Multi-interface Communication: Interface Selection under Statistical Performance Constraints

Sounak Kar
Multimedia Communications Lab
Technische Universität Darmstadt
sounak.kar@kom.tu-darmstadt.de

Amr Rizk
Multimedia Communications Lab
Technische Universität Darmstadt
amr.rizk@kom.tu-darmstadt.de

Markus Fidler
Institute of Communications Technology
Leibniz Universität Hannover
markus.fidler@ikt.uni-hannover.de

***Abstract*—Recent advancements of multipath communications have made a new dimension of resource management accessible focusing on balanced and efficient utilization of available communication alternatives. Dynamic strategies in this context leverage the most up to date information on available interfaces to outperform comparable static versions. In this work, we propose an adaptive strategy that seeks to improve the worst-case performance of a multipath communication system by periodically looking at waiting time bounds of respective subsystems. We compare the performance of the proposed adaptive assignment with established round-robin and join-the-shortest-queue strategies numerically where we observe that our strategy tends to be superior in heterogeneous environments.**

I. Introduction

Following the recent evolution of multipath communications, modern mobile devices have been endowed with a new degree of freedom, i.e., the ability to utilize combinations of multiple network interfaces for communications. This ability, which has been present in data center networks among others, now carries over to mobile systems that are able to leverage different wireless technologies such as LTE, WiFi, Bluetooth, and mmWave. Each wireless technology is, however, exposed to strongly varying and fundamentally different channel conditions and features significant differences in the PHY and MAC layer implementations. Hence, we seek an adaptive system that leverages these different communication technologies to benefit from the diversity of these technologies that provide statistically independent service.

A system that takes this approach is a transition-based adaptive system (TBS) that switches the actively used communication technology during run time, as depicted in Fig. 1. On a flow level, we consider each of the depicted subsystems to provide a time-varying service S_i to the flow arrivals A that are to be transmitted. Given an optimization metric for the communication quality, e.g., the delay, the adaptive system continuously chooses the most appropriate subsystem. As a consequence, each subsystem receives a thinned stream from the common system input process.

In this work, we present an analytical approach to such a transition-based system that builds on a service curve description [5] of different subsystems while taking into account the queue lengths at the different subsystems simultaneously. More precisely, the TBS takes a decision in the beginning of every epoch of length Δ based on delay bounds that are calculated for the different subsystems $i \in \{1, \ldots, N\}$ to determine the most suitable interface. In case the properties of the service processes are not known beforehand, the TBS takes into account inference results of the service it has witnessed on different subsystems so far.

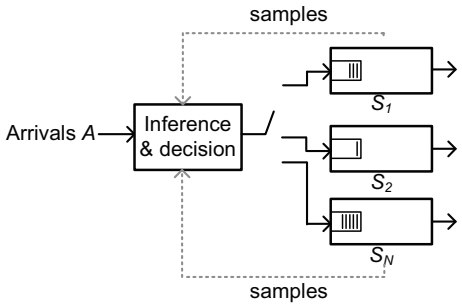


Fig. 1: Transition-based system: Subsystems provide different service $S_1, \ldots, S_N$ and the system decides at the beginning of every epoch which subsystem to transmit the arrivals A on.

The contributions of this work are twofold: First, we provide an envelope for general discrete time Markov modulated service processes. Secondly, we provide an analytical description of the adaptive assignment strategy showing a formal criterion to choose a specific subsystem i.

Our results particularly show that for heterogeneous multi-interfaces, the adaptive assignment strategy provides better worst case performance, i.e., in terms of the maximum waiting time across the subsystems. We note, however, that the performance differences between the assignment strategies in terms of the maximum and average waiting times diminish with lower system utilization.

The remainder of the paper is structured as follows: In Sect. II we provide the queueing model that lays the groundwork for the transition-based system. In Sect. III we first review service models for wireless systems before introducing the transition-based system together with the inference module. We provide numerical results in Sect. IV before discussing related work on adaptive queueing systems in Sect. V and conclude by summarizing our findings in Sect. VI.

