Computer Communications 000 (2016) 1–10

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Queue-aware uplink scheduling with stochastic guarantees

Amr Rizk[a,*], Markus Fidler[b]

[a] ECE Department, University of Massachusetts Amherst, 151 Holdsworth Way, Amherst, MA 01003, USA
[b] Leibniz Universität Hannover, Institut für Kommunikationstechnik, Appelstraße 9A, Hannover 30167, Germany

ARTICLE INFO

Article history:
Received 6 September 2015
Revised 16 February 2016
Accepted 23 February 2016
Available online xxx

Keywords:
Scheduling
Stochastic network calculus
Resource allocation
Mobile uplink

ABSTRACT

Adaptive resource allocation arises naturally as a technique to optimize resource utilization in communication networks with scarce resources under dynamic conditions. One prominent example is cellular communication where service providers seek to utilize the costly resources in the most effective way. In this work, we investigate an uplink resource allocation scheme that takes into account the buffer occupation at the transmitter to retain a given level of quality of service (QoS). First, we regard exact results for the class of Poisson traffic where we investigate the sensitivity of the resource adaptation and QoS level to the actuating variables. We show relevant resource savings in comparison with a static allocation. Further, we regard a queueing setting with general random arrival and service processes. In particular, we consider the service of wireless fading channels. We show two different resource adaptation mechanisms that depend on the strictness of different assumptions. Finally, we present simulation results that show substantial resource savings using the queue-aware scheduling scheme, where we provide insight on the implementation and operation of such an adaptive system.

1. Introduction

Many components of communication networks are subject to variability. This includes the usage behavior of communicating parties, as well as, the service provided by the network. While the user behavior translates to a variable resource demand, the provisioned service is constricted by expenditure and the technological state-of-the-art. This inherent variability is the raison d'être for many optimizations found in communication networks. An intrinsic difficulty in cellular wireless communication is the fading nature of the channel which causes the transmission rate to vary over time. Hence, to better utilize the wireless channel, respectively, to provide quality of service guarantees in cellular communication networks, a base station has to estimate the statistical properties of the wireless fading channel. For example in LTE this estimate is captured in the channel quality indicator (CQI) [1].

In addition to channel quality estimates, current LTE systems offer a valuable source of information, i.e., buffer status reports (BSR) [2], that can be exploited for adaptive resource allocation [3]. In Fig. 1(a) we depict a user equipment (UE) that transmits BSRs in uplink direction to signal the buffer occupancy to the base station. The base station takes the buffer occupancy into account when updating the resource allocation to the UE. This is illustrated in Fig. 1(a) as a variable amount of (shaded) time–frequency resource blocks that are granted to the UE. In addition, Fig. 1 comprises the scheduling epoch Δ, i.e., the recurrence period of the resource scheduling operation.

Promising applications of adaptive resource allocation include jitter control, substantial radio resource savings, as well as, battery savings on the UE side. Jitter, i.e., high delay variations, may arise in wireless communications due to the fading characteristics of the channel. It is known that jitter has a strong adverse influence on the quality of experience. Adaptive resource allocation can mitigate the impact of the channel fading to reduce jitter at the receiver. Further, adaptive resource control may achieve substantial resource savings compared to static resource grants due to an effective use of available information.

Despite the expected benefits and the recent significant progress in the analysis of QoS metrics, few strategies are derived that use analytical models to consider adaptive resource optimization under QoS constraints. In this work we provide an analytical approach to adaptive resource allocation based on buffer occupancy. We present a queue-aware scheduling scheme that adapts the amount of resources provided to a single UE under probabilistic QoS constraints.

Consider the scenario in Fig. 1(a) where traffic denoted A arrives at a UE transmit buffer. The UE regularly signals BSRs that

* Corresponding author. Tel.: +1 4134041316.
E-mail addresses: arizk@umass.edu, arizk@engin.umass.edu (A. Rizk), markus.fidler@ikt.uni-hannover.de (M. Fidler).

http://dx.doi.org/10.1016/j.comcom.2016.02.014
0140-3664/

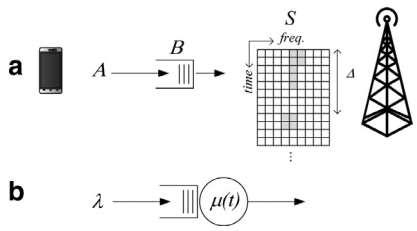


Fig. 1. (a) Example of queue-aware scheduling in cellular networks. The base station decides on the amount of uplink service S depicted as a varying number of resource blocks (gray) granted to a UE depending on its transmit buffer filling B. The scheduling epoch is denoted Δ. (b) Abstraction of queue-aware scheduling with a single user as a queueing system with an adaptive mean service rate $\mu(t)$. The service rate is adjusted at scheduling epochs of length Δ, to maintain a small queue.

include the transmit buffer filling B to the base station, which in turn seeks to adapt the service S, i.e., the uplink bandwidth resource grants, based on the knowledge of BSR and CQI. First, we regard the abstraction in Fig. 1(b) with a queuing system fed by Poisson traffic arrivals of mean rate λ and a time-varying mean service rate $\mu(t)$. We present a study of exact results for Poisson traffic that clearly shows resource savings when queue-aware scheduling is deployed. One desired property of adaptive resource allocation is robustness with respect to variations of the actuating variables. Hence, we present a sensitivity study that shows the impact of actuating variables, as well as, the system robustness with respect to misadaptation. In a practical scenario this would, for example, capture imperfect CQI. For general arrival and service processes we present an analytical framework to implement queue-aware scheduling that is based on the stochastic network calculus. We distinguish two regimes for the adaptive system that we denote frequent and infrequent adaptation. Consequently, we provide a detailed analysis of two resource adaptation schemes showing evaluation results and insight on the implementation and operation of such systems. We include a compact investigation of the adaptive system in multi-user scenarios. The main contributions of this paper are:

- For the class of Poisson processes, we present exact results to quantify best-case resource savings, i.e., given full knowledge of the traffic and service statistics.
- Our model reveals an important relation of the average traffic arrival rate, the scheduling epoch length, and the target queue constraint. We identify two regimes, one where adaptive scheduling is effective and one where it is not. The result is significant as it shows in a mathematical, exact framework that there are relevant cases where an adaptive system cannot benefit from the additional information provided by BSRs.
- Our results show that the adaptive system can stabilize the queue even in case of a systematic service rate misadaptation. This robustness is important, since in practice an adaptive system can only *estimate* the number of radio resource blocks that are required to achieve a target service rate.
- Our mathematical treatment of queue-aware scheduling is applicable to a broad class of arrival and service processes known in the stochastic network calculus.

This work is an extended version of the work in [4]. Here we provide a fundamentally different bounding method that is particularly adapted for the considered wireless channel model. We apply the new methodology to the so called infrequent adaptation scheme in Section 5.2 and provide a comparison of the respective results showing that the performance using the new technique matches the target criterion more closely and hence enables saving more resources compared to [4]. Further, we expand here the description of queue aware scheduling techniques in multi-user scenarios in Section 5.3 and provide analytical formulations for the resource share that is given to each mobile user given a certain scheduling discipline.

The rest of this paper is structured as follows. In Section 2 we discuss related work on the analysis of adaptive resource allocation techniques and queueing systems with variable service rates. Section 3 presents a study of exact results for Poisson traffic. In Section 4 we introduce a model for wireless systems and provide an introduction to the analytical framework. Sections 5.1 and 5.2 present a description of the implementation of frequent and infrequent adaptation including evaluation results and insight on the implementation. In Section 5.3 we include simulation results for multi-user scenarios under different scheduling policies. We conclude the paper in Section 6.

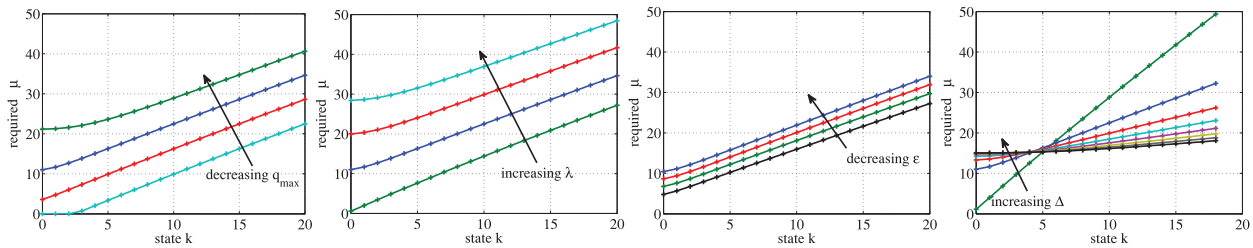
2. Related work

We find that studies related to this work were mainly conducted in the context of (i) the optimization of service policies for queueing systems and (ii) the optimization of power and rate control in cellular networks. First, we will review works with the first objective (i) showing the main difference to the work at hand.

