# Utility-Based Resource Allocation and Scheduling in OFDM-Based Wireless Broadband Networks

Guocong Song, Motorola Inc. and Ye (Geoffrey) Li, Georgia Institute of Technology
Email: guocongsong@motorola.com and liye@ece.gatech.edu

**Abstract**

This article discusses downlink resource allocation and scheduling for *orthogonal frequency division multiplexing* (OFDM)-based broadband wireless networks. We present a cross-layer resource management framework leveraged by utility optimization. It includes utility-based resource management and QoS architecture, resource allocation algorithms, rate-based and delay-based multichannel scheduling that exploits wireless channel and queue information, and theoretical exploration of the fundamental mechanisms in wireless resource management, such as capacity, fairness, and stability. We also provide a solution that can efficiently allocate resources for heterogeneous traffic with diverse QoS requirements.

## I. Introduction

The allocation and management of resources are crucial for wireless networks, in which the scarce wireless spectral resources are shared by multiple users. In the current dominate layered networking architecture, each layer is designed and operated independently. However, wireless channels suffer from time-varying multipath fading; moreover, the statistical channel characteristics of different users are different. The suboptimality and inflexibility of this architecture result in inefficient resource use in wireless networks. We need an integrated adaptive design across different layers.

Recently, the principles of multiuser downlink and MAC designs have been changed from the traditional point-to-point view to a multiuser network view. For instance, channel-aware scheduling strategies are proposed to adaptively transmit data and dynamically assign wireless resources based on *channel state information* (CSI). The key idea is to choose one user with good channel conditions to transmit packets [1]. Taking advantage of the independent channel variation across users, channel-aware scheduling can substantially improve the network performance through *multiuser diversity*, whose gain increases with the number of users [1], [2]. From a user point of view, packets are transmitted in a stochastic way in the system using channel-aware scheduling, which is also called opportunistic communications [3]. Based on its concept, channel-aware dynamic packet scheduling is applied in 1x *evolution* (1xEV) for

The view and opinions expressed in this article are those of the authors and do not reflect Motorola's current position.

*code division multiple access 2000* (CDMA2000) and *high speed downlink packet access* (HSPDA) for wideband CDMA.

The growth of Internet data and multimedia applications requires high-speed transmission and efficient resource management. To avoid *intersymbol interference* (ISI), *orthogonal frequency division multiplexing* (OFDM) is desirable for high-speed wireless communications. OFDM-based systems are traditionally used for combating frequency-selective fading. From a resource allocation point of view, however, an OFDM system naturally has a potential for more efficient *median access control* (MAC) since subcarriers can be assigned to different users [4], [5]. Another advantage of OFDM is that adaptive power allocation can be applied for a further performance improvement.

In this article, we present a cross-layer framework for the downlink resource management of *Internet protocol* (IP)-based OFDM wireless broadband networks that is able to effectively enhance spectral efficiency and guarantee *quality of service* (QoS) built on a utility-optimization-based architecture. In this architecture, exploiting knowledge of the CSI and the characteristics of traffic, the network aims to maximize the total utility, which is used to capture the satisfaction levels of users. This article focuses on both developing resource allocation algorithms and exploring the fundamental mechanisms in resource management, such as capacity, fairness, and stability, in multiuser frequency-selective fading environments. Another contribution of this article is to develop an effective solution for handling multiple types of traffic (non-real time and real time) with diverse QoS requirements.

## II. Resource Management in OFDM-Based Broadband Wireless Networks

In this section, we will briefly introduce resource management in OFDM-based broadband networks, including the system model, challenges and techniques of adaptive resource management, and related standardization activities.

### *A. Adaptive Resource Allocation Techniques in OFDM-Based Networks*

The architecture of a downlink data scheduler with multiple shared channels for multiple users is shown in Figure 1. OFDM provides a physical basis for the multiple shared channels, where the total bandwidth $B$ is divided into $K$ subcarriers. The OFDM signaling is time-slotted, and the length of each time slot is $T_s$. The base station simultaneously serves $M$ users, each of which has a queue to receive its incoming packets. Let $\mathcal{M} = \{1, 2, \cdots, M\}$ denote the user index set. To achieve high efficiency, both frequency and time multiplexing are allowed in the whole resource. The scheduler makes a resource assignment once at every slot.

OFDM-based networks offer more degrees of flexibility for resource management compared to single-carrier networks. Taking advantage of knowledge of the CSI at the transmitter (base station), OFDM-based systems can employ the following adaptive resource allocation techniques:

- *Adaptive modulation and coding* (AMC): the transmitter can send higher transmission rates over the subcarriers with better conditions so as to improve throughput and simultaneously to ensure an

acceptable *bit-error rate* (BER) at each subcarrier. Despite the use of AMC, deep fading at some subcarriers still leads to low channel capacity.

- *Dynamic subcarrier assignment* (DSA): the base station dynamically assigns subcarriers according to CSI or/and QoS requirements. Channel characteristics for different users are almost mutually independent in multiuser environments; the subcarriers experiencing deep fading for one user may not be in a deep fade for other users; therefore, each subcarrier could be in good condition for some users in a multiuser OFDM wireless network. Besides, frequency multiplexing provides fine granularity for resource allocation.
- *Adaptive power allocation* (APA): the base station allocates different power levels to improve the performance of OFDM-based networks, which is called multiuser water filling.

Employing these adaptive techniques at each subcarrier results in a large control overhead. In practice, serval subcarriers can be grouped into a cluster (subchannel), in which we apply those adaptive techniques. The size of a cluster determines the resource granularity. Obviously, the resource allocation schemes or algorithms designed for a subcarrier-based adaptive OFDM system can be directly used in a cluster-based system.

The major issue is how to effectively assign subcarriers and allocate power on the downlink of OFDM-based networks by exploiting knowledge of the CSI and the characteristics of traffic to improve spectral efficiency and guarantee diverse QoS. Three main challenges for cross-layer design for resource management in OFDM-based networks are present as follows:

- DSA belongs to the matching or bin packing problems in discrete optimization, which are mostly NP-hard or NP-complete.
- Unlike a single-carrier network, a multicarrier network can serve multiple users at the same time; hence, the design of multicarrier scheduling for bursty traffic is a new and interesting problem.
- The general relationship among spectral efficiency, fairness, and the stability property of wireless scheduling are not clear for wireless networks with time-varying fading.

All above problems are crucial for establishing high-speed and efficient wireless Internet networks.

