scheduling algorithm maximizes the throughput (number of call requests) serviced without unduly affecting the response time for call requests. Chapter 5 provides detailed discussions on how aggregate scheduling helps improve the CH efficiency and reduce the overall response time of reservation requests.

4.4.3 Admission Control

Whenever a sender wants to make a call to a receiver, there should be sufficient resources (e.g., link capacity and buffer space) along the particular path from the sender to the receiver to avoid packet losses and delays. Since on-line resource reservation is very costly, the goal of our design is to minimize the amount of per-link reservation that needs to be made for a particular call. Based on the reservation status within a domain, a particular path is chosen such that the number of new per-link resource reservations is minimized. If the LCH fails to locate any links with sufficient resources reserved to complete a chosen path, the ER will block the new call. The admission control decisions involve some trade-offs in the QoS assurance and the number of rejected calls.

So far, we have presented a general CH-architecture that can deploy both parameter-based and measurement-based admission control. We concentrate on the latter approach in this dissertation. The CH architecture allows an ISP to dynamically infer the traffic matrix within its own domain based on a collection of partial views from its sub-domains. With

this knowledge, an ISP can make a better admission control decision for the incoming flows from its customer networks. A detailed description of the proposed scheme, Traffic-Matrix based Admission Control scheme (TMAC), can be found in Chapter 5.

4.4.4 Traffic Policing and Malicious Flow Detection

Another equally important task is to police the admitted flows to make sure that each flow uses its right share of allocated bandwidth and not more than that. To reduce the amount of state information maintained at the edge routers, we propose to aggregate flows for group policing. Chapter 6 presents the Malicious Flow Detection via Aggregate Policing (MDAP) scheme that is designed to uniquely identify malicious flows without keeping per-flow state at the edge router.

4.5 Summary and Discussions

We have designed a Clearing House architecture that coordinates intra- and inter-domain resource reservations for high-priority traffic. The scalability of the architecture is attributed to its hierarchical CH-tree structure and the aggregation of reservation requests at multiple levels of the logical tree. We use a Gaussian predictor to estimate bandwidth usage and set up reservations in advance to reduce the overall reservation setup time. The details of the predictive reservation scheme and evaluation of its effectiveness is presented in the next chapter (Chapter 5). Further analysis on how the throughput can be improved using aggregate scheduling algorithms is also reported.

In summary, the basic strengths of the Clearing House approach are:

- The state information that needs to be maintained by the entire ISP domain is shared between various CH-nodes in the local hierarchy. Since every CH-node maintains only the state for its own domain, this allow the entire CH-tree to scale better to a larger

user base.

- The hierarchical model with aggregate reservations provides scalability of the architecture. Core routers do not need to maintain per-flow state information. The architecture supports easy insertion and deletion of the domains from the *CH-tree*. If a particular CH-node gets overloaded due to the growth of user-base, it is possible to split a logical domain (LD) into two or more sub-LDs, and create a new CH-node for the newly created LDs.
- The queue size of call requests in every CH is bounded due to aggregation of call requests at the children CH-nodes. This architecture is optimized for making decisions based on locality. End-to-end resource reservations can be set up quickly through the CH architecture, and therefore reducing the call setup time.
- The CH leverages knowledge of traffic matrix and routing topology within an ISP for admission control.
- Caching of inter-domain paths can enhance the performance of the system considerably.

4.5.1 VPN: An Example Application

Although we present the CH as a general architecture, one specific example where CH will be useful is for IT managers to manage a WAN (wide-area network) that interconnects corporate offices, remote and mobile employees. Corporations have turned to Internet VPNs to deliver performance, security and manageability to their various sites scattered across the country. However, existing SLAs⁴ [10] between service providers (ISPs) and customers have focused on backbone performance guarantees, and do not reflect the end-to-end performance of individual applications. In addition, some fraction of the traffic may

⁴A service level agreement (SLA) is an explicit statement of the expectations and obligations that exist in a business relationship between two organizations: the service provider and the customer

traverse multiple routing domains that belong to different ISPs. IT managers still face the challenge of provisioning the total capacity (VPN tunnels) efficiently among the various types of traffic to meet application requirements such as latency and reliability characteristics. A CH-architecture can be deployed in this case to handle intra- and inter- domain resource allocation. For example, IT managers can treat each corporate site as a basic domain, and introduce a CH-node at each site to monitor the traffic flow, adapt resource allocation, and re-negotiate SLAs with the corresponding ISPs when necessary. Various sites can be aggregated to form a larger LD, or several LDs, depending on the layout of the corporate network. The CH-nodes associated with these LDs can coordinate the aggregate resource allocation between domains that reflect on end-to-end performance requirements.

4.5.2 Other Resource Control Problems

In this thesis, we focus on using CH to allocate link capacity to provide statistical packet loss and delay guarantees to high-priority traffic. However, CH can be easily modified to address other resource allocation problems. The following are some examples:

Disk space allocation in storage area network: In today's high-technology economy, the storage needs for companies are growing exponentially due to the increasing dependence on "information". A storage area network (SAN)⁵ is designed to meet this challenge by enabling any-to-any interconnection of servers and storage systems. It supports centralized management of both local and remote storage resources within an enterprise. The CH architecture can be used to perform load-balancing between different servers and control disk space allocations. For example, CH can exploit the data access patterns to determine how the storage system should be partitioned among a group of users. This can improve the overall disk utilization. CH can also be used to manage the sharing of link capacity between users in SAN to ensure rapid

⁵<http://www.storage.ibm.com/ibmsan/>

access to data and to avoid losses.

Distributed cache management for content distributions: Content distribution network (CDN) providers, such as Akamai⁶, address the needs for e-commerce companies to deliver streaming media, rich Web content and Internet applications to end customers with high performance and reliability. Akamai deploys thousands of servers (web caches) across hundreds of access networks worldwide. By placing the Web objects and applications at these caches that are close to the end users, Akamai eliminates the impact of congestion points in wide-area network on the performance of content delivery. The CH framework can be applied in this case to choose the optimal placement of these caches to achieve the required performance with the minimum number of caches. Using real-time measurements of cache performance, traffic load, and user access patterns, CH can perform load balancing to improve the overall efficiency of CDN networks.

⁶<http://www.akamai.com>