

# Chapter 1

## Introduction

Historically, different types of networks were developed to handle specific types of media. For example, telephony networks handled voice, cable networks supported only video, and computer networks transmitted discrete data. Today, advances in computer network technology have resulted in integrated networks that must accommodate a large number of users and a variety of media. Applications range from simple data transfer programs to sophisticated multimedia applications. Multimedia applications are of special interest since their presence is expected to increase in the future. For this reason, more functionality is needed from the network to support these applications.

### 1.1 Multimedia Applications

Multimedia applications incorporate various media such as, voice, video and data information. Example multimedia applications include, teleconferencing, video on demand, and broadcast video. These applications can be classified based on their interaction requirements [101]. Broadcast video is non-interactive, since the user typically has limited control over the audio, video and text presented. In contrast, teleconferencing is an example of an interactive application. Furthermore, video oriented applications can be classified based on whether the video is stored or live. Video on Demand (VoD) is the transmission of stored video to users, while teleconferencing is an example of live video transmission. Different combinations of these classifications are possible, for example VoD can be considered a

stored interactive multimedia application. Users can control the play of the stored video via rewind and fast-forward commands. Regardless of the classification, an increasing number of multimedia applications are video oriented. For this reason, video encoding techniques are needed to successfully transmit video over computer networks.

### 1.1.1 Video Encoding

The development of video encoding standards, such as H.263, MPEG-1 and MPEG-2, have made the transmission of video over computer networks possible [63, 87]. Compression techniques can reduce the bandwidth requirements of digital video from Mbps to a few hundred Kbps. This bandwidth reduction makes the transmission of video feasible over current computer networks. Video encoders can be classified as either Constant Bit Rate (CBR) or Variable Bit Rate (VBR). CBR encoders target a specific bit rate for the encoded video, which is achieved by varying the video quality. Such a method simplifies the network transmission requirements; however, the varying quality is unsuitable for certain multimedia applications (high definition video).

VBR encoding differs from CBR, since the objective is a certain video quality. Based on the frame contents of the video, the encoder will generate different bit rates in order to provide a constant picture quality. For this reason, rates produced by VBR encoders can have high, peak-to-mean ratios and autocorrelations [41, 89]. This signifies that, periods of time may exist when the bit rate of the encoded video is much higher than the average rate. For example, the encoded video given in figure 1.1 has a peak-to-mean ratio of 18.36 and a 2.35 coefficient of variation. As depicted in the figure, these burst periods are difficult to predict (both in amount and duration) and can occur at various times. Furthermore, these bit rates have been reported to have self-similar behavior [41, 89]. Self-similar behavior exhibits Long Range Dependence (LRD) or slowly decaying autocorrelation; however, implications of LRD on network transmission are still debated [25, 31]. In summary, the benefit of constant quality compressed video is at the expense of highly variable (bursty) and unpredictable bit rates. The efficient and reliable transmission of such sources is difficult, which is discussed next.

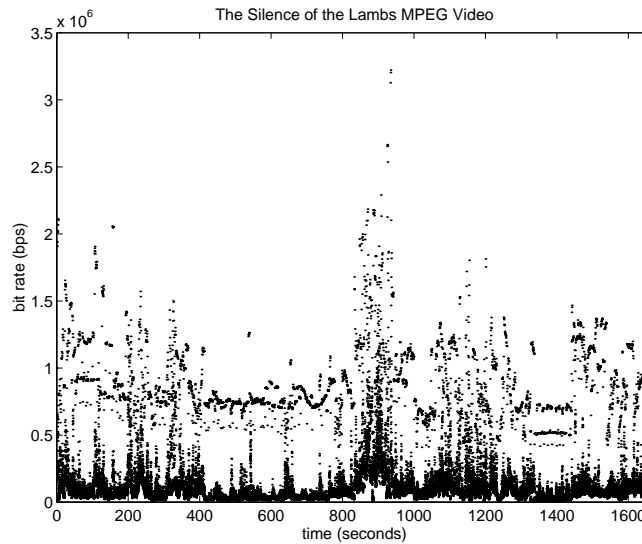


Figure 1.1: Bandwidth requirements for the transmission of an example MPEG video.

