Scheduler Design for Heterogeneous Traffic in Cellular Networks with Multiple Channels

Karthikeyan Sundaresan
Broadband & Mobile Networking
NEC Labs America
karthiks@nec-labs.com

Xiaodong Wang
School of ECE
Columbia University
wangx@ee.columbia.edu

Mohammad Madihian
Broadband & Mobile Networking
NEC Labs America
madihian@nec-labs.com

ABSTRACT

The design of an efficient base station scheduler with the ability to support different kinds of IP traffic, ranging from conventional data to real-time IP services plays a crucial role in the *all-IP* convergence goal of next-generation cellular systems. In this context, we first consider the basic network utility based data (rate) scheduler and extend it to a more generic, unified scheduler, capable of handling heterogeneous (data and voice) traffic types and their respective parameters (rate, delay and jitter). More importantly, we then consider the case where both the base station and mobile users are equipped with multiple sub-channels as in OFDM systems. The introduction of multiple sub-channels exponentially increases the search space for scheduling decisions as well as significantly increases the feedback overhead. To reduce the complexity of scheduling and the feedback overhead without sacrificing appreciably on performance, we propose a *parameter-based optimization*. While such an optimization does not reduce the complexity and overhead in single channel systems, we show that it has the potential to significantly reduce complexity and overhead in multiple channel systems. This is achieved in the form of a unified scheduling algorithm that identifies and exploits specific transmission strategies that optimize individual parameters. This is verified through comprehensive evaluations in a packet-level event-driven network simulator.

Categories and Subject Descriptors

C.2.1 [**Network Architecture and Design**]: Wireless Communication

General Terms

Design, performance, algorithms

1. INTRODUCTION

Next generation cellular systems are aimed at adopting advanced technologies like multiple-input multiple-output (MIMO) and orthogonal frequency division multiplexing (OFDM) for providing improved network services and aiding in the convergence towards an *all-IP* network. One of the key requirements in such an endeavor, is to have an efficient scheduler at the base station that is capable of effectively using the available network resources to satisfy heterogeneous user requirements ranging from data to VoIP, streaming, etc. This in turn, characterizes the two main components of an efficient scheduler design, namely, (i) accommodation of heterogeneous traffic types (parameters), and (ii) exploitation of physical layer degrees of freedom (DoF) like OFDM, MIMO, etc. for improved network performance. In this context, we focus on OFDM-based systems that carry data and voice traffic with an emphasis on rate, delay and jitter parameters.

There exist several scheduling rules for individual traffic types and their respective parameters like rate, delay, etc. [7, 11, 8, 3, 2]. However, when it comes to heterogeneous traffic (eg. mixed data and voice flows), there are few schedulers available [4, 12, 5]. Even among these heterogeneous schedulers, most of them do not provide the ability to flexibly trade-off resources between the different traffic types. This critically limits the ability to provide fine-grained QoS, which is a key requirement for next-generation cellular networks. Further, while the packet-level schedulers exploit the inherent multi-user diversity (MUD) [6] present in the system, they fail to exploit the physical layer DoF like multiple sub-channels, antennas, etc. On the other hand, there are several scheduling rules designed in the physical layer that operate at a much finer granularity of bits and are focused primarily on exploiting the multi-user diversity as well as the physical layer capabilities [15, 13, 16]. However, they are limited in the parameters they optimize to predominantly rate and losses; and do not account for the increased complexity and overhead resulting from exploitation of DoF or sophisticated fairness models across users. Thus, the goals of this work and hence the contributions are two-fold:

- Design of a unified scheduler capable of handling heterogeneous traffic.
- Design of a scheduler capable of efficiently exploiting multiple sub-channels at the physical layer, albeit at a lower cost of complexity and overhead.

Specifically, the unified framework is designed on the basis on network utility maximization. It extends the existing utility based throughput optimization for best effort data

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
WICON 2007 October 22-24, 2007, Austin, Texas USA
Copyright 2007 ACM 987-963-9799-04-2/07/10 ...$5.00.

traffic [10, 13] to a more unified framework, capable of accommodating multiple traffic types that belong to different QoS classes, with the ability to effectively tradeoff resources between them. It is also capable of optimizing multiple parameters (rate, delay, jitter, etc.) in tandem along with desired fairness model. The scheduler is then extended to the case where both the base station and mobile users are equipped with multiple channels. The ability to leverage the advantages of OFDM sub-channels also exponentially increases the search space for scheduling decisions as well as significantly increases the communication overhead. To reduce the complexity of scheduling and the communication overhead without sacrificing appreciably on performance, we propose a *parameter-based optimization*. We show that while such an optimization does not reduce the complexity and overhead in single channel systems, it has the potential to significantly reduce complexity and overhead in multiple channel systems (MCS). This is achieved in the form of a unified scheduling algorithm that identifies and exploits specific transmission strategies that optimize individual parameters. The proposed scheduler for MCS significantly reduces complexity and communication overhead, while performing reasonably close to the optimal scheme. This is verified through comprehensive evaluations in a packet-level event-driven network simulator.

The rest of the paper is organized as follows. Section 2 presents the system model that will used for reference in the rest of the paper. Section 3 presents the design of the unified scheduling framework. The scheduler design for multiple channels is presented in Section 4. The design of both the scheduling framework and the efficient algorithms are evaluated in Section 5. Related work in literature is presented in Section 6, followed by concluding remarks in Section 7.

2. SYSTEM MODEL

We consider an OFDM-based single-cell cellular system. While the design of the scheduler applies to both downlink and uplink systems, we present the scheduler design in the context of a downlink system. The users are uniformly located within the cell radius. The distribution of users automatically generates asymmetry in user channel rates, while the mobility of users induces Rayleigh fading and consequently provides potential for multi-user diversity. The presence of multiple sub-channels at the BS and MS provides an additional degree of frequency diversity.

Multiple users, each capable of receiving data or voice flows and hence interested in one or more parameters (rate, delay, jitter) are considered. Users could belong to different traffic QoS classes, and even within a single traffic class, users could have different priorities based on their traffic type. We assume that the BS has a buffer for each of the users that it services. In addition to this, we assume that each user has a finite *virtual* buffer at the base station for each of the different traffic types that it services. If there are multiple flows for a user of a single traffic type, they will share the same virtual buffer. Further, the voice flows are assumed to be time-sensitive with a maximum tolerance level for packet delay and jitter. Any packet that violates these requirements is considered to be equivalent to a lost packet.

With multiple channels at both the base station and user terminals, single user (operating on all channels) as well multi-user (operating on a subset of channels) transmission strategies are considered. A K-user frequency selective broadcast fading channel is considered for the system. P represents the maximum power used by BS for its transmission, which is split equally across all sub-channels. The fading processes are assumed to be independent across users. Note that a sub-channel could correspond to a single carrier or a bunch of contiguous carriers as considered in practical systems. Since our scheduler is not specific to any particular receiver processing scheme, we consider the Shannon capacity for the different transmission strategies as the model for instantaneous channel rate estimation, namely

$$r_\ell(k) = \log_2\left(1 + \frac{P}{A}\frac{|H_k(f_\ell)|^2}{N_k(f_\ell)}\right)$$

where $H_k(f_\ell)$ and $N_k(f_\ell)$ represent the channel and noise frequency response for user k on sub-channel ℓ.

