Chapter 1
Introduction

Over the last decade, the growth and uses of computer networking have increased at an explosive rate, and the global Internet has become an infrastructure of importance equal to that of the power grid and the transportation system. The success of the Internet has to do with the advances in the underlying technology as well as the innovations in the services available to users. The last few years have witnessed dramatic and continuing improvements in networking technology, including optical and wireless transmission systems and link capacities, router and switch speeds, and network protocol and software capabilities. These improvements have made it possible to operate and manage effectively networks that may support performance differentiation and guarantees in serving heterogeneous users with a wide range of requirements. As the Internet has evolved into a ubiquitous and reliable communication system, it has changed the way people interact, access information, or conduct business.

Computer networking has enabled new and exciting applications that have transformed every aspect of business and commerce, education and scientific exploration, entertainment and social interactions, and government and military services. The increased reliance on, and utility of, computer networks has created an insatiable user demand for better and faster Internet connectivity and services, and an entire industry has emerged to develop related products. Whereas workplace and University networked environments have traditionally provided Internet access to employees, researchers, educators, and students, recent years have seen a proliferation of home networks with high-speed broadband access, public wireless hot-spots, and smartphones with Web browsing and data networking capabilities that further extend the reach of the Internet in terms of both geography and user diversity.

Technology advances typically lead to the emergence of new business models and service offerings, which in turn influence the direction for further research and development in the underlying technologies, and so on. This interplay between technology and business models creates a virtuous cycle of innovation of which users are the primary beneficiaries. Interestingly, however, there has been a notable discrepancy between the designers and developers of networking hardware and software included in the equipment making up the Internet infrastructure, on the one hand,

and the Internet service providers (ISPs) that devise the products to make available to customers, on the other. This discrepancy manifests itself in terms of the *granularity* at which users may access the services offered by the network. Specifically, networks are typically designed with fine granularity in mind, whereas ISPs market products developed around models of coarser-grained services, as we explain in the following.

1.1 Continuous-Rate Packet-Switched Networks

Historically, packet-switched computer networks, including the Internet and legacy networks based on Asynchronous Transfer Mode (ATM) or Frame Relay (FR) technologies, are designed to be *continuous-rate*. In a continuous-rate network, users may request any rate of service (bandwidth), and the network must be able to accommodate arbitrary requests. Theoretically speaking, continuous-rate networks may allocate bandwidth at very fine granularities; for instance, one client may request a rate of 98.99 Megabit per second (Mbps), while another customer may ask for 99.01 Mbps. Taken to the limit, bandwidth in such networks could potentially be allocated at increments of 1 bit per second (bps). Clearly, the option of requesting arbitrary rates offers clients maximum flexibility in utilizing the available network capacity.

On the other hand, supporting bandwidth allocation at such extremely fine granularity may seriously complicate the operation and management of the network. Based on the above example, the network provider faces the problem of designing mechanisms to distinguish between the two rates (i.e., 98.99 Mbps vs. 99.01 Mbps) and enforce them in an accurate and reliable manner. However, the task of differentiating between the two users on the basis of these two rates may be extremely difficult, or even impossible to accomplish for traffic of finite duration, undermining the network's ability to support important functions such as robust traffic policing or accurate customer billing. Furthermore, given the unpredictability of future bandwidth demands in terms of their size, arrival time, and duration, link capacity across a continuous-rate network may become fragmented. Such fragmentation poses significant challenges in terms of traffic engineering, and may compromise the ability to achieve an acceptable level of utilization or meet users' quality of service (QoS) requirements.

To illustrate the challenges associated with operating a continuous-rate network, consider packet fair scheduling algorithms such as the weighted fair queueing (WFQ) [90] or its variants. The WFQ discipline can be used to allocate the capacity of a link to any number of competing flows in proportion to the weights assigned to each flow, as well as to guarantee a strict upper bound on the delay of each packet of a flow under certain conditions. However, existing implementations of the WFQ discipline or its variants have been designed under the assumption that flow weights can be arbitrary; in other words, they are designed to allocate the link bandwidth at the finest possible granularity. As a result, these implementations suffer from severe

scalability challenges which have impeded their wide adoption in Internet routers. We discuss packet fair queueing disciplines in more detail in Part III of the book, where we present a scalable implementation based on the concept of tiered services that is introduced next.

