# 1

# Introduction

## 1.1 Convergence of Wireless Systems and the Internet

The explosive growth of the Internet in the last two decades has escalated it as the key information transport platform. Meanwhile, with the rapid advances of microelectronics and radio technologies, wireless systems are firmly increasing their popularity. The Internet Protocol (IP)-based Internet and wireless systems are converging into a ubiquitous all-IP information transport infrastructure, allowing mobile and stationary users to access the wireless Internet for multimedia services anywhere, anytime, as shown in Figure 1.1. Wireless access networks can be infrastructure-based or infrastructureless. Examples of infrastructure-based networks are wireless cellular systems/Worldwide Interoperability for Microwave Access (WiMAX), Wireless Metropolitan Area Networks (WMAN), Wireless Local Area Networks (WLAN) and Wireless Personal Area Networks (WPAN); infrastructureless ones are wireless mesh and ad hoc networks. In this text, we are particularly interested in infrastructure-based wireless cellular systems, which can provide good mobile services and currently support more than one billion customers worldwide [1]. We also address resource management and control issues in other wireless systems.

Traditionally, the IP-based Internet and wireless systems have different design principles and resource management approaches. In wireless cellular systems, since the wireless spectrum is limited and expensive,[1] to maintain the QoS of existing and handoff calls in a cell, centralized resource management and allocation schemes are used. Wireless resources are channelized, and dedicated channels are allocated to handle multimedia traffic when the connection is established, e.g., allocating certain time-slots, frequency bands, codes, or carriers in Time Division Multiple Access (TDMA), Frequency Division Multiple Access (FDMA), Code Division Multiple

[1] In 2001, the German government raised around $46 billion for four licenses of $2 \times 10$ MHz and two licenses of $2 \times 5$ MHz 3G spectrum.

*Multimedia Services in Wireless Internet: Modeling and Analysis* Lin Cai, Xuemin Shen and Jon W. Mark

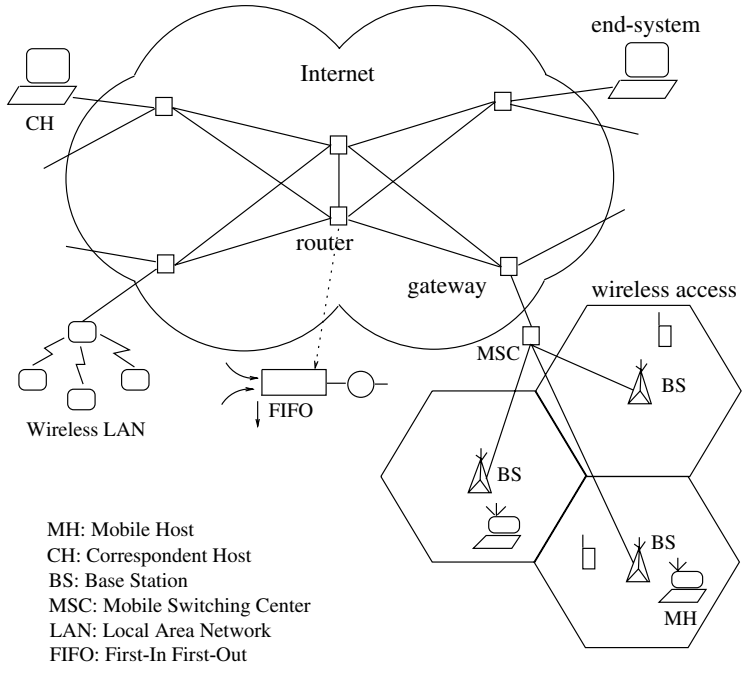


Figure 1.1 Wireless Internet.

Access (CDMA), or Orthogonal Frequency-Division Multiplexing (OFDM) systems, respectively.

Unlike wireless cellular systems, the Internet is based on the simple, robust, scalable IP protocol, which provides minimal best-effort datagram delivery services without centralized resource management and traffic control functions. As portrayed diagrammatically in Figure 1.2, when the offered traffic load is larger than the network capacity, the network power (ratio of throughput to delay) will decrease sharply and the network will be driven to deadlock and congestion collapse [2]. To efficiently and fairly share network resources in a distributed manner, the end-systems voluntarily deploy the TCP, a transport layer protocol that adjusts the sending rate according to network conditions. TCP's flow and congestion control can make the network operate near the desired operational region and maintain fairness among coexisting TCP flows. Therefore, TCP's flow and congestion control is vital for stability and integrity of the Internet.