### *B. Standardization Activities*

OFDM is already widely adopted in IEEE 802.11 *wireless local area networks* (WLANs) and the digital audio and video broadcasting systems in Europe. However, these standards do not support frequency multiplexing for multiple access. The IEEE 802.16 standard [6], which is developed for *broadband wireless access* (BWA) networks, specifies two flavors of OFDM systems: OFDM and *OFDM access* (OFDMA). In the OFDMA mode, 1536 data subcarriers out of 2048 ones are equally divided into 32 subchannels, which can be assigned to different users. Thus, DSA is an important function for improving the efficiency of resource allocation in the OFDMA mode. Note that the standard only specifies the system structure to guarantee inter-operability among multiple vendors' equipment and allows them to differentiate their equipment. In summary, the IEEE 802.16 standard supports DSA, but the details

of DSA and scheduling algorithms are left unstandardized for vendors' selection. Therefore, advanced resource management and scheduling are a crucial part that determines the spectral efficiency and the QoS capability of equipment from different vendors.

## III. Utility-Based Resource Management and QoS Framework

An effective trade-off among spectral efficiency, fairness, and QoS is desired in wireless resource allocation. The issues on efficient and fair resource allocation have been well studied in economics, where utility functions are used to quantify the benefit of usage of certain resources. Similarly, utility theory can be used in communication networks to evaluate the degree to which a network satisfies service requirements of users' applications, rather than in terms of system-centric quantities like throughput, outage probability, packet drop rate, power, etc. [7]. The basic idea of utility-pricing structures is to map the resource use (bandwidth, power, etc.) or performance criteria (data rate, delay, etc.) into the corresponding utility or price values and optimize the established utility-pricing system. In this article, we introduce the use of a utility-based architecture for cross-layer resource management of OFDM-based networks.

### *A. Utility Functions*

Representing the level of customer satisfaction received for the system, utility functions play a key role in resource management and QoS differentiation. Different applications have different utility function curves or even different parameters. For instance, the utility functions of best-effort applications are with respect to throughput, whereas those of delay-sensitive applications are with respect to delay. There are usually two approaches to obtaining utility functions. For a specific type of application, the utility function may be obtained by sophisticated subjective surveys. Another method is to design utility functions based on the habits of the traffic and appropriate fairness in the network. Therefore, a utility function for an application characterizes its corresponding QoS requirements.

### *B. Utility-Based Optimization Formulation*

Since resource management in the downlink has a central controller – the base station, we formulate the cross-layer optimization as *one that maximizes the aggregate utility in the system subject to the capacity limit determined by the physical-layer techniques*. Note that in this formulation, the optimization constraints come only from the physical layer. We put other constraints regarding QoS into the optimization objective – utility functions through mathematical methods or in light of the physical meanings of those constraints. This formulation has two advantages:

- For a new application, we only need to change the corresponding utility function and still use the existing optimization algorithms since the optimization structure is not changed.
- Because the optimization constraints determine the system capacity region, this mechanism makes fairness, stability, and QoS tractable and even analytical, which will be shown in the next section.

### C. DSA and APA Algorithms

Usually, the formulated utility optimization problems cannot be directly solved since most optimization objectives are in terms of long-term performance criteria such as long-term average data rates. However, DSA or/and APA is performed at each time slot. Thus, the original optimization objective needs converting into an instantaneous optimization objective that is related to instantaneous data rates. In most cases, it is easy to do that. Then, the instantaneous optimization objective can be regarded as a summation of utility functions with respect to instantaneous data rates, which is defined as $\sum_i U_i(r_i[n])$, where $r_i[n]$ is the instantaneous data rate for user $i$ at time $n$, and $U_i(\cdot)$ is the corresponding utility function.

As mentioned before, those algorithms that maximize $\sum_i U_i(r_i[n])$ serve as an important part in this utility-based optimization architecture. However, developing algorithms is very challenging since the utility functions are usually nonlinear, the DSA problem can be regarded as a nonlinear combinatorial optimization. Algorithm development has been carried out in [8]. The major results are summarized as follows:

*1) DSA Algorithms:* If the utility functions are all linear, the utility function of user $i$ is given by $U_i(r_i[n]) = U_i' r_i[n]$, where $U_i'$ is the marginal utility function (derivative) of $U_i(\cdot)$. In this case, subcarrier assignment is independent for different subcarriers, which means that the assignment of a subcarrier does not affect assigning other subcarriers. DSA can be done by a simple gradient scheduling algorithm $m(k,n) = \arg\max_{i\in\mathcal{M}}\{U_i' \cdot c_i[k,n]\}$, where $m(k,n)$ represents that subcarrier $k$ is assigned to user $m(k,n)$ at time $n$, and $c_i[k,n]$ is the achievable data rate for subcarrier $k$ at time $n$, which is fully determined by the CSI at that time. For a more general scenario in which the utility functions are nonlinear, assigning different subcarriers is not independent anymore, and DSA becomes very complicated. In the case of concave utility functions, which are applicable to most applications, we have developed sorting-search algorithm, in which sequence sorting and binary search are mainly used. This algorithm has several advantages. The search-sorting algorithm has no convergence problem since the total utility increases or stay unchanged after each step. Furthermore, the computational complexity of this algorithm is very low.

*2) APA Algorithms:* Theoretically, utility-based multi-level water filling is the optimal solution for the optimization with concave utility functions when continuous-rate AMC is used. However, since AMC only employs several rate levels in practice, this water filling cannot work in this scenario. We have developed a greedy bit-power allocation algorithm, which is able to achieve the optimality when the utility functions are concave. If subcarrier and power can both be adaptively adjusted, we can implement the DSA and APA algorithms iteratively.

In brief, if the utility functions are concave with respect to the instantaneous data rates, the above DSA and APA algorithms can solve the utility-based optimization effectively. Note that in a real system, DSA requires much less feedback information than APA. This is because APA requires the SNR value of each subcarrier while DSA only needs to know the achievable data rate at each subcarrier. Therefore, DSA is more practical for commercial systems. We will only discuss DSA in the following content; however, most results are applicable to APA or joint DSA and APA.

### D. Design Examples

Most network applications can be classified into two types: best-effort (non-real-time) and delay-sensitive (real-time) traffic. Here we introduce examples for each types of traffic.

*1) Rate-Based Utility Optimization for Best-Effort Traffic:* From a subject perspective, best-effort or elastic applications have no specific QoS requirements. From a system perspective, the throughput of a best-effort connection is controlled by its transport layer according to the level of network congestion. A well-accepted utility function for best-effort traffic is $\ln(\bar{r}_i[n])$ [9], where $\bar{r}_i[n]$ is the long-term average throughput for user $i$. For generality, we assume user $i$ to have a concave utility function $U_i(\bar{r}_i[n])$. It is shown in [8] that the instantaneous optimization maximizing $\sum_i U_i'(\bar{r}_i[n])r_i[n]$ leads to a long-term optimization that maximizes $\sum_i U_i(\bar{r}_i[n])$. Therefore, the gradient scheduling algorithm is

$$m(k,n) = \arg\max_{i\in\mathcal{M}}\{U_i'(\bar{r}_i[n]) \cdot c_i[k,n]\}. \tag{1}$$

A more mathematically rigorous proof is provided in [10]. Although it focuses on single-carrier systems, the method can be extended to multicarrier systems.

