

Chapter 4

Multi-User Allocation and Pricing Techniques

As described in chapter 1, computer networks must contend with a diverse variety of network applications. Given a finite supply of network resources and ever changing demands, networks need to allocate resources in a fair and efficient manner to provide the QoS these applications require. Different from single-user allocation (chapters 2 and 3), this chapter concerns the allocation of resources to all users in the network. Objectives in the multi-user environment include efficient and fair allocation of network resources.

This chapter reviews and categorizes various multi-user allocation methods. Since this thesis concerns microeconomic approaches to multi-user resource allocation, a brief discussion of basic microeconomic principles is also provided. This is followed by a review of previous microeconomic-based allocation work. Finally, multi-user allocation performance objectives are presented and discussed.

4.1 Multi-User Allocation Objectives and Classifications

There are two objectives associated with multi-user resource allocation, fairness among applications and the balance between throughput and QoS. Fairness can be defined in various manners. For example, fairness can refer to the amount of resource allocated

to each application (for example, the max-min fairness criterion). The max-min fairness criterion provides all users a “fair” share of the resource [6]. Alternatively, fairness can be defined with respect to the QoS observed by each application (social welfare criterion) [38, 76]. Now the objective is to provide applications equivalent QoS, which is better suited for multimedia applications [39]. The second objective of resource allocation, the balance between throughput and QoS, is the concept that the network should seek high utilization but not at the expense of QoS (and vice versa).

Resource allocation methods can be classified as either static or dynamic in nature. Static resource allocation techniques reserve a single amount for the duration of the session. In general, static methods can be too conservative for bursty traffic such as MPEG-compressed video, or rely heavily on statistical models to predict resource requirements. In either case, static methods require a priori source information (for example, peak transmission rate or bounds on the maximum burst size), which is not available for live or interactive applications. Alternatively, dynamic allocation methods adjust resource amounts based on network conditions and application requirements. This responsiveness attempts to provide the desired QoS while maintaining high utilization; however, providing strict QoS guarantees is difficult, since contention for resources may occur.

Methods that perform dynamic resource allocation can be generally classified on whether they maintain per-connection state information [51]. Methods that maintain per-connection state information that is directly used in the calculation of the allocations will be referred to as *state-maintaining*. Alternatively, if per-connection state information is not required for the calculation of the allocations, it will be referred to as *state-less*. Of these two categories, a state-less method is preferred. Such a method does not require the overhead (storage and computational) of connection tables when computing allocations. Also, state-less implementations are scalable to larger networks since additional data structures are not required.

Recently, economic theory has been applied to multi-user resource allocation. Modeling the network as an economy, economic theory can be used to allocate network resources. A simple network economy consists of two types of agents, consumers (applications) and producers (switches). Consumers require resources to satisfy their QoS. Produc-

ers own the resources sought by consumers, and maximize their satisfaction by renting or selling. Economic-based allocation methods have several advantages. Many of economic-based techniques offer a distributed allocation method, eliminating the need for a central controlling entity. Economic-based methods can also achieve optimal allocations (Pareto-optimal, proportionally fair per unit charge [53] and weighted max-min fair). These methods are typically able to scale to large networks and provide a framework for economic goals (such as, cost recovery and profit maximization).

4.2 Economics and Resource Allocation

Economics is often defined as the study of ‘the allocation of scarce resources among competing ends’ [76]. This definition contains two fundamental concepts of economics: scarcity of resources and allocation choices. Scarcity of resources signifies that there is never enough of a resource to satisfy all wants all the time. Choices must be made concerning possible allocations and their impact. Therefore, economic theory examines the interaction of agents to understand how resources are allocated.

Economic theory can be divided into two categories, *microeconomics* and *macroeconomics*. Macroeconomics concerns large aggregate behavior (group of agents), instead of individual actions (a single agent). In contrast, microeconomics, also known as price theory, concerns the behavior of individual agents and their interaction in the market. Microeconomic paradigms will be the focus of this thesis, because the attention to individual behavior is appropriate considering the need for individual QoS in a computer network.