With the rapid advances in optical and wireless communication, the Internet is becoming a more heterogeneous and disparate system: link capacity varies from several Kbps to several Gbps, with six orders of magnitude; transmission bit error rates vary from $< 10^{-9}$ to $10^{-3}$, also with about six orders of magnitude; and end-to-end delay varies from several milliseconds to several seconds. An immediate question

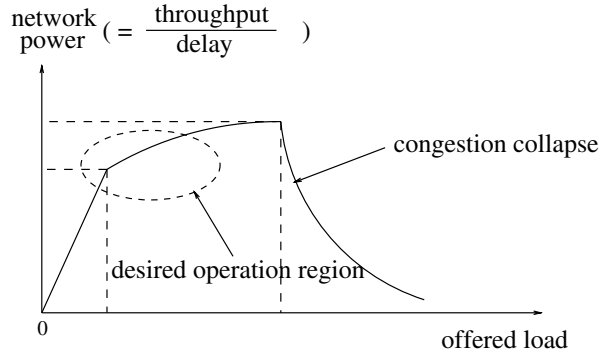


Figure 1.2 Network power.

is how to ensure the efficiency, fairness and stability of the wireless Internet, and provide satisfactory QoS for various multimedia applications, independent of the heterogeneity of communication links.

Salient background information on TCP flow and congestion control and active queue management in the Internet is summarized in Appendix A. A more comprehensive introduction can be found in [3–5].

## 1.2 Main Challenges in Supporting Multimedia Services

The driving force for Internet growth in the future is from emerging multimedia applications, such as audio and video streaming, IPTV and online gaming, which have a wide variety of QoS requirements. Many of these have much tighter requirements on delivery timeliness rather than object integrity. On the other hand, wireless access networks can provide convenient anytime, anywhere Internet services in a cost-effective manner (without truck-roll or rewiring costs). However, there exist many challenges to the efficient support of heterogeneous multimedia applications over wireless networks with satisfactory QoS.

First of all, a wireless channel has its inherent impairments. It suffers time-varying and location-dependent fading, shadowing, interference and ambient noise, etc., which lead to higher transmission error probabilities. Efforts from both the physical (PHY) and link layers have been conducted to mitigate the errors perceived by the upper layer protocols and applications, e.g., using coding, power control, rate control and local retransmission schemes. Although the error rate can be controlled below a certain threshold, the time to successfully transmit a packet becomes a random variable. Thus, even with a dedicated wireless channel, the service rate of a wireless link is time-varying, which introduces random delay. In addition, the wireless medium is broadcast in nature. To share the premium wireless resources, either deterministic resource allocation schemes or random access medium access

control (MAC) protocols are needed. For wireless networks (e.g., IEEE 802.11 WLANs) using random access MAC protocols, competition among neighboring users that may lead to collisions can bring more uncertainty for timely transmissions. Therefore, the service rate in wireless networks is highly variable.

Different multimedia applications also have different levels of QoS requirements and traffic characteristics. For instance, IPTV users may pay for and require cable-TV-comparable quality of experience (QoE); YouTube users can tolerate much lower-quality videos and longer latency, and they pay a minimal price for the service.

For high-cost multimedia applications, dedicated wired and wireless resources can be allocated, or given a higher priority to use network resources. These high-cost (typically non-elastic) applications are typically not regulated by end-to-end flow and congestion control. Thus, the research problems are relatively simple: how to effectively reserve resource for them, and how to use admission control mechanisms to guarantee their QoS. Chapter 3 will address these issues, and Chapters 4–8 will discuss more challenging research problems for multimedia applications under flow and congestion control.

For low-cost (typically elastic) multimedia applications, the flows will share the network resources with other data traffic, so they should be regulated by end-to-end flow and congestion control to ensure network integrity and stability. However, the dominant transport layer protocol in the Internet, TCP, has been designed and engineered mainly for bulk data transfer over wireline links. It encounters the following challenges for supporting multimedia applications in wireless Internet. First, TCP provides reliable data transfer service: a TCP sender detects and retransmits corrupted or lost segments. Retransmissions by the TCP sender may introduce intolerable delay and delay jitter for delay-sensitive multimedia applications. Second, the TCP sender uses a congestion window (*cwnd*) to probe for available bandwidth[2] and respond to network congestion using an AIMD mechanism: the TCP sender additively increases its *cwnd* by one segment per round-trip time (*rtt*) if no congestion indicator is captured, and decreases the *cwnd* by half otherwise. With the increase-by-one or decrease-by-half control strategy, even an adaptive and scalable source coding scheme cannot hide the flow throughput variation, so the user-perceived audio and video quality may degrade ungracefully. Third, TCP has been designed for applications in wireline networks, where packet losses are mainly due to network congestion. Thus, packet losses seen by the TCP receiver are assumed as due to network congestion and the TCP sender is notified by congestion notifications. However, in wireless networks, packet losses may be due to transmission errors or link-layer collisions, so cutting the rate by half may underestimate the available bandwidth in wireless networks.