3. UNIFIED SCHEDULING FRAMEWORK

We resort to a *network utility* based framework for designing a unified scheduling algorithm. *Utility* refers to the amount of satisfaction perceived by an end-user/application for a certain amount of network resources received and hence directly captures end-user performance. Further, when maximizing a set of concave utility functions of all increasing/decreasing preferences, the nature of the utility function chosen automatically determines the nature of the fairness model that results in the system [10]. Utility based scheduler designs have been popular for throughput (rate) optimization of data traffic [10, 13]. In this section, we extend such designs to a unified utility framework for scheduling heterogeneous IP traffic in tandem, where the utility functions for different traffic types can be captured by concave, differentiable functions. This is a reasonable assumption for most of the common application requirements of data, voice and video for the parameters of rate, delay, and jitter. We consider the following notations in the rest of our discussions.

- F: number of traffic types in the system
- S: a potential set of users to be scheduled in a given slot
- f_k: f^{th} traffic flow type of user k
- $U_{k,f}(x)$: utility of user k for parameter x of traffic f
- $r_k, d_k, \jmath_k$: instantaneous rate, delay and jitter for user k
- $\overline{r_k}, \overline{d_k}, \overline{\jmath_k}$: average rate, delay and jitter for user k
- L_k: packet size for user k
- α_f: priority weight for traffic f within any class
- β_k: priority weight of user k based on the traffic class

3.1 Unified Scheduling Rule

The goal is to perform scheduling such that it maximizes the aggregate utility of the network.

$$\max\left\{\sum_{k=1}^{K}\beta_k\sum_{f=1}^{F}\alpha_f U_{f,k}\right\}$$

Given the nature of utility functions, it can be shown that the system has a unique optimum, which in turn can be

achieved by the scheduler following the steepest gradient [1] or the maximum marginal utility in every slot. Thus, the schedule for each time slot (Δt) can be given in terms of the net marginal utility as,

$$S_{max} = \arg\max_S \{\Delta U_s\} = \arg\max_S \left\{ \sum_{k \in S} (\Delta U)_k^+ - \sum_{l \in \bar{S}} (\Delta U)_l^- \right\}$$

Note that the net marginal utility gain for a schedule is captured by not just the utility gains of the schedules users but also by the utility losses of the non-scheduled users.

Without loss of generality, we assume that a set of users can be scheduled in a given slot depending on the number of sub-channels available at the physical layer. Further, each user has a set of queues at the network layer, one each for the different traffic types that could belong to different QoS classes. We assume that as long as a user has a waiting packet in its queues, the network layer first determines which of its different traffic types will contend for the time slot. Then, comes the second phase where the MAC layer determines the specific set of users for transmission on the sub-channels from the contending set of users. Thus, arbitration of resources follows a two-step scheduling process:

Step 1: Intra-user Scheduling

Every user servicing multiple flow types, identifies the specific flow type that will generate the maximum marginal utility if scheduled and allows it to contend for the slot.

$$f_k^* = \arg\max_f \{\alpha_f (\Delta U)_{k,f}\} = \arg\max_f \left\{ \frac{\alpha_f}{P_k} \sum_{p=1}^{P_k} (\Delta U)_{k,f,p} \right\}$$

where P_k corresponds to the total number of parameters relevant to the chosen flow f of user k.

Step 2: Inter-user Scheduling

Having chosen the traffic type for every backlogged user, the BS then identifies the specific set of users to be scheduled that will generate the maximum marginal utility in a similar manner. On simplification, we have (see [14]),

$$S_{max} = \arg\max_S \left\{ \sum_{k \in S} \frac{\beta_k \alpha_k^* r_k}{L_k P_k} \sum_{i=1}^{P_k} U_k(p_i)'(\delta p_i) \right\} \qquad (1)$$

where δp_i corresponds to the change in parameter p_i for a single packet transmission duration at time t; $\alpha_k^* = \alpha_{f_k^*}$ corresponds to that of traffic type f_k^* of user k from intra-user scheduling.

3.2 Individual Scheduling Rules

The unified scheduling rule for heterogeneous traffic can be used to derive individual scheduling rules for specific traffic types/parameters, namely average rate (throughput), packet delay and packet jitter after some analysis (see [14]). All users are assumed to contain flows of the same traffic type with a single parameter of interest ($\alpha_f = 1$).

$$\begin{aligned}
S_{rate} &= \arg\max_S \left\{ \sum_{k \in S} \beta_k U_k'(\bar{r_k}) r_k(t) \right\} \\
S_{delay} &= \arg\max_S \left\{ \sum_{k \in S} \beta_k U_k'(\bar{d_k}) \frac{r_k(t)}{L_k} \bar{d_k}(t) \right\} \\
S_{jitter} &= \arg\max_S \left\{ \sum_{k \in S} \beta_k U_k'(\bar{\mathrm{J}_k}) \frac{r_k(t)}{L_k} \bar{\mathrm{J}_k}(t) \right\}
\end{aligned}$$

Further, substituting back the appropriate $\delta p(t)$ for the different traffic types into Equation 1, we can obtain the final simplified form of the unified scheduling rule as well.

For the scheduler to make decisions, information on the following parameters will be required for each user k, namely $r_k(t)$, $\bar{r_k}(t)$, $\bar{d_k}(t)$, and $\bar{\mathrm{J}_k}(t)$. While, the instantaneous rate for each user is fed back in the form of channel quality indicator (CQI), the average values of the parameters can be kept track at the BS itself as follows.

$$\begin{aligned}
\bar{r_k}(t) &= (1-\gamma)\bar{r_k}(t-1) + \gamma r_k(t) \\
\bar{d_k}(t) &= (1+\gamma)\bar{d_k}(t-1) - \gamma d_k(t) \\
\bar{\mathrm{D}_k}(t) &= (1+\gamma)\bar{\mathrm{J}_k}(t-1) - \gamma \mathrm{J}_k(t)
\end{aligned}$$

While instantaneous rate and jitter can be measured at the BS, instantaneous delay can be estimated from queuing delay, assuming the wireless hop forms the bottleneck in flow's path.

3.3 Fairness

While the nature of the utility functions helps provide fairness across flows of the same type, β_k provides service differentiation across users. However, neither of them account for fairness across flows of different traffic types, thereby causing the scheduler to bias towards flows of certain types. This is because, different parameters provide different levels of marginal utility for the same amount of network resources allocated. α_f can help address this issue. α_f as a priority weight, should not just account for service differentiation across traffic types but also for fairness between them. This is achieved by considering α_f to be composed of two components, $\alpha_{f,q}$ and $\alpha_{f,x}$. $\alpha_{f,q}$ would capture the service differentiation (QoS) required by the flow type, and $\alpha_{f,x}$ would capture the normalization of the flow's parameter (x) to the basic network allocation resource (say, bandwdith), with the parameter that requires a higher network resource to achieve a certain desired utility being provided a higher $\alpha_{f,x}$ to eliminate scheduler bias. Now, by maximizing the aggregate weighted utility of all the flows, service differentiation is incorporated, while a system-wide notion of fairness is also provided both within and across traffic types. For eg., if a logarithmic utility function is considered for the different traffic parameters, then weighted proportional fairness is achieved in the sytem.

4. DOWNLINK SCHEDULING WITH MULTIPLE CHANNELS

While the unified and individual scheduling rules were generically designed and hence would directly apply to multiple channels as well, the complexity of the solution and the communication overhead involved plays a vital role in such multiple channel systems (MCS). In the case of single channel systems (SCS), the feedback overhead for making scheduling decisions is $O(K)$; also the number of possibilities for the schedule is equal to just the number of users, $\binom{K}{1}$. This keeps the running time complexity of the scheduling algorithms within reasonable limits to allow for repeated execution at every frame transmission. However, for MCS (with A channels), this is not the case.

