

Chapter 8

Conclusions and Future Work

Current and future networks must support a variety of applications that differ widely in terms of their traffic characteristics and Quality of Service (QoS) requirements. Network resources (such as buffer space and link bandwidth) must be allocated efficiently and fairly to these applications in order to provide QoS. Two important resource allocation issues were addressed in this thesis: single-user resource allocation and multi-user resource allocation.

8.1 Single-User Allocation

The service provider and the user are interested in allocating network resources as efficiently as possible, while providing a desired QoS. Efficiency refers to the amount of resources required, as well as the number of renegotiations. Minimizing the amount of resources allocated can increase the utilization, while reducing the number of renegotiations reduces the signaling strain on the network. Similarly, the user is interested in reducing costs, which can be achieved by reducing the amount of resources and the number of renegotiations required. Determining an efficient allocation is difficult due to unpredictable nature of certain traffic sources (such as live or interactive video). While there are many different types of network resources and QoS measures, this part of the thesis (single-user allocation) focuses on the allocation of link bandwidth to provide a desired cell loss probability.

Many different methods of single-user bandwidth allocation have been proposed.

These can be categorized as either off-line or on-line. Off-line techniques require complete source information before transmission begins in order to determine the appropriate allocation amounts. While these methods are able to achieve efficient allocations, they are not applicable to live or interactive sources. In contrast, on-line methods allocate link bandwidth using real-time measurements. Due to the limited a priori information required, on-line methods are suitable for allocating bandwidth for stored or live media. However, none of the on-line methods have the ability to reduce the bandwidth and number of renegotiations required for various VBR sources.

This thesis presented an on-line algorithm, DSA+, which efficiently allocates resources to provide a required QoS. DSA+ was used to manage link bandwidth to achieve a desired cell loss probability for MMBP generated traffic and MPEG-compressed video traces. Reducing the bandwidth allocated and the number of renegotiations are the goals of this allocation mechanism. For MMBP traffic, DSA+ allocated the same (slightly less) bandwidth than the effective bandwidth value. However unlike the effective bandwidth calculation, DSA+ does not require prior knowledge of statistics of the underlying traffic generation process. For the MPEG experiments, fifteen actual MPEG traces were collected and used. As compared to an off-line peak-rate allocation, DSA+ saved 13–58% in bandwidth. On average 36 renegotiations were required, but only 44% were for more bandwidth, which seems acceptably low. Other methods which were compared, either over-allocated bandwidth or required up to 47 times more renegotiations. The effect of multiplexing was investigated and showed DSA+ has no problem guaranteeing QoS to such a traffic source.

Experiments also indicate the algorithm is fairly insensitive to the choice of initial parameter values. For all the experiments performed the same initial parameters were used and showed excellent results. DSA+ requires limited information about the source, however any a priori information can, of course, benefit the performance of any on-line algorithm.

Multiple hop connection allocation was also addressed. In this case, a connection of four nodes was simulated to evaluate the performance of DSA+ for end-to-end CLP. Two implementations were investigated; each-node and first-node. Both methods were able to provide the end-to-end QoS, however each method may suffer from some possible disadvantages.

8.1.1 Future Work

While the focus of this thesis was the allocation of bandwidth for network applications, DSA+ may be useful for other real-time applications. Examples include CPU scheduling and disk bandwidth management. In both cases the central idea is to provide guaranteed service to variable traffic, with the minimum amount of resources and user input.

For the single-user allocation experiments, no limit on the availability of resources was made. When any allocation method renegotiated for more resources, they were instantly granted. However in an actual implementation this assumption can not be made. In the case of network overload, where contention for more resources is high, resources may not be available. This was the primary purpose for reducing the number of renegotiations for more resources. Nevertheless, if more resources are required yet not available the users QoS will suffer. If the QoS manager has access to the MPEG compression rate, the shortage of resources can be compensated by altering the Q factor of the compression [86]. The result is a loss of picture quality, until resources are available.

Finally, how DSA+ manages resources for other QoS measures should be investigated. For example, it may be possible to allocate bandwidth to provide some desired delay bound. Such a method could be integrated into a Weighted Fair Queue scheduler [27], where the weights of each class/user is dynamically adjusted based on measured delay.

8.2 Multi-User Allocation

Multi-user allocation differs from single-user allocation since it concerns the allocation of resources for all users in the network. Multi-user allocation typically has two goals: fairness, and the balance between throughput and QoS. Defining 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 were 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 allocation remains a challenging problem.

Microeconomics can be used to allocate network resources in an efficient and fair manner. A simple network economy consists of finite resource (link bandwidth) and two types of agents: consumers (applications) and producers (switches). Consumers purchase resources to satisfy their QoS. Producers maximize their satisfaction by renting or selling resources to consumers. Economic-based allocation methods have several advantages. Many economic-based techniques offer a distributed allocation method, eliminating the need for a central controlling entity. Economic-based methods can also achieve Pareto-optimal and fair allocations (weighted max-min fair, proportionally fair per unit charge and equitable). These methods are typically able to scale to large networks and provide a framework for economic goals (such as, cost recovery and profit maximization). However, to date none of the microeconomic methods are able to allocate resources under network dynamics (changing user demands and users entering/exiting) or provide guarantees of price stability. This thesis introduced two microeconomic-based methods for allocating link bandwidth to multiple users: the spot market approach and the multi-market approach.

