Chapter 2

QoS on All-IP Networks

Telecommunication is moving toward a converged network which uses a single global IP based packet switched network to carry all types of networks to replace the traditional separated packet switched and circuit switched networks. With the increasing commercial deployment of wireless networks, the issue of providing multiple services (including voice, video, data, etc.,) is becoming more and more important [1]. So the concept of "Quality of Service" ([10], [36],) is being widely discussed and implemented.

2.1 Quality of Service

The Internet Protocol (IP) [36], and the architecture of the Internet itself, is based on the simple concept that datagrams with source and destination addresses can traverse a network of (IP) routers independently, without the help of their sender or receiver. The Internet was historically built on the concept of a dumb network, with smarts at either end (at the sender and receiver) [4]. There is a price to pay for this simplicity, however. Routers are allowed to discard IP datagrams en route, without notice to sender or receiver [19]. IP relies on upper-level transports (e.g.

TCP) to keep track of datagrams, and retransmit as necessary. And these reliability mechanisms can only assure data delivery; neither IP nor its high-level protocols can ensure timely delivery or provide any guarantees about data throughput. IP provides what is called a best effort service [20]. It can make no guarantees about when data will arrive, or how much it can deliver. This limitation has not been a problem for traditional Internet applications like web, email, file transfer, and the like [4]. But the new breed of applications, including audio and video streaming, demand high data throughput capacity (bandwidth) and have low-latency requirements when used in two-way communications (i.e. conferencing and telephony) [5]. Public and private IP Networks are also being used increasingly for delivery of mission-critical information that cannot tolerate unpredictable losses. Unlike pure virtual circuit technologies like ATM and Frame Relay, IP does not make hard allocations of resources [5]. This provides much more efficient use of the available bandwidth, and it is more flexible also. Typical network traffic is bursty rather than continuous. IP is datagram-based so it uses the available bandwidth most efficiently, by sharing what is available as needed. This also allows IP to adapt more flexibly to applications with varying needs. However, it also leads to some unpredictability in service. The capacity to tolerate this unpredictability relates to the level of guarantee they require.

In the simplest sense, Quality of Service (QoS) [36] means providing consistent, predictable data delivery service, in other words, satisfying customer application requirements. QoS is to the ability of a network element (e.g. an application, host or router) to have some level of assurance that its traffic and service requirements can be satisfied. To enable QoS requires the cooperation of all network layers from top-to-bottom, as well as every network element from end-to-end. Any QoS assurances are only as good as the weakest link in the chain between sender and receiver. QoS does not create bandwidth. QoS only manages bandwidth according to application demands and network management settings, and in that regard it cannot provide certainty if it involves sharing. Hence, QoS with a guaranteed service level requires

bandwidth allocation to individual data streams [20]. A priority for QoS designers has been to ensure that best-effort traffic is not starved after reservations are made. QoS-enabled (high-priority) applications must not disable the low-priority Internet applications. Among many challenges yet to be overcome, the QoS problem is one of the most critical problems to be solved [10], [12], [25], [26], etc.

2.2 Resource Allocation & Routing with QoS Guarantees

Resource allocation decisions ([25], [31]) are concerned with the allocation of limited bandwidth so as to achieve the best system performances. QoS routing concerns the selection of a path satisfying the QoS requirements of users. The path selected most likely is not the traditional shortest path. Depending on the specifics and the number of QoS metrics involved, computation required for path selection can become prohibitively expensive as the network size grows [12]. The path selection process ([3], [7], [26]) involves the knowledge of the connections' QoS requirements and characteristics on the network (,e.g., limited available bandwidth and delay).

Broadband integrated services networks are expected to support multiple and diverse applications, with various QoS requirements. Accordingly, a key issue in the design of broadband architectures is how to provide the bandwidth in order to meet the requirements of each connection, and, moreover, how to meet that goal in an efficient manner.