1.2 Tiered-Service Networks

Due to the issues involved in making fine-granular services available to a large, heterogeneous user population cost-effectively, in practice, most network operators have developed a variety of *tiered service* models in which users may select only from a small set of service *tiers* (levels) which offer progressively higher rates (bandwidth). The main motivation for offering such a service is to simplify a wide range of core functions (including network management and equipment configuration, traffic engineering, service level agreements, billing, and customer support), enabling the providers to scale their operations to hundreds of thousands or millions of customers. Returning to the previous example, a tiered-service network might assign both users requesting 98.99 Mbps and 99.01 Mbps to the next higher available rate, say, 100 Mbps. In this case, there is no need to handle the two customers' traffic differently; furthermore, the network operator only needs to supply policing mechanisms for a small set of rates, independent of the number of users.

For a more formal definition, consider a network that offers a service characterized by a single parameter, e.g., the bandwidth of the user's access link. A tiered-service network is one that offers p levels (tiers) of service, where typically p is a small integer, much smaller than the number n of (potential) network users (i.e., $p \ll n$). Let

$$Z = < z_1, z_2, \cdots, z_p > \tag{1.1}$$

denote the vector of service tiers offered by the network provider. Without loss of generality, we make the assumption that the service tiers are distinct and are labeled such that

$$z_1 < z_2 < \cdots < z_p. \tag{1.2}$$

Users wishing to receive service are limited to only these p tiers, and may subscribe to any tier depending on their needs and their willingness to pay the corresponding service fee. In particular, z_1 is the minimum and z_p the maximum amount of service that a user may receive. In the case of residential Internet access, for instance, z_1 may correspond to a minimum bandwidth for the service to be considered "broadband," while z_p may correspond to the capacity of the access link, e.g., as determined by limitations imposed by Asymmetric Digital Subscriber Line (ADSL) technology.

According to this definition, traditional telephone networks and transport networks based on Synchronous Optical Network or Synchronous Digital Hierarchy (SONET/SDH) technology [43] belong to the class of tiered-service networks. Indeed, such networks allocate bandwidth in discrete tiers that are multiples of a basic

unit rate that corresponds to the slot size in the underlying Time Division Multiplexing (TDM) system.

Motivated by the discrete nature of bandwidth allocation in TDM systems, an early study by Lea and Alyatama [71] investigated the benefits of "bandwidth quantization" in packet-switched networks. In their terminology, "bandwidth quantization" refers to sampling the (effectively) continuous range of possible rates to select a small set of discrete bandwidth levels that are made available to users; in essence, these levels correspond to the service tiers we defined earlier. This work was carried out with the goal of reducing the number of states required to analyze broadband ATM networks under the assumption of Poisson traffic, and it presented a heuristic based on simulated annealing to obtain a sub-optimal set of discrete bandwidth levels of service[1]. Because of the significantly reduced state space, it is possible to apply elegant and exact theoretical models to analyze the performance of a tiered-service network efficiently. The main contribution of this study was to demonstrate for the first time that this benefit comes almost for free, as even with a sub-optimal set of tiers the performance degradation (e.g., in terms of call blocking) compared to a continuous-rate network is negligible.

Current tiered service offerings by major ISPs can be broadly classified in two categories based on the tiering structure. The structure of one class of service tiers for Internet access, especially those targeted to business customers, is based on the bandwidth hierarchy of the underlying transport network infrastructure (e.g., DS-1, DS-3, OC-3, etc.), While this is a natural arrangement for the service provider, it is unlikely that hierarchical rates designed decades ago for voice traffic would be a good match for today's business data applications. The second class employs *exponential tiering* structures in which each tier offers twice the bandwidth of the previous one. The various ADSL tiers (e.g., 384 Kbps, 768 Kbps, 1.5 Mbps, 3 Mbps, 6 Mbps, etc.) available through several ISPs are an example of such an exponential structure. While such simple tier structures may be an appropriate choice for marketing purposes, the relationship between these exponentially increasing levels of service (and their price) and the usage patterns (and willingness or ability to pay) of the population of potential subscribers is open to debate.