978-0-9883045-5-0/18/$31.00 ©2018 ITC
DOI 10.1109/ITC30.2018.10054

II. Queue Aware Scheduling

First, we describe the notation of the queue aware scheduling system where we assume time is discrete. Arrivals to the system between time points $\tau+1$ and t are denoted by $A(\tau,t)$ whereas the departures of the system in the same period are denoted by $D(\tau,t)$. We model the time varying service offered by the system in the period $[\tau+1,t]$ using the random process $S(\tau,t)$. If we define $A(t)=A(0,t)$ and $D(t)$ in the same vein, the backlog at time t can be expressed as $B(t)=A(t)-D(t)$. Using the notion of dynamic server [4], the departures from the system can be expressed in terms of the arrivals and the time varying service as

$$D(t) \geq \min_{\tau\in[0,t]} \{A(\tau)+S(\tau,t)\}. \tag{1}$$

This, in turn, gives us a bound on backlog of the form

$$B(t) \leq \max_{\tau\in[0,t]} \{A(\tau,t)-S(\tau,t)\}. \tag{2}$$

The stochastic analogue of the above bound introduces a notion of violation probability, i.e., the bound can be breached with probability ϵ, which is typically small. For the discrete time model, this probability can be computed making a sample path argument that makes use of union bound [5], [7], [9].

In this work, we look at a system (cf. Fig. 1) offering alternative service guarantees and devise a scheduling strategy that sways between available subsystems each defined by a service curve, with the objective of minimizing the waiting time. More specifically, at epoch beginnings $l\Delta$, with $l\in\mathbb{N}$ and epoch length Δ, the decision module of the system chooses the subsystem i among alternate subsystems with service processes $\{S_1,S_2,...,S_n\}$, such that $i=\arg\min_{k=1,2,...,n} w_k$ where w_k's are smallest reals satisfying $P[W_k>w_k]\leq\epsilon$ for a predefined service quality ϵ. Here, W_k denotes the waiting time at the end of the epoch if the k^{th} subsystem is chosen.

The assignment problem becomes more complex when the decision module does not possess a characterization of the service processes $\{S_1,S_2,...,S_n\}$. Here, it takes samples from the observed service to infer about the parameters that describe the candidate service curves as depicted in Fig. 1.

III. Transition-based System

Our construction is based on the assumption that the arrivals and service processes adhere to the envelopes similar to [9], [7], [21], [13], [5]; i.e., we consider arrivals admitting an (b_A,ρ_A) envelope of the form

$$\mathsf{P}[\exists\tau\in[0,t]: A(\tau,t)>\rho_A(t-\tau)+b_A]\leq\epsilon(b_A). \tag{3}$$

Further, we consider service processes that adhere to $(-b_i,\rho_i)$ envelopes in the sense of [9], i.e.,

$$\mathsf{P}\left[\exists\tau\in[0,t]: S(\tau,t)<\rho_i(t-\tau)-b_i\right]\leq\epsilon(b_i), \tag{4}$$

for all subsystems $i\in\{1,\ldots,N\}$. Note that the service processes envelopes above satisfy the notion of service curves. Further, the service envelope can be expressed in a tighter form as $\max\{\rho_i(t-\tau)-b_i,0\}$ as the service process is always non-negative.

A. Transition Scheme

Given that the arrival and service processes admit envelopes (b_A,ρ_A) and $(-b_i,\rho_i)$, respectively, with violation probability $\frac{\epsilon}{2}$, where $\rho_i>\rho_A,\forall i$, it holds for the delay $W(t)$ at time t:

$$P[W(t)>w]\leq\epsilon,$$

where $w=(b_A+b_i)/\rho_i$ [9].

Following this, our strategy aims to achieve shorter waiting times by making transitions to the subsystem which has the smallest waiting time bound at each decision epoch. Once we determine the envelope parameters $\{(-b_i,\rho_i)\}_{i=1}^N$ for each subsystem, together with the arrival envelope (b_A,ρ_A) and the existing backlog at each subsystem $\{B_i\}_{i=1}^N$, the waiting time bound can be written as: $w_i=(b_A+B_i+b_i)/\rho_i$. This is due to the fact that the arrival envelope can be reconstructed incorporating existing backlog of a subsystem as additional arrivals at the first time point in the epoch. To sum up, at the beginning of the epoch we choose the m^{th} subsystem as

$$m=\underset{i=1,2...,N}{\arg\min}\quad\frac{b_A+B_i+b_i}{\rho_i}. \tag{5}$$

Here we assume each backlog B_i is observable in the beginning of each epoch. In the special case of arrival and service processes being governed by independent Poisson processes with parameters λ and μ_i respectively, it can be shown that $\big(-\log\epsilon,\lambda(e^\theta-1)/\theta\big)$ and $\big(\log\epsilon,-\mu_i(e^{-\theta}-1)/\theta\big)$ are valid envelopes, where ϵ denotes the violation probability for each envelope and $\theta>0$ is a free parameter to optimize.