## 1.2 Quality of Service

To accommodate the variety of current and future applications, networks must now provide more functionality than previously offered. For example, many applications require Quality of Service (QoS) guarantees for their proper execution. QoS performance measures include bounds on the packet delay, delay variation and loss rate [3, 101].

### 1.2.1 Delay

The majority of real-time multimedia applications are delay sensitive because the information transmitted needs to be re-played at the receiver in real-time. A small average delay is desirable; yet, a more important delay measurement is the delay bound. The delay bound is the maximum network delay experienced by any packet. For a live video transmission, the delay bound is associated with the playback time of the receiver. If a packet arrives after this delay bound, it is useless to the receiver; therefore, the application requires a delay bound for each packet transmitted. If a packet in transit exceeds the delay bound, it is possibly better to immediately drop the packet and reduce congestion than to continue to transmit the packet to the receiver. Delay bounds will vary depending on

application type, due to playback environments. Non-interactive multimedia applications (VoD) may allow a higher delay bound than interactive applications (teleconferencing). Therefore, networks must expect a wide range of delay bounds from applications.

### 1.2.2 Delay Variation

The delay variation, also referred to as jitter, is the relative difference in delay that packets may experience in the network. A large delay variation can decrease the smoothness of the traffic, causing burstier traffic along the transmission path. Methods that smooth or reshape bursty traffic have been proposed; however, the implementation cost may be prohibitive [9, 42]. High delay variation also causes the receiver to maintain a large play-back buffer in order to store packets until the information can be re-played.

### 1.2.3 Loss Rate

Loss rate is the result of congestion occurring in the network, where packets are dropped due to buffer overflow. The dropped packets directly impact the quality at the receiver; however, the degree of quality degradation depends on the application. For example, a teleconferencing application may tolerate a higher loss rate than the transmission of high definition video. Applications may also use error recovery to reduce the effects of packet loss [55]. In addition, research has also addressed the need for applications to “gracefully” adapt to packet losses [85, 96]. For example, a video application can alter the encoding of the video to offset any packet losses that may occur. This prevents any further quality degradation due to packet loss.

## 1.3 Quality of Service Management

Many different network components must work in concert to provide QoS in computer networks. These components include, packet scheduling, connection admission control, and source policing.

### 1.3.1 Scheduling

Packets arrive at a switch where they are switched to out-going links based on their destinations. Packets waiting to be transmitted are stored in a transmission queue for their output link. A switch typically has multiple logical queues associated with each out-going link. Each logical queue may represent packets from a single flow or a certain class of flows. Given a scheduling discipline, a packet scheduler at the switch determines the transmission order of the packets from these logical queues. Since the out-going link has a maximum speed, the scheduling algorithm directly affects the QoS.

The scheduler can provide the desired QoS by managing and allocating resources, such as link bandwidth and buffer space. In addition, the scheduler should maximize utilization and allocate resources in a fair manner. One approach is to allocate resources per flow or to a class of flows (aggregate flow) [55]. Such a scheme would then ensure packets are serviced with respect to their allocated resource share; thus, providing the desired QoS. However, determining the resource share that provides a desired QoS efficiently is a challenging problem. For VBR sources (especially live video), determining the appropriate allocation amounts is difficult due to their bursty behavior and limited a priori information.

### 1.3.2 Admission Control

Admission control determines if a new flow should be admitted into the network. The acceptance or denial of a new flow affects the amount of traffic in the network; therefore, proper admission control is essential for providing QoS [82]. It is important for an admission control mechanism to determine or predict the impact of a new flow in the network. Accurately determining whether the new flow can be accommodated and how it would impinge on existing flows is essential. Admission control requires traffic descriptors, that are used to describe the traffic characteristics of the flow. Typical descriptors may include the peak and mean bit rate of the source [35] or the parameters of a token bucket [78]. These traffic descriptors and the desired QoS are then used to determine if sufficient resources are available.