While both single-user (SU) and multi-user (MU) strategies exploit MUD from user mobility, the SU strategy requires the feedback of only $\sum_{\ell=1}^{A} r_\ell(k)$ from each user k,

thereby resulting in an overhead of only $O(K)$ and a running-time complexity of $\binom{K}{1}$. But the SU strategy lacks in performance, since in addition to MUD, MU strategies also provide frequency (channel) diversity. However, to obtain the maximum gain possible from diversity, every MS k would have to send back, $CQI_\ell = r_\ell(k)$, $\forall \ell \in \{1,..,A\}$, which is a significant overhead of the order of $O(KA)$. Also, the complexity of the scheduling algorithm at the BS would now require a search over a much larger space of $\binom{KA}{A}$ potential schedules, diminishing its ability to be executed at the granularity of frame transmissions. Hence, the focus is now to exploit the derived scheduling rules through an efficient, unified scheduling algorithm that aims to retain the performance reasonably close to that of the ideal scheme although at significantly reduced complexity and communication overhead.

4.1 Parameter-based Optimization

Recall the two-stage optimization process for scheduling. Extending this to multiple channel systems, we have,

Step 1: Intra-user Scheduling

$$f_k^* = \arg\max_f \left\{ \frac{\alpha_{f,q} r_{k,SU}}{L_k P_k} \sum_{i=1}^{P_k} \alpha_{f,p_i} U_{k,f}(p_i)'(\delta p_i) \right\}, \quad \forall k$$

$r_{k,SU}$ represents the rate corresponding to all the channels being used (SU strategy). In deciding the single contending flow for each user (at the network layer), since the channel conditions for the different flows are the same for a given user, one does not need to consider multiple strategies in this decision. However, once the contending flow from each user has been decided, it then becomes the responsibility of the MAC layer scheduler to perform *inter-user scheduling*, taking into account multiple communication strategies to optimize network performance.

Step 2: Inter-user Scheduling

$$S_{UO} = \arg\max_S \left\{ \sum_{(k,v_k)\in S} \frac{\beta_k r_{k,v_k} I_{k,v_k}}{L_k P_k} \sum_{i=1}^{P_k} \alpha^*_{k,p_i} U_{k,v_k}(p_i)'(\delta p_i) \right\}$$

$$\sum_{k=1}^{K} I_{k,v_k} v_k(m) \le 1, \quad \forall m \in \{1,2,...,A\}$$

$$where \;\; S \leftarrow \{(k,v_k)\} \;\; k \in \{1,..,K\} \;\; v_k \in V = \{(0..0),..,(1..1)\}$$

and $\alpha^*_{k,p_i} = \alpha_{f_k^*,q} \cdot \alpha_{f_k^*,p_i}$. V represents the set of all possible assignments of A channels to a user; v_k represents a vector of A boolean variables denoting the specific assignment of the A channels to user k with r_{k,v_k} representing the aggregate rate for the specific channel assignment. Note that, $\sum_{k=1}^{K} I_{k,v_k} v_k(m) \le 1$, where I_{k,v_k} is a boolean indicator, ensures the orthogonality constraint, whereby no channel is assigned to more than one user. Thus, a schedule S consists of a vector of (user, channel assignment) pairs that specifies the set of users to be scheduled along with their respective strategy for transmission (number of channels and specific channel indices for each scheduled user). Solving the above formulation would provide us the optimal schedule that maximizes the aggregate network utility taking into account all the parameters in tandem. However, it requires that the scheduler first get feedback from all K users on their A channels each (overhead of $O(KA)$). Then the scheduler needs to search a space of $\binom{KA}{A}$ potential schedules to identify the optimal one.

We now propose an alternative approach to obtaining a good schedule at a reduced complexity and overhead. We refer to this approach as the *parameter-based optimization* (PO) as opposed to the aggregate utility based optimization (UO) above. In PO, we focus on optimizing the specific parameter that provides the maximum marginal utility in every slot. This is in contrast to UO, where all parameters are optimized in tandem. While PO does consider marginal utility of user as the metric for deciding the specific parameter to optimize in each slot, the focus is still on isolating and optimizing an individual parameter in each slot as opposed to aggregate utility of all parameters in UO. The formulation for PO can be given as,

$$S_{PO} = \arg\max_{(S_{p_i})_{max}} \left\{ \Delta U_{(S_{p_i})_{\max}} \right\}, \quad i \in \{1,2,...,P\}$$

$$\Delta U_{(S_{p_i})_{\max}} = \max_{S_{p_i}} \left\{ \sum_{(k,v_k)\in S_{p_i}} \frac{\beta_k r_{k,v_k} I_{k,v_k}}{L_k P_k} \sum_{i=1}^{P_k} \alpha^*_{k,p_i} U_{k,v_k}(p_i)'(\delta p_i) \right\}$$

$$\sum_{k=1}^{K} I_{k,v_k} v_k(m) \le 1, \quad \forall m \in \{1,2,...,A\}$$

$$S_{p_i} \leftarrow \{(k_{p_i}, v_{k,p_i})\} \;\; k_{p_i} \in \{p_i \; relevant \; users\} \;\; v_{k,p_i} \in V$$

Note that, in the PO formulation, we have decomposed the problem into sub-problems that are specific to individual parameters (p_i). The entire search space (set) of users is now decomposed into subsets of users based on their relevance to individual parameters. The individual parameters are then optimized over their respective subset of users. $S(p_i)$ now represents a schedule that is a vector of (user, channel assignment) pairs, where the users considered are only those that require p_i as a parameter of optimization. P represents the total number of parameters in the system. A single user can also be a part of multiple sub-problems by virtue of requiring optimization of multiple parameters. The formulation, thus identifies the maximum marginal utility for each of the parameters along with the corresponding schedule ($S(p_i)_{max}$) in the first step. In the next step, it selects the specific parameter along with its corresponding schedule that provides the maximum marginal utility among all the parameters considered. This in turn forms the final schedule for transmission in each slot. Due to the presence of system and channel dynamics, PO's optimization of an individual parameter in each slot, will make it sub-optimal to UO. But the key focus here is to identify *if the sub-optimality of PO has the potential to reduce the complexity and overhead incurred in UO in MCS.*

In PO, since the problem is decomposed into optimizing individual parameters, it is possible to reduce the complexity and overhead *if we can identify specific communication strategies that optimize specific traffic parameters.* In this case, the assignment vector v_{k,p_i} for each of the parameter optimizations does not have to span the entire set V but will have to span only subsets of it, depending on the specific strategies. This would reduce not just the complexity but also the feedback overhead. In the rest of this work, we focus on applications involving data and voice flows over IP, and hence consider the parameters of average rate (data), and delay and jitter (voice) for optimization.

4.2 Strategies Optimizing Individual Parameters

Consider the two extremes of strategies possible, SU with lowest complexity-overhead but also lowest diversity gain,

and MU_A (A users with one channel each) with largest diversity gain but also largest complexity-overhead.