B. Markov Modulated Service

Following [17], we adopt the Finite State Markov Chain framework for modelling the service increments in wireless fading channels. We assume that the service increment X_j at time slot j is modulated by a Markov chain $\{Z_j\}_j$, with $Z_j\in\{0,1\}$. If the service increments are given by $(X_j|Z_j=0)\sim Y_0$ and $(X_j|Z_j=1)\sim Y_1$ with moment generating functions $\mathbf{E}[e^{\theta Y_k}]=M_k(\theta)$, $k\in\{0,1\}$; we see that

$$\begin{aligned}\mathbf{E}[e^{\theta S(\tau,\tau+t)}|\mathbf{Z}] &= \prod_{j=\tau+1}^{\tau+t}\big(M_0(\theta)\big)^{1-Z_j}\big(M_1(\theta)\big)^{Z_j}\\ &= \big(M_0(\theta)\big)^{t-\sum_{j=\tau+1}^{\tau+t}Z_j}\big(M_1(\theta)\big)^{\sum_{j=\tau+1}^{\tau+t}Z_j},\end{aligned}$$

where $\mathbf{Z}$ denotes the random vector $(Z_{\tau+1},\ldots,Z_{\tau+t})$. Thus,

$$\begin{aligned}&\mathbf{E}\big[e^{-\theta S(\tau,\tau+t)}\big]=\mathbf{E}\Big[\mathbf{E}\big[e^{-\theta S(\tau,\tau+t)}|\mathbf{Z}\big]\Big]\\ &=\mathbf{E}\big[\big(M_0(-\theta)\big)^{t-\sum_{i=\tau+1}^{\tau+t}Z_i}\big(M_1(-\theta)\big)^{\sum_{i=\tau+1}^{\tau+t}Z_i}\big]\\ &=\sum_{z\in\{0,1\}^t}\big(M_0(-\theta)\big)^{t-|z|}\big(M_1(-\theta)\big)^{|z|}p_t(z)\\ &\leq\Big(\sum_{z\in\{0,1\}^t}\big(M_0(-\theta)\big)^{lt-l|z|}\big(M_1(-\theta)\big)^{l|z|}\Big)^{1/l}\\ &\cdot\Big(\sum_{z\in\{0,1\}^t}p_t^m(z)\Big)^{1/m}\\ &\leq\Big(M_0^l(-\theta)+M_1^l(-\theta)\Big)^{t/l}\Big(\tilde{p}_t\Big)^{1/l}.\end{aligned} \tag{6}$$

Here $|z| = \sum_i z_i$, $p_t(z) = P[(Z_{\tau+1}, \ldots, Z_{\tau+t}) = z]$, and $\tilde{p}_t = \max_z P[(Z_{\tau+1}, \ldots, Z_{\tau+t}) = z]$. The derivation uses Hölder's inequality in the second last step with $1/m+1/l = 1$ and $m > 1, l > 1$. Further, the final step exploits the fact that $p_t^m(x) \leq p_t(x)(\tilde{p}_t)^{m-1}$ and substitutes m in terms of l. There are two free parameters that can be optimized, i.e., $\theta > 0$ and $l > 1$.

One can see that the bound (6) can be further generalized when the modulating variable Z has K states. Assuming respective moment generating functions $\mathbf{E}[e^{\theta X}|Z = j] = M_j(\theta)$, $j \in \{1, 2, ..., K\}$; one can deduce

$$\mathbf{E}\big[e^{-\theta S(\tau,\tau+t)}\big] \leq \left(\sum_{j=1}^{K} M_j^l(-\theta)\right)^{t/l} \left(\tilde{p}_t\right)^{1/l}, \qquad (7)$$

where θ, l and $\tilde{p}_t$ bear the same meaning as in (6).

Application to MIMO Channels: The bound in (7) can be used to derive envelopes of the form $(-b_i, \rho_i)$ as in (4) for MIMO wireless systems under spatial multiplexing. Here, the K states correspond to the available degrees of freedom due to the multiple antennas.