As mentioned in the previous section, determining the appropriate amount of

resources that provide a desired QoS may be difficult for certain applications. Traditional approaches to admission control use a priori source information to calculate the worst-case (peak rate) behavior of all flows, in addition to the new flow. While guaranteed QoS can be provided, such a method may lead to under utilization of resources (consider VBR sources). Measurement-based admission control attempts to address this problem by measuring the actual network traffic and performance [47, 48]. These methods provide predictive QoS, which allows occasional QoS degradation. A renegotiation process may be required to update the desired QoS and the traffic descriptors over time [85]. Renegotiations increase the network signaling load, so a balance between efficient allocations and few renegotiations is needed. When accepted, the user is expected to conform to the negotiated traffic descriptors. Source policing is performed to ensure user compliance.

### 1.3.3 Source Policing

Source policing is used to ensure flows comply with their negotiated traffic descriptors; otherwise, congestion and poor QoS could occur in the network. A disadvantage of using mean and peak bit rate traffic descriptors (as described in the previous section) is the period of time samples must be taken in order to reliably determine compliance. If too few samples are taken, then the policing mechanism may incorrectly detect non-compliance. An alternative method uses a leaky bucket mechanism to describe and police the traffic [55]. The traffic would have to agree to a set of leaky bucket parameters (a form of admission control) before transmission begins. Once the parameters are determined, the mechanism can detect violating packets and immediately drop or tag them. Tagged packets can be dropped at a switch if congestion occurs.

## 1.4 Resource Allocation

As described in the previous section, QoS guarantees can be provided if network resources are available. However, providing QoS guarantees efficiently is complicated by the diversity of applications and the network performance they require. Furthermore, network resources are expected to have costs associated with their usage (amount and renegotiation)

[24, 65, 77, 97, 98]. Users must consider the cost of transmitting traffic across the network, and will prefer to do this as cheaply as possible. As a result, an important issue for providing QoS guarantees is the proper allocation of network resources.

Resource allocation can be viewed as a single-user as well as a multi-user issue, and both contexts are addressed in this thesis<sup>1</sup>. Regardless, the objective throughout this thesis is to provide the requested level of QoS, while maintaining high utilization of resources and staying compatible as possible with current technology.

#### 1.4.1 Single-User Allocation

As previously discussed, proper allocation of network resources is essential for QoS guarantees. Due to the large number of users in the network, the service provider is interested in providing QoS guarantees as efficiently as possible. Efficiency refers to the amount of resources required for transmission, as well as the the number of renegotiations. Minimizing the amount of resources allocated can increase the utilization of the network by providing resources to more users. Reducing the number of renegotiations reduces the signaling strain on the network. In addition, the user is interested in reducing costs, which is also achieved by reducing the amount of resources and the number of renegotiations. For these reasons, it is important to determine the appropriate amount of resources required to provide a desired QoS. However, determining an efficient allocation is difficult for certain traffic sources (such as live or interactive video), due to their unpredictable nature [37]. Therefore, the diversity of applications and their QoS requirements makes efficient resource allocation a challenging problem.

#### 1.4.2 Multi-User Allocation

Multi-user allocation differs from single-user allocation in that it concerns the allocation of resources for all users in the network. However, as described in [55] single-user and multi-user allocation decisions are related since, each connection should limit its resource allocation to the smallest single-user allocation along its path. Multi-user allocation typically has two goals: fairness, and the balance between throughput and QoS [6]. Defining

---

<sup>1</sup>In [55], this is referred to as local and global allocation.

fairness is difficult because of the various types of applications and their desired QoS. In this thesis, standard network-oriented and microeconomic-based definitions of fairness are used. The balance between throughput and QoS is the concept that the network should seek high resource utilization, but not at the expense of poor QoS (and vice versa). Hence, due to heterogeneous networks, diverse resource requirements and the goals associated with multi-user allocation, proper multi-user allocation remains a challenging problem.