The authors of [5–7] consider a dynamic control approach (speed scaling) of the service rate of $M|M|1$, respectively $M|GI|1$ processor sharing queues, that depends on the queue state at each time instant. The service rate is optimized with respect to service costs that are defined as a function of the queue length at each time point, as well as, the instantaneous service rate. The result is a service policy, i.e., an optimization for entire service sample paths with respect to a given criterion. For example, the authors of [5] provide recursive algorithms to minimize the average service costs. General tradeoffs in the design of speed scaling controllers for queues are shown in [8], e.g., combining the response time with job energy consumption. The authors show that for certain schedulers only two of the three attributes “optimality, fairness and robustness” can be achieved. The work in [9] studies multi-class $M|G|1$ queues with variable service rates. The authors show scheduling policies that minimize service costs associated with the instantaneous service through convex functions. The authors of [10] consider an $M|M|1$ queue with time varying externally Markov modulated server speed. Although not explicitly given, the authors show a method to numerically obtain the average waiting time. In [11] the authors straightforwardly employ the Pollaczek–Khinchine formula in conjunction with a power model, that is known for networks on-chip to minimize the average power consumption in an $M|G|1$ queue.

The work at hand differs basically from the related work above in the analysis of an epoch based adaptation scheme that takes *general* arrival and service processes into account. We consider a probabilistic QoS constraint as optimization metric in contrast to service cost functions.

The second category of related works comprises rate and power optimization in cellular networks such as [12–15]. Typically, the criterion for optimization is the average queueing delay. In [12] the authors regard a transmitter with variable rate that serves a queue filled at a constant rate. The authors perform optimizations over power and rate policies for a single user scenario to minimize the average delay under power constraints. The technique used is dynamic programming which provides numerical solutions for a predefined cost function that consists of a weighted sum of the buffer length and the transmission power. Using a similar approach the authors of [14] provide an optimal service policy for a finite service sample path length. They assume a channel of Gilbert–Elliot type and a linear relationship of transmission power and rate. The work in [15] considers a scenario with arrivals and service processes given by Markov chains where data arriving from higher layers is buffered until transmission. The authors provide



(a) $q_{\max} \in \{5, 10, 15, 20\}$ (b) $\lambda \in \{5, 10, 15, 20\}$ (c) $\varepsilon \in 1/2^{\{3,2,1,0\}} \cdot 10^{-2}$ (d) $\Delta \in \{0.5, 1, 1.5, ...4\}$

Fig. 2. Required service rate μ to satisfy the bound (1) depending on the initial queue state k at the beginning of the scheduling epoch. Influence of parameters: (a) bound on the queue length at the end of the epoch $q_{\max}$, (b) arrival rate λ, (c) violation probability ε, (d) scheduling epoch length Δ. Baseline (blue curve): $q_{\max} = 10, \lambda = 10, \varepsilon = 10^{-2}, \Delta = 1$. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

results on regulating the user transmission rate and power to control the average transmission power and average delay using concepts from Markov decision theory. Further, BSRs, respectively, the transmit queue length, have been used for scheduling optimization in [3,16,17]. In [3] BSRs are used to improve packet drop rates in OFDM downlink transmissions. In [16] the authors model the polling service of an IEEE 802.16 network using a Markov model to show the impact of queue length aware rate and bandwidth control on the average delay. In [17] BSRs are used in a scheduling metric to distribute physical resource blocks over different UEs at relay nodes.

Key differences to the related work above are that we regard an online scheme that enables adapting the parameters for a scheduling epoch during the runtime of the system. In contrast to objectives in the related work such as minimizing the average delay or a weighted sum of buffer length and the transmission power, we adapt the resource allocation with respect to the tail of the queue size distribution. This provides a natural relation of the provided quality of service during a scheduling epoch to the adaptively allocated resources.

3. Exact results for Poisson traffic: Tradeoffs and sensitivity

In this section, we develop a model of queue-aware scheduling for Poisson traffic. We use this basic model to provide exact results that yield relevant insights. We will relax the assumptions in Section 4 where we consider general arrival and service processes.

3.1. Epoch-based resource allocation

Next, we use the model of a single queue to express the adaptation of the mean service rate $\mu(t)$ at multiples of the scheduling epoch length Δ to provide a probabilistic bound on the queue length at the end of the scheduling epoch. The service rate $\mu(t)$ is chosen for a scheduling epoch Δ depending on the initial queue length at the start of the epoch, as well as, the arrival rate λ. With respect to the wireless application scenario in Fig. 1, the adaptive system models a base station that decides on the amount of resources it will provide to a UE during a scheduling epoch given knowledge of the UE transmit buffer filling and its average arrival rate. In the following, we investigate the tradeoffs and fundamental limits of such a system and conduct a sensitivity analysis with respect to misadaptation. We will show that queue-aware scheduling achieves target QoS constraints and provides significant resource savings.

The service rate $\mu(t)$ is adjusted based on the queue state k at the beginning of the epoch. During the epoch Δ the average service rate is fixed such that the probability that the queue is in a state higher than $q_{\max}$ after Δ is bounded by ε, i.e.,

$$\sum_{l=q_{\max}+1}^{\infty} p_{kl}(\Delta) \leq \varepsilon \tag{1}$$

with $p_{kl}(\Delta)$ being the probability that the queuing system is in state l at time Δ after initially being in state k. The transient behavior of the $M|M|1$ queuing system has been investigated in [18,19] leading to the closed form solution

$$p_{kl}(\Delta) = e^{-(\lambda+\mu)\Delta}\Big[\varrho^{\frac{l-k}{2}}\mathcal{I}_{l-k}(z\Delta) + \varrho^{\frac{l-k-1}{2}}\mathcal{I}_{l+k+1}(z\Delta) + (1-\varrho)\varrho^{l}\sum_{j=l+k+2}^{\infty}\varrho^{-\frac{j}{2}}\mathcal{I}_{j}(z\Delta)\Big] \tag{2}$$

with utilization $\varrho = \lambda/\mu$, $z = 2\mu\sqrt{\varrho}$ and the modified Bessel function of the first kind $\mathcal{I}_{(\cdot)}(\cdot)$. We denote this system as the adaptive system, where we compute μ for the next epoch from (2) given the queue size at the beginning of the epoch is k. External parameters are $q_{\max}$, λ, Δ, ε.

First, we illustrate the operation of the adaptive system. Fig. 2 shows the required service rate μ given the queue state k at the beginning of the scheduling epoch. This adaptive system may be viewed as a controller with input parameter k and an actuating variable μ. Next we evaluate how the parameters of the adaptive system, $q_{\max}$, λ, ε and Δ, influence the adaptation of the service rate μ. Fig. 2(a)–(c) shows the required service rate, that increases with the initial state k. It also shows the increase of the required service rate μ with tighter constraints, i.e., with decreasing $q_{\max}$, increasing arrival rate λ or decreasing violation probability ε from (1). Note that the curves are equidistant with linear change in $q_{\max}$ and λ, respectively, with logscale change in ε. Also note the nonlinear behavior for boundary scenarios, i.e., small $q_{\max}$ and small initial k. Fig. 2(d) shows the impact of the length of the epoch Δ on the required service rate. Smaller Δ cause a stronger adjustment.