C. Markov Modulated On-Off Service

The case of Markov modulated On-Off service is a special case of the derivations from Sect. III-B where we have $Z_i = 0$ and $Z_i = 1$ denoting the *Off* and the *On* states, respectively. In these two states, respective constant rate service increments provided by the channel are $Y_0 = 0$ and $Y_1 = r$. Further, plugging in $l = 2$ in (6) gives

$$\mathbf{E}\big[e^{-\theta S(\tau,\tau+t)}\big] \leq \big(1 + e^{-2\theta r}\big)^{t/2} \sqrt{\tilde{p}_t}. \qquad (8)$$

Further, we have $\tilde{p}_t \leq \max(\mathbf{P})^{t-1}$, where $\mathbf{P}$ is the one-step transition matrix of $\mathbf{Z}$. Therefore, we can write

$$\begin{aligned} &P[S(\tau, \tau + t) < E_s(t)] \\ &\leq e^{\theta E_s(t)} (1 + e^{-2\theta r})^{t/2} p_*^{(t-1)/2}. \end{aligned} \qquad (9)$$

Here, we have used the shorthand notation p_* for $\max(\mathbf{P})$. Now, if ϵ_s denotes the violation probability on the right hand side of (9), we can derive the service envelope $E_s(t)$ as

$$E_s(t) = \frac{1}{\theta}\left(\log \epsilon_s + \frac{\log p_*}{2} - t\left[\frac{\log(1 + e^{-2\theta r})}{2} + \frac{\log p_*}{2}\right]\right), \qquad (10)$$

where $\theta > 0$ is a free parameter that can be optimized. The parameter p_* above can be inferred from observed service information if it is not known a-priori. We describe the estimation procedure in the appendix.

Application to Gilbert-Elliot Channels: The On-Off service model described in Sect. III-C is known in the literature as the Gilbert-Elliott [11] channel model. Here, a wireless fading channel is modeled as being in either a good (*On*) or a bad state (*Off*). Successful transmission is possible in the good state while the bad state denotes no possible transmission. The channel memory is modeled using the transition matrix $\mathbf{P}$ which together with the *On* rate r characterizes the model.

IV. Numerical Evaluation

In the following, we illustrate the performance of our adaptive algorithm that uses the subsystem selection rule given in (5) in comparison to *Round-Robin* (RR) and *Join the Shortest Queue* (JSQ) subsystem assignment. For JSQ the expression in (5) reduces to taking the subsystem $m = \arg\min_{i=1,\ldots,N} B_i$ at the beginning of the epoch. We consider two distinct cases: *i)* subsystems characterized by memoryless Poisson service and *ii)* subsystems that possess channels with memory captured using the MMOO model. For simplicity, we assume Poisson arrivals to the entire system. If not stated otherwise, we set the number of simulation runs to 10^3, each spanning 10^4 time slots and fix the decision epoch length at 10 time slots.

A. Poisson Service

Here we consider the case when both the arrival and subsystem service increment processes follow a Poisson distribution. In particular, the arrival and service increments are iid with parameters λ and $\mu = [\mu_1, \ldots, \mu_N]$, respectively. Here, μ_i denotes average service increment rate of i^{th} subsystem.

Next, we explore the effects of the following factors on system performance: (i) heterogeneity of subsystems (ii) the number of subsystems N, and (iii) the epoch length Δ. We present CCDFs of the waiting times at the different subsystems, as well as, the CCDF of the maximum waiting time across different subsystems. The rationale here is to show the impact of the assignment strategy on the worst case waiting time within the system.

To begin with, Fig. 2 depicts the CCDF of the maximum waiting time as well as the subsystem-specific waiting times for a system with $N = 2$ subsystems with arrival rate $\lambda = 1.5$ and heterogeneous service rates $\mu = [1, 2]$. This heterogeneous scenario captures the case of two subsystems, e.g., two technologies such as WiFi and LTE, having different average bandwidths with convenient iid assumptions on the service increments. We observe that in this heterogeneous case the adaptive strategy strongly enhances the performance in the *worst case*, measured by the maximum of the waiting times of the two subsystems. Looking at Fig. 2c we observe that this improvement comes at the cost of higher waiting times at the faster system. We observe here that for the individual subsystems the adaptive system is *not* generally superior.

Next, we aim to quantify the performance benefit of including a new subsystem. To this end, we show in Fig. 3 CCDFs of average waiting times for an increasing number of *homogeneous* subsystems N with mean service rate $\mu_i = 1$ each. Interestingly, we observe a comparable performance of the average waiting times in the system for all three assignment strategies, i.e., RR, JSQ and our adaptive assignment. We infer, first, that differences between the assignment strategies diminish with decreasing utilization and that the adaptive strategy performs comparably to the other assignment strategies, especially JSQ, in homogeneous settings.

To assess the impact of heterogeneity on how the adaptive assignment strategy utilizes the available subsystems we consider $N = 2$ subsystems and fix the mean service rate

of subsystem 1 at $\mu_1 = 1$ and vary $\mu_2 \in [1.5, 16]$. We plot the average waiting time in Fig. 4a and 4b showing that the adaptive system clearly favors the faster subsystem. This preference becomes more pronounced with increasing heterogeneity.