4.2.1 Delay

Inherent Potential: Initially assume that there is no fading across channels and hence no additional gain of frequency diversity for MU_A over SU. Now, it can be shown that the SU strategy has an inherent potential for optimizing packet delay as opposed to MU_A. Consider K packets of the same size L meant one each for K users (Figure 1(a)). Let the BS and MS have A channels each and let the transmission rate on individual channels be the same r_k for user k. For simplicity, assume $K = A$ and that users have the same resulting rate ($r_k = r, \forall k$). In SU, the packets for the K users are sent sequentially but using all channels, resulting in the following average delay per packet: $\frac{\sum_{k=1}^{K} \frac{kL}{Ar}}{K} = \frac{L(K+1)}{2Ar}$. On the other hand, in MU_A, the packets for the K users are sent in parallel using (in this case) only one channel each, resulting in an average delay of: $\frac{\sum_{k=1}^{K} \frac{L}{r}}{K} = \frac{L}{r}$. Since $K = A$, we have $D_{SU} = \frac{(A+1)}{2A}\frac{L}{r}$ and $D_{MU} = \frac{L}{r}$, thus resulting in an inherent potential for the sequential transmission in SU to favor average delay over the parallel transmissions in MU_A ($D_{SU} \leq D_{MU_A}$).

Alleviation of Buffer Effects: More importantly, the scheduling decision for the head-of-line (*hol*) packet impacts the delay of not just itself but all the other packets behind it in the queue. This is a key difference from rate and jitter. Hence, the cumulative effect of buffered packets plays a key role in the optimization of the delay parameter. Consider a system of K users in steady state, where the delays of all users are kept low with almost empty buffers. Now assume, that the instantaneous rate of user k, r_k goes down below the incoming traffic rate R for a period of time due to a deep fade and then comes up. This causes user k's buffer to accumulate rapidly during that period. Let the buffer size of user k after the fade be B packets. User k now needs to be scheduled to help optimize its increased delay. Consider a delay specific scheduling rule. In MU_A, in addition to user k, $A-1$ other users will be scheduled in every slot, though they might not need optimization as critical as user k. The $A-1$ users will help contribute to increased system diversity although at the cost of k's optimization. On the other hand, in SU, all the available channels are allocated to k for optimization. While this does not increase system diversity, it will optimize k's delay significantly better than MU_A and hence outweigh the benefits of increased diversity in MU_A.

Once again consider no fading across channels initially. Let B_{MU_A} indicate the number of packets for which the user k needs to be scheduled in succession to empty the buffer in the MU_A strategy. Note that, the corresponding number of packets in the SU strategy would be much lesser due to the usage of all channels for user k. Considering the reduction in delay in the two strategies over a period of B_{MU_A} packet transmissions, it can be shown (see [14]) that for reasonably moderate-large number of channels,

$$D_{MU_A} \approx W_{hol} - \left(\frac{L_k B_{MU_A}}{2}\right)\left(\frac{r_k - R}{r_k R}\right)$$

$$D_{SU} \approx D_{MU_A}\left(1 - \frac{R}{r_k}\right) + \left\{\frac{L_k}{Ar_k}\right\}\left(\frac{R}{r_k}\right)$$

where W_{hol} represents the waiting delay of the *hol* packet. Comparing the two average delays, it can be seen that the larger the number of channels available, the SU strategy will be able to empty out the buffer much faster, after which the incoming packets will only experience transmission delay. Further, whenever the instantaneous rate falls close to the traffic rate, SU has the ability to reduce delay accumulation much better than MU_A.

Exploitation of Diversity: Now, if we were to consider fading across the channels, the MU_A strategy will obtain an advantage over the SU strategy in the form of channel diversity. However, the channel diversity gains will be limited due to the incorporation of fairness through utility optimization (as opposed to Max-Rate optimization). Even the improved instantaneous rate from limited channel diversity only helps the transmission delay but does not help the accumulated queuing delay, which in turn forms the significant portion of the packet delay in practical (moderate-high) load conditions. Further, while the additional channel diversity helps improves the net instantaneous rate of the system in MU_A, it does not directly help optimize the user in need; consequently only making the scheduling decisions in MU_A deviate more from the optimum. Also, if the channel rate of the user goes into a deep fade for a long time, such that $r_k < R$, then the MU_A strategy will face an exponentially increasing buffer and hence packet drops. On the other hand, the SU strategy might still be able to optimize delay as long as $Ar_k > R$. Thus, though the MU_A strategy has the additional gain from limited channel diversity, this is not sufficient enough for it to overcome the inherent potential of SU strategy along with its ability to alleviate buffer effects, in trying to optimize the *delay* parameter.

Thus, *SU helps optimize packet delay better than MU_A.*

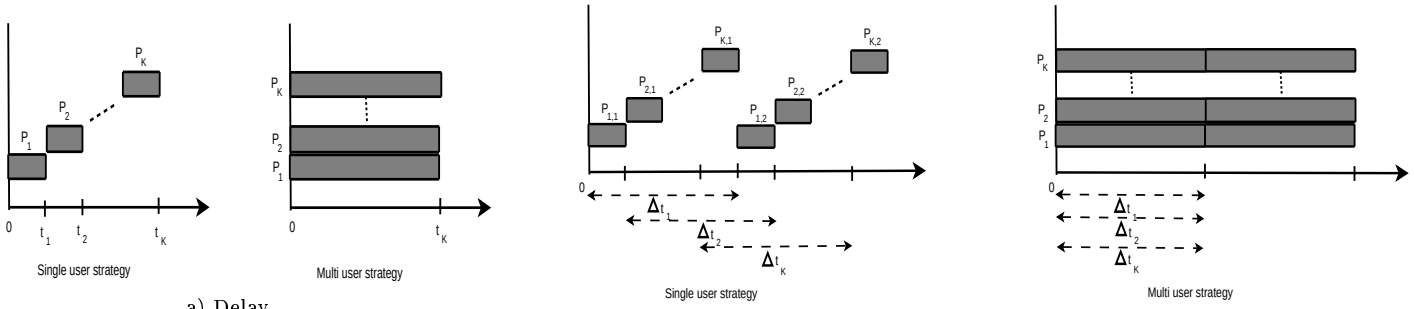
4.2.2 Jitter and Rate

Let us first consider jitter optimization.

Inherent Potential: Consider the same scenario as in packet delay discussion. Since packet jitter corresponds to the (inter-packet) delay with respect to the previous packet, consider the sequential and parallel transmission of K packets in SU and MU_A strategies respectively to occur in periodic rounds (Figure 1(b)). In SU, the sequential transmission of the K packets, results in an inter-packet separation of K transmission delays, each at a rate of Ar. Thus, the average packet jitter in SU can be given as $J_{SU} = \frac{\sum_{k=1}^{K} K\frac{L}{Ar}}{K} = \frac{KL}{Ar}$. On the other hand, in MU_A, the packets for the K users are sent in parallel, thereby resulting in consecutive transmissions for each user, resulting in an inter-packet separation of one transmission delay at a rate of r. Thus, the average packet jitter in MU_A can be given as, $J_{MU_A} = \frac{L}{r}$. With $A = K$, we have $J_{MU_A} = J_{SU}$. Hence, there is no inherent potential for one strategy over the other in optimizing packet jitter.

Alleviation of Buffer Effects: Packet jitter is an instantaneous parameter unlike delay which is a cumulative parameter. The scheduling decision for the *hol* packet impacts only itself and not the subsequent packets in the buffer. Thus, there is no necessity for alleviation of buffer effects.