3.2. Improvement on the static system

Next, we compare the adaptive system to a static $M|M|1$ system with identical arrival rate λ and a fixed equivalent service rate $\overline{\mu} = \mathrm{E}[\mu(t)]$. We denote this system the static system and show results from discrete event simulations that compare its performance to the adaptive system. Figs. 3 and 4 use the basic parameter set $\lambda = 10$, $\Delta = 1$, $\varepsilon = 10^{-2}$ and for the simulation results we considered 10^4 epochs. First, consider the case $q_{\max} = 10$ in Fig. 3(a). The figure shows the complementary cumulative distribution function (CCDF) of the queue length at multiples of the scheduling epoch Δ for the adaptive and for the static system. The adaptive system attains the QoS requirement, i.e., the queue length exceeds $q_{\max} = 10$ at most with probability $\varepsilon = 10^{-2}$, and outperforms the static system in terms of the queue length distribution. The reason behind

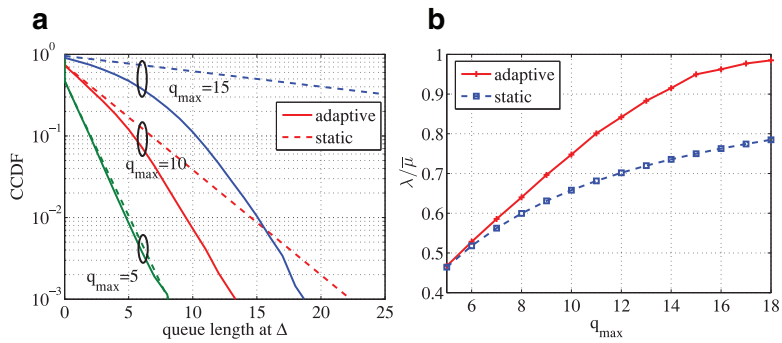


Fig. 3. (a) The regimes of adaptation: Different $q_{\max} \in \{5, 10, 15\}$ with corresponding static equivalents. The adaptive $M|M|1$ system outperforms the equivalent system with constant μ. (b) Resource saving with adaptive allocation. The utilization increases with $q_{\max}$ as the adaptive $M|M|1$ system makes efficient use of the variations of the service rate μ. In comparison we show the utilization for static scenarios that attain the same probabilistic bound on the queue length. The adaptive system runs at a higher utilization than the comparable static system and hence saves resources.

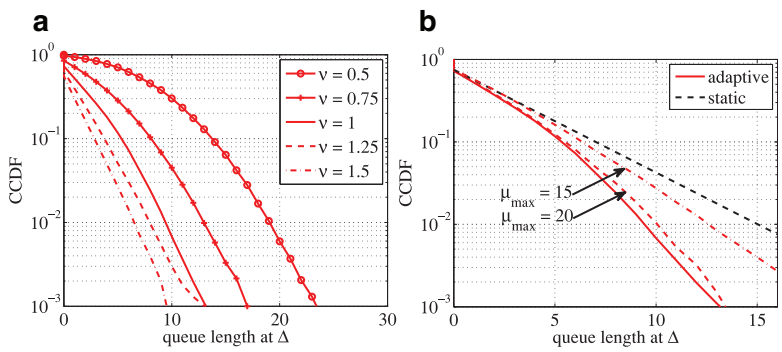


Fig. 4. (a) Robustness with respect to service rate misadaptation. The provided service rate at each scheduling epoch is a scaled version $\nu\mu$ of the required service μ. For $\nu \neq 1$ the CCDF is shifted but the queue does not grow unbounded. (b) Robustness with respect to service rate limitation. The provided service rate is bounded by $\mu_{\max}$. The CCDF is shifted with respect to the unconstrained scenario.

this performance difference is that the adaptive system reduces the service rate for epochs with small initial queue fillings (and vice versa for epochs with large initial queue fillings). Hence, the adaptive system continuously minimizes idle times where resources would be effectively wasted. A downside is that data may wait longer in the queue if the queue length is much smaller than $q_{\max}$. The adaptive system allocates only low μ in case of a small initial queue length, e.g., $\mu = 0$ for $k \leq 2$ and $q_{\max} = 20$ (see Fig. 2(a)). Data arriving during such a scheduling epoch wait longer in the queue than in case of the static system.

Next, we inspect the utilization of the adaptive system defined as $\lambda/\overline{\mu}$, i.e., the arrival rate divided by the average service rate of the adaptive system. Fig. 3(b) shows that the adaptive system runs at a higher utilization with increasing $q_{\max}$. We find that a key relation is the ratio of $\lambda\Delta$ to $q_{\max}$ for a given ε, i.e., the average amount of arrivals in one epoch vs. the bound on the queue length at the end of the epoch. The figure shows that for increasing $q_{\max}$ with respect to $\lambda\Delta$ the adaptive system may run under very high utilization, while still maintaining the probabilistic bound (1). The figure also depicts the required utilizations for the static system to provide the same probabilistic bound on the queue distribution. The difference in Fig. 3(b) reveals the substantial resource saving provided by the adaptive system. Observe that the difference between the adaptive system and the static system is apparent for $q_{\max} \gg \lambda\Delta$. The adaptive system is aware of the queue length at the beginning of the scheduling epoch, yet, the actual arrivals are unknown in advance and may vary significantly. The information on the initial queue length becomes less helpful if the unknown, i.e., the traffic amount in Δ, predominates, i.e., if $q_{\max} \ll \lambda\Delta$. This is also reflected by the CCDFs for $q_{\max} = \{5, 15\}$ in Fig. 3(a). Hence, the adaptive system is favorable for $q_{\max} \gg \lambda\Delta$.

3.3. Robustness

One desired property of such an adaptive system is robustness with respect to misadaptation, which we define as a queue that does not grow unbounded if the actual service rate is only a scaled version $\nu\mu$ of the required service rate μ. This robustness property is important in practice. For example, consider the cellular system from Fig. 1, where the base station uses CQI to estimate the channel condition. It is desirable that an adaptive resource allocation scheme is robust with respect to deviations of these estimates from actual channel conditions. For a static system an allocation of $\nu\mu$ could lead to instability, hence, to an unbounded queue. Fig. 4(a) shows the impact of misadaptation for different values of ν and $q_{\max} = 10$. We observe that the queue length distribution is shifted for $\nu \neq 1$ and that the probabilistic bound is violated as expected for $\nu < 1$. However, the queue length does not grow unbounded.

A second requirement of practical implementations is that the adaptive service rate $\mu(t)$ is upper bounded by some finite $\mu_{\max}$. We simulate the operation of the adaptive system under service limitation and show the results in Fig. 4(b). Note that the queue length distribution is shifted away with respect to the unconstrained scenario with stricter $\mu_{\max}$.

In this section, we provided a proof of concept for queue-aware scheduling for the example of Poisson traffic. The adaptive system retains a given probabilistic bound on the queue length while it may substantially save resources. Interesting though are constellations, which we showed, that hardly comprise resource savings. This reveals that the operation of queue-aware scheduling is non-trivial and requires a careful analysis. The question of how to exploit the potential resource savings in a wireless system that

deviates from the Poisson assumption is a difficult challenge. In the next sections we will relax the Poisson traffic assumption and present implementations of queue-aware scheduling for wireless fading channels and general traffic arrivals.

4. Modeling wireless systems

Next, we will formulate a basic queueing model from the network calculus to include general arrival and service processes. We will adopt a basic channel model for wireless communication systems that is known from [20].

4.1. Queueing model

We apply concepts of the framework of the stochastic network calculus [21–25], and consider a discrete time, lossless and work-conserving queueing system. Cumulative traffic arrivals to the system are denoted $A(\tau, t)$, i.e., the cumulative amount of bits arriving in the time interval $(\tau, t]$ for $t \geq \tau \geq 0$. Hence, $A(t,t) = 0$ for all $t \geq 0$ and there are no arrivals for $t \leq 0$. By convention we use $A(t)$ to denote the arrivals between $(0, t]$, where $A(t)$ is a non-negative non-decreasing random process that passes through the origin. Further, we use λ to denote the average arrival rate, i.e., $\lambda = \lim_{t\to\infty} A(t)/t$. The cumulative departures of the queuing system up to time t denoted $D(t)$ are related to the arrivals through the service provided by the system. The queuing model considers the service in $(\tau, t]$ as a random process $S(\tau, t)$ which is non-increasing in τ and non-decreasing in t. Further, note that $S(t,t) = 0$ for all $t \geq 0$.