We combine the previous two scenarios in Fig. 4c and 4d where we increase the number of subsystems N each with a service rate i being $\mu_i = i$. The figures show diminishing differences between the assignment strategies with overprovisioning. We note that at tightly provisioned systems significant differences between the assignment strategies are apparent.

Finally, we focus on the reactiveness of the assignment system which we capture through the length of the decision epoch Δ. In Fig. 5 we show the CCDF of the maximum waiting time as well as the subsystem-specific waiting times for a system with $N = 2$ subsystems with arrival rate $\lambda = 1.5$ and heterogeneous service rates $\mu = [1, 2]$. Varying the epoch length in this heterogeneous scenario, we see that the adaptive strategy outperforms RR and JSQ. We attribute this to the fact that in contrast to RR and JSQ, the adaptive strategy *takes the statistical characteristics of the arrival and service processes* into account for the assignment decision as defined in (5).

B. MMOO Service

Next, we consider the case when the arrivals follow a Poisson distribution with rate $\lambda = 0.5$ and the subsystems in Fig. 1 possess channels with memory which we model through service processes S_i that are given as MMOO processes with service increment rates at the *On* state $r = [5, 50]$. The underlying transition matrix at the first subsystem is given by $[0.8, 0.2; 0.25, 0.75]$, and at the second subsystem by $[0.85, 0.15; 0.2, 0.8]$. Corresponding initial probabilities of being in the *On* state are 0.5 and 0.3 respectively. Note that the service envelope is dynamically estimated in the beginning of each decision epoch by inferring the transition matrices through service feedback from respective subsystems as described in the appendix. The performance at both subsystems together with the worst case is illustrated in Fig. 6 through the CCDF of waiting times. We observe that the adaptive system provides a higher decay rate of the worst case waiting times.

V. Related work

Although, in this work, we have presented a service-curve based strategy for scheduling, our approach is essentially different than the well-known SCED and SCED+ algorithm described in [19] and [8]. While SCED and SCED+ discuss the way multiple sessions should share a server given a set of delay bounds and burstiness of arriving traffic, our work focuses on choosing one among alternative subsystems with an objective to reduce occurrence of longer waiting times. Similarly, the authors in [12] prescribe a strategy that seeks to minimize buffer overflow probabilities for individual sessions and thereby improve the system throughput in wireless networks. A closely related work is [15] which describes a buffer based uplink strategy in relay-assisted LTE networks with similar objective in mind. In this context, authors of [16] also demonstrate a queue-aware uplink scheduling that optimizes resource utilization that simultaneously asserts a given level of quality of service (QoS).

Further related works deal with systems with adjustable service rate that is dynamically determined to optimize some cost. The authors in [10] devise an optimal strategy with respect to the average long term cost. Such an approach of dynamic control is also seen in [1] and [20]. While the service rate is adjusted in the beginning of each decision epoch in [1], [20] optimizes a linear combination of the expected cost and the expected job response time.

In [14], the authors compute the average queue length and waiting time for queues with Markov modulated service using a matrix geometric method and demonstrate its application in the context of web-servers with multi-class requests. Additionally, there is a class of related work that specifically focuses on rate and power optimization. Typically, the optimization criteria in this sphere of work is the average queuing delay. In [3], the authors optimize over power and rate policies to minimize the average delay under power constraints. Assuming Gilbert-Elliot type channels and a linear relationship of power and rate, the authors in [6] prescribe an optimal service policy for a service path of finite length using a similar approach. Further, in [2], results on regulating rate and power to control average delay and transmission power have been provided using concepts of Markov decision theory. Finally, as our adaptive assignment was shown to be particularly effective in case of heterogeneous systems, we note that some inherent aspects of load balancing in such systems have been explored in [18] using Reinforcement learning.