*2) Delay-Based Utility Optimization for Delay-Sensitive Traffic:* The incoming rate of a delay-sensitive stream is usually determined by its source. Assume that user $i$ is associated with an average waiting time $W_i[n]$ and the corresponding utility is $U_i(W_i[n])$. Obviously, with a long delay, the user has a low level of satisfaction (utility). It is reasonable to assume that $U_i(W_i[n])$ is decreasing. The long-term optimization objective with respect to average waiting times leads to an instantaneous optimization objective, which is given in [11] by

$$\max \sum_{i\in\mathcal{M}} \frac{\left|U_i'(W_i[n])\right|}{\bar{r}_i[n]} \min(r_i[n], \frac{Q_i[n]}{T_s}), \tag{2}$$

where $Q_i[n]$ is the queue length of user $i$. The $\min(x,y)$ function is to make sure that the service bits of each user should be less than or equal to the accumulated bits in its queue to avoid bandwidth wastage. The average waiting time of each user can be estimated by utilizing the information about the queue length and the service rate. We call this scheduling *Max-Delay-Utility* (MDU) scheduling. Obviously, the MDU is a joint channel- and queue-aware scheduling scheme. Since function $\min(x,y)$ is concave, the MDU scheduling needs the sorting-search algorithm as a solution.

## IV. Capacity, Fairness, Stability, and QoS of Utility-Based Optimization Framework

To fully understand this utility-based cross-layer design for OFDM-based wireless networks, we discuss capacity, fairness, stability, and QoS issue in this section. Furthermore, this study directly leads to an effective and simple technique for QoS differentiation, which is also introduced in this section.

### A. Capacity Region

As mentioned before, the degrees of freedom of resource allocation determine the long-term average capacity. Figure 2 demonstrates how adaptive physical-layer techniques result in capacity improvement

in a two-user scenario. A fixed modulation and coding scheme must consider the worst case; as a result, each user has the same transmission efficiency (bps/Hz). Adjusting transmission data rates according to users' channel conditions, AMC significantly enlarges the system capacity. With AMC, different users achieve different transmission efficiencies, but the efficiency ratio of two users (one user to another user) is still constant. The combination of AMC and DSA is able to further improve the capacity through multiuser diversity [2]. In this scenario, even the efficiency ratio of two users varies. Therefore, it is not easy to handle QoS guarantees if DSA is used although DSA increases the capacity.

### *B. Fairness and Rate-Based Utility Optimization*

Since best-effort traffic has no specific QoS requirements, fairness among those users sharing the same bandwidth is one of the most important criterion. Proportional fairness [9] provides each connection a priority inversely proportional to its long-term average throughput and lets each connection have an equal access chance. It can be effectively used in Internet networks since it is desirable to best-effort traffic. Using convex analysis, we have revealed the relationship between a concave utility function with respect to the long-term average data rate and a certain type of fairness in [8]. In other words, a concave utility function is associated with its corresponding fairness. Since the utility-based gradient scheduling (1) leads to maximizing the sum of the utility functions, it directly achieves the fairness related to the used utility function. The relationship directly shows that the logarithmic utility function is associated with the proportional fairness for the utility-based optimization. The corresponding scheduling is called *proportionally fair* (PF) scheduling.

### *C. Stability and Delay-Based Utility Optimization*

Unlike best-effort traffic, a necessary condition for guaranteeing the QoS requirements of a delay-sensitive stream is that the service rate must be larger than the incoming rate of the stream. Therefore, the study of stability issue is the key to analyze the performance of scheduling algorithms for delay-sensitive traffic. This is because the stability issue is essential for QoS provisioning and admission control. Moreover, the stability issue is mathematically tractable in many cases. For a queueing system, the system is stable if each queue length reaches a steady state and does not go to infinity. There are two important methods to deal with the stability issue: Foster-Lyapunov drift [12], [13] and fluid limit [14]. The Foster-Lyapunov mothed is classical for stability and harmonic analysis. The fluid limit technique establishes the equivalency on stability between the orignal network and the associated fluid model with deterministic and continuous arrival streams.

The *stability region* of a policy is defined as the set of all possible arrival rate vectors for which the system is stable with the policy [12]. Note that the capacity region is concerned with the service data rates, whereas the stability region is with regard to the arrival rates. The *maximum stability region* (MSR) is defined as the largest stability region that can be achieved by some scheduling schemes. Similarly, a policy is called an MSR policy if the stability region of the policy covers all stability regions under all other policies. Two main results made in [15] regarding the MSR are stated as follows:

- The MSR covers any point within the long-term average capacity region.
- The MDU scheduling can guarantee the MSR when the marginal utility functions (with respect to average delays) are polynomials. Actually, the conditions for the MDU scheduling to achieve the MSR can be further loosed [15]. In reality, polynomial delay-based utility functions are enough to quantify QoS in most cases.

Generally, channel-aware-only scheduling schemes cannot reach the MSR since they are unable to sense the queueing information, which reflects the status of networks. Figure 3 illustrates the stability regions of MDU and PF scheduling. It is seen from the figure that the MDU scheduling can fully exploit the capacity enhanced by AMC and DSA. Note that the concept of MSR policy is interchangeable with the concept of throughput-optimal policy in [14]. As a result, *modified largest weighted delay first* (M-LWDF) scheduling [14] is also an MSR policy. Its multichannel version [15] is to schedule a user with the highest value of $T_{\mathrm{HOL},i}c_i[k,n]/\bar{r}_i[n]$ for subcarrier $k$, where $T_{\mathrm{HOL},i}$ is the delay of the head-of-line packet of user $i$.

### *D. Diverse QoS Guarantee*

Guaranteeing QoS for multiple types of traffic is challenging to resource allocation and scheduling, especially for wireless data networks. Due to a small stability region, channel-aware-only scheduling such as PF scheduling is inefficient for delay-sensitive applications. Although MSR scheduling schemes have the largest stability region, they may not guarantee good QoS provisioning since the MSR is only a necessary condition for QoS guarantee. On the other hand, the trend of MSR scheduling to stabilize all connections may cause best-effort connections to aggressively occupy the bandwidth in a scenario in which the scheduler serves both best-effort and delay-sensitive traffic. This is because the sources of best-effort connections may increase transmission rates, resulting in competing more resources from delay-sensitive connections. In addition, handling complicated QoS requirements is really a challenge.