For a work-conserving system with a time-varying service $S(\tau, t)$ it holds for all $t \geq \tau \geq 0$ where τ, t fall into the same busy period that $D(t) \geq D(\tau) + S(\tau, t)$ [26,27]. This is referred to as strict service. Systems offering strict service also provide a so-called adaptive service curve [23,27] such that for all $t \geq \tau \geq 0$ it holds that

$$D(t) \geq \min\Big[D(\tau)+S(\tau,t),\ \inf_{u\in[\tau,t]}\{A(u)+S(u,t)\}\Big]. \tag{3}$$

The stochastic evaluation of queueing systems with respect to performance metrics, e.g., backlog and delay, frequently uses non-random lower bounding functions $\mathcal{S}(t-\tau)$ of the service process $S(\tau, t)$ defined for all $t \geq \tau \geq 0$ as

$$\mathrm{P}[S(\tau,t) \geq \mathcal{S}(t-\tau)] \geq 1-\varepsilon_p, \tag{4}$$

with a violation probability ε_p. For systems providing adaptive service (3) we require, however, a bound on $S(\tau, t)$ that is valid for an entire interval to derive a probabilistic extension of (3) as defined in [23] for $t \geq \tau \geq 0$ as

$$\mathrm{P}\Big[D(t) \geq \min\Big[D(\tau)+\mathcal{S}(t-\tau),\ \inf_{u\in[\tau,t]}\{A(u)+\mathcal{S}(t-u)\}\Big]\Big] \geq 1-\varepsilon_s. \tag{5}$$

A bound on $S(\tau, t)$ for an entire interval is given as

$$\mathrm{P}[S(u,t) \geq \mathcal{S}(t-u),\ \ \forall u \in [\tau,t]] \geq 1-\varepsilon_s \tag{6}$$

for $t \geq \tau \geq 0$. It is known as an ε-effective service curve in [23,28]. For a function $\mathcal{S}(t)$ satisfying (4) with ε_p we find that it satisfies (6) with $\varepsilon_s = (t-\tau)\varepsilon_p$. This is directly obtained by using Boole's inequality as

$$\mathrm{P}[\exists u \in [\tau,t] : S(u,t) < \mathcal{S}(t-u)] \leq \sum_{u=\tau}^{t-1} \mathrm{P}[S(u,t) < \mathcal{S}(t-u)] = \sum_{u=\tau}^{t-1} \varepsilon_p = (t-\tau)\varepsilon_p = \varepsilon_s. \tag{7}$$

We used this technique (7) in [4] to provide stochastic bounds for general service processes $S(\tau, t)$. While the general technique above is applicable to a number of service processes, we illustrate a refined method in Section 5.2 that provides better bounding accuracy. In the following, we review a known model of the Rayleigh fading channel including a corresponding bound in the sense of (6) on its service process $S(\tau, t)$.

4.2. Wireless channel model

We adopt the basic concept of a wireless transmission over a fading channel from [20]. We estimate the capacity in one time slot in a block fading model from the Shannon capacity formula as $C = W\log_2(1+\gamma_i)$ with channel bandwidth W and γ_i denoting the signal-to-noise ratio (SNR) in the ith block. We consider a time slotted service process with a slot length that is congruent with the block length in the block fading model. The service process is composed of iid increments c_i that are given, using the shorthand notation $\beta = W\delta/\ln 2$, as

$$c_i = \beta\ln(1+\gamma_i), \tag{8}$$

with random γ_i and a fixed time slot length that is denoted δ. The iid assumption is reasonable if the time slot duration is large enough compared to the channel coherence time [20]. Here, c_i presents the number of bits that can be served in the ith slot. Considering a Rayleigh fading channel, it follows that γ_i is exponentially distributed with parameter η and average SNR $E[\gamma] = 1/\eta$. The service process $S(\tau, t)$ is given as $S(\tau,t) = \sum_{i=\tau+1}^{t} c_i$, for $0 \leq \tau < t$. Given the service process $S(\tau, t)$ for a Rayleigh fading channel with SNR parameter $1/\eta$, the function

$$\mathcal{S}(t) = \frac{1}{\theta}(\ln(\varepsilon_p) - t[\eta + \theta\beta\ln(\eta) + \ln(\Gamma(1-\theta\beta,\eta))]) \tag{9}$$

satisfies the condition (4) with violation probability ε_p. Here, $\theta > 0$ is a free parameter that can be optimized and $\Gamma(\cdot,\cdot)$ denotes the incomplete gamma function. The derivation of (9) is given as

$$\begin{aligned}\mathrm{P}[S(\tau,t) < \mathcal{S}(t-\tau)] &\leq e^{\theta\mathcal{S}(t-\tau)}\mathrm{M}_S(-\theta,t-\tau)\\ &= e^{\theta\mathcal{S}(t-\tau)}(\mathrm{M}_{c_i}(-\theta))^{t-\tau} := \varepsilon_p.\end{aligned} \tag{10}$$

In the first step, we used Chernoff's lower bound with the Laplace transform $\mathrm{M}_S(-\theta, t-\tau)$ of $S(\tau, t)$ for $\theta \geq 0$. In the second step, we used the iid property of the increments of $S(\tau, t)$ and equated the expression with ε_p. We obtain $\mathcal{S}(t)$ by inserting the Laplace transform of one increment $\mathrm{M}_{c_i}(-\theta)$ [29] which is evaluated as

$$\begin{aligned}\mathrm{M}_{c_i}(-\theta) &= \int_0^\infty e^{-\theta\beta\ln(1+t)}\eta e^{-\eta t}dt\\ &= e^{\eta}\eta^{\theta\beta}\Gamma(1-\theta\beta,\eta),\end{aligned} \tag{11}$$

where the result follows by transformation of the integration variable. Solving for $\mathcal{S}(t)$ in (10) yields the function in (9). In the next section we will use the channel and queueing model to implement two queue-aware scheduling schemes.

5. Implementation of queue-aware scheduling

In this section we will show two implementations of queue-aware scheduling that we denote frequent and infrequent adaptation, respectively. We will draw conclusions on the requirements and operation of such adaptive systems. In addition, we will present a study of deploying the adaptive system in a multi-user cellular network showing performance results for different types of schedulers.

5.1. Frequent adaptation

First, we present queue-aware scheduling with frequent adaptation. Here, we assume a small epoch length Δ in the sense of $\lambda\Delta \ll b_{\max}$, where $b_{\max}$ is the desired backlog bound at the end

of the scheduling epoch that may be exceeded at most with probability ε. The system is the Rayleigh fading channel as described in Section 4.

Generally, the backlog of a queueing system at time t is defined as $B(t) = A(t) - D(t)$. In the following, we will denote the start of the current epoch by τ if not stated otherwise. Given the epoch $(\tau, \tau + \Delta]$ with arrivals $A(\tau, \tau + \Delta)$ and initial backlog $B(\tau)$ we calculate the required resources, e.g., bandwidth, such that $\mathrm{P}[B(\tau + \Delta) \le b_{\max}] \ge 1 - \varepsilon$ is attained. Taking the definition of backlog, it follows directly for systems offering strict service $D(t) \ge D(\tau) + S(\tau, t)$ for $t \ge \tau \ge 0$ where τ, t fall into the same busy period that

$$B(t) \le B(\tau) + A(\tau, t) - S(\tau, t),$$

i.e., a basic relation of the backlog at time τ and at time t given the arrivals $A(\tau, t)$ and the service $S(\tau, t)$. Let τ be the beginning of the epoch and $t = \tau + \Delta$ be the end of it. We substitute $\mathcal{S}(\Delta)$ from (4) for $S(\tau, \tau + \Delta)$ to find

$$\mathrm{P}[B(\tau + \Delta) \le B(\tau) + A(\tau, \tau + \Delta) - \mathcal{S}(\Delta)] \ge 1 - \varepsilon,$$

with ε being the violation probability from (4). Given the non-trivial case $B(\tau) + A(\tau, \tau + \Delta) > b_{\max}$, we equate $B(\tau) + A(\tau, \tau + \Delta) - \mathcal{S}(\Delta)$ with $b_{\max}$ and solve for

$$\mathcal{S}(\Delta) = B(\tau) + A(\tau, \tau + \Delta) - b_{\max} \tag{12}$$

that is the required service to ensure $b_{\max}$ with violation probability ε. Finally, we substitute the service characterization of the Rayleigh fading channel (9) for $\mathcal{S}(\Delta)$ in (12) to compute the required resource allocation β given average SNR $1/\eta$.

A refinement of the implementation above is to include an additional statistical delay constraint. As a secondary effect, such a delay constraint ensures that small backlogs which may not endanger the backlog bound $b_{\max}$ will eventually be cleared. Overall the system allocates the service to fulfill both of the following conditions:

§1 The backlog at the end of the epoch is statistically bounded by $b_{\max}$, i.e., $\mathrm{P}[B(\tau + \Delta) \le b_{\max}] \ge 1 - \varepsilon$.
§2 The backlog at the beginning of the scheduling epoch is cleared within a given delay bound $d = \nu\Delta$ with $\nu \ge 1$.

In a practical implementation of a cellular uplink transmission the base station possesses the required information, i.e., BSR and the received data amounts $D(t)$, to implement the above rules for uplink resource allocation. Using the backlog definition, the base station is able to infer arrivals within any epoch $A(\tau, \tau + \Delta)$ for all epoch starts τ using $B(\tau), B(\tau + \Delta)$ together with $D(\tau, \tau + \Delta)$ to enforce §2.