VI. Conclusions and Outlook

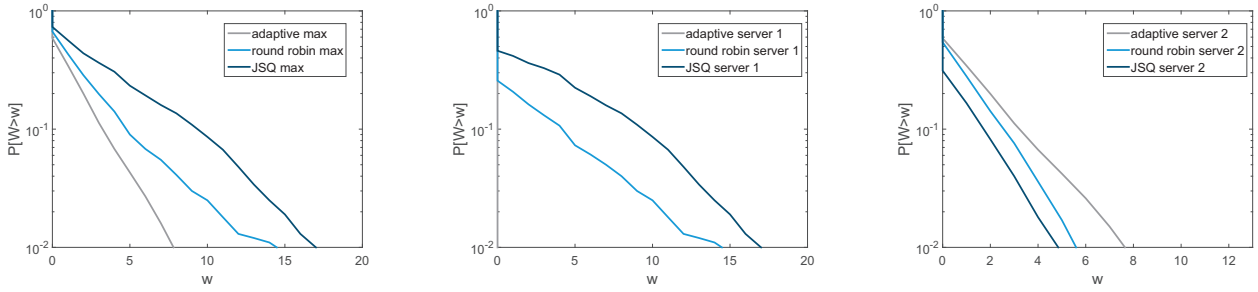
In this work, we proposed a queue-aware scheduling strategy for multi-interface communications that switches between available subsystems by looking at the respective service guarantees together with the statistical properties of the data arrivals. We illustrated the performance of this adaptive assignment strategy for different statistical subsystem characteristics, especially for subsystems providing bursty service. The numerical comparison with the established Round-Robin and Join-the-Shortest-Queue strategies showed that the adaptive strategy enhanced the worst case system waiting times, especially under subsystem heterogeneity. A logical next step could be comparing this approach to other scheduling algorithms such as choosing the subsystem with the lowest expected backlog after one epoch or using weighted RR assignments with weights proportional to average subsystem service rates. Lastly, looking into interdependent subsystem service processes will broaden the scope of the results.

Appendix

A. MMOO Inference

The parameters of the Markov chain $\mathbf{Z}$ in Section III-C can be estimated in a straightforward manner since $Z_i = X_i/r$ i.e., not hidden. Let $\mathbf{P}$ be the 2×2 transition matrix of Z, i.e.,

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}.$$

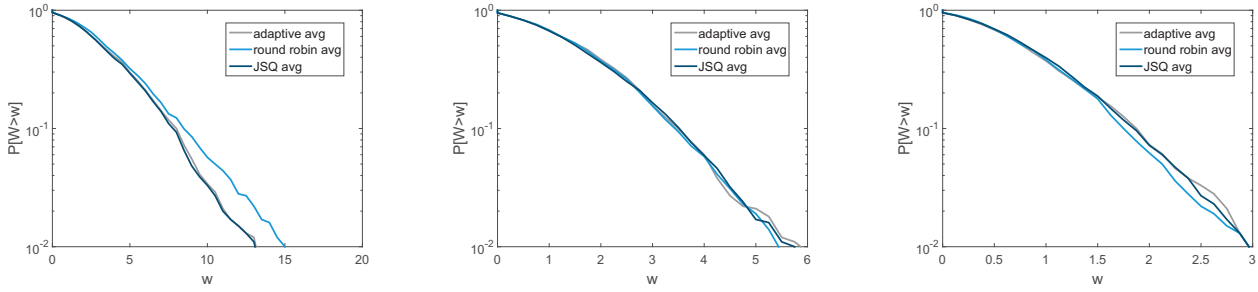


(a) Empirical CCDF of the maximum of the waiting times at two subsystems.

(b) Empirical CCDF of the waiting times at the slower subsystem.

(c) Empirical CCDF of the waiting times at the faster subsystem.

Fig. 2: Two subsystems with mean service rates $\mu = [1, 2]$. The empirical CCDF of the adaptive policy coincides with the y-axis at the slower subsystem. Parameters: arrival rate $\lambda = 1.5$, violation probability $\epsilon = 10^{-4}$.

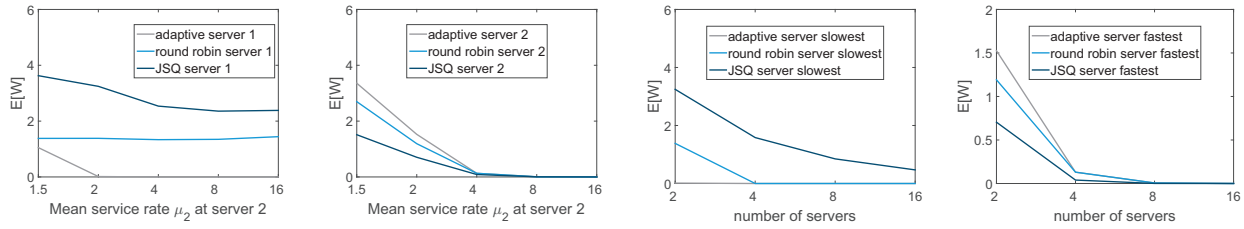


(a) $N = 2$ subsystems.

(b) $N = 4$ subsystems.

(c) $N = 8$ subsystems.