4.2.1 Economic Models

An economic model is composed of a finite amount of resources, a set of agents, and rules specifying their interaction. Given this framework, agents in the economy acquire resources in an attempt to optimize some metric. For example, an agent may seek to maximize their utility, which is a measurement of satisfaction. The utility obtained from an amount of a resource is determined from a utility function. The utility function maps a resource amount to a real number, that corresponds to a satisfaction level. Assuming $u(\cdot)$

is a utility function, if the agent prefers an allocation amount a over \hat{a} (this is represented using the notation $a \succ \hat{a}$) then $u(a) > u(\hat{a})$. Using a utility function, an agent can rank possible allocations and acquire resources that maximize their utility. Once the structure is defined, the performance of the economy can be analyzed.

Various criteria can be used to measure the distribution of resources in the economy. Two “economic-oriented” optimal criteria are used in this thesis: *Pareto-optimal allocation* and an *equitable allocation*. A Pareto-optimal allocation is one in which no agent can increase their utility with out decreasing the utility of another. Many different Pareto-optimal allocations exist; however only a few can be considered equitable [76]. Social welfare economics defines an equitable allocation as one where all agents achieve the same level of utility. In addition, this thesis will use the “network-oriented” criterion weighted max-min fair to define optimal allocations. All of these criteria are important for determining the success of an economic-based network allocation method.

4.3 Microeconomic-Based Multi-User Allocation

Microeconomic-based allocation methods use microeconomic theory to allocate network resources and have several advantages. Many of economic-based techniques offer a distributed allocation method, eliminating the need for a central controlling entity. Economic-based methods can also achieve optimal allocations (Pareto-optimal, proportionally fair per unit charge and weighted max-min fair). These methods are typically able to scale to large networks and provide a framework for economic goals (such as, cost recovery and profit maximization). Pricing network resources also provides a disincentive to over-allocate network resources.

Microeconomic-based methods can be categorized based on how the allocations are determined (centralized or distributed). In addition, some techniques are designed for certain networks, such as ATM or the Internet. A review of these categories is provided next.

4.3.1 Constrained Maximization

One approach of applying microeconomics to computer networks uses optimization techniques to maximize utility [49, 50, 53, 67, 70]. The maximization process determines the optimal resource allocation such that the utility of users is maximized subject to budget and resource availability constraints [15, 76]. Since the computation required for the maximization process increases as the number of users increases, these methods are not scalable to networks with a large number of users. To provide scalability, some approaches group users and use a single utility curve to represent the group. The maximization process is then performed for the smaller number of groups instead of individual users. Groups can be created based on desired QoS [49, 50] or on traffic types (or service classes) [67]. Accurately grouping users together may be problematic due to the wide variety of applications and their diverse resource requirements. Another problem is that these approaches generally require a centralized entity to determine the optimal allocation amount. This is undesirable because the economy relies on one entity, which is not reliable or fault tolerant. Furthermore, a centralized entity can also become a source of congestion in the network when demands or prices change.

4.3.2 Congestion Pricing

Congestion pricing [70] is another approach that charges users for their consumption of resources [2, 19, 33, 34, 53, 70, 73, 90]. Users act independently, attempting to maximize their own utility and prices are set in a distributed fashion based on local resource conditions (supply and demand). It has been shown that pricing based on supply and demand results in higher utilization than traditional flat (single) pricing [19, 70]. These methods are able to achieve Pareto-optimal allocations [34], proportional fairness per unit charge [53], max-min fair, and equitable (QoS-fair) allocations [38]. For these reasons, many of the following allocation methods use congestion pricing as a means to efficiently allocate resources. However, one important disadvantage of congestion pricing, in its original form, is the inability to handle changing demands on various time scales. For this reason, many of these methods can not easily handle VBR sources.