[2]Bandwidth is defined as the spectral width or capacity of a communication channel or network. Analog bandwidth is measured in Hertz (Hz) or cycles per second. Digital bandwidth is the amount or volume of data that may be sent through a channel or network, measured in bits per second. In this book, 'bandwidth' refers to digital bandwidth.

Some emerging multimedia applications use the unreliable and unresponsive UDP without flow and congestion control. However, using unresponsive transport layer protocol is not a long-term solution for high data rate and long-lived multimedia traffic: (i) it is difficult to send data using UDP in the presence of firewalls that are configured to block inbound UDP traffic; (ii) more importantly, future IP-based networks will continue to be decentralized such that network stability depends on end-to-end flow and congestion control. When unresponsive traffic occupies a large percentage of the traffic mixture, IP-based networks may be driven to congestion collapse [6]. To avoid congestion collapse, the core of the network will deploy efficient and scalable schemes to punish unresponsive traffic, e.g., dropping packets from unresponsive UDP traffic harshly when the network is under congestion [7–9]. On the other hand, although it is possible to provide congestion control at the application layer, below UDP, or by modifying other non-congestion-controlled transport layer protocols (e.g., Stream Control Transmission Protocol (SCTP) [10] and Real-time Transport Protocol (RTP) [11]), these solutions have their disadvantages and limitations: application layer congestion control has difficulties in handling Explicit Congestion Notification (ECN); congestion control below UDP needs congestion feedback from either the application or the layer below UDP, which complicates the system design; modifying SCTP or RTP may not be desirable for applications requiring minimal overhead, as discussed in [12].

Therefore, to support multimedia applications in wireless Internet, new transport layer protocols with minimal overhead, unreliable datagram delivery services, and flow and congestion control are needed. The design objectives of the new transport layer protocols for cross-domain multimedia traffic are: (i) to efficiently and fairly share highly multiplexed wired links in a distributed manner; (ii) to efficiently utilize the premium bandwidth of lightly multiplexed or dedicated wireless links; and (iii) to satisfy the QoS requirements of multimedia applications. Since TCP traffic is dominant in the Internet, the new protocol should also be TCP-friendly. *TCP-friendliness* is defined as the average throughput of non-TCP-transported flows over a large timescale does not exceed that of any conformant TCP-transported ones under the same circumstances [6].

One major difficulty for TCP-friendly congestion control is that the end-systems do not have complete knowledge of the internal network conditions, e.g., network topology, bottleneck capacity, volume of traffic sharing the bottleneck and wireless channel conditions, which are highly dynamic and difficult to anticipate. In wireless Internet, is it possible for the end-systems to obtain a fair share of the network resources and efficiently utilize them, especially the premium wireless bandwidth, with satisfactory user-perceived QoS? If so, why and how?

To meet the challenges, efforts from several layers of the protocol stack are required, and each layer tries to utilize the service from a lower layer to meet the requirements of an upper layer. Here, we use the TCP/IP reference model and focus attention on the components enclosed by the dotted ellipse shown in Figure 1.3.

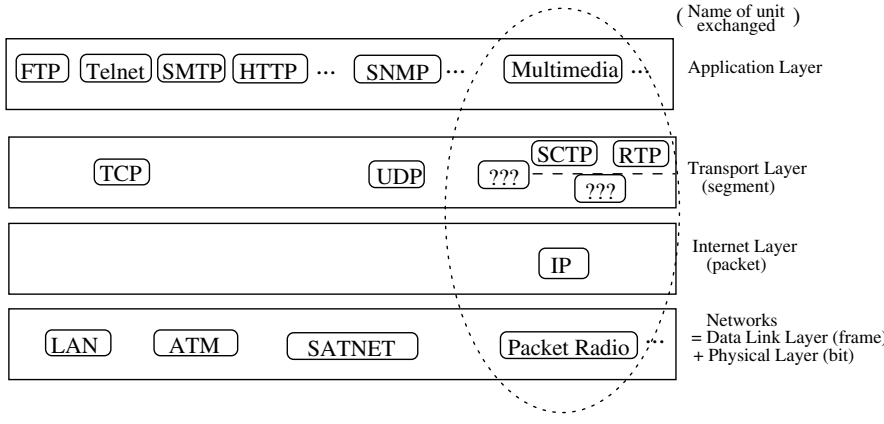


Figure 1.3 TCP/IP reference model.