In this article, we employ the MDU scheduling for a mixture of delay-sensitive and best-effort traffic by exploiting the powerful and flexible control capability of the utility-based architecture. To apply the MDU scheduling, we need to design utility functions with respect to average waiting times for the corresponding QoS requirements. Since the marginal utility functions are proportional to the scheduling weights, the marginal utility functions, the $U_i'(\cdot)$'s, play a crucial role in scheduling. Therefore, we can directly design the marginal utility functions rather than the utility functions themselves. They can be designed based on both certain objective and subjective performance criteria. The objective consideration here is the system stability. Designing marginal utility functions is based on the following two rules:

- Let the marginal utility functions of delay-sensitive applications satisfy the MSR conditions, which are discussed in the previous subsection. The more specific design of the marginal utility functions is based on the subjective performance criteria of certain applications.
- Make the marginal utility functions of best-effort applications bounded to control the greed of their connections. If a best-effort connection is not stable, packet losses will happen, then its transport-layer mechanisms (such as TCP) will reduce its data rate to make the connection stable again.

It follows from the design that

$$\lim_{W\to\infty} \frac{U'_{\text{best effort}}(W)}{U'_{\text{delay sensitive}}(W)} = 0. \tag{3}$$

The above equation indicates that the MDU scheduling can sense the level of network congestion. If the network is congested, best-effort connections hardly obtain resources to transmit packets according to (3). Therefore, the MDU scheduling does not allow those best-effort connections to affect the stability of delay-sensitive connections. If the network load is low, the scheduler can automatically assign more resources to those best-effort connections.

## V. Simulation Results

In this section, we demonstrate simulation examples that take into account the impacts of different traffic types and average SNRs on scheduling performance. More details are referred to [15].

### *A. Marginal Utility Functions*

We can design the marginal utility functions according to the corresponding required QoS for packet-switched voice, streaming, and best-effort traffic, which are shown in Figure 4. For packet-switched voice or *voice over IP* (VoIP), end-to-end delays are usually required less than 100 ms. Good-quality streaming transmission needs end-to-end delays between 150-400 ms. For best-effort traffic, we can still assign the marginal utility function in terms of average waiting time. In fact, the MDU scheduling for best-effort traffic becomes the PF scheduling if average waiting times are large.

### *B. Simulation Conditions*

For comparison, we assume that the number of each traffic type is an even integer. For each type of traffic, half of users, called good users, have an average SNR of 15 dB; the rest, called bad users, have an average SNR of 8 dB. In the simulation, each bad user's channel suffers multipath Rayleigh fading with the delay profile of Channel B for outdoor to indoor and pedestrian environments of International Mobile Telecommunications-2000 (IMT-2000), and each user is assumed to be stationary or slowly moving so that the maximum Doppler shift is 10 Hz. Each good user experiences Rician fading with a factor of 0.5 whose delay profile and Doppler shift are the same as those of bad users' channels. In the OFDM network, there are 256 subcarriers in a total channel bandwidth of 2.048 MHz. These subcarriers are grouped into 32 clusters, each of which can be dynamically assigned to a user during a time slot. Assume that a set of achievable transmission rates in bps/Hz is $\{0,1/2,1,2,3,4\}$.

The traffic model for voice traffic is the on-off voice activity model with exponentially distributed duration of voice spurts and gaps. The average talk spurt is 1.00 s, and the average silent interval is 1.35 s. Within each talk spurt interval, a 32 kbps digital voice coding is assumed. In the model of video streaming, the duration of each state is exponentially distributed with a mean of 160 ms. The minimum, maximum, and average data rates in each state are 64, 256, and 180 kbps, respectively. A full-buffer model in which there are infinite data packets in the queues is applied to best-effort traffic. Although this model may not be realistic, it can obtain the maximum achievable throughput for best-effort traffic.

### C. Simulation Results

We show two examples in the simulation and compare the performance of the MDU scheduling and that of the multichannel version of a combination of M-LWDF and PF scheduling, which is called M-LWDF-PF [15]. The performance of delay-sensitive traffic is evaluated in terms of 95th percentile delay, and that of best-effort traffic is measured in terms of average throughput. We focus on the properties of the MDU scheduling at first.

*1) Increase of streaming users:* In this examples, we fix the numbers of voice and best-effort users both to be 20 and increase the number of streaming users. We can clearly see the performance in both less-congested and congested situations in Figure 5. When the network is less-congested (the number of streaming users does not exceed 16), the MDU scheduling can maintain high-quality delay performance for those delay-sensitive applications and provide a high data rate for the best-effort users. When the network is congested, e.g. in the 20-streaming-user case, the throughput for the best-effort users becomes extremely small, and the delay for the streaming users has a dramatical increase. However, the performance of the voice users is still very good.

*2) Increase of best-effort users:* In this example, we fix the numbers of voice and streaming users to be 20 and 10, respectively, and increase the number of best-effort users. It is seen from Figure 6 that as the number of the best-effort users increases, the performance of the voice and the streaming users remains very well with the MDU scheduling, and the throughput for the best-effort connections increases, which results from multiuser diversity.

Therefore, we can in these two examples see the excellent mechanisms of the MDU scheduling: high spectral efficiency by taking advantage of knowledge of CSI and good diverse QoS provisioning by exploiting utility functions. We also compare the MDU with the M-LWDF-PF in Figures 5 and 6. Note that the M-LWDF scheduling is also a scheduling scheme that can adjust resource allocation according to users' channel and queue state information and has the MSR. All examples show that both scheduling schemes offer similar delay performance for the voice users, and that in most of the cases, the MDU scheduling provides considerably smaller delays for streaming traffic than the M-LWDF-PF while the MDU allows the best-effort users to achieve higher throughput than the M-LWDF-PF at the same time. This is mainly because the MDU scheduling more appropriately captures the required QoS compared to other scheduling schemes. In addition, the MDU scheduling does not need statistical information about incoming traffic, and its implementation complexity is very low.

## VI. Conclusion

In this article, we investigate resource management and scheduling in a wireless OFDM-based downlink that serves multiple users and supports various applications based on a cross-layer approach. The current standardization activities of broadband wireless networks provide a chance for DSA and advanced multichannel scheduling to be implemented in commercial systems. We not only present a utility-based cross-layer wireless resource management architecture and corresponding scheduling algorithms that substantially improve spectral efficiency and satisfy diverse performance objectives of heterogeneous

traffic, but also provide a theoretical framework that allows us to understand the fundamental mechanisms in state-of-the-art wireless resource management, including capacity, fairness, and stability. Even though much efforts are required to fully understand the theories behind these advanced adaptive resource management techniques, their implementation is quite simple and effective.

## ACKNOWLEDGMENT

This work was supported by the National Science Foundation under Grand CCR-0121565, by the U.S. Army Research Laboratory under the Collaboration Technology Alliance Program, Cooperative Agreement DAAD19-01-20-0011, and by Motorola Inc.

## REFERENCES

[1] P. Viswanath, D. N. C. Tse, and R. L. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inform. Theory*, vol. 48, no. 6, pp. 1277–1294, June 2002.