4.3.3 ATM Virtual Circuit and Virtual Path Allocation

One application of microeconomic theory to network resource distribution is for allocating ATM Virtual Circuits (VC) [2, 33, 34, 73]. Prices are iteratively determined at each link based on user demand and capacity. In [33, 34], users purchase bandwidth at each link along their path, attempting to achieve a certain minimum throughput and to minimize the average delay (assuming a M/M/1 queueing model). In this model users are considered “selfish” since purchasing decisions are based only on the interests of the user. When a price is determined that causes demand to equal supply (equilibrium price), Ferguson et al. proved the Nash equilibrium is achieved. A Nash equilibrium, normally associated with game theory, is a point where the strategy chosen by each player is the best considering the choices of all other players. This model was extended to include Virtual Paths (VP) by Anerousis et al. [2]. In this paper, pricing is done to control VP demand and VC blocking probability experienced by users. At equilibrium, VC prices are higher per unit capacity than VP prices, which is expected due to the increase signaling costs. An alternative method for VC pricing was presented by Murphy et al. [73]. Again, prices for link bandwidth are iteratively determined based on supply and demand; however, users determine the amount of bandwidth that will maximize their marginal benefit (benefit minus cost). The benefit function of the user is assumed to be concave increasing. Normally assumed in economics, the user has a higher value of the initial amount of a resource, and has a diminishing value as the amount increases [73]. Once the equilibrium prices are determined, they and the demand are fixed for a period of time. This procedure repeats for the next segment of time. Therefore, demands change over a long time-scale, which is unsuitable for bursty traffic. This marginal benefit model is also used by Kelly, et al. for generic bandwidth pricing (not necessarily ATM) [53].

4.3.4 Microeconomic ABR Rate Control

Microeconomic-based techniques designed specifically for ABR rate control include [20, 21]. In [20], switches allocate ABR bandwidth in a proportionally fair manner based on the “willingness-to-pay” (wealth declaration) submitted by each user. When conditions

change, users determine a new willingness-to-pay via a curve fitting process which relies on a history of previously optimal decisions. In the ABR rate control method of [21], users bid for some amount of effective bandwidth. While effective bandwidth allocates over a longer time scale, these techniques are difficult to apply to sources with little or no a priori information (for example, live and interactive video) and can be considered too conservative [21].

4.3.5 Effective Bandwidth Pricing

Another method of resource pricing uses effective bandwidth [52] to measure resource usage [22, 24, 54, 95, 97]. Charging considers the “static traffic contract” (a priori traffic contract and information) and dynamic traffic measurements. Users are then charged for their session length as well as their traffic volume [54]. This method guarantees that users will provide truthful estimates of their traffic parameters, which avoids the need for traffic policing. Effective bandwidth pricing has been proposed to control ABR bandwidth (as described in the previous section) [22, 24]. Users bid for an effective bandwidth and the network controls the flow by adjusting the allowed transmission rate of the user. These approaches to rely on heavy multiplexing to provide accurate effective bandwidth calculations [22]. Otherwise the effective bandwidth values can be considered too conservative.

4.3.6 Smart Market

A method for pricing services in the Internet called the “smart market” was proposed by MacKie-Mason and Varian [69]. In this congestion control strategy, the user attaches a bid to the header of each packet sent. Routers in the network calculate a price based on the equilibrium price or the marginal cost of sending one more packet. The marginal cost consists of a non-congestion cost associated with transmission, and a congestion cost. Packets with bid amounts that exceed the current price for the link are transmitted. As a result, users have an incentive to quickly reveal their true value of each packet.

4.3.7 Game Theory

Game theory can be applied to network flow control and resource allocation. In this approach, each application is considered a player in a cooperative or non-cooperative game. A cooperative game requires players communicate information about their strategies. Alternatively, a non-cooperative game requires players to work individually without information from others. In either type of game, the goal of each player is to optimize their performance. Non-cooperative games have the advantage of less player-to-player communication overhead [94]. Nonetheless, the use of this information in cooperative games can result in a Pareto-optimal allocation [60, 72].