Fig. 3: Empirical CCDF of the average waiting time at N subsystems: Strong impact of the number of subsystems N on the average waiting time in the system. All three assignment strategies have comparable performance. Arrival rate $\lambda = 1.5$, mean service rates $\mu_i = 1$ for every subsystem i, violation probability $\epsilon = 10^{-4}$.



(a) Slower subsystem

(b) Faster subsystem

(c) Slowest subsystem

(d) Fastest subsystem

Fig. 4: Fig. 4a and 4b show the expected waiting time at subsystems 1 and 2 for a varying mean service rate of the second subsystem. Fig. 4c and 4d show the expected waiting time at the slowest / fastest subsystem when increasing the number of subsystems N. Here, subsystem i possesses a mean service rate $\mu_i = i$. We fix $\mu_1 = 1$, $\lambda = 1.5$, $\epsilon = 10^{-4}$, and the number of simulation runs to 10^4 .

Note that here we deal with the inference of the modulating transition matrix for a certain subsystem. Since our algorithm uses possibly different subsystems in different epochs we have observations as $\{X_j^i\}_{i,j}$ where X_j^i denotes the j^{th} observation of an epoch when the relevant subsystem was chosen for the i^{th} time. If there are n_i observations in the i^{th} batch, we have

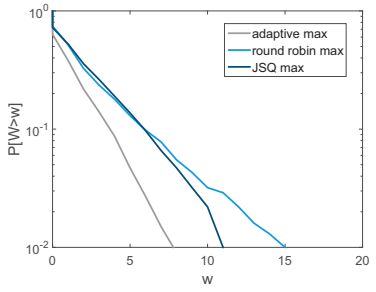
$$\hat{p_{00}} = \frac{\sum_i \sum_{j=1}^{n_i - 1} \mathbf{1}_{X_j^i > 0} \mathbf{1}_{X_{j+1}^i > 0}}{\sum_i \sum_{j=1}^{n_i - 1} \mathbf{1}_{X_j^i > 0}}.$$

Other elements of $\mathbf{P}$ can be estimated in a similar fashion. The maximum of these elements then can be plugged in (10).

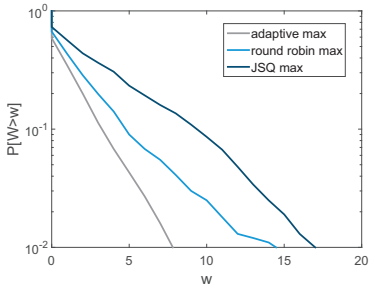
REFERENCES

[1] K. M. Adusumilli and J. J. Hasenbein, "Dynamic admission and service rate control of a queue," *Queueing Systems*, vol. 66, no. 2, pp. 131–154, Oct 2010.

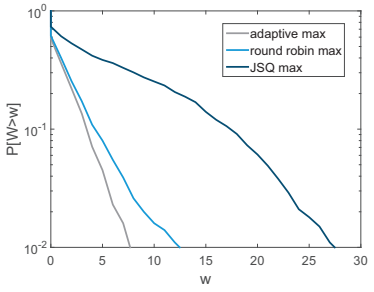
[2] R. A. Berry and R. G. Gallager, "Communication over fading channels with delay constraints," *IEEE Transactions on Information Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.



(a) Empirical CCDF of maximum waiting time when epoch length $\Delta = 5$

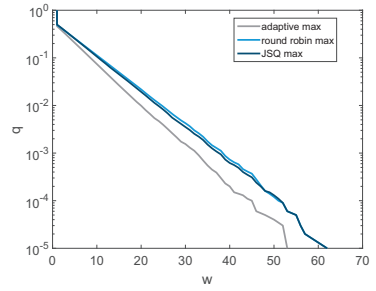


(b) Empirical CCDF of maximum waiting time when epoch length $\Delta = 10$

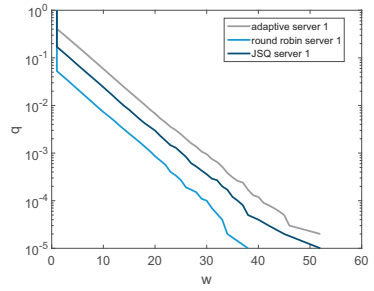


(c) Empirical CCDF of maximum waiting time when epoch length $\Delta = 20$

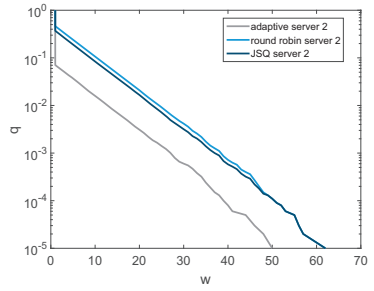
Fig. 5: Impact of varying the epoch length Δ on the worst-case waiting time for two subsystems with mean service rates $\mu = [1, 2]$. The adaptive strategy tends to outperform throughout whereas the performance of JSQ deteriorates with increasing Δ. Parameters: arrival rate $\lambda = 1.5$, violation probability $\epsilon = 10^{-4}$.