[2] R. Knopp and P. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc., IEEE Int. Conf. on Commun.*, Seattle, WA, June 1995.

[3] X. Liu, E. Chong, and N. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE J. Select. Areas Commun.*, vol. 19, no. 10, pp. pp. 2053–2064,, Oct. 2001.

[4] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE J. Select. Areas Commun.*, vol. 17, no. 10, pp. 1747–1758, Oct. 1999.

[5] J. Chuang and N. Sollenberger, "Beyond 3G: Wideband wireless data access based on OFDM and dynamic packet assignment," *IEEE Commun. Magazine*, pp. 78–87, July 2000.

[6] IEEE Std 802.16-2004, "IEEE Standard for Local and Metropolitan Area Networks – Part 16: Air Interface for Fixed Broadband Wireless Access Systems," Oct. 2004.

[7] S. Shenker, "Fundamental design issues for the future Internet," *IEEE J. Select. Areas Commun.*, vol. 13, no. 7, pp. 1176–1188, Sept. 1995.

[8] G. Song and Y. (G). Li, "Cross-layer optimization for OFDM wireless network – part I and part II," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 614–634, March 2005.

[9] F. Kelly, "Charging and rate control for elastic traffic," *European Trans. On Telecommunications*, vol. 8, pp. 33–37, 1997.

[10] A. L. Stolyar, "On the asymptotic optimality of the gradient scheduling algorithm for multi-user throughput allocation," *Operations Research*, vol. 53, no. 1, pp. 12–25, 2005.

[11] G. Song, Y. (G). Li, L. J. Cimini, and H. Zheng, "Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels," in *Proc., IEEE Wireless Commun. Networking Conf.*, Mar 2004.

[12] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling for maximum throughput in multihop radio networks," *IEEE Trans. Automatic Control*, vol. 37, no. 12, pp. 1936–1949, Dec. 1992.

[13] A. Eryilmaz, R. Srikant, and J. Perkins, "Stable scheduling policies for fading wireless channels," *IEEE/ACM Trans. Networking*, vol. 13, no. 2, pp. 411– 424, Apr. 2005.

[14] M. Andrews, S. Borst, F. Dominique, P. Jelenkovic, K. Kumaran, K. Ramakrishnan, and P. Whiting, "Dynamic bandwidth allocation algorithms for high-speed data wireless networks," Bell Labs Technical Memorandum, Tech. Rep., 2000.

[15] G. Song, "Cross-layer resource allocation and scheduling in wireless multicarrier networks," Ph.D. dissertation, Georgia Institute of Technology, 2005.

PLACE PHOTO HERE

**GUOCONG SONG** (guocongsong@motorola.com) received his B.S. and M.S. degrees form Tsinghua University, China, in 1997 and 2000, respectively, and his Ph.D. degree from Georgia Institute of Technology in 2005, all in Electrical Engineering. Since September 2005, he has been with Motorola Inc., Libertyville, IL. His research interests are in wireless communications and networking, with a current focus on cross-layer design and optimization for wireless networks, system design and performance analysis for broadband wireless networks, and multiple antenna systems.

PLACE PHOTO HERE

**YE (GEOFFREY) LI** (liye@ece.gatech.edu) received his B.S.E. and M.S.E. degrees in 1983 and 1986, respectively, from the Department of Wireless Engineering, Nanjing Institute of Technology, Nanjing, China, and his Ph.D. degree in 1994 from the Department of Electrical Engineering, Auburn University, Alabama. After spending several years at AT&T Labs - Research, he joined the School of Electrical and Computer Engineering at Georgia Tech as an Associate Professor in 2000. His general research interests include statistical signal processing and wireless communications. In these areas, he has contributed over 100 papers published in referred journals and presented in various international conferences. He also has over 10 USA patents granted or pending. He once served as a guest editor for two special issues on Signal Processing for Wireless Communications for the *IEEE J-SAC* and an editorial board member of EURASIP Journal on Applied Signal Processing. He is currently serving as an editor for Wireless Communication Theory for the *IEEE Transactions on Communications*. He organized and chaired many international conferences, including Vice-Chair of *IEEE 2003 International Conference on Communications*.