## 1.5 Network Service Models

In this section three important network service models, designed to provide end-to-end QoS, are discussed: Internet Integrated Services, Internet Differentiated Services and Asynchronous Transfer Mode. Each of these service models presents a different framework for providing QoS.

### 1.5.1 Internet

Currently, the existing Internet provides only best-effort datagram service, with no guarantees of delay, delay variation or loss rate. Since this best-effort service is inappropriate for multimedia applications, the IETF has proposed two enhancements to provide predictable and reliable services: Internet Integrated Services (IIS) and Internet Differentiated Services (IDS).

#### **Internet Integrated Services**

The Internet Integrated Services architecture (IIS), proposed by [10, 18], was the first attempt to provide QoS in the Internet. IIS defines three service classes: guaranteed, predicted and best-effort. The guaranteed and predicted service classes are connection oriented and rely on admission control mechanisms to ensure sufficient resources are available. The guaranteed service class provides a worst-case bounds on network delay for all packets. A connection requesting guaranteed service must declare its peak rate and the desired minimum delay. Admission control mechanisms use this information to determine if enough resources are available and whether the connection should be admitted. The predicted ser-

vice also uses admission control, but is more lenient than guaranteed service [47]. Admission control mechanisms for predicted services predict (estimate) traffic on existing connections based on actual measurements, instead of using the worst-case traffic descriptors. Predictive service can allow more connections than guaranteed service; however, this is at the risk of violating delay guarantees. For this reason, predicted services are better suited for applications that can adapt to congestion. The last category, best-effort, is the same as currently found in the Internet today.

Resource ReSerVation Protocol (RSVP) is a static resource reservation protocol for individual connections, proposed for supporting integrated services [11, 105]. RSVP sends information concerning the requesting connection, the desired service type and the token bucket parameters. First a PATH message is sent from the sender to receiver indicating the traffic characteristics. The receiver then returns a RESV message, which attempts to reserve the desired resources at each router in the path. If the resources were allocated, then transmission can proceed. The Real Time Protocol (RTP) can also be used to provide timing information that is helpful for transmitting multimedia traffic [93]. Unfortunately, IIS/RSVP models require that all routers, along the path, keep a record of all current service requests and check packets to determine if they need special handling. Therefore, IIS/RSVP does not scale well with the Internet, due to the high overhead given to the routers. Furthermore, RSVP may not be feasible for applications that send only a few packets (WWW browsing). Even RFC 2208 recommends that RSVP not be deployed at the network backbone [71].

### **Internet Differentiated Services**

In 1998, the “Differentiated Services” working group at the IETF was created to develop simple methods for providing differentiated services for the Internet. The Internet Differentiated Services (IDS) architecture is based on a model where traffic entering a network is classified (and conditioned) at the boundaries, then assigned to different service classes [5, 8]. Inside the network, per-hop QoS is given to traffic aggregates which have been appropriately marked. To differentiate packets, IDS relies on the Differentiated Service field (DS field, which is the IPv4 ToS or IPv6 Traffic Class field). This field indicates the need

for low delay, high throughput or low loss rate; however, unlike IIS/RSVP, the type service choices are limited. For this reason, the amount of state information required at each router is proportional to the number of classes rather than the number of connections (IIS/RSVP). This yields a more scalable solution than IIS/RSVP.

Different service classes can be provided using the IDS classification, policing, shaping and scheduling mechanisms. For example, three service classes were defined in [75]: premium service, assured service and best-effort. Premium service is for applications requiring low delay and delay variation, while assured service provides better reliability than best-effort. In addition, pricing the service classes with respect to the service they provide (differentiated pricing, where a higher QoS has a higher price) is expected to encourage users to determine the most cost-effective service class for their connection.

