Computer Communications 84 (2016) 63–72

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 3 March 2016

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

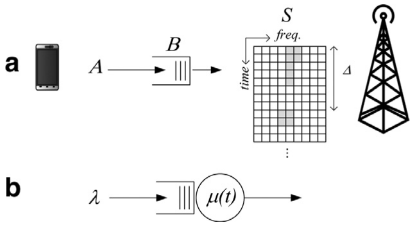

Fig. 1. (a) Example of queue-aware scheduling in cellular networks. The base station decides on the amount of uplink service S depicted as a varying number of resource blocks (gray) granted to a UE depending on its transmit buffer filling B. The scheduling epoch is denoted Δ. (b) Abstraction of queue-aware 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