Channel and Delay Margin Aware Bandwidth Allocation for Future Generation Wireless Networks

Quang-Dung Ho, Mohamed Ashour, and Tho Le-Ngoc
Department of Electrical and Computer Engineering, McGill University
3480 University Street, Montreal, Quebec, H3A 2A7, Canada
Emails: {quang.ho, mohamed.ashour, tho.le-ngoc}@mcgill.ca

Abstract—This paper studies the problem of adaptive resource allocation for downlink transmission at the base station (BS) of future multi-service IP-based networks. A Delay Margin based Scheduling (DMS) approach is proposed to determine the queuing priority of an incoming packet by taking into consideration the distance between its currently experienced delay and its maximum allowed limit since such an instantaneous delay margin can timely and faithfully quantify its urgency. A number of queue architectures that aim to increase the bandwidth efficiency by integrating DMS with opportunistic User Channel based Scheduling (UCS) are also suggested. Wireless channel bandwidth efficiency and delay outage probability of the proposed approaches are extensively investigated by simulations.

I. INTRODUCTION

Next-generation wireless communication systems aim to support a wide range of multimedia and real-time services with strict Quality-of-Service (QoS) requirements in terms of delay, packet loss and delay jitter. Although numerous scheduling schemes are available for wireline networks, they are not directly applicable to the wireless setting because the link capacity of a mobile user is location-dependent and time-varying. The time-varying link capacity may in turn change the capacity allocated to each user from that intended by the scheduler. Therefore, in order to efficiently provide QoS, a wireless scheduler needs to consider both the user QoS requirements of each traffic flow and the channel conditions (e.g., propagation, interference, spectrum) of each mobile user.

Initial approaches to wireless scheduling in [1] tried to resolve this problem by providing means to compensate users that are scheduled to transmit when the channel is bad. The main disadvantage of those approaches is that they did not exploit the wireless multiuser diversity. In a wireless multiuser environment, each user experiences a different channel condition. Multiuser scheduling approaches make use of this diversity by opportunistically scheduling the user with the best channel for transmission. Opportunistic schedulers rely on Adaptive Modulation and Coding (AMC) at the physical layer to maximize the channel utilization. Although opportunistic approaches maximize the network transmission rate in the presence of infinite data backlog, in a multiuser environment, they may cause QoS violations for some users. Since users can have diverse QoS requirements, a user with a strict QoS may have a bad channel condition and hence low data rate. Efficiently supporting QoS in this case requires the use of cross-layer scheduling approaches [2]–[4]. These approaches depend on formulating the scheduling problem in the form of a linear utility function which is a combination of the user rate and the user QoS. The linearity of the problem allows separating the problem into a per user scheduling objective functions and choosing the users with the maximum/minimum objective function for transmission. A survey of these formulations can be found in [5]. However, by including the QoS in the objective function, these approaches aim to enhance the overall QoS of the system without ensuring that the QoS of each user is satisfied. These approaches are even unable to provide QoS differentiation for users requiring different services. They assume that each user supports a single QoS requirement and users with multiple QoS requirements are modeled as separate users. Due to the lack of QoS constraints, those differentiated services will receive a similar QoS or their QoS will be proportional to their input rates.

This paper introduces a scheduling approach that is based on the instant delay margin of each packet. Delay margin is the difference between the delay requirement of a service and the delay that the packet has encountered when traveling from the source to the BS. Delay margin provides information about the QoS constraint for each packet as well as the current states seen by each packet. The use of delay margin is particularly important in wireless networks where a larger (more positive) margin can compensate for unexpected changes of wireless channel characteristics. The main objectives of bandwidth allocation scheme proposed in this paper are (i) to satisfy delay requirements, and (ii) to utilize wireless channel bandwidth efficiently. In order to obtain the first objective, Delay-Margin-based Scheduling (DMS) considers the delay margin of each packet. The latter objective is obtained by User-Channel-based Scheduling (UCS) that considers the instantaneous channel states of each user. In the queuing architecture proposed in this paper, DMS and UCS are nested and the scheduler gives higher priorities to packets (i) whose delay margins are smaller (i.e., they are more urgent), and (ii) that are destined to mobile users whose instant channel conditions are better.