Exploitation of Diversity: Since there is no inherent potential for a specific strategy to optimize jitter as well as there is no impact of decisions on buffered packets, it is the transmission delay that forms the main factor in optimizing jitter. Thus, both SU and MU have the same impact if the instantaneous rates are the same. However, the



a) Delay

b) Jitter

Figure 1: Illustration of Strategies

MU strategies have a higher instantaneous channel rate and hence lower transmission delay than the SU strategy, owing to exploitation of channel diversity. Further, it is possible that in certain instances more than one channel needs to be assigned to a single user for the maximum exploitation of channel diversity depending on the instantaneous channel rates, and hence all possible MU strategies need to be considered. However, it must be noted that our purpose is not to maximize the aggragate rate of all users but instead their utility, whereby fairness is also taken into account. Hence, some diversity and hence rate (transmission delay) is sacrificed to ensure fairness. The concave nature of the utility functions for jitter/rate provides diminishing returns in utility when multiple channels are assigned to a single user ($U(R_1+R_2) < U(R_1)+U(R_2)$), thereby automatically providing incentives for assigning lesser channels but to more users, which tends to one when there are large number of users. This also helps exploit channel diversity much better than SU.

The same arguments are applicable to rate as well. Since the instantaneous transmission rate determines the amount of data delivered to the user(s), the strategy that has a better instantaneous channel rate will help optimize rate better. Both SU and MU$_A$ strategies exploit MUD, but it is only the MU$_A$ strategy that also exploits channel diversity, while also considering fairness and hence optimizing utility to the maximum extent.

Thus, considering the concave nature of the utility functions as well as the potential for maximum exploitation of channel diversity, *it is the MU$_A$ strategy that helps optimize packet jitter and rate better than SU.*

Note that, we have considered only the two extreme strategies of SU and MU$_A$. Any 'in-between' strategy, transmitting to multiple users on multiple channels, will provide a performance in-between those of SU and MU$_A$ for the respective parameters, but will not provide the best optimization for any of the individual parameters. Since our goal is to isolate the different parameters and identify the corresponding best optimizing strategy, it is sufficient to consider the two extreme strategies in identifying the best one.

4.3 Scheduling Algorithm

Note that, in the PO formulation, we had decomposed the search space of all users into subsets that are specific to the individual parameters. However, the channel assignment space was still considered to be the entire V space for all the parameters. Now, having identified the communication strategies that optimize specific parameters, we can reduce the complexity and overhead by reducing the channel assignment search space for each of the parameters ($p_x \leftarrow \{p_r, p_d, p_j\}$) and hence their schedule search space $S_{p_x} \leftarrow \{(k_{p_x}, v_{k,p_x})$, in a system of data and voice users as follows.

$$k_{p_d} \in \{Voice\ users\},\ \ v_{k,p_d} \in V_{p_d}\ s.t.\ \forall k,\ \sum_{m=1}^{A} v_{k,p_d}(m) = A$$

$$k_{p_j} \in \{Voice\ users\},\ \ v_{k,p_j} \in V_{p_j}\ s.t.\ \forall k,\ \sum_{m=1}^{A} v_{k,p_j}(m) = 1$$

$$k_{p_r} \in \{Data\ users\},\ \ v_{k,p_r} \in V_{p_r}\ s.t.\ \forall k,\ \sum_{m=1}^{A} v_{k,p_r}(m) = 1$$

Thus, we find that the channel assignment search space for delay is restricted to vectors where the voice users are assigned all the A channels as in SU strategy. For jitter (and rate), the assignment space is restricted to vectors where the voice (and data) users are assigned only one of the A channels as in the MU$_A$ strategy. This reduces the complexity of the scheduling algorithm. If K_d and K_v represent the number of data and voice users respectively, with $K_d + K_v = K$, then for delay, the scheduler has to consider only K_v potential schedules since no channel assignment vector is needed for the users. For jitter and rate, the scheduler has to consider $\left[\binom{K_v}{A} \cdot A!\right]$ and $\left[\binom{K_d}{A} \cdot A!\right]$ potential schedules respectively. Thus, the total schedule search space reduces to $\left[K_v + \left\{\binom{K_v}{A} + \binom{K_d}{A}\right\} \cdot A!\right]$, which is much lesser than $\binom{KA}{A}$ incurred in the optimal scheduler.

However, each user still needs to feed back rate information for all A channels, resulting in a feedback overhead of $O(KA)$. Also, the jitter and rate parameters still incur an appreciable amount of complexity. We now alleviate both the complexity and overhead with assistance from the mobile stations. Our optimizing strategy for jitter and rate, namely MU$_A$ plays a crucial role in this regard. Since MU$_A$ requires only one channel to be eventually assigned to a user atmost, this is further exploited to allow each mobile user, k to feed back the rate information of only the strongest of its A channels along with the channel index as opposed to the rate information of all A channels. Thus, the scheduler would then consider only the strongest of the A channels for each user for the schedule search space, thereby significantly reducing its complexity, while also reducing the feedback overhead from $O(KA)$ to just $O(K)$. This would however reduce the feasible (user set with orthogonal choice of antennas) number of schedules and consequently tradeoff a slight degradation in performance. But the degradation is kept

Algorithm 1 Unified Scheduling Algorithm at BS: UNI

for all $k \in K$ **do**
  Obtain feedback $\{\langle C_{SU}(k)\rangle, \langle C_{MU}(k), X_k\rangle_1\}$
  if $k \in K_v$ **then**
    $f_k^* = \arg\max_f \left\{\frac{\alpha_{f,p_d}(\Delta U)_{k,f,p_d}+\alpha_{f,p_j}(\Delta U)_{k,f,p_j}}{2}\right\}$ using $C_{SU}(k)$
  end if
  if $k \in K_d$ **then**
    $f_k^* = \arg\max_f \{\alpha_{f,p_r}(\Delta U)_{k,f,p_r}\}$ using $C_{SU}(k)$
  end if
end for
$S_{p_d} \rightarrow$ {Set of K_v voice users}
for all $s \in S_{p_d}$ **do**
  $\Delta U_{p_d}(s) = \sum_{k\in s} \beta_k \left(\frac{\alpha^*_{k,p_d}(\Delta U)_{k,p_d}+\alpha^*_{k,p_j}(\Delta U)_{k,p_j}}{2}\right)$
end for
$(S_{p_d})_{\max} = \arg\max_{s\in S_{p_d}} \{\Delta U_{p_d}(s)\}$
$(\Delta U_{p_d})_{\max} = \max_{s\in S_{p_d}} \{\Delta U_{p_d}(s)\}$
$S_{p_j} \rightarrow$ {All feasible combinations of A distinct voice users using v^*_{k,p_j}, $\forall k \in K_v$}
for all $s \in S_{p_j}$ **do**
  $\Delta U_{p_j}(s) = \sum_{k\in s} \beta_k \left(\frac{\alpha^*_{k,p_d}(\Delta U)_{k,p_d}+\alpha^*_{k,p_j}(\Delta U)_{k,p_j}}{2}\right)$
end for
$(S_{p_j})_{\max} = \arg\max_{s\in S_{p_j}} \{\Delta U_{p_j}(s)\}$
$(\Delta U_{p_j})_{\max} = \max_{s\in S_{p_j}} \{\Delta U_{p_j}(s)\}$
$S_{p_r} \leftarrow$ {All feasible combinations of A distinct data users using v^*_{k,p_r}, $\forall k \in K_d$}
for all $s \in S_{p_r}$ **do**
  $\Delta U_{p_r}(s) = \sum_{k\in s} \beta_k \alpha^*_{k,p_r}(\Delta U)_{k,p_r}$
end for
$(S_{p_r})_{\max} = \arg\max_{s\in S_{p_r}} \{\Delta U_{p_r}(s)\}$
$(\Delta U_{p_r})_{\max} = \max_{s\in S_{p_r}} \{\Delta U_{p_r}(s)\}$
$S_{\max} = \arg\max_{s\in(S_{p_i})_{\max};\, i\in\{r,d,j\}} \{(\Delta U_{p_i})_{max}\}$