The ability to provide end-to-end guarantees [14], such as delay, depends to a large extent on the scheduling policy and service discipline employed in the nodes. Such disciplines are characterized by bounds on the maximal delay that any node can incur, and hence a corresponding bound on the end-to-end delay can be derived. Such a bound provides a valuable tool for quantifying the quality of a path in terms

of its ability to meet the QoS delay requirement. The corresponding routing problem is, therefore, to identify the path that has the best performance, according to that bound and with respect to the QoS requirement. Typical schedulers map delay guarantees into rate guarantees, and have nodes advertise the residual rate they have available [19]. In particular, the Guaranteed Service class proposed for the Internet is based on such rate-based principles.

Network design today often considers the problem of designing networks that carry elastic traffic (see [2], [8], [14], [15], etc.) If the network is also used for other types of communication that require guaranteed quality of service, the network design problem can be decomposed into two parts: first, design the network to carry non-elastic traffic in such a way that all demands for that communication are satisfied. Next, use the spare capacity to carry elastic traffic of the IP protocol. Resource allocation models may be used to solve network design problems (see [9], [17], [25], [27]).

Network management must stay within a budget of expenses for purchasing link bandwidth. Network dimensioning with elastic traffic may be thought of as a search for such network flows that will maximize the network throughputs (the sum of all flows in the network) while staying within a budget constraint for the costs of link bandwidth. However, such a problem of formulation would lead to the starvation of bandwidth between certain network nodes. Looking at the problem from the user perspective, the bandwidth between different users should be treated as fairly as possible. Whatever the user's preference, it would be expressed in terms of fairness for a certain set of criteria which depend on the individual connection. Let us first consider providing fairness for all connections between competitive activities. Network management must consider two goals: increasing throughput and providing fairness [25]. These two goals are clearly conflicting, if the budget constraint has to be satisfied.

Consider a fixed network topology $G = (V, E)$, where V and E denote the set

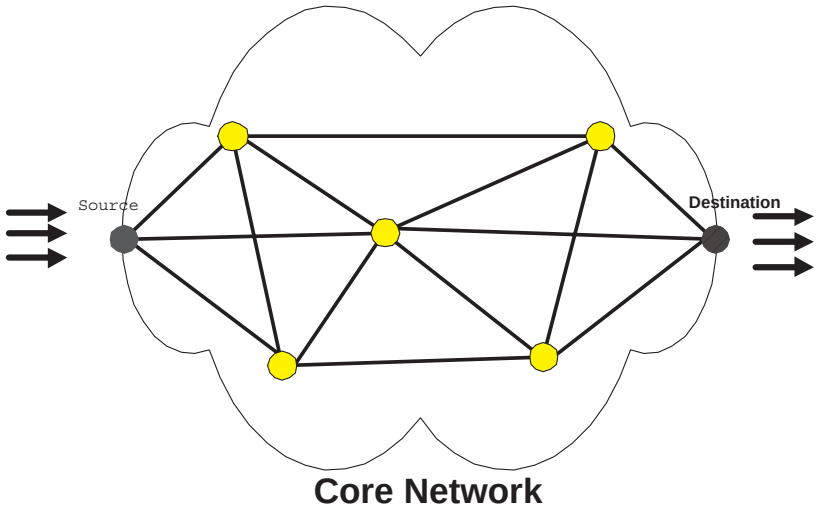

Figure 2.1: A Simplified Network Architecture

of nodes and the set of links in the network respectively. The network topology investigated in this work is like Figure 2.1. Given the maximal possible capacity of each link. Suppose we know that, for each link, the cost taking account of delay and the purchasing cost of bandwidth. In this network, there are m different classes which have their own QoS requirement. In each class, every connection is allocated the same bandwidth and has the same QoS requirement. Suppose each connection is delivered between the same source and destination in this (core) network. Under a limited available budget, we want to allocate the bandwidth in order to provide each class with maximal possible QoS. The purpose of this work is to show that a methodology that allows the decision maker to explore a set of solutions could satisfy preferences with respect to throughput and fairness, and choose the solution which the decision maker finds best.