As the base station cannot know the exact arrivals a priori, we make use of the condition $\lambda\Delta \ll b_{\max}$ that permits neglecting the arrivals $A(\tau, \tau + \Delta)$ in (12) such that we can approximate the required service during Δ. In this case, the obtained bound for $B(\tau + \Delta)$ would comprise an error of roughly the ε-quantile A^{ε} of $A(\tau, \tau + \Delta)$. We denote this queue-aware scheduling without knowledge of the arrivals as "blind adaptation." It shows how the lack of arrival information impacts the system performance. Given information on $A(\tau, \tau + \Delta)$, e.g., a bound on its distribution, or given A^{ε}, the adaptive system can compute a more precise estimate of the service required in the next scheduling epoch. Similar considerations are made in Section 5.2.

Next, we consider an implementation of our frequent adaptation scheme (§1 and §2) in a baseline scenario of an LTE cellular system with 10 MHz channel bandwidth comprising 50 available resource blocks each of 180 kHz width and $\delta = 0.5$ ms length [1,2]. We use the Rayleigh wireless channel model with average SNR of $1/\eta = 3$ dB. The base station receives BSRs $B(n\Delta)$ with $n \in \mathbb{N}$ and adapts β which is the bandwidth (amount of resource blocks) granted to the UE for the upcoming scheduling epoch $n + 1$. Using CQI, the base station has channel state information that permits estimating the SNR.

For a numerical evaluation, we consider $\Delta = 10$ slots and memoryless arrivals. We normalized the system parameters such that $\mathrm{E}[c_i] = 1.33$ with $\beta = 1$ and average arrival rate of $\lambda = 0.65$. The backlog bound is $b_{\max} = 50$ and the delay bound for the combined algorithm is $d = 5\Delta$, both with violation probability $\varepsilon = 10^{-2}$. Fig. 5 shows backlog and delay CCDFs with a sole backlog constraint §1 compared to the backlog and delay constraint combination §1 and §2. The simulation length is 10^5 slots. Observe that due to "blind adaptation" the CCDF for the system using only §1 deviates at $\varepsilon = 10^{-2}$ by roughly the ε-quantile of $A(\tau, \tau + \Delta)$, i.e., $A^{\varepsilon} = 25$. Fig. 5(b) shows larger delays if only using §1 compared to the combination of §1 and §2. Adding a delay constraint substantially improves the performance. The additional constraint leads to increased resource grants as the base station complies with the tighter condition of §1 and §2.

Fig. 5 (c) shows the resource savings of queue-aware scheduling given a fixed QoS constraint, i.e., $b_{\max}$ with $\varepsilon = 10^{-2}$. First, we run a static version of the queue-aware scheduling system to find the amount of *fixed* resource blocks β that attains the QoS constraint, i.e., $b_{\max}$ at ε (dashed line). We compare the static system to the adaptive one given the same QoS constraint, i.e., $b_{\max}$ and ε. We plot the average amount of resource blocks granted (average β) for different $b_{\max}$ (solid line). The adaptive system is efficient as it provides substantial resource savings (high utilizations) for a wide range of QoS constraints.

5.2. Infrequent adaptation

In this section we regard large scheduling epochs Δ, in the sense that $\lambda\Delta$ is in the order of $b_{\max}$. Here, the amount of arrivals during the epoch Δ is non-negligible. Hence, we use bounds on the arrivals together with (6) to obtain a probabilistic bound on the backlog at the end of the epoch.

First, we use the formulation (5) together with an ε-effective service curve (6) that is violated with probability ε_s and some algebraic manipulations to express the backlog at the end of a scheduling epoch $B(\tau + \Delta)$ given the backlog at the beginning of the scheduling epoch $B(\tau)$ as

$$\mathrm{P}\Big[B(\tau + \Delta) \le \max\Big[B(\tau) + A(\tau, \tau + \Delta) - \mathcal{S}(\Delta), \sup_{u \in [\tau, \tau + \Delta]} \{A(u, \tau + \Delta) - \mathcal{S}(\tau + \Delta - u)\}\Big]\Big] \ge 1 - \varepsilon_s. \tag{13}$$

Equipped with (13) we implement a queue-aware scheduling that regulates $\mathcal{S}(t)$ to ensure that

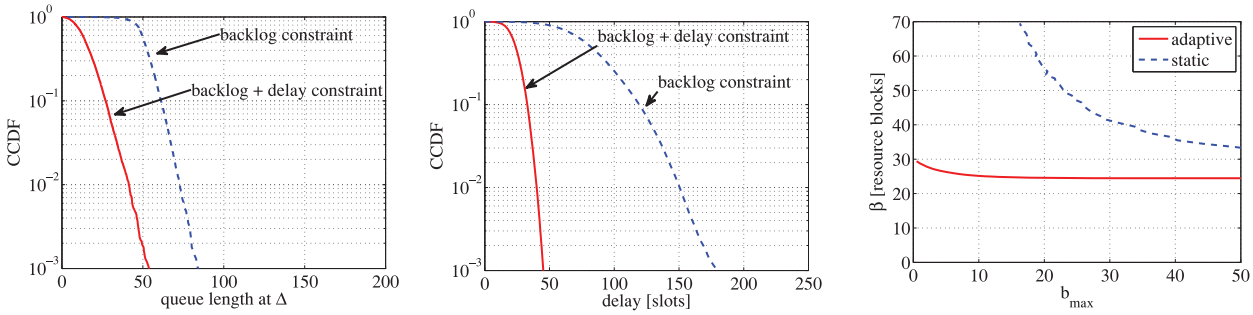
$$\mathrm{P}[B(\tau + \Delta) \le b_{\max}] \ge 1 - \varepsilon_s. \tag{14}$$

Equation (13) establishes the following requirements on $\mathcal{S}(t)$ for the scheduling epoch:

$$\mathcal{S}(\Delta) \ge B(\tau) + A(\tau, \tau + \Delta) - b_{\max}, \quad \text{and} \tag{15}$$

$$\mathcal{S}(\tau + \Delta - u) \ge A(u, \tau + \Delta) - b_{\max}, \quad \forall u \in [\tau, \tau + \Delta]. \tag{16}$$

The adaptive system takes the following input: (i) the queue size at the beginning of the scheduling epoch $B(\tau)$, (ii) the target backlog bound $b_{\max}$ with violation probability ε_s, and (iii) the arrivals between τ and $\tau + \Delta$. In the cellular scenario $B(\tau)$ is available through BSRs. The arrivals of the upcoming epoch are, however, not known a priori. Since in case of infrequent adaptation the arrivals cannot be neglected, we use upper envelope functions $E(t)$ as an estimate. These arrival envelopes can be either deterministic



(a) Backlog (b) Delay (c) Resource requirement

Fig. 5. Performance of frequent adaptation: (a) and (b) The adaptive system uses either a backlog constraint or a combined backlog and delay constraint to grant resources to the transmitter in every scheduling epoch. (c) Amount of resource blocks β required to retain a given backlog bound $b_{\max}$ and $\varepsilon = 10^{-2}$. The static system has a fixed β. For the adaptive system we plot the average β. The adaptive system saves resources by running at a high utilization.

[27], i.e., $A(u, \tau+\Delta) \le E(\tau+\Delta-u)$ for all $u \in [\tau, \tau+\Delta]$ or probabilistic of the form [21,22,30]

$$\mathsf{P}\left[\sup_{u\in[\tau,\tau+\Delta]} \{A(u,\tau+\Delta) - E(\tau+\Delta-u)\} > 0\right] \le \varepsilon \tag{17}$$

with a violation probability ε. Arrival envelopes can be constructed for a wide range of traffic models [21,24], they can be computed from traffic traces, or they can be enforced, e.g., by a traffic shaper. We substitute the arrivals in (15), and (16) by the envelope $E(t)$ to obtain valid requirements on $\mathcal{S}$. In case we use a probabilistic bound on the arrivals as in (17) we can upper bound the violation probability in (13) by the sum of ε_s of the ε-effective service curve (6) and ε of (17).