Algorithm 2 Optimal Scheduling Algorithm: CENT

for all $k \in K$ **do**
  Obtain feedback $\langle C_{MU}(k), X_k\rangle_\ell$, $\forall \ell \in \{1, ..., A\}$
  if $k \in K_v$ **then**
    $f_k^* = \arg\max_f \left\{\frac{\alpha_{f,p_d}(\Delta U)_{k,f,p_d}+\alpha_{f,p_j}(\Delta U)_{k,f,p_j}}{2}\right\}$ using $C_{SU}(k)$
  end if
  if $k \in K_d$ **then**
    $f_k^* = \arg\max_f \{\alpha_{f,p_r}(\Delta U)_{k,f,p_r}\}$ using $C_{SU}(k)$
  end if
end for
$S \leftarrow$ {Set of all feasible combinations of A distinct users (data+voice) out of K}
for all $s \in S$ **do**
  $s \rightarrow s_d \cup s_v$
  $\Delta U_d(s) = \sum_{k\in s_d} \beta_k \alpha^*_{k,p_r}(\Delta U)_{k,p_r}$
  $\Delta U_v(s) = \sum_{k\in s_v} \beta_k \left(\frac{\alpha^*_{k,p_d}(\Delta U)_{k,p_d}+\alpha^*_{k,p_j}(\Delta U)_{k,p_j}}{2}\right)$
  $\Delta U(s) = \Delta U_d(s) + \Delta U_v(s)$
end for
$S_{\max} = \arg\max_{s\in S} \{\Delta U(s)\}$

minimal because, whenever a potential schedule of A users is identified to be infeasible, a good feasible schedule with a slightly lesser number of users than A is constructed out of the original schedule, although in a greedy manner to avoid increase in complexity. Further, this performance degradation becomes very small when the number of users is large compared to the number of sub-channels, as is the case in most practical systems. Thus, the channel assignment vector for jitter and rate for a given user k is now given by a single vector $v^*_{k,p}$,

$$v^*_{k,p} = \arg\max_{v_{k,p}} \{r_{k,v_{k,p}}\}, \quad v_{k,p} \in V_p, \ and \ \sum_{m=1}^{A} v_{k,p}(m) = 1$$

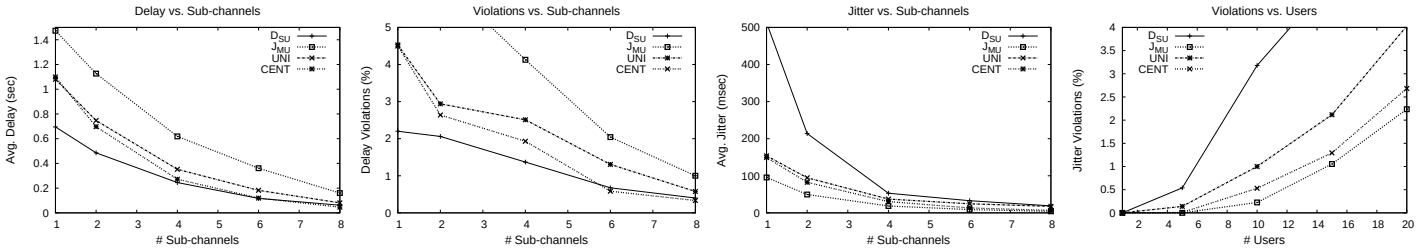
This in turn reduces the potential schedule search space for jitter and rate to $\binom{K_v}{A}$ and $\binom{K_d}{A}$ respectively, thereby resulting in a final schedule search space of $\left[K_v + \binom{K_v}{A} + \binom{K_d}{A}\right]$, which is significantly lesser than the $\binom{K_A}{A}$ incurred in the optimal scheduler. The final scheduling algorithm can be presented as,

- **Step 1:** Every MS k obtains rate for SU and MU_A strategy along with its best channel index for MU_A strategy. Three elements are fed back to the BS, namely $\left\{\sum_{\ell=1}^{A} r_\ell(k), \max_\ell\{r_\ell(k)\}, \arg\max_\ell\{r_\ell(k)\}\right\}$.
- **Step 2:** BS uses the feedback to generate its potential schedule set based on the parameters considered and then determines the best schedule.
  - BS calculates the contending flow for each user using the rate feedback of $\sum_{\ell=1}^{A} r_\ell(k)$.
  - BS calculates parameter specific schedules that provide maximum marginal utility.
  - BS calculates the final schedule as the one that provides maximum marginal utility among the parameter specific schedules already determined.

Algorithms 1 and 2 present our low complexity-overhead scheduling algorithm (UNI) and the optimal algorithm (CENT) respectively. The optimal scheduler requires that the MS feedback rate information for all possible MU and SU strategies, namely $r_\ell(k)$, $\forall \ell \in [1, A]$. The algorithms not only accommodate users belonging to multiple QoS classes, but also accommodate heterogeneous traffic within a QoS class. The marginal utilities for the different parameters can be obtained from the expressions in Section 3.

4.4 Complexity, Overhead and Performance

The schedule search space of the proposed UNI algorithm is $\left[K_v + \binom{K_v}{A} + \binom{K_d}{A}\right]$ while that of the optimal scheme is $\binom{K_A}{A}$. Since $\binom{K_A}{A} \geq \binom{K}{A} \cdot A^A$, UNI thus brings about a significant reduction of the order of A^A in the running-time complexity of the scheduler, making it functional at the granularity of frame transmissions. Further, to obtain the optimal schedule, CENT also requires significant feedback from the users. Let B_r and B_a represent the number of bits allocated for rate and antenna index feedback respectively. CENT would require a feedback of $r_\ell(k)$, $\forall \ell \in [1, A]$ from every user, resulting in a net feedback of KAB_r bits. UNI, on the other hand, incurs a feedback of only three elements from every user, $\left\{\sum_{\ell=1}^{A} r_\ell(k), \max_\ell\{r_\ell(k)\}, \arg\max_\ell\{r_\ell(k)\}\right\}$, resulting in an overhead of $K(2B_r + B_a)$ bits. Given that B_a



a) Average packet delay b) Delay violations c) Average packet jitter d) Jitter violations

Figure 2: Multiple Sub-channel Scheduler: Delay and Jitter optimization

is typically much smaller than B_r, UNI also provides almost an $O(A)$ reduction in feedback overhead when compared to the optimal scheme.

Given that UNI provides significantly lower complexity and performance, it is necessary to understand its sub-optimal performance. Since UNI optimizes an individual parameter at every slot as opposed to all parameters in tandem and owing to system and channel dynamics, it will be sub-optimal to CENT. However, it must be noted that the gap in performance will be reasonably small. This is because at each time slot, UNI optimizes a single parameter to a much larger extent that CENT. However, in doing so, the other parameters tend to deviate from their optimal value in CENT. But in the subsequent time slots, UNI will come back to optimize the other parameters that have deviated from the optimum. Its mechanism of using the best strategy for individual parameters, helps it optimize an individual parameter to a significant extent in each slot, thereby giving it the ability to reduce the magnitude of deviations from the optimum. This will also be evident from the evaluation results in Section 5.

5. PERFORMANCE EVALUATION

An event-driven packet level simulator written in C++, named *queuing network simulator* [9] is considered for implementation and evaluation of the proposed solutions.