II. ASSUMPTIONS AND NOTATIONS

This paper considers downlink multiuser resource allocation over a wireless time-division multiplexing system with N mobile users. The system is assumed to be part of a larger system where the wireless BS is used to transport packets that
originated in other networks to wireless end-users. The BS is assumed to have perfect knowledge of downlink channel states of each user. The user channel states are statistically distributed and are independent of each other. The variation of the channel state is such that the channel can be assumed constant during the time slot. At each time slot, the BS uses this perfect channel state to calculate the maximum rate $r_i[n]$ that a user $i$ can achieve in time slot $n$. The calculation assumes fixed power allocation, and uses the maximum applicable channel coding and modulation for this user channel state at time slot $n$. 

Each user is assumed to receive three traffic classes: (i) Real-time (Rt), (ii) non-Realtime (nRt), and (iii) Best-Effort (BE). The downlink traffic is queued at the BS and at each time slot, a packet (or more) from each user service is scheduled for transmission. The resource allocation algorithm is responsible for assigning the transmission turns to each user. Each packet $p$ in the system has maximum allowed end-to-end delay $D_{max}[p]$, which is determined by delay requirement of the traffic class that the packet belongs to. A packet exceeding its maximum delay limit is dropped. Each packet $p$ in the system is assumed to encounter an instantaneous end-to-end delay $D^q_d[p]$ from source $s$ to destination $d$. The delay accounts for (i) the propagation delay $D^q_d[p]$ when packet $p$ travels from its source $s$ to the BS, (ii) the queuing delay $D^q_s[p]$ of packet $p$ at the BS, and (iii) the propagation delay $D^q_{p,s} [p]$ when packet $p$ travels from the BS to its destination. It is noted that $D^q_{bs}[p]$ in this paper may be used to represent the delay that a packet encountered in networks other than the downlink wireless network. $D^q_{bs}[p]$ is therefore calculated as follows:

$$D^q_{bs}[p] = D^q_{bs,s}[p] + D^q_s[p] + D^q_{p,s}[p] \tag{1}$$

The delay constraint of a packet can be stated as follows: when a packet arrives to its destination, its end-to-end delay must not exceed its maximum allowable value. Otherwise, the packet is expired and thus dropped, i.e.,

$$P \text{ packet is delay-satisfied if } D^q_d[p] \leq D_{max}[p] \tag{2}$$

III. THE PROPOSED BANDWIDTH SCHEDULING SCHEME AND QUEUING ARCHITECTURES

A. Delay-Margin-based Scheduling (DMS)

Although packets arriving to the BS are classified at their origin into a limited number of QoS services, the delay they encounter until they get to the BS, $D_{bs}[p]$, will vary based on their origins. This may result in realtime packets that have encountered very small delays because they originated in nodes that are close to the BS. Although belonging to the realtime service, such packets can tolerate larger values of $D^q_d[p]$ and $D^q_{p,s} [p]$ at the BS. On the other hand, non-realtime packets that have suffered long delays will require small values of $D^q_d[p]$ and $D^q_{p,s} [p]$. This means that the packet instant delay margin, i.e., the maximum additional delay that the packet can tolerate for traveling from the BS to destination, is significantly more important than its delay requirement. The delay margin can quantify exactly how urgent the packet is, respectively to its delay requirement, and thus directly relates to the priority that the packet should be given in the queue. Therefore, we propose Delay-Margin-based Scheduling (DMS) as follows.

Packets received by the BS and destined to downlink are classified by their delay margins (rather than simply by their delay requirements). It is assumed that there are three different classes of delay margins, i.e., Extremely Urgent (EU), Moderately Urgent (MU), and Not-Urgent (NU), therefore, there are three corresponding queues. Note that because BE traffic does not have any delay requirement, packets belonging to this type of traffic are always placed in BE queue.