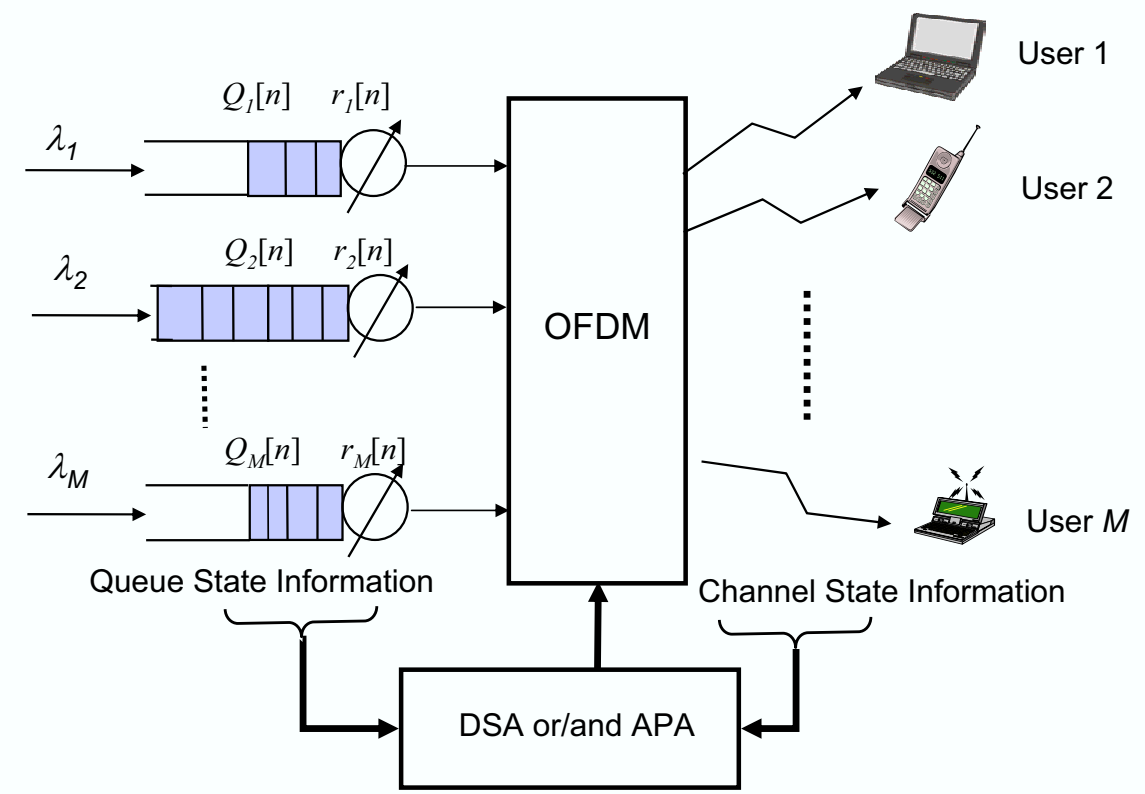



Fig. 1. Downlink data scheduling over multiple shared channels based on OFDM

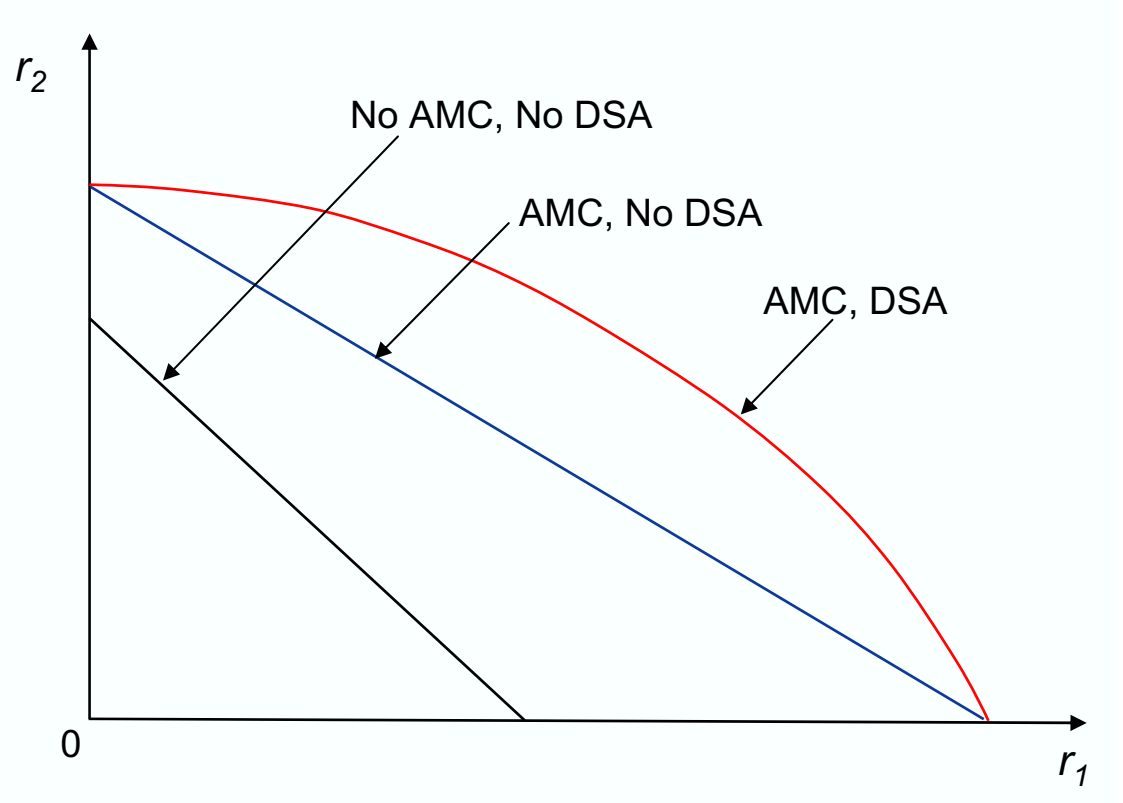



Fig. 2. Long-term average capacity region