A single cell downlink environment is considered, where the users are distributed uniformly in a cell of radius 600m. The received SNR and hence the instantaneous rate at a user is determined by its distance with respect to the BS as well as the user's Rayleigh and shadowing loss channel model. Further, each user's Rayleigh channel has a Doppler fading equivalent to a velocity ranging from 3-10 Km/hour. A time slot is consider to be of 5 ms duration; carrier frequency is assumed to be 2 GHz. The peak rate for a sub-channel is about 250 Kbps. The flows for the different users are assumed to originate in the Internet backbone for simplicity. UDP is considered to be the transport protocol generating traffic. We consider data and voice as the two different traffic types in the network. The backbone links emulate the delay (mean = 50 ms, deviation = 10 ms) and losses ($< 1\%$) in the backbone Internet. The number of users vary from 1 to 20, while the number of sub-channels at the BS and MS vary from 1 to 8. We focus on only throughput (rate) for the data flows, while we focus on packet delay and jitter, for the voice flows. We consider the following exponential utility functions,

$$U(r) = \frac{1-\exp(\frac{-r}{\rho_r})}{1-\exp(\frac{-r_{max}}{\rho_r})},\ U(p) = \frac{1-\exp(\frac{-(p_{max}-p)}{\rho_p})}{1-\exp(\frac{-p_{max}}{\rho_p})},\ p = d, j$$

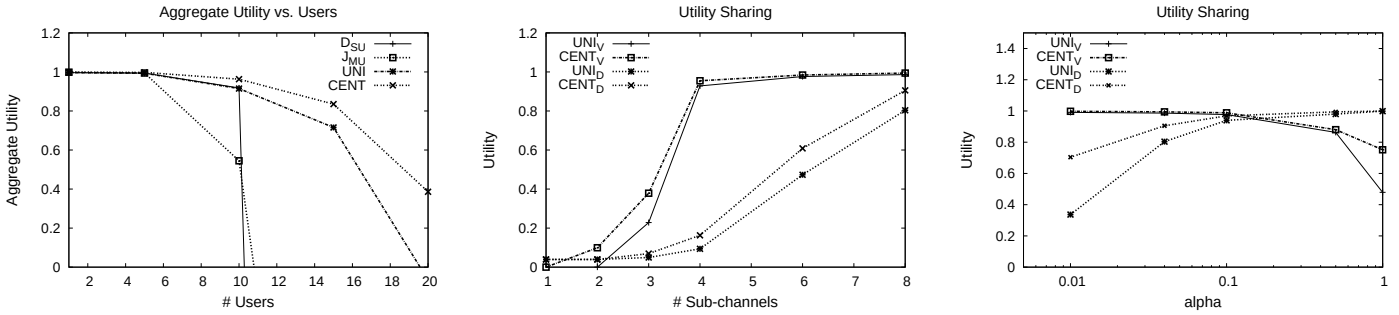
We also consider maximum thresholds for delay and jitter (d_{max} and j_{max}), that are fixed at 500 ms and 50 ms respectively. ρ_p, $p \in \{r, d, j\}$ determine the degree of concavity of the utility functions and are set to $\frac{p_{max}}{5}$. The data flows are sent at 125 Kbps and the voice flows at 64 Kbps. Note that, no call-admission control mechanism is currently considered, although the solutions are perfectly inter-operable with any call admission control mechanism. We consider loads ranging from low, moderate to even high to stress test the mechanisms. The metrics used for measurement are the average per-user utility, average parameter (rate/delay/jitter) value and violations (equivalent packet losses). Results are measured either as a function of increasing users or increasing sub-channels (bandwidth). When not varied, the fixed values for users and sub-channels are maintained at 10 and 4 respectively. First, we evaluate the optimal version of the individual (parameter) and unified scheduling rules, followed by the low complexity-overhead version of the unified rule.

5.1 Heterogeneous Traffic Parameters

We first consider a system of voice users alone. The importance of having parameter-specific rules can be clearly observed from the results in Figure 2, where the delay (D-rule) and jitter (J-rule) rules best optimize their respective parameters. The performance trends are consistent across not just average parameter values but also for violations. This indicates that while specific rules optimize specific parameters, they cannot optimize multiple parameters of varying characteristics in tandem, therbey necessitating the need for a unified rule. It can be clearly seen that the optimal unified rule (CENT) presents a fair balance between the two parameters of delay and jitter, while staying close to the best optimizing rule for the specific parameter, both in terms of average value and violations. The superiority of the unified rule over the individual rules is captured by the significant gains in aggregate voice (delay+jitter) utility in Figure 3(a).

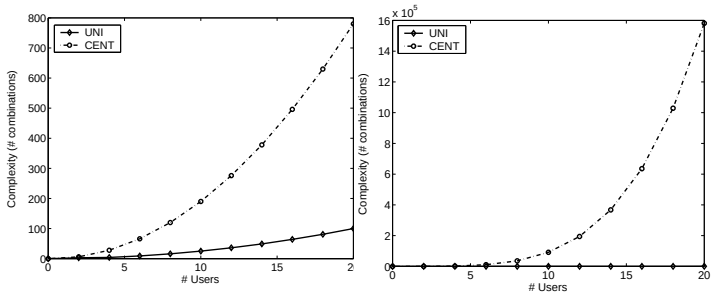
5.2 Heterogeneous Traffic Types

We now consider both data and voice flows. First we assume that both of them belong to the same traffic class. Since the voice flows have a higher priority over the data flows from the same traffic class due to their time-sensitive nature, we want the data flows to utilize network resources only after all the requirements of the voice flows have been satisfied. This can be clearly seen in Figure 3(b), where the voice flows first attain their maximum utility value with increasing number of sub-channels and only then do the data flows use up the additional network resources to increase their utility.



a) Delay + Jitter b) Voice > Data c) Voice + Data

Figure 3: Multiple Sub-channel Scheduler: Resource Sharing



a) 2 sub-channels b) 4 sub-channels

Figure 4: Complexity

We also consider the case where the priority of the traffic class of data flows is gradually increased with respect to voice flows. The combined weight of the class and flow type of data flows is integrated into the α in Figure 3(c). For lower values of α, priority goes to the voice flows which attain a per-user utility of close to one. But when α increases, the data flows start obtaining higher priority, thereby using more network resources to increase their utility at the cost of the voice flows. This in turn, clearly indicates the flexibility of our unified scheduler to arbitrate resources across multiple traffic parameters, types and QoS classes.

5.3 Low Complexity-Overhead Unified Scheduler

We now turn our attention to our low complexity unified scheduling algorithm (UNI) in Figures 2 and 3.While the ideal scheduler (CENT) performs the best in optimizing multiple parameters, we find that our proposed UNI algorithm performs reasonably close to that of the CENT scheme. It suffers a performance degradation of about 20% with respect to the CENT scheme under moderate to high load conditions and the degradation is much lower in low load conditions. Only when the load is significantly higher than available capacity, the worst-case degradation increases to about 40%, although the performance is still significantly better than the individual rules. However, in the presence of call-admission control (CAC) algorithms, this increase in degradation will be eliminated.