Park et al. describe a non-cooperative provisioning game to provide multiple levels of QoS [14, 80]. In [14, 80], QoS agents are installed at each switch in the network. The QoS agent intercepts the packets entering the switch and determine the appropriate QoS class for the packet in order to satisfy the end-to-end QoS (specified by the user). For this reason, the end-to-end QoS must be mapped to a local QoS at each switch along the route. Once the local QoS responsibility has been determined, the agent determines the QoS class that satisfies this value at the lowest cost [14]. It is assumed that classes with a proportionally higher QoS (for example, lower loss rate) also have a proportionally higher price (linear price differential); however, exactly how these prices are determined is not discussed.

Another non-cooperative game approach to bandwidth allocation was proposed by Korilis et al. [59]. Users split their traffic across multiple paths in the network so to minimize its individual costs. Prices associated with a link are proportional to the congestion at the link. This acts as an incentive for users to route traffic away from expensive (congested) paths.

4.3.8 Allocating Multiple Resource Types

Methods for pricing bandwidth and buffer space are described by [67, 68, 90]. In these methods, buffer space and link bandwidth are priced independently to reflect supply and demand. In [90], users compete for buffer space and link capacity to achieve a certain loss rate by employing certain queueing models (M/M/1/B). When the markets are in

equilibrium (price where supply equals demand), the resulting allocations (both buffer and bandwidth) are proven to be Pareto-optimal. However, the application of this method to actual VBR sources is not addressed.

4.4 Chapter Summary

This chapter discussed the various goals and categories of multi-user allocation. Goals normally associated with multi-user allocation include, fairness among applications and the balance between throughput and QoS. Fairness can be defined in various manners and is difficult due the various applications types and demands. The balance between throughput and QoS, is the concept that the network should seek high utilization but not at the expense of QoS (and vice versa). Methods can be categorized as either static or dynamic. Static methods allocate a single amount for the duration of the session, while dynamic methods can reallocate resources depending on network conditions. Static methods have the advantage of simple connection admission and control, but tend to be too conservative resulting in low utilization. Dynamic allocation methods achieve higher resource utilization, but can not guarantee resource availability. In addition methods can be classified based on the information required to determine allocation amounts. A state-maintaining approach requires per-connection state information to calculate allocation amounts, while a state-less approach uses only aggregate information. State-less approaches are preferred since less overhead (storage and computational) is required.

This chapter also reviewed microeconomic-based allocation methods. These techniques use microeconomic theory to allocate network resources and have several advantages. Many of economic-based techniques offer a distributed allocation method, eliminating the need for a central controlling entity. Economic-based methods can also achieve optimal allocations (Pareto-optimal, proportionally fair per unit charge and weighted max-min fair). These methods are typically able to scale to large networks and provide a framework for economic goals (such as, cost recovery and profit maximization). In addition, pricing network resources provides a disincentive to over-allocate network resources.

None of the microeconomic methods discussed in this chapter are able to han-

dle network dynamics over multiple time scales (changing user demands and users entering/exiting the network). When demands change new prices must be calculated, typically, off-line. Such methods are not suitable for allocating resources for VBR traffic. Furthermore, no method of wealth distribution is provided in order to achieve any desired measure of fairness in the economy. Therefore, a microeconomic-based allocation method should: achieve the advantages associated with microeconomic approaches (distributed technique and fair allocations), handle network dynamics, and have a reasonable implementation cost.

In the next chapter, a review of the competitive market is given. This market model will serve as the basis of the microeconomic-based allocation methods presented in chapters 6 and 7. The competitive market model has the ability to achieve efficient and fair resource allocations. How to distribute wealth in order to achieve certain measures of fairness is also provided. A modified version of this market model is then used in chapters 6 and 7 to allocate network resources.