Denote $\tilde{D}^q_{bs,d}[p]$ as delay margin of packet $p$ when it arrives at the BS. It is the maximum allowed delay for packet $p$ to travel from the BS to its destination when it arrives at the BS, i.e.,

$$\tilde{D}^q_{bs,d}[p] = \max\{D^q_d[p] + D^q_{p,s}[p]\} \tag{3}$$

By combining (3) with (1), and (2), $\tilde{D}^q_{bs,d}[p]$ can be easily calculated as follows:

$$D^q_d[p] \leq D_{max}[p] \Rightarrow \tilde{D}^q_{bs,d}[p] = D_{max}[p] - D^q_{bs,d}[p] \tag{4}$$

It is important to note that when packet $p$ arrives at the BS, its queuing delay and propagation delay from the BS to its destination are not known in advance. Let $D^q_d[p]$ and $\tilde{D}^q_{bs,d}[p]$ denote the estimated values of $D^q_d[p]$ and $D^q_{p,s}[p]$, respectively. They can be determined based on their historical values. For simplicity, they can be the weighted average of their actual values experienced by packets scheduled before packet $p$, i.e.,

$$D^q_d[p] = \sum_{\forall p' \in H(p)} \alpha[p'] \times D^q_d[p'], \quad \text{and} \quad \tilde{D}^q_{bs,d}[p] = \sum_{\forall p' \in H(p)} \beta[p'] \times D^q_{bs,d}[p'],$$

where $H(p)$ is the set of all packets scheduled before packet $p$, $\alpha[p']$ and $\beta[p']$ are the weights to determine the contributions of actual historical queuing delay and propagation delay of packet $p'$ to the respective estimated values of packet $p$ ($\sum_{\forall p' \in H(p)} \alpha[p'] = 1$, and $\sum_{\forall p' \in H(p)} \beta[p'] = 1$). Those weights can be arbitrary constants between 0.0 and 1.0 in order to control how rapidly the estimated delays adapt to changes.

Then, the estimated delay for packet $p$ to travel from the BS to its destination, calculated when packet $p$ arrives at the BS, denoted by $\tilde{D}^q_{bs,d}[p]$, includes the estimated queuing delay and the estimated propagation delay, i.e.,

$$\tilde{D}^q_{bs,d}[p] = D^q_d[p] + \tilde{D}^q_{p,s}[p] \tag{5}$$

Next, denote $\tilde{D}^q_{bs,s}[p]$ as the normalized delay margin of packet $p$ when it arrives at the BS. It is the delay margin of packet $p$ when it arrives at the BS normalized to the estimated delay for packet $p$ to travel from the BS to its destination, i.e.,

$$\tilde{D}^q_{bs,s}[p] = \frac{\tilde{D}^q_{bs,d}[p]}{\tilde{D}^q_{bs,d}[p]} \tag{6}$$

From (1), (2), (4), and (6), it is observed that: (i) if $\tilde{D}^q_{bs,d}[p] \leq 0$, or in other words, $\tilde{D}^q_{bs,s}[p] \geq D_{max}[p]$, then $D^q_d[p] + D^q_s[p] + D^q_{p,s} = D^q_d[p] > D_{max}[p]$ since $D^q_d[p] + \tilde{D}^q_{bs,s}[p] > 0$. Therefore, definitely, packet $p$ should be dropped at the
For example, if packets \( p \) has its maximum allowed end-to-end delay at its destination is classified based on user identification of their destinations changed accordingly to the rates of arrival traffic, queue states, or (iv) not urgent if its delay margin is more than two times of its estimated value. Those thresholds can be dynamically or (iii) moderately urgent if its delay margin is more than 20% but less than 80% of its estimated value, or considered as (ii) extremely urgent if its delay margin is immediately dropped if its delay margin is less than 20% of \( T_{\text{drop}}[p] \), then packet \( p \) is dropped, extremely urgent, moderately urgent, or not urgent.