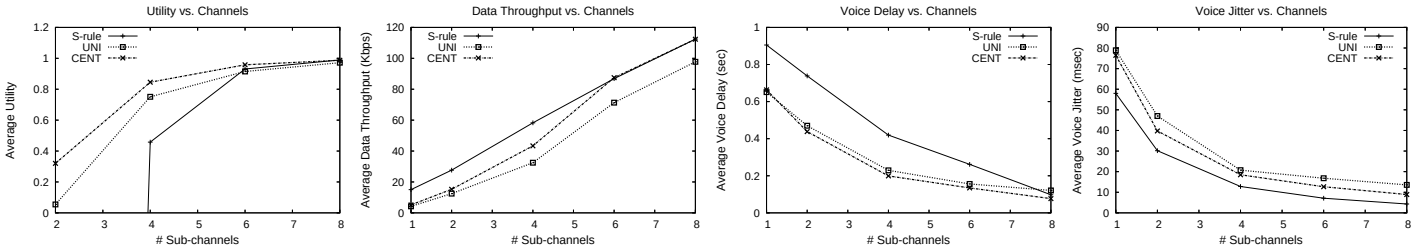
While these results indicate that our UNI scheme performs close to the ideal CENT scheme under typical load conditions (moderate to high), we now show that these large gains are obtained at a significantly lower complexity. The number of potential schedules that need to be evaluated by the CENT and UNI shemes before making a scheduling decision at every time slot, is presented in Figures 4(a) and (b) as a function of increasing number of users for 2 and 4 sub-channels respectively. It can be clearly seen that the complexity of CENT over UNI increases exponentially with increasing number of sub-channels and users (more with sub-channels due to the complexity factor of $O(A^A)$). Note that, in addition to the reduction in complexity, we also have a reduction factor of almost $O(A)$ in feedback overhead.

5.4 Comparison against OFDM Allocations

We now compare our proposed scheduling algorithms with the conventional dynamic sub-channel allocation algorithm (for a given power allocation) in OFDM literature. While there exist several sub-channel allocation mechanisms [13, 15, 16] that aim to optimize throughput performance, we consider the centralized (optimal) version of a utility-based solution proposed in [13]. We refer to this scheduling algorithm as *S-rule*. Figure 5 presents the results of the comparison of our proposed centralized and low complexity-overhead algorithms with that of the S-rule. The scenario consists of a mixture of five data and ten voice flows. Since the S-rule identifies a schedule, such that aggregate throughput is optimized, it is incapable of handling heterogeneous traffic with varying characteristics in tandem. This can be seen from the aggregate utility result in Figure 5(a), where the CENT and UNI schemes provide a much higher aggregate utility than the S-rule. Since rate and jitter have similar characteristics, S-rule helps reduce the transmission delay and hence jitter indirectly, as seen in Figures 5(b) and 5(d). But since delay has significantly different characteristics, S-rule fails to accommodate it as seen in Figure 5(c). On the other hand, UNI and CENT provide a good balance of the available sub-channel resources between the different parameters by sacrificing performance slightly on rate and jitter when their values are well below the threshold so as to gain in performance on delay and consequently improve the net utility, while avoiding threshold violations. Thus, while dynamic sub-channel allocation schemes work well for throughput/rate optimization, when it comes to heterogeneous traffic we need a unified scheme such as CENT or its efficient, low complexity-overhead version, UNI.

6. RELATED WORK

There have been several works on channel-dependent packet scheduling that exploit MUD to improve throughput per-



a) Aggregate Utility b) Data Throughput c) Packet Delay d) Packet Jitter

Figure 5: Evaluation against OFDM Allocation

formance [11, 8, 3, 2]. Some works have extended this to incorporate QoS requirements of users, where delay requirements of voice flows are taken into account in addition to the throughput of data flows [7, 12, 4, 5]. However, these works still do not provide an elegant framework where multiple traffic parameters can be optimized simultaneously as well as network resources can be flexibly arbitrated between the different traffic types. Also, they do not exploit physical layer DoFs like multiple OFDM sub-channels.

Among the works on resource allocation, [13, 15, 16] propose mechanisms to efficiently allocate sub-carriers to users in OFDM systems. The focus of these works is to allocate resources efficiently to meet user requirements that are predominantly related to rate or losses. They do not consider other QoS parameters like delay, jitter either in isolation or in tandem. Further, they also do not have the flexibility to consider sophisticated fairness models and also do not focus primarily on the increased complexity and overhead in scheduling multiple sub-channel resources to multiple users.

7. CONCLUSIONS

We have designed a unified scheduling framework for accommodating heterogeneous traffic with the ability to effectively tradeoff resources between them. A unified scheduling rule for handling heterogeneous traffic, as well as individual scheduling rules for optimizing individual traffic parameters have also been derived from the framework. The framework has also been extended to the case where both the base station and mobile users are equipped with multiple channels. Here, specific transmission strategies that optimize individual parameters have been identified and exploited in our unified scheduling rule to significantly reduce complexity and communication overhead, while performing reasonably close to the optimal scheme. Results indicate the superior ability of the proposed algorithms to efficiently handle heterogeneous traffic in multiple channel cellular networks over existing works. Further, the significant reduction in complexity and overhead achieved over the optimal scheme, enables the adoption of the algorithms in future high-speed cellular networks.

8. REFERENCES

[1] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, USA, 2004.

[2] S. Das, H. Viswanathan, and G. Rittenhouse. Dynamic load balancing through coordinated scheduling in packet data systems. In *IEEE INFOCOM*, Mar 2003.

[3] A. Eryilmaz, R. Srikant, and J. R. Perkins. Stable Scheduling Policies for fading wireless channels. *IEEE Transactions on Networking*, 13(2), Apr 2005.

[4] V. Huang and W. Zhuang. QoS-oriented packet scheduling for wireless multimedia CDMA communications. *IEEE Transactions on Mobile Computing*, 3(1), Jan-Mar 2004.

[5] H. Jiang and W. Zhuang. Cross-layer resource allocation for integrated voice/data traffic in wireless cellular networks. *IEEE Transactions on Wireless Communications*, 5(2), Feb 2006.

[6] R. Knopp and P. A. Humblet. Information capacity and power control in single cell multiuser communications. In *IEEE ICC*, Jun 1995.

[7] P. Liu, R. Berry, and M. L. Honig. Delay-sensitive packet scheduling in wireless networks. In *IEEE WCNC*, Mar 2003.

[8] X. Liu, K. P. Chong, and N. B. Shroff. Opportunistic transmission scheduling with resource sharing constraints in wireless networks. *IEEE Journal on Selected Areas in Communications*, 19(10), Oct 2001.

[9] R. G. Mukthar. Qns: Queuing network simulator. http://www.cubinlab.ee.mu.oz.au/rgmukht/qns.

[10] T. Nandagopal, T.-E. Kim, X. Gao, and V. Bhargavan. Achieving MAC Layer Fairness in Wireless Packet Networks. Proceedings of ACM MOBICOM, Aug 2000.

[11] A. Sang, X. Wang, M. Madihian, and R. Gitlin. Coordinated load balancing, handoff/cell-site selection, and scheduling in multi-cell packet data systems. In *ACM Mobicom*, Oct 2004.

[12] S. Shakkottai and A. L. Stolyar. Scheduling algorithms for a mixture of real-time and non-real time data in hdr. In *Bell Labs Technical report*, 2000.

[13] G. Song and Y. Li. Cross-layer optimization for OFDM wireless networks - Part I: Theoretical Framework. *IEEE Transactions on Wireless Communications*, 4(2), Mar 2005.

[14] K. Sundaresan, X. Wang, and M. Madihian. Scheduler design for heterogeneous traffic in next generation cellular networks. In *NEC Labs Technical Report*, Jan 2007.

[15] Y. J. Zhang and K. B. Letaief. Adaptive resource allocation for multiaccess MIMO/OFDM systems with matched filtering. *IEEE Transactions on Communications*, 53(11), Nov 2005.

[16] Z. Zhang, Y. He, and K. P. Chong. Oppotunistic downlink scheduling for multiuser ofdm systems. In *IEEE WCNC*, Mar 2005.