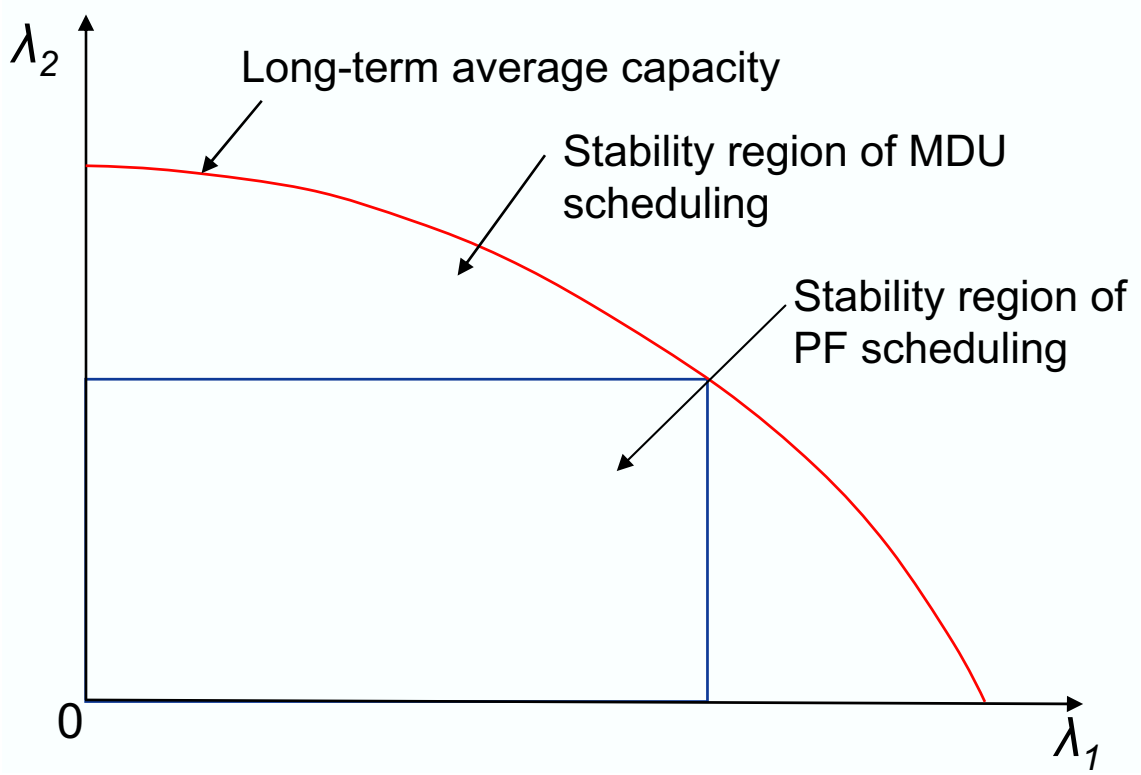



Fig. 3. Stability regions for MDU and PF scheduling

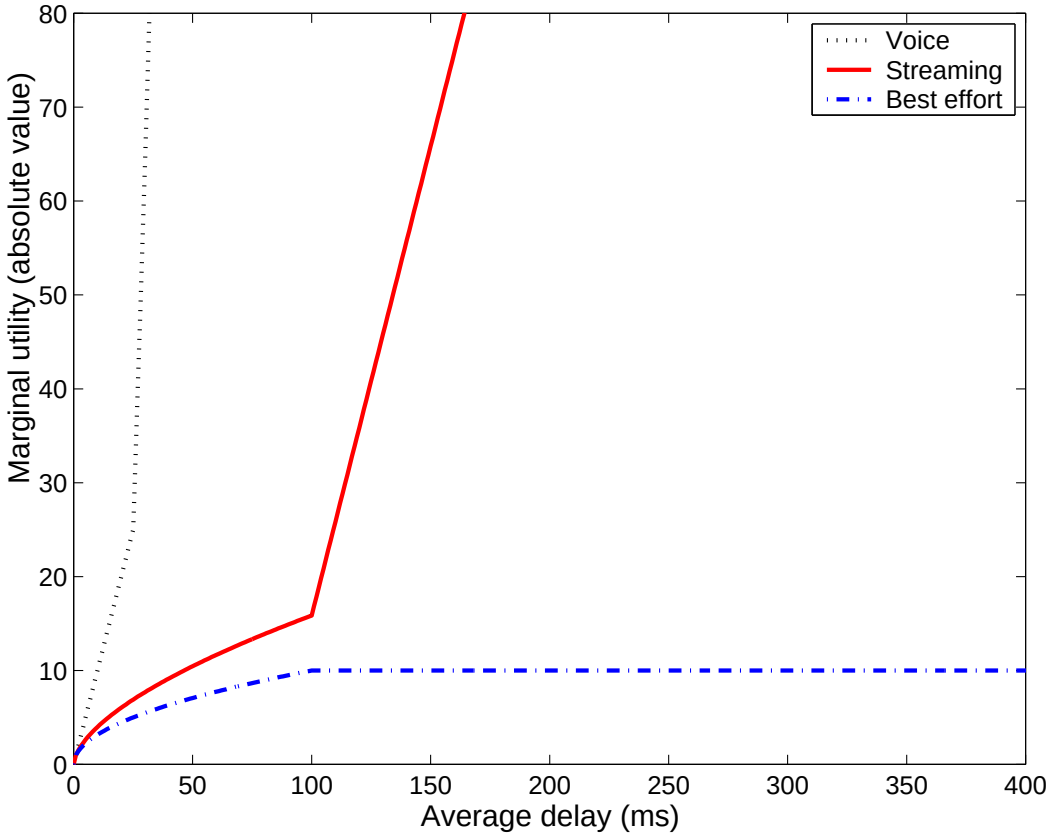
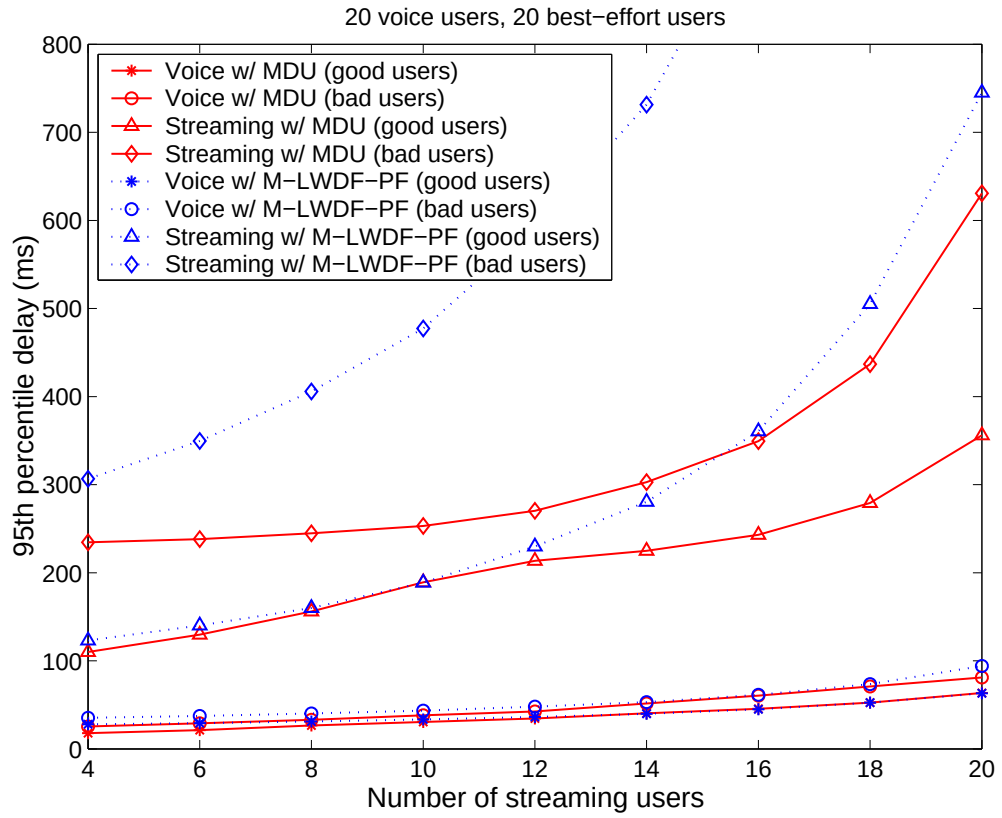
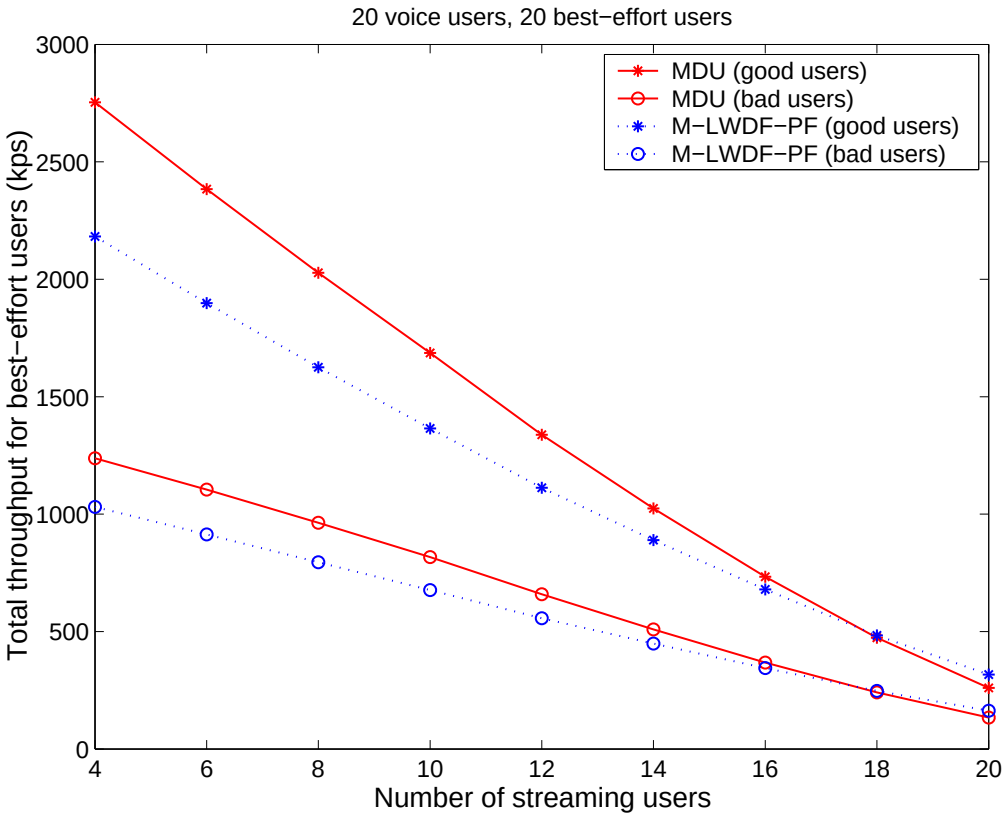