One example of those thresholds can be \( D_{\text{bs},d}^q[p] \), normalized delay margins to determine if the packet \( p \) arrives at the BS: 
- If \( D_{\text{bs},d}^q[p] \) is dropped, immediately dropped.
- If \( T_{\text{drop}}[p] \leq D_{\text{bs},d}^q[p] < T_{\text{EU}}[p] \), then packet \( p \) is considered as extremely urgent, and placed in EU queue.
- If \( T_{\text{EU}}[p] \leq D_{\text{bs},d}^q[p] < T_{\text{NU}}[p] \), then packet \( p \) is considered as moderately urgent, and placed in MU queue.
- If \( D_{\text{bs},d}^q[p] \geq T_{\text{NU}}[p] \), then packet \( p \) is considered as not urgent, and placed in NU queue.

Note that transmission if it has the highest achievable data transmission rate: 
\[
k = \arg \max \{ r_i[n] \}.
\]

C. Proposed Queuing Structures

In this paper, DMS and UCS can be arranged in various structures as follows.

1) DMS-then-UCS (DMS-t-UCS) Structure: DMS and UCS are done sequentially where DMS is carried out first. When a packet arrives at the BS, it is first classified based on its delay margin and placed in its respective queue. DMS regulates traffic by considering the delay margin of each packet and queue states. Then, each packet forwarded by DMS is classified based on its destination and placed in its respective queue. UCS regulates traffic by considering channel states of each user. DMS-t-UCS structure is shown in Fig. 1. The (whose channel states are changed over time) and buffered into one of the queues in the BS. Opportunistic scheduling is employed, i.e., at time slot \( n \), user \( k \) is selected for transmission if it has the highest achievable data transmission rate: 

B. User-Channel-based Scheduling (UCS)

Packets received by the BS and destined to downlink are classified based on user identification of their destinations.
service rate of DMS, namely $R$, is tunable. It is observed that $R$ can be used to control the trade-off between delay-based scheduling and channel-based scheduling. In section IV, an appropriate value of $R$ will be determined by simulations.

2) DMS-and-UCS (DMS-n-UCS) Structure: DMS and UCS are done jointly where DMS is given a higher priority. When a packet arrives at the BS, it is first classified based on its delay margin and managed by its corresponding queue group. In each delay-margin-based queue group, the packet is placed in the queue respective to its destination. Traffic is regulated by jointly considering delay margins of each packet, queue states, and channel states of each user. DMS-n-UCS structure is shown in Fig. 2.

3) UCS-and-DMS (UCS-n-DMS) Structure: DMS and UCS are done jointly where UCS is given a higher priority. When a packet arrives at the BS, it is first classified based on its destination and managed by its corresponding queue group. In each user-based queue group, the packet is placed in the queue respective to its delay margin. Traffic is regulated by jointly considering channel states of each user, delay margins of each packet, and queue states. UCS-n-DMS structure is shown in Fig. 3.

IV. SIMULATION RESULTS

The simulated network consists of a server generating packets that are delivered to the BS by an IP-based core network. The BS then broadcasts traffic to $N = 20$ mobile users over Rayleigh fading channels. On-off traffic model is assumed for each connection from the server to mobile users. The total offered traffic load is defined by $\rho = \frac{\lambda}{C_{\text{avg}}}$, where $\lambda$ and $C_{\text{avg}}$ are the total offered load and the average capacity of the wireless channel, respectively. The maximum allowed delays $D_{\text{max}}$ of realtime traffic and non-realtime traffic are 300.0 milliseconds and 2.0 seconds, respectively. AMC is employed at the BS to guarantee a bit error rate performance of $10^{-6}$ or better for all wireless links. The performance of proposed scheduling algorithms is studied in terms of the queue length survivor function (i.e., the probability that the queue length exceeds a given queue size), and the packet delay survivor function (i.e., the probability that the end-to-end packet delay exceeds a given maximum tolerated value).

As mentioned in section III, the service rate $R$ of DMS in DMS-t-UCS structure can be tuned to variate the trade-off
between delay-based scheduling and channel-based scheduling. In order to determine an appropriate value of $R$, the performance of DMS-t-UCS structure is measured when $R$ is varied from a very low value to a very high value, as can be seen in Fig. 4. It is noted that $C_{\min}$, $C_{\text{avg}}$, and $C_{\text{max}}$ are the minimum, average, and maximum data transmission rate of the wireless channel from the BS to end-users, i.e., $C_{\min} = \min_{i,n} \{r_i[n]\}$, $C_{\text{avg}} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{n=1}^{T} r_i[n]$, and $C_{\text{max}} = \max_{i,n} \{r_i[n]\}$, where $r_i[n]$ is the maximum rate that a user $i$ can achieve in time slot $n$. The total offered load $\rho$ is 90% (moderate load).