In the following, we show the calculation for an LTE cellular system assuming a Rayleigh wireless channel model as given in Section 4. The formulation (9), which satisfies (4), has two parameters, the average SNR $1/\eta$ and the granted bandwidth β. We consider an adaptive system that manipulates the bandwidth grants β to retain the backlog bound (14). A direct extension based on adaptive power regulation through η is also possible. For the evaluation, we assume leaky bucket constrained arrivals with known envelope $E(t) = \sigma + \varrho t$. In the sequel, we will use the envelope $E(t)$ instead of the actual arrivals $A(\tau, \tau+t)$ for the requirements (15) and (16). Given $b_{\max}$ and the epoch Δ we fix $\mathcal{S}(t) = \varrho[t-\zeta]_+$ as a latency-rate function with latency term $\zeta = \frac{b_{\max}-\sigma}{\varrho}$, where $[x]_+$ denotes $\max\{x, 0\}$. Note that $b_{\max} > \sigma$. The latency-rate shape of $\mathcal{S}(t)$ is chosen in congruence with the shape of $E(t)$ such that the vertical deviation between both is constant and equal to $b_{\max}$. The rationale is that we are looking for the minimum resource allocation that retains the specified QoS bound. First, we consider the condition on $\mathcal{S}(\Delta)$ that follows from (15). We use an envelope formulation $E(t) = B(\tau) + \sigma + \varrho t = \sigma'' + \varrho t$ set $t = \tau + \Delta$ and calculate $\mathsf{P}[S(\tau, \tau+\Delta) < \mathcal{S}(\Delta)] \le \varepsilon''_p$ similar to (10). Then, we insert the latency rate function $\mathcal{S}(\Delta) = \varrho[t-\zeta'']_+$ with $\zeta'' = \frac{b_{\max}-\sigma''}{\varrho}$ and the Laplace transform $\mathsf{M}_{c_i}(-\theta)$ from Section 4.2 to relate β to ε''_p.

Now, we turn to condition (16) and calculate the violation probability in (6) given channel resources β. In [4] we provided a calculation that was based on a classical combination of Boole's inequality and Chernoff's bound as illustrated in (7). The following bounding technique is more adapted to the considered wireless channel model while providing better bounding accuracy. The following theorem provides a solution to (6) for the introduced Rayleigh wireless channel model. The proof is given in Appendix and it goes along a technique known from [31,32].

Theorem 1 (Bound on the finite interval Rayleigh wireless channels service)**.** *Given a Rayleigh fading channel that is described by a service process $S(\tau, t) = \sum_{i=\tau+1}^{t} c_i$ with iid increments c_i from (8) and parameters β and η. A bound on the service provided by the channel for a finite interval Δ is given as*

$$\mathsf{P}[\exists u \in [\tau, \tau+\Delta] : S(u, \tau+\Delta) < \mathcal{S}(\tau+\Delta-u)] \le e^{-\theta\kappa}, \tag{18}$$

with $\mathcal{S}(t) = [\varrho t - \kappa]_+$, the free parameters $\varrho, \kappa > 0$ and with θ equal to the unique positive solution of

$$\mathsf{M}_{c_1}(-\theta)e^{\theta\varrho} = 1. \tag{19}$$

The stability condition

$$\varrho < \mathsf{E}[c_1] = \beta e^{\eta}\Gamma(0, \eta) \tag{20}$$

guarantees the existence of a unique solution to (19).

Now, given $b_{\max}$ and Δ, we substitute κ in Theorem 1 by $b_{\max} - \sigma$, to obtain a bound that relates to the requirement (16). Recall the definition of $E(t)$ as $\sigma + \varrho t$. Hence, the events $\{\varrho k - \sum_{i=1}^{k} c_i > (b_{\max} - \sigma)\}$ and $\{E(k) - \sum_{i=1}^{k} c_i > b_{\max}\}$ are identical. The relation between β and the violation probability ε'_s in (18) is established through θ that is the solution of (19). An expanded version of this relation is given in (26) in Appendix. Further, we choose β to ensure that the stability condition (20) holds for the Rayleigh wireless channel. Finally, we obtain a bound on the violation probability ε_s of the backlog bound in (14) as the sum of ε''_p and ε'_s from above, i.e., using the combination of the requirements (15) and (16).

In the following we present simulation results for queue-aware scheduling with infrequent adaptation in an LTE scenario as depicted in Fig. 1. The baseline scenario remains unchanged with respect to Section 5.1 except for $\Delta = 100$ slots and $b_{\max} = 65$, i.e., $\lambda\Delta = b_{\max}$ where $\lambda = 0.65$ as before and the violation probability $\varepsilon_s = 10^{-2}$. An arrival envelope with parameters $\sigma = 10$ and $\varrho = 0.66$ is enforced on Poisson traffic with mean rate λ. We apply a numerical binary search to find β that satisfies (18) for a given ε_s.

Fig. 6 (a) shows the adaptive system successfully providing the configured probabilistic bound on the backlog at the end of the scheduling epoch. The figure also compares the bound provided by the calculation in [4] and the martingale bounding technique in Section 5.2. The performance of queue-aware scheduling using the new bounds is closer to the target criterion of a queue length of 65 at $\varepsilon = 10^{-2}$ and hence enables saving resources compared to [4]. It is worth noting that the difference between both calculations grows with longer periods Δ. In Fig. 6(b) we observe lower delays using the technique from [4] as it allocates slightly more resources than the new bounding methodology from above. We also observe in Fig. 6(b) a base level of delays after which the CCDF shows a sharp bend. The intuition behind this is that the adaptation algorithm saves resources by leaving a residual amount of backlog not

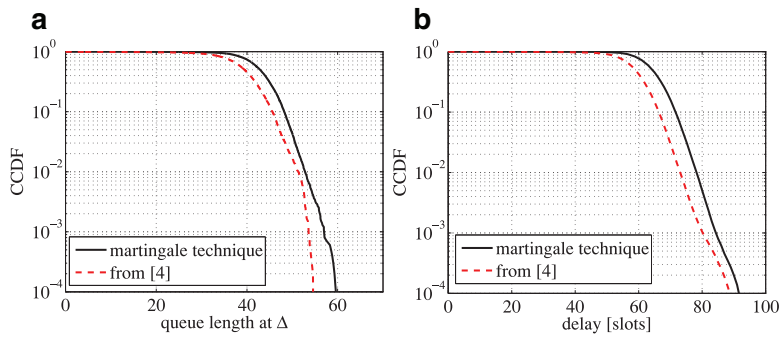


Fig. 6. Infrequent adaptation: The system retains the probabilistic backlog bound. (a) The martingale technique provides a more accurate performance bound, i.e., queue length of 65 at $\varepsilon = 10^{-2}$. The technique in [4] allocates slightly more resources per scheduling period leading to some degree to shorter queues in (a) and smaller delays in (b).

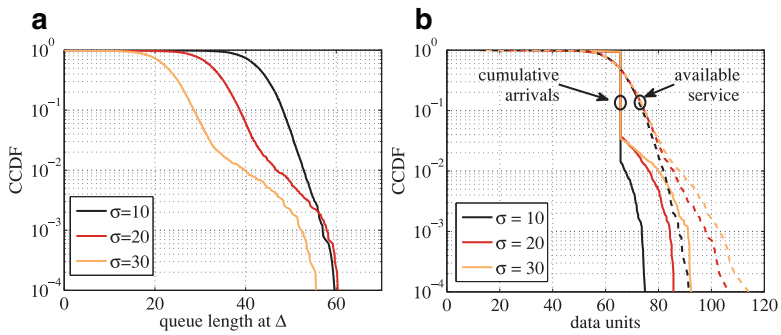


Fig. 7. Impact of the burstiness constraint σ. Bursts occur only rarely, as can be seen from (b) that shows the CCDF of the amount of data arrivals in each scheduling epoch (solid lines). The adaptive system provisions resources for bursts of up to σ. Hence, most of the time it maintains a smaller queue size if σ is increased, see the staggered CCDFs of the queue length in (a). The effect of actual burst arrivals is visible in the tail of the CCDFs in (a). (b) also shows the amount of service that is available in each scheduling period (dashed lines). Since for larger σ the system mostly maintains a correspondingly smaller queue size, the resources that are required for the next scheduling epoch are similar for different σ and differ only in the tail, if the queue size increased due to an actual burst arrival.

cleared if it does not threaten to violate the QoS constraint. Observe that in Fig. 6(b) the delay variation (jitter) is small with respect to the base level of delays. In Fig. 7(a) we fix the Poisson arrival traffic and the previously noted leaky bucket rate ρ while varying its maximum burst size σ. The adaptive system allocates comparable resources in the majority of the scheduling periods as larger bursts arrive rarely. This is shown in Fig. 7(b) where we depict the CCDF of the cumulative arrivals and the amount of service that is available per scheduling period.

The decision whether to use infrequent or frequent adaptation strongly depends on the length of the scheduling epoch Δ and the relation of $b_{\max}$ to the amount of traffic which is expected in Δ. Given base stations that do not have any information on the arrivals at the UE (except for the average rate) the choice would be frequent adaptation. Given more information on the arrivals, e.g., a probabilistic/deterministic bound for the time span Δ, the base station can deploy the more refined algorithm of infrequent adaptation over longer scheduling epochs, which reduces signaling and can save computational resources at the base station.