Fig. 4. Marginal utility functions

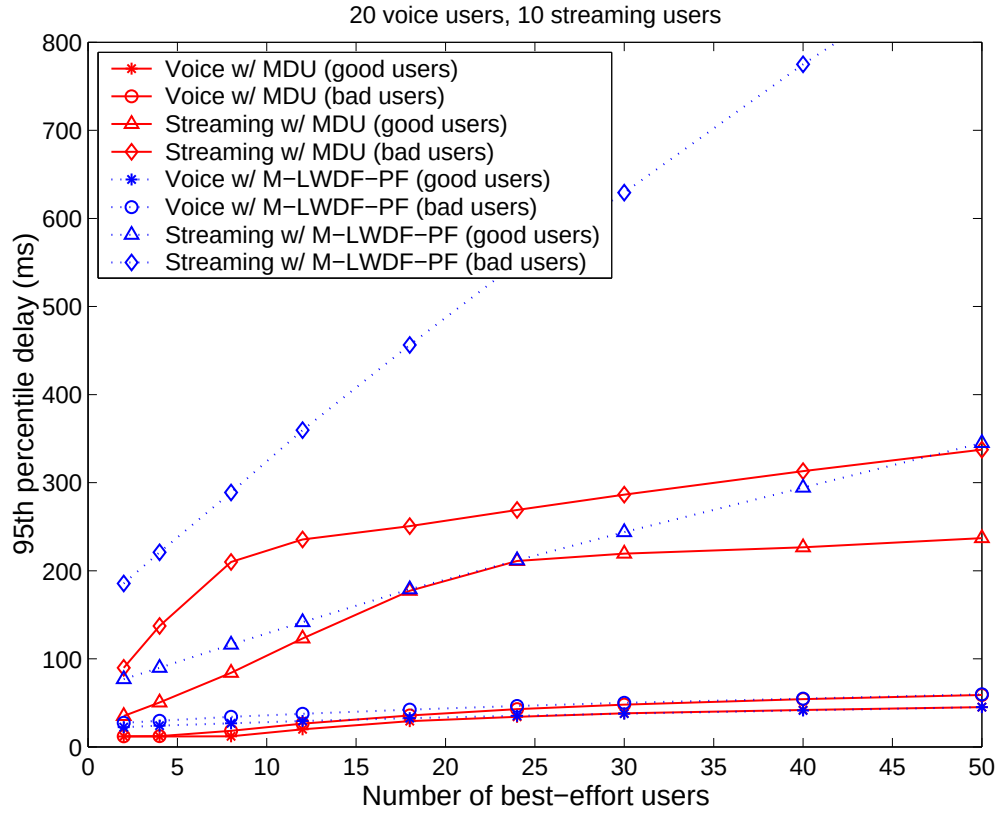



(a) 95th percentile delay for voice and streaming traffic

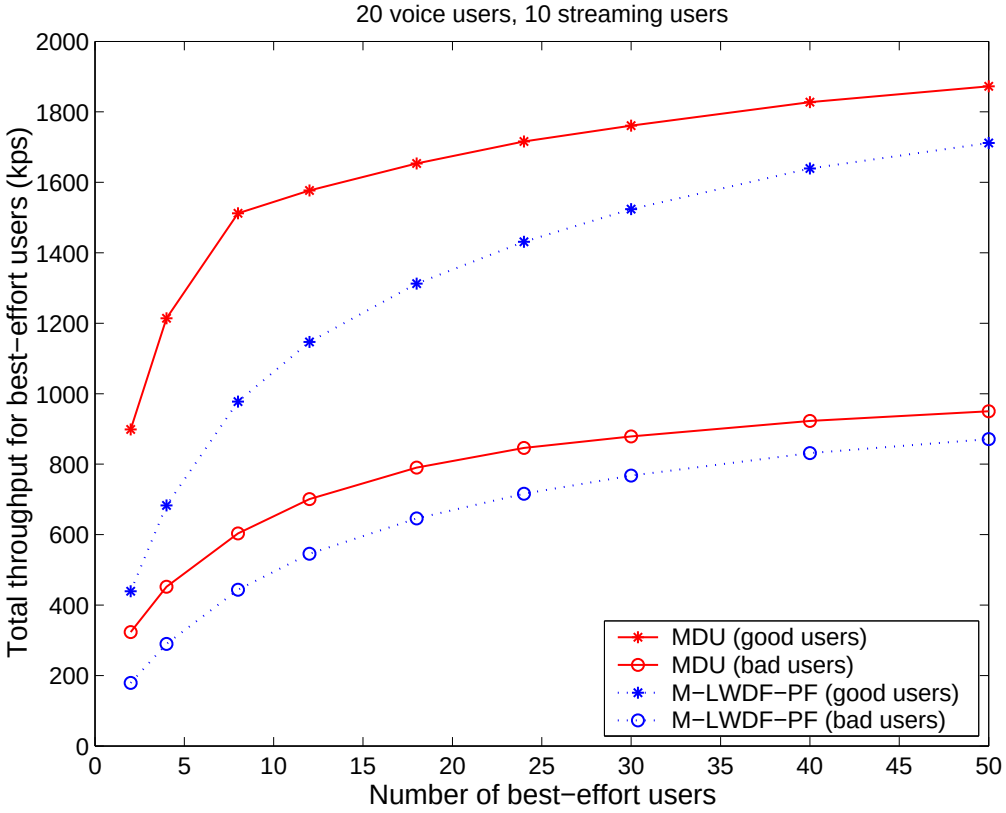



(b) Average total throughput for best-effort traffic

Fig. 5. Heterogeneous traffic performance versus the number of streaming users




(a) 95th percentile delay for voice and streaming traffic




(b) Average total throughput for best-effort traffic

Fig. 6. Heterogeneous traffic performance versus the number of best-effort users