So far, we have discussed tiering in the context of broadband Internet access services that are generally characterized by the bandwidth available on the customer's access link. However, the concept of tiering is equally applicable to other parameters that may characterize the QoS experienced by a customer's traffic and may be included in the service level agreement (SLA) negotiated between the customer and provider. Consider, for instance, a provider offering a service that guarantees an upper bound on the delay experienced by its customers' packets. On the one hand, the provider is unlikely to be able to support fine-granularity delay bounds (e.g., at the level of nanoseconds) within a network of realistic size even with the most sophisticated (and expensive) QoS mechanisms. On the other hand, users are unlikely to require (or afford) delay bounds at such a level of precision. A more reasonable approach would be to offer a small set of delay bound tiers that are tailored to specific

[1] In Chapter 3 we show that the problem considered by Lea and Alyatama in [71] can in fact be solved optimally, and we present an efficient algorithm for obtaining an optimal solution.

applications, e.g., voice, (stored) video on demand, (live) video conferencing, etc. Such delay bounds are likely to be tied to human perception abilities that allow for coarse granularities of tens of milliseconds, making it unnecessary for the network to have to distinguish packet delays at extremely fine precision.

Similar observations apply to other QoS parameters, e.g., level of protection of user traffic. In this case, it may be possible to define and offer a set of discrete grades of service from which customers may select based on the level of quality of protection [41] appropriate for their traffic. Tiered structures may be also be employed when the offered service is not characterized by bandwidth but by amount of traffic generated. As an example, in early 2008, Time-Warner, a major cable ISP in the United States, started a pilot program in Beaumont, Texas, under which it charges customers based on how much data they transfer (i.e., upload *and* download) [50,51]. For the pilot program, Time Warner put in place an exponential structure with tiers at 5 GB, 10 GB, 20 GB, and 40 GB of monthly traffic.

1.3 Multi-Tiered Pricing Schemes

ISPs around the world have introduced several forms of tiered services along the lines of the model we described above, with each tier associated with a higher level (amount) of service than the previous one with a corresponding increase in price. Multi-tiered price systems are prevalent for both business and residential Internet access, and arise naturally under both pricing schemes, capacity-based or usage-sensitive, that are prevalent for Internet services [62].

1. **Capacity-Based Pricing.** Capacity-based schemes relate pricing to usage by setting a price based on the bandwidth or speed of the user's connection link. This is accomplished by charging for the configuration (i.e., bandwidth) of the connection, but not the actual bits sent or received. This scheme relates to the tiered service model as follows: the service is characterized by the amount of access bandwidth, each of the service tiers $z_1, \cdots, z_p$, corresponds to a certain access speed, and users are charged based on the tier to which they have subscribed. Capacity-based pricing is the prevailing pricing policy for residential broadband Internet access services, although the pilot program by Time-Warner [50, 51] may be the beginning of a shift towards usage-sensitive pricing for residential markets.
2. **Usage-Sensitive Pricing.** Usage-sensitive pricing policies charge users for the actual amount of traffic they generate, hence price is a function of the actual bytes transferred over a certain time period, usually one month. In current practice, ISPs charge business customers (e.g., a video-on-demand provider) based on their traffic volume using a *95-th percentile rule* [52, 114][2] designed to disregard the low probability peak-load periods. Specifically, the ISP measures the

[2] Time-Warner's pilot program discussed above is an exception as the ISP charges based on the total amount of bytes uploaded or downloaded, not the 95-th percentile.

user's traffic volume over 5-minute intervals during each billing period (e.g., one month), and charges the user based on the 95-th percentile value among these measured values. Typically, ISPs have a tiered pricing structure [114] in which each of the service tiers $z_1, \cdots, z_p$, corresponds to a certain traffic volume and higher tiers are mapped to higher prices. Such a structure can be mapped to the tiered service model by considering a customer with a 95-th percentile value x such that $z_{j-1} < x \leq z_j$ as having "subscribed" to tier z_j and charging the customer accordingly.

Note that with capacity-based pricing, the tier (e.g., access speed) to which a user subscribes does not change over time (except, for instance, when a user upgrades to a higher speed). With usage-sensitive pricing, on the other hand, a user may be charged according to a different tier every billing period, i.e., depending on the actual traffic volume generated during each period.

If designed and applied appropriately, tiered services and corresponding multi-tiered pricing schemes have the potential to be a catalyst for Internet service innovation and penetration. On the provider side, tiered structures can be an effective tool for ISPs to optimize and specialize their offerings so as to capitalize on the increasing sophistication and requirements of various segments of Internet users, as well as to differentiate themselves from the competition. On the user side, tiered pricing is likely to spur adoption by providing a wide menu of customized services from which users may select based on needs and affordability. To realize this potential, it is crucial that both the service tiers and the corresponding prices be determined in a manner that takes into account simultaneously the (usually conflicting) objectives of users and providers. The purpose of this book is to provide insights into the selection and pricing of tiered structures for Internet services and offer solutions that consider the perspectives of both users and ISPs.