### 1.5.2 ATM

Asynchronous Transfer Mode (ATM) is a packet-switched technology that supports different QoS requirements for different types of traffic. The ATM service model includes the following five service classes [3].

- Constant Bit Rate (CBR). The CBR service class allocates a static quantity of bandwidth for each connection. For this reason, CBR service provides bounds on delay and delay variation to traffic that can be characterized by its peak.
- Real-Time Variable Bit Rate (rt-VBR). rt-VBR relies on more complex traffic characterization to provide tight constraints on delay and delay variation. This service can provide higher network utilization than CBR due to multiplexing.
- Non-Real-Time Variable Bit Rate (nrt-VBR). The nrt-VBR service class provides guarantees on the average delay and maximum loss rate of a connection.
- Available Bit Rate (ABR). The ABR service class enforces a bound on the minimum throughput of a connection and divides the unused portion of bandwidth fairly among its connections.

- Unspecified Bit Rate (UBR). Similar to the Internet best-effort service class, UBR does not provide any QoS guarantees.

Of these service classes, CBR and rt-VBR are intended to transport real-time traffic (for example real-time multimedia). However, it is possible and advantageous to transmit video using the ABR service class [62, 88].

## 1.6 Thesis Contributions

This thesis concerns the allocation and pricing of network resources to provide QoS in computer networks. Contributions of this thesis address single-user and multi-user resource allocation.

### 1.6.1 Single-User Allocation

A single-user allocation technique called Dynamic Search Algorithm (DSA+) is presented, that allocates resources to provide a desired QoS. Specifically, DSA+ is used to allocate link bandwidth to provide a desired loss rate.

Simulation results demonstrate that DSA+ performs better than other comparable single-user allocation mechanisms. In addition, DSA+ is applied to multiple hop allocation and its robustness to initial parameter selection is addressed. This single-user resource allocation technique has the following unique features.

- An on-line design requiring minimal a priori source information, which has the ability to allocate resources for stored or live/interactive VBR sources. DSA+ performance is demonstrated to be better than other comparable on-line allocation methods.
- Provides efficient allocations with few renegotiations while providing a desired QoS. Reducing the allocation amount increases the utilization while reducing the number of renegotiations lowers the signaling strain on the network.
- DSA+ has a simple design that requires little processing time and storage ( $O(1)$ ), which is beneficial for implementation in high speed networks.

### 1.6.2 Multi-User Allocation

Microeconomic-based methods are presented in this thesis for multi-user allocation. These methods are based on the competitive market model (and a variation). Network resources are priced based on supply and demand, and users purchase resources to maximize their own QoS. This results in distributed allocation methods that provide high resource utilization as well as fair allocations. In this thesis, microeconomic approaches are used to allocate link bandwidth for a variety of sources. Two unique microeconomic methods, the spot market and the multi-market, are described next.

#### Spot Market Approach

A method for multi-user resource allocation is presented based on a modified competitive market model (called the spot market). Switches own the network resources in the economic model and price these resources based on supply and demand. In the spot market resources are considered non-storable (similar to residential electricity). Users purchase these resources to maximize their own QoS.

The spot market method is proven to achieve high utilization and fair allocations (Pareto-optimal, weighted max-min and equitable) for constant demand sources. Simulation is used to measure the performance of the spot market approach under network dynamics (VBR sources and users randomly entering and exiting). Simulation results, presented in this thesis, will indicate the spot market performs better than max-min (optimal implementation) and demand-based weighted max-min [62]. The spot market method has the following unique properties.

- The dynamic competitive market model (spot market) allows network dynamics (VBR sources and users entering/exiting) to occur. This is due to the modified tâtonnement process, that is unique to this market. When demands change, the modified competitive market adapts (changes prices) on-line. Users have the advantage of immediate resource availability (no reservation overhead is required) unlike other price-based allocation schemes. This yields a more QoS aware environment than other strategies.