5.3. Multi-user scheduling

We conclude this section with a concise evaluation of queue-aware scheduling for a system serving multiple users with overall resource constraints. We utilize the infrequent adaptation algorithm with unchanged parameters as above. For ease of exposition, we consider M homogeneous and statistically independent UEs in a cell, each signaling BSRs to the base station. The heterogeneous case follows at the expense of additional notation. The base station deploys the infrequent adaptation system to provide each UE with resource blocks for every scheduling epoch. For a given epoch, the amount of resource blocks that are required for user j according to the adaptive system is denoted β_j. For convenience, we drop the epoch index in the following notation. Naturally, the base station is constrained by the overall amount of resource blocks β_s that are available at each epoch. Hence, the resource amount received by a user j, i.e., $\hat{\beta}_j$, depends on the required amount β_j and a scheduling policy implemented at the base station to distribute the available amount of resource blocks β_s. In general, the overall amount of resources received by the users is constrained by β_s, i.e.,

$$\sum_{j=1}^{M} \hat{\beta}_j \le \beta_s,$$

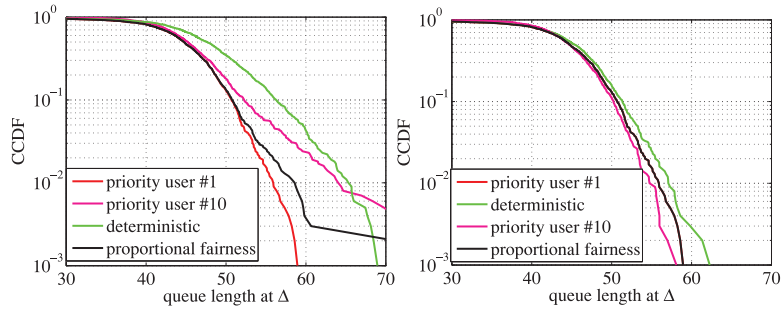
while the resource amount the individual users receive is constrained by β_j, i.e.,

$$\hat{\beta}_j \le \beta_j \quad \text{for } j \in \{1, \dots, M\}.$$

In the sequel, we consider three different scheduling algorithms that run on top of the queue-aware scheduling. The rationale behind this is to additionally provide a mechanism for service differentiation. We consider the following notions of scheduling: (i) deterministic (FDMA), (ii) priority, and (iii) proportional fair scheduling.

In the first case, the deterministic scheduler divides the available resources β_s evenly over M users while taking into account their respective resource requirements β_j that are provided by the adaptive system. The resource share of user j of the available resources β_s is hence

$$\hat{\beta}_j = \min\{\beta_j, \beta_s/M\}, \tag{21}$$



(a) 95% utilization (b) 90% utilization

Fig. 8. Multi-user scenario: Backlogs in the adaptive system under different scheduling algorithms. Notable difference only at very high utilizations.

i.e., the minimum of the amount required by the adaptive system and the even share of β_s. This rule is simply implementable into the base station logic.

In the second case, we consider a priority scheduler with M ordered priority classes. For notational simplicity, we consider here a one-to-one mapping of M users to M priority classes. The extension of multiple users per class is, however, straightforward. Under priority scheduling, the user in class j receives

$$\hat{\beta}_j = \min\left\{\beta_j, \beta_s - \sum_{k=1}^{j-1}\hat{\beta}_k\right\}, \tag{22}$$

where we adopt the convention $\sum_{k=1}^{0}\hat{\beta}_k = 0$. Here, the higher the priority of a user, i.e., smaller j, the more resources may it receive of the overall amount β_s given that the higher priority user requirements are satisfied.

The last case concerns what we denote as proportional fair scheduling where we basically employed the resource distribution of the priority scheduling given in (22), however, we reorder the priority list at the beginning of every scheduling epoch based on a proportionality score that is calculated for each user j similar to the definition in [33]. Here, the score of the jth user reflects the proportion of the expected service during the upcoming epoch in relation to the user average send rate over the last epochs. In particular, the score of the jth user is given as $S_j(\tau, \tau+\Delta)/(D_j(\tau)/\tau)$, i.e., the amount of service that user j expects in the scheduling epoch divided by the average transmission rate of the user up to the beginning of the scheduling epoch τ.

We consider a simulation of the multiuser system with the following parameters: $\sigma = 10$, $b_{\max} = 65$, $\varepsilon = 0.05$, $\Delta = 100$ slots, $M = 10$ users, and a simulation length of 10^5 slots. Fig. 8 shows the performance of the multi-user system under different utilizations. For the priority scheduler the CCDF of the first priority user remains unchanged through all considered utilizations. The priority scheduler may starve low priority classes to provide high priority classes with enough resources to attain the QoS constraint. In case of proportional fair scheduling and deterministic scheduling the resources are "fairly" distributed such that either none or all UEs are provided with the QoS constraint. The backlog CCDFs for all UEs are identical such that we display only one for the proportional fair case and one for the deterministic case. An interesting observation is that the CCDF of the backlog for a single user scenario without the overall resource constraint β_s matches the CCDF of the priority user #1 in Fig. 8. The adaptive system shows strong performance providing the QoS constraint to all UEs. The deterministic scheduler behaves similarly to the single user case in Section 5.2 as the users' uplinks can be regarded as parallel independent systems. Yet, it relates through the condition on $\hat{\beta}_j$ to the study of fixed service rate constraints in Section 3. For the priority scheduler the system benefits from statistical multiplexing effects when distributing the overall available resources β_s. In case of proportional fair scheduling β_s is evenly distributed such that all UEs are provided with comparable QoS level.

6. Conclusions

In this work, we presented an adaptive resource allocation scheme that provides probabilistic quality of service guarantees based on transmit buffer occupation. Adaptive resource allocation enables the optimization of the resource utilization in communication networks under dynamic conditions. First, we used exact formulae for the class of Poisson traffic to show substantial resource savings under certain conditions compared to static resource allocations. We also showed the robustness of the adaptive system with respect to misadaptation and resource limitation. Motivated by the exact results we provided a general framework for implementing queue-aware scheduling that takes as input general traffic arrival and service processes. We considered a wireless channel model and described two algorithms for adaptive resource allocation. Using the example of a cellular network we presented simulation results that show the performance gain with queue-aware scheduling. The adaptive system saves resources, while retaining a given QoS level. We showed a brief example of the performance of the adaptive system in multi-user scenarios together with insight and recommendations for the operation of queue-aware scheduling.

Acknowledgments

The research leading to these results has received funding from the European Research Council under an ERC Starting Grant "UnIQue". The work by A.R. has been supported in parts by the DAAD Postdoc program.

Appendix

Proof of Theorem 1. We assume a common filtered probability space $\left(\Omega, \mathcal{F}, (\mathcal{F}_k)_k, \mathrm{P}\right)$ and that the service increment process $(c_k)_k$ is adapted. Next, we insert the expressions for $S(u, \tau+\Delta)$ and for $\mathcal{S}(\tau+\Delta-u)$ from the theorem into (18). Since $S(u, \tau+\Delta) \geq 0$ for all u, τ, Δ we rewrite (18) as

$$\mathrm{P}\left[\max_{0\leq k\leq\Delta}\left\{\varrho k - \sum_{i=1}^{k} c_i\right\} > \kappa\right] \leq \varepsilon_s', \tag{23}$$

where we made use of the reversibility property of the increment process and used a variable substitution for the time index. We adopt the convention $\sum_{i=1}^{0} c_i = 0$.

First, we prove that the process $e^{\theta(k\varrho-\sum_{i=1}^{k}c_i)}$ is a martingale with respect to $\mathcal{F}_k$ which is the filtration corresponding to the history of the process $(c_k)_k$. This is directly obtained given the iid property of the increment process, i.e.,

$$\begin{aligned}\mathsf{E}[e^{\theta(k\varrho-\sum_{i=1}^{k}c_i)}|\mathcal{F}_{k-1}] &= \mathsf{E}[e^{\theta(\varrho-c_k)}]e^{\theta((k-1)\varrho-\sum_{i=1}^{k-1}c_i)}\\ &= e^{\theta((k-1)\varrho-\sum_{i=1}^{k-1}c_i)},\end{aligned} \tag{24}$$

under the condition (19) on θ. The existence and uniqueness of the solution of (19) given the stability condition (20) is proven through the monotonicity of the Laplace transform $\mathsf{M}_{c_1}(-\theta)$ and the function $e^{-\theta\varrho}$ similar to a technique in [32].