Multimedia applications provide users with satisfactory visual and sound quality, based on the service provided by the transport layer protocol. The QoS provided by the transport layer protocol to the applications is measured in terms of flow throughput, packet loss rate, and end-to-end delay distribution. The transport layer protocol should probe for available bandwidth and efficiently utilize it with QoS provisioning. The link layer protocol tries to *hide* the link impairments from the upper layer protocols. The PHY layer protocol decides how information bits can be efficiently delivered through the channel.

The functionality of the relevant layers of the protocol stack can be succinctly summarized as follows. The PHY layer uses certain modulation and coding schemes to efficiently transmit information bits and ensure that the bit error rate (BER) is below a certain threshold. So long as the delay bound for delay-sensitive traffic is not violated, the link layer decides when and at what transmission power level to transmit or retransmit packets over the wireless link, so the link impairments can be as invisible to the upper layer protocols as possible. The transport layer determines when packets can be inserted into the network according to estimated network conditions. The application layer decides which information should be put in a packet to achieve the best user-perceived quality.

In the literature, scalable and error-resilient source coding schemes for multimedia applications [13–18], and link layer error correction and recovery schemes for wireless links have been extensively studied and widely deployed [19–21]. However, how the transport layer protocols appropriately support multimedia applications in ever-diversified wireless Internet and ensure their QoS needs further investigation.

## 1.3 Organization of the Text

Having presented an overview of the challenging issues in supporting multimedia services in wireless Internet, in Chapters 2 through 8 we provide in-depth discussions of wireless channel modeling, multimedia traffic modeling, TCP-friendly congestion control, stability and performance bounds of the Internet, QoS performance of non-congestion-controlled and congestion-controlled multimedia flows in wireless networks with dedicated resource allocation schemes and random access schemes. In the following, we briefly summarize the rationale behind the material presented in each of these chapters.

Since wireless channel characteristics and statistics determine the distribution of the service time in the wireless domain, reasonably accurate and mathematically tractable wireless channel models are essential for quantifying the performance in wireless networks. Chapter 2 introduces packet-level channel models based on Markov chains, which can be easily incorporated in the analytical frameworks and network simulation tools.

In Chapter 3, we investigate the multimedia traffic characteristics and study the traffic models for voice and video applications in the Internet. For non-congestion-controlled multimedia flows, an analytical framework is given that combines the traffic model and channel model to quantify the QoS parameters of multimedia services in wireless networks.

Chapter 4 studies the AIMD protocol, which inherits the congestion control mechanism of TCP and allows the adjustment of a pair of parameters to provide more flexible services for multimedia traffic sharing the network resources with existing TCP flows. The necessary and sufficient TCP-friendly condition for AIMD controlled flows is derived. The effectiveness and responsiveness properties of the AIMD algorithms and the practical implications are also studied.

It has been discovered that TCP and AIMD congestion control cannot ensure network asymptotic stability. However, so long as the end-systems do not overshoot the available bandwidth too severely, the overall system efficiency can still be very high, and the packet loss rate and queueing delay can still be well bounded. In this context, it is critical to investigate the theoretical bounds of the system if the network were to operate at states away from the desired equilibrium state. Chapter 5 derives an upper bound and a lower bound of the flow's congestion window size, and an upper bound of queue length. Numerical results with Matlab and simulation results with NS-2 are given to validate the correctness and demonstrate the tightness of the derived bounds. The analysis provides important insights on which system parameters contribute to higher oscillations of the system and how to effectively control system parameters to ensure system efficiency with bounded delay.

Chapter 6 develops an analytical framework for quantifying the QoS performance of window-controlled flows (e.g., TCP and AIMD) in wireless Internet. The end-to-end delay distribution and packet loss rate of window-controlled flows over

hybrid wireless and wireline networks are analytically obtained, and a delay control scheme is introduced. Based on the QoS performance analysis the AIMD protocol parameters can be appropriately selected. Simulation results are given to validate the analysis, demonstrate the feasibility of the approach and show that the TCP-friendly AIMD protocol can even outperform the unresponsive UDP protocol for supporting multimedia applications in wireless Internet.

Another important TCP-friendly transport layer protocol is the TCP-Friendly Rate Control (TFRC) protocol introduced in Chapter 7. The performance of TFRC-controlled flows in wireless Internet is studied via analysis and simulation.

In Chapter 8, we further study the QoS performance in WLANs and integrated WWANs/WLANs. Since WLANs use a random access MAC protocol, the access control mechanism and collisions observed in the link layer further complicate the performance study of congestion-controlled multimedia flows. We first study the link layer performance of WLANs, and then address the more challenging research issues of how to quantify the end-to-end performance of congestion-controlled flows over WLANs and integrated WWANs/WLANs.