- Allocation computations are performed only at the edge of the network, which is unique. This greatly reduces the computation and storage requirements of switches. Therefore the implementation cost of the spot market is reasonable.
- The spot market approach has the flexibility to easily achieve various types of fairness (Pareto-optimal, weighted max-min and equitable). A wealth distribution algorithm is given in this thesis to achieve an equitable (QoS-fair) allocation. In addition, an approximation is defined that determines the wealth distribution that achieves an equitable allocation with reduced source information.

## **A Multi-Market Approach**

Two types of markets are used in the multi-market approach, the spot market and the reservation market. The spot market has the unique advantage of immediate availability of resources, but the disadvantage of no guarantees. The reservation market provides guarantees of resource availability, for a period of time, but incurs reservation overhead (signaling and waiting for the reservation period to begin). The multi-market attempts to combine the advantages of both market types in a single economic model. In the multi-market approach, users have the flexibility to purchase various amounts from either market type, and can value reserved and spot bandwidth differently.

The multi-market is proven to achieve fair distributions (considering resources in the spot and reservation markets separately). In addition to the advantages described for the spot market approach, the multi-market approach has the following unique properties.

- The multi-market approach is a multi-user allocation method that effectively combines the advantages of traditional static and dynamic allocation techniques. For this reason, the multi-market approach provides immediate resource availability, resource guarantees as well as high utilization of resources.
- Users have the flexibility to purchase various resources and weigh the benefits and risks associated with the different market types. Cautious users can prefer the reservation market, while other users may prefer the immediate availability of the spot market. These are only two possibilities within a large range of choices.

- In the multi-market economy, users can dynamically change from one resource type to another during their session. This allows users to react quickly to market and source changes, reducing QoS degradation. No other microeconomic allocation method provides this capability.

## 1.7 Thesis Outline

This thesis discusses and presents allocation techniques for providing QoS in computer networks. As previously described, resource allocation can be addressed as a single-user allocation issue (chapters 2 - 3) and multi-user allocation issue (chapters 4 - 7). The following describes each chapter of this thesis in detail.

Chapter 2 reviews single-user allocation techniques and classifications. Methods will be differentiated based on when information is collected, as well as what type of information is used for making allocation decisions. Also, performance objectives for single-user allocation techniques are defined.

In chapter 3, a new on-line algorithm for resource allocation called DSA+ is presented. A detailed description of the algorithm is provided as well as performance goals. A comparison of DSA+ with peak rate, effective bandwidth, Hsu's algorithm and RED-VBR is then presented. The robustness of initial parameter selection is discussed, as well as multiplexed source and multiple-hop allocation.

Chapter 4 reviews the goals and types of multi-user allocation. Multi-user allocation methods are categorized based on the how and where allocation decisions are made. Since this thesis concerns microeconomic-based allocation, a review of microeconomics is also provided. This is followed by a review of previous microeconomic-based allocation work. Finally, goals and objectives for microeconomic-based resource allocation methods are given.

The competitive market model will serve as the basis for the microeconomic-based allocation approaches presented in this thesis. Chapter 5, reviews the competitive market structure. Proofs that an economy consisting of competitive markets will achieve optimal and fair allocations are given. A wealth distribution algorithm that distributes wealth

for an equitable allocation is described. In addition, simple example allocations and their associated fairness are provided.

In chapter 6, a spot market approach to resource allocation is described. A detailed description of the spot market approach is provided as well as simulation results. The performance of the spot market approach is compared and contrasted with other rate control methods.

Chapter 7 introduces another microeconomic-based allocation method that uses two types of markets: spot market and the reservation market. Details of the multi-market design are provided as well as simulation results that demonstrate its performance. Finally, chapter 8 reviews the allocations methods presented in this thesis. Future work and some open questions are also discussed.