In the sequel, we use the following stopping time T for the process $e^{\theta(k\varrho-\sum_{i=1}^{k}c_i)}$ to derive a bound as in (6):

$$T := \min\left\{\min\left\{k\geq 0 : k\varrho-\sum_{i=1}^{k}c_i>\kappa\right\},\Delta\right\},$$

where $\min\{k\geq 0 : k\varrho-\sum_{i=1}^{k}c_i)>\kappa\}$ is also the first point in time where the event $\{e^{\theta(k\varrho-\sum_{i=1}^{k}c_i)}>e^{\theta\kappa}\}$ occurs. Using T we invoke Doob's maximal inequality [34] to find the upper bound ε_s' in (23) as

$$\begin{aligned}\mathsf{P}\left[\max_{0\leq k\leq\Delta}e^{\theta(k\varrho-\sum_{i=1}^{k}c_i)}>e^{\theta\kappa}\right] &\leq \mathsf{E}\left[e^{\theta(\varrho-c_1)}\right]e^{-\theta\kappa}\\ &= e^{-\theta\kappa} := \varepsilon_s',\end{aligned} \tag{25}$$

where we used the condition (19) in the second line. Now, we can determine θ in (25) after inserting ϱ and the Laplace transform of c_1 from (11) into the condition (19) as the solution of

$$\eta+\theta(\beta\ln(\eta)+\varrho)+\ln(\Gamma(1-\theta\beta,\eta))=0. \tag{26}$$

The stability condition (20) ensures the existence of the unique solution to (19). We calculate $\mathsf{E}[c_1]$ as follows

$$\begin{aligned}\mathsf{E}[c_1] &= \beta\eta\int_0^\infty \ln(1+x)e^{-\eta x}dx\\ &= \beta\eta\left[\left[-\frac{e^{-\eta x}}{\eta}\ln(1+x)\right]_0^\infty+\int_0^\infty\frac{e^{-\eta x}}{(1+x)\eta}dx\right]\\ &= \beta e^{\eta}\Gamma(0,\eta).\end{aligned}$$

In the first step, we did an integration by parts. In the second step, we evaluate the first term and use a variable transformation in the second term to obtain the incomplete gamma function. □

References

[1] C. Cox, An Introduction to LTE: LTE, LTE-Advanced, SAE and 4G Mobile Communications, Wiley, 2012.
[2] 3GPP specification TS 36.321, Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification, 2012, Release 8, version 8.12.
[3] J. Huang, Z. Niu, Buffer-aware and traffic-dependent packet scheduling in wireless OFDM networks, in: Proceedings of IEEE Wireless Communications and Networking Conference (WCNC), IEEE, 2007, pp. 1554–1558.
[4] A. Rizk, M. Fidler, Queue-aware uplink scheduling: Analysis, implementation, and evaluation, in: Proceedings of IFIP Networking Conference, IEEE, 2015, pp. 1–9.
[5] J.M. George, J.M. Harrison, Dynamic control of a queue with adjustable service rate, Oper. Res. 49 (5) (2001) 720–731.
[6] K. Adusumilli, J. Hasenbein, Dynamic admission and service rate control of a queue, Queueing Syst. 66 (2) (2010) 131–154.
[7] A. Wierman, L. Andrew, A. Tang, Stochastic analysis of power-aware scheduling, in: Proceedings of the 46th Annual Allerton Conference on Communication, Control, and Computing, IEEE, 2008, pp. 1278–1283.
[8] L.L. Andrew, M. Lin, A. Wierman, Optimality, fairness, and robustness in speed scaling designs, SIGMETRICS Perform. Eval. Rev. 38 (1) (2010) 37–48.
[9] C.-p. Li, M.J. Neely, Delay and rate-optimal control in a multi-class priority queue with adjustable service rates, in: Proceedings of IEEE INFOCOM, IEEE, 2012, pp. 2976–2980.
[10] S.R. Mahabhashyam, N. Gautam, On queues with Markov modulated service rates, Queueing Syst. Theor. Appl. 51 (1-2) (2005) 89–113.
[11] A. Bianco, M. Casu, P. Giaccone, M. Ricca, Joint delay and power control in single-server queueing systems, in: Proceedings of IEEE Online Conference on Green Communications (GreenCom), IEEE, 2013, pp. 50–55.
[12] I. Bettesh, S. Shamai, Optimal power and rate control for minimal average delay: The single-user case, IEEE Trans. Inform. Theor. 52 (9) (2006) 4115–4141.
[13] E. Yeh, A. Cohen, Throughput and delay optimal resource allocation in multiaccess fading channels, in: Proceedings of IEEE International Symposium on Information Theory, IEEE, 2003, p. 245.
[14] B.E. Collins, R.L. Cruz, Transmission policies for time varying channels with average delay constraints, in: Proceedings of Allerton Conference on Communication, Control, and Computing, 1999, pp. 709–717.
[15] R.A. Berry, R.G. Gallager, Communication over fading channels with delay constraints, IEEE Trans. Inform. Theor. 48 (5) (2002) 1135–1149.
[16] D. Niyato, E. Hossain, Queue-aware uplink bandwidth allocation and rate control for polling service in IEEE 802.16 broadband wireless networks, IEEE Trans. Mobile Comput. 5 (6) (2006) 668–679.
[17] M. Mehta, S. Khakurel, A. Karandikar, Buffer-based channel dependent UpLink scheduling in relay-assisted LTE networks, in: Proceedings of IEEE WCNC, IEEE, 2012, pp. 1777–1781.
[18] L. Kleinrock, Queueing Systems, Theory, Vol. 1, Wiley & Sons, 1975.
[19] J. Abate, W. Whitt, Calculating time-dependent performance measures for the M/M/1 queue, IEEE Trans. Commun. 37 (10) (1989) 1102–1104.
[20] H. Al-Zubaidy, J. Liebeherr, A. Burchard, Network-layer performance analysis of multihop fading channels, IEEE/ACM Trans. Netw. 24 (1) (2016) 204–217.
[21] F. Ciucu, A. Burchard, J. Liebeherr, Scaling properties of statistical end–to-end bounds in the network calculus, IEEE/ACM Trans. Netw. 14 (6) (2006) 2300–2312.
[22] R.L. Cruz, Quality of service management in integrated services networks, in: Proceedings of Semi-Annual Research Review, Center of Wireless Communication, UCSD, 1996, pp. 1–9.
[23] A. Burchard, J. Liebeherr, S. Patek, A min-plus calculus for end-to-end statistical service guarantees, IEEE Trans. Inform. Theor. 52 (9) (2006) 4105–4114.
[24] M. Fidler, A survey of deterministic and stochastic service curve models in the network calculus, IEEE Commun. Surv. Tutor. 12 (1) (2010) 59–86.
[25] M. Fidler, A. Rizk, A guide to the stochastic network calculus, IEEE Commun. Surv. Tutor. 17 (1) (2015) 92–105.
[26] C.-S. Chang, Performance Guarantees in Communication Networks, Springer-Verlag, 2000.
[27] J.-Y. Le Boudec, P. Thiran, Network Calculus a Theory of Deterministic Queuing Systems for the Internet, LNCS, vol. 2050, Springer-Verlag, 2001.
[28] R. Lübben, M. Fidler, J. Liebeherr, Stochastic bandwidth estimation in networks with random service, IEEE/ACM Trans. Netw. 22 (2) (2014) 484–497.
[29] M. Fidler, R. Lübben, N. Becker, Capacity-delay-error-boundaries: A composable model of sources and systems, IEEE Trans. Wireless Commun. 14 (3) (2015) 1280–1294.
[30] O. Yaron, M. Sidi, Performance and stability of communication networks via robust exponential bounds, IEEE/ACM Trans. Netw. 1 (3) (1993) 372–385.
[31] J.F.C. Kingman, Inequalities in the theory of queues, J. R. Stat. Soc. Ser. B. Stat. Methodol. 32 (1) (1970) 102–110.
[32] F. Poloczek, F. Ciucu, Service-martingales: Theory and applications to the delay analysis of random access protocols, in: Proceedings of IEEE Conference on Computer Communications (INFOCOM), IEEE, 2015, pp. 945–953.
[33] F.P. Kelly, A.K. Maulloo, D.K. Tan, Rate control for communication networks: Shadow prices, proportional fairness and stability, J. Oper. Res. Soc. 49 (3) (1998) 237–252.
[34] G. Grimmett, D. Stirzaker, Probability and Random Processes, third, Oxford University Press, 2001.

Please cite this article as: A. Rizk, M. Fidler, Queue-aware uplink scheduling with stochastic guarantees, Computer Communications (2016), http://dx.doi.org/10.1016/j.comcom.2016.02.014