Fig. 4 shows that when $R$ is very low, both delay performance and channel efficiency of DMS-t-UCS are very poor. This can be explained as follows. When $R$ is very low (compared to the average channel capacity of the wireless channel), packets arriving at DMS are forwarded to UCS very slowly. Since the rate of traffic arriving at UCS is much slower than the serving rate of UCS which is actually the channel capacity of wireless channel, UCS approximately works as a simple First-Come-First-Serve (FCFS) queue. As a result, DMS-t-UCS can be seen as a single DMS which attempts to satisfy delay requirement without exploiting the multiuser gain of the wireless channel. On the other hand, when $R$ is very high (compared to the average channel capacity of the wireless channel or the arrival rate of traffic from the core network), when a packet arrives at DMS, it is forwarded to UCS very quickly (before the next packet arrives), and DMS simply works as a simple FCFS queue. DMS-t-UCS then can be seen as a single UCS which attempts to exploit the wireless channel very opportunistically without considering the packet delay requirement. Therefore, although the channel bandwidth efficiency is high, the delay performance is very poor. From Fig. 4, it can be seen that when $R$ is set to be the maximum data transmission rate of the wireless channel, both delay performance and channel efficiency are very high. Hence, hereafter, $R = C_{\text{max}}$ is used for DMS-t-UCS.

The overall channel bandwidth efficiency and packet delay outage probabilities of realtime traffic and non-realtime traffic given by the traditional Service-Requirement-based Scheduling (SRS: packets received by BS are classified and scheduled by simply considering their service requirements, i.e., Rt, nRt, or BE), DMS-t-UCS, DMS-n-UCS, and UCS-n-UCS are compared in Fig. 5, Fig. 6, and Fig. 7, respectively. A wide range of total offered load is considered ($\rho$ is varied from 70% to 95%).

Fig. 5 shows that UCS-n-UCS can obtain the highest channel bandwidth efficiency. Fig. 6 shows that, for realtime traffic, (i) at low loads ($\rho$ is from 70% to 85%) DMS-n-UCS offers the lowest delay outage probability, and (ii) at high loads ($\rho$ is 90% or 95%) UCS-n-UCS outperforms other schemes in terms of delay outage. Fig. 7 shows that, for delay performance of non-realtime traffic, UCS-n-UCS provides the best delay performance as compared to other schemes over the whole range of loads of interest.

The UCS-n-UCS scheme is further investigated by plotting its delay survival function, i.e., the delay outage probabilities against different maximum allowed delays, of three different types of traffic in Fig. 8. The total offered load $\rho$ is 90%. The figure shows that while opportunistically using the wireless channel, UCS-n-UCS can provide a clear QoS differentiation between realtime, non-realtime, and best-effort traffic.

V. CONCLUSIONS

This paper provides a queueing architecture that integrates delay margin based scheduling with opportunistic resource allocation. By giving higher transmission priorities to packets whose delays are closer to their maximum tolerated values, Delay Margin based Scheduling (DMS) attempts to guarantee the end-user quality of service. Furthermore, the overall bandwidth efficiency is improved by incorporating the opportunistic scheduling, namely User Channel based Scheduling (UCS), with DMS. The paper presents three different possible approaches to combine DMS and UCS, and extensively investigates the performances of those approaches by simulations. The DMS-t-UCS queueing architecture that serially combines DMS and UCS exhibits a very interesting capability: it can easily move between delay-based scheduling and channel-based scheduling by varying the service rate of DMS. The simulation results also show that by combining DMS and UCS with a higher priority assigned to UCS, the resulting UCS-n-UCS scheme can offer the highest bandwidth efficiency at the lowest delay outage probability.

REFERENCES