(a) Empirical CCDF of maximum waiting time



(b) Empirical CCDF of waiting time at the slower subsystem



(c) Empirical CCDF of waiting time at the faster subsystem

Fig. 6: Two MMOO subsystems with *on* state service rates $r = [5, 50]$, arrival rate $\lambda = 0.5$. The adaptive assignment strategy achieves lower worst case waiting times.

[3] I. Bettesh and S. S. Shamai, "Optimal power and rate control for minimal average delay: The single-user case," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 4115–4141, Sept 2006.

[4] C.-S. Chang, *Performance Guarantees in Communication Networks*. Springer-Verlag, 2000.

[5] F. Ciucu, A. Burchard, and J. Liebeherr, "Scaling properties of statistical end-to-end bounds in the network calculus," vol. 14, no. 6, pp. 2300–2312, Jun. 2006.

[6] B. Collins and R. L. Cruz, "Transmission policies for time varying channels with average delay constraints," in *Proceedings of the Annual Allerton Conference on Communication Control and Computing*, vol. 37. The University; 1998, 1999, pp. 709–717.

[7] R. L. Cruz, "Quality of service management in Integrated Services networks," in *Proc. of Semi-Annual Research Review, Center of Wireless Communication, UCSD*, Jun. 1996.

[8] ——, "Sced+: efficient management of quality of service guarantees," in *INFOCOM '98. Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 2, Mar 1998, pp. 625–634 vol.2.

[9] M. Fidler and A. Rizk, "A guide to the stochastic network calculus," *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 92–105, Firstquarter 2015.

[10] J. M. George and J. M. Harrison, "Dynamic control of a queue with adjustable service rate," *Operations Research*, vol. 49, no. 5, pp. 720–731, 2001.

[11] E. Gilbert, "Capacity of a burst-noise channel," *Bell Systems Technical Journal*, vol. 39, no. 5, pp. 1253–1265, 1960.

[12] J. Huang and Z. Niu, "Buffer-aware and traffic-dependent packet scheduling in wireless ofdm networks," in *2007 IEEE Wireless Communications and Networking Conference*, March 2007, pp. 1554–1558.

[13] C. Li, A. Burchard, and J. Liebeherr, "A network calculus with effective bandwidth," vol. 15, no. 6, pp. 1442–1453, Dec. 2007.

[14] S. R. Mahabhashyam and N. Gautam, "On queues with markov modulated service rates," *Queueing Systems*, vol. 51, no. 1, pp. 89–113, Oct 2005.

[15] M. Mehta, S. Khakurel, and A. Karandikar, "Buffer-based channel dependent uplink scheduling in relay-assisted lte networks," in *2012 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2012, pp. 1777–1781.

[16] A. Rizk and M. Fidler, "Queue-aware uplink scheduling with stochastic guarantees," *Computer Communications*, vol. 84, pp. 63–72, 2016.

[17] P. Sadeghi, R. A. Kennedy, P. B. Rapajic, and R. Shams, "Finite-state markov modeling of fading channels - a survey of principles and applications," *IEEE Signal Processing Magazine*, vol. 25, no. 5, pp. 57–80, September 2008.

[18] F. Samreen and M. S. H. Khiyal, "Q-learning scheduler and load balancer for heterogeneous systems," *Journal of Applied Sciences*, vol. 7, no. 11, pp. 1504–1510, 2007.

[19] H. Sariowan, R. L. Cruz, and G. C. Polyzos, "Sced: a generalized scheduling policy for guaranteeing quality-of-service," *IEEE/ACM Transactions on Networking*, vol. 7, no. 5, pp. 669–684, Oct 1999.

[20] A. Wierman, L. L. H. Andrew, and A. Tang, "Stochastic analysis of power-aware scheduling," in *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, Sept 2008, pp. 1278–1283.

[21] Q. Yin, Y. Jiang, S. Jiang, and P. Y. Kong, "Analysis of generalized stochastically bounded bursty traffic for communication networks," in *Proc. of IEEE LCN*, Nov. 2002, pp. 141–149.