8.2.1 Spot Market Approach

In the spot market approach, a computer network was viewed as an economy consisting of three entities; users, Network Brokers (NB) and switches. Switches own the resources sought by users, and price their resources based on local supply and demand using a *modified* tâtonnement process. The modified tâtonnement process allows demands for bandwidth to change dynamically and is a unique feature of the spot market approach. In addition, price updates are performed using only aggregate information. For this reason, the spot market approach is state-less and should have a reasonable implementation cost. Bandwidth is sold as a non-storable resource, so users are charged based on their consumption (similar to residential electricity). A user requires link bandwidth to maximize their individual QoS. Representing the user in the economy, the NB makes the resource purchasing decisions based on current needs of the user and prices. Once a new allocation amount has been determined, the user can send using this amount immediately. There is no reservation overhead required. Therefore, users are able to dynamically change their demands based on their application requirements and link prices. Users and switches act indepen-

dently, which yields a distributed allocation method. This competitive market structure is also proven to achieve Pareto-optimal and fair (weighted max-min and equitable) allocations when demands are constant. The spot market approach is flexible, easily achieving a variety of fair allocations. In addition, pricing bandwidth in the spot markets provides a disincentive to over-allocate bandwidth.

The spot market approach was simulated to measure the performance under network dynamics. Simulation results demonstrate the ability of the spot market approach to successfully allocate bandwidth of a network to a large number of diverse users, each transmitting an actual MPEG-compressed video trace. The economy also provided substantially better control of QoS than max-min or demand-based weighted max-min [62]. To date, no other microeconomic-base allocation method has been tested under such conditions.

A limitation of the spot market approach is the inability to provide resource guarantees (price stability). It is possible that a user, who enters the network when prices are low, may be forced out of the economy if prices become too high. In addition, price “spikes” may cause a degradation in QoS. For this reason, a method that provides resource guarantees is needed. This is the motivation for the multi-market economy.

8.2.2 Multi-Market Approach

To address the need for resource guarantees, a multi-market economy was introduced in this thesis. Similar to the spot market approach, a computer network is viewed as an economy consisting of three entities (users, Network Brokers and switches) and two different markets/resources (reserved and spot bandwidth). Switches own the bandwidth, which is sold in the reservation and spot markets. Reserved bandwidth has the advantage of ownership over a period of time, providing the user with some predictability of their expected QoS. In contrast, spot bandwidth has the advantage of immediate availability without reservation overhead. Both market types are modeled as competitive markets; therefore efficient as well as Pareto-optimal and fair allocations are possible. Users require link bandwidth for their applications and are represented in the economy by a Network Broker (NB). The NB buys bandwidth to maximize the utility (QoS) of the user and considers the risks and benefits associated with the two bandwidth types (demonstrated in

the simulation results). This multi-market approach uniquely integrates the benefits of the spot market (such as Pareto-optimal and equitable allocations) with the price stability offered with the reservation market. This is done in a distributed and state-less manner. To date, no other microeconomic approach has integrated these two types of markets into one allocation method. Users are able to dynamically change bandwidth amounts in response to market and source requirements. This is another unique feature of the multi-market approach.

8.2.3 Future Work

The market-based approaches to multi-user allocation presented in this thesis assumed the user only had one path available from source to destination. However, in reality multiple paths may exist. One area of research is price based routing. Given a set of possible paths the user could reroute traffic based on current prices or availability of reserved bandwidth. Such a method gives the user more choices when sending traffic. In addition, the wealth rate was assumed to be equally divided among all the switches in the route. An alternative approach would allow the user to dynamically adjust the amount of wealth spent on these switches. This would allow the user to save money over time.

Another issue not directly addressed is price-based admission control. The spot market and the multi-market approaches provide a unique method of admission control. Given the prices associated with a route (or set of routes), the NB can determine quickly whether the user should be allowed to enter the network. Alternatively, the NB/user could require a minimum amount of reserved bandwidth be purchased before transmission begins.

A generic implementation of the spot market approach was provided in this thesis. However, a more urgent application is to the Internet. Future work should address the compatibility of the spot market with the current Internet. For example, price distribution methods should be addressed. Such a method should have low overhead and work with legacy networks. In addition, future work should address the integration of the spot and reservation markets with differentiated services. Differentiated services relies on a pricing mechanism to help users select the appropriate service class. The pricing mechanisms presented in this thesis may be suitable for differentiated services.

Finally, the reservation market sells bandwidth for fixed periods of time. If the session of the user finishes before the end of the current segment, the user returns the bandwidth to the switch. Since the user does not pay for the returned bandwidth, there is an incentive to return unused bandwidth. However, an alternative approach would allow the user to re-sell this bandwidth to other users. In this environment users could increase their wealth by purchasing reserved bandwidth at low prices, then re-selling at higher prices.