On the Delays in Time-Varying Networks: Does Larger Service-Rate Variance Imply Larger Delays?

Sébastien Henri
EPFL, Switzerland
sebastien.henri@epfl.ch

Seva Shneer
Heriot-Watt University, UK
V.Shneer@hw.ac.uk

Patrick Thiran
EPFL, Switzerland
patrick.thiran@epfl.ch

ABSTRACT

In all networks, link or route capacities fluctuate for multiple reasons, e.g., fading and multi-path effects on wireless channels, interference and contending users on a shared medium, varying loads in WAN routers, impedance changes on power-line channels. These fluctuations severely impact packet delays. In this paper, we study delays in time-varying networks. Intuitively, we expect that for a given average service rate, an increased service rate variability yields larger delays. We find that this is not always the case. Using a queuing model that includes time-varying service rates, we show that for certain arrival rates, a queue with larger service rate variance offers smaller average delays than a queue with the same average service rate and lower service rate variance. We also verify these findings on a wireless testbed. We then study the conditions under which using simultaneously two independent paths helps in terms of delays, for example, in hybrid networks where the two paths use different physical layer technologies. We show that using two paths is not always better, in particular for low arrival rates. We also show that the optimal traffic splitting between the two paths depends on the arrival rate.

CCS CONCEPTS

• **Networks** → **Network performance modeling**; **Network performance analysis**; *Network dynamics*;

KEYWORDS

Delay analysis; Queueing; Time-varying networks.

ACM Reference Format:
Sébastien Henri, Seva Shneer, and Patrick Thiran. 2018. On the Delays in Time-Varying Networks: Does Larger Service-Rate Variance Imply Larger Delays?. In *Mobihoc '18: The Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing, June 26–29, 2018, Los Angeles, CA, USA.* Los Angeles, CA, USA, 10 pages. https://doi.org/10.1145/3209582.3209603

1 INTRODUCTION

The demand for high-throughput connectivity is increasing at a very fast pace. This has led to the publication of new standards for high throughput, such as 802.11n and 802.11ac for WiFi, LTE for cellular networks, and IEEE 1901 for power-line communications (PLC). However, the throughput that these technologies offer individually is arguably close to reaching a maximum. For this reason, the next natural step to increase throughput is to simultaneously use multiple technologies. This is well illustrated by the standardization of hybrid networks by the IEEE 1905 working group [3], and by the development of multipath TCP (MPTCP) [27], in particular with WiFi and LTE. When the technologies employed do not interfere, having several technologies rather than a single one is in principle always beneficial in terms of throughput. But throughput is rarely the only metric that needs to be optimized, and hybrid networks open interesting perspectives in terms of delays, power consumption, reliability. In this paper, we focus on delays. Delays in today's networks are of paramount importance and are acknowledged to remain fundamental in tomorrow's networks, as illustrated by the standardization efforts towards ultra-reliable low-latency communications (URLLC) in wireless networks [4].

The time variability of the service rate has a strong impact on the delays: If the service rate of a link (i.e., the link capacity) decreases, packets accumulate in the queue of the network interface, which increases the delays. Because of varying signals (fading, multi-path effects, etc.), WiFi typically presents a behavior with time-varying service rates. This is illustrated by Figure 1 (left), where we show the instantaneous physical rate, obtained by the modulation and coding scheme (MCS) index employed, averaged over ten packets, in a 25 seconds WiFi trace, with no other traffic. Clearly, the MCS index is not constant; it varies between (mostly) four areas with two dominant states, a *high-rate* state (about 80 Mb/s) and a *low-rate* state (about 60 Mb/s). These service rate variations with high-rate and low-rate states can be observed in many other contexts, e.g., when several users employ the same channel and interfere with each other: If Alice is streaming a video or downloading a file with WiFi or LTE and no other user is active, Alice has a high throughput: she is in a high-rate state. If another user Bob is active (e.g., he is surfing the Web), Alice's throughput decreases: she is in a low-rate state. This can also be observed on WAN routes if one WAN router alternates between high-traffic loads (low-rate state) and low-traffic loads (high-rate state); or in data-centers where servers typically alternate between high-rate and low-rate states, that can be caused by a higher load of the server, but also by many external reasons such as garbage collection, network interrupts, or background work [12]; or at home with PLC, because switching on and off appliances changes the electrical impedance and can cause link-capacity drops [30].

In Section 2, we describe a model that captures the variability of the service rate and that enables us to study the impact on delays of this variability, in two different settings: (*i*) For a given arrival rate, we want to choose between two paths (for example,

Mobihoc '18, June 26–29, 2018, Los Angeles, CA, USA

ACM ISBN 978-1-4503-5770-8/18/06…$15.00
https://doi.org/10.1145/3209582.3209603

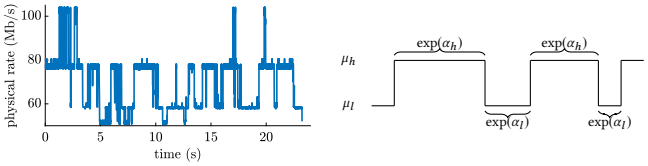

Figure 1: Left: physical rate of WiFi, averaged over ten packets. Right: sample path of the service rate.

two routes in a WAN or two technologies in a LAN, e.g., WiFi and LTE or WiFi and PLC). Increasing the average service rate clearly decreases the average delays. To isolate the effect of time variability, we hence assume that the two paths have the same *average* service rate, but different *variances* of the service rate, and we want to find which path yields the smallest average delays. (*ii*) We next study the second setting where two paths, potentially with different average service rates, can be used simultaneously, and traffic be split between them. In Section 4, this setting is described and its analysis is provided. We prove that the optimal load balance depends on the arrival rates, and that for low arrival rates, it is better to use only one path. We validate our analyses with experiments on a wireless testbed in Section 5. We discuss related work in Section 6 and give concluding remarks in Section 7.

2 QUEUE MODEL AND BACKGROUND WITH TIME-VARYING SERVICE RATES

We model a network interface by a queue with i.i.d. exponentially-distributed arrivals with rate parameter λ. The server operates in two different regimes: A *low-rate* state in which the packets are sent (or the jobs are completed) with exponential service rate μ_l, and a *high-rate* state in which the packets are sent with exponential service rate μ_h, with $0 < \mu_l < \mu_h$. From now on, we simply write low state and high state for low-rate state and high-rate state. The sojourn times in low and high states are exponentially distributed, with parameter α_l when in low state and α_h when in high state. The service model is thus a continuous-time Markov chain, with a sample path illustrated in Figure 1 (right). If $(Q(t), S(t))$ denotes the number of packets and the state of the server at time t, and if $r((q_1, s_1), (q_2, s_2))$ denotes the transition rates from state (q_1, s_1) to state (q_2, s_2), then the only non-zero transition rates are

$$\begin{aligned} r((q,s),(q+1,s)) &= \lambda, \quad \text{for all} \quad q \geq 0, \quad s \in \{h,l\}, \\ r((q,s),(q-1,s)) &= \mu_s, \quad \text{for all} \quad q \geq 1, \quad s \in \{h,l\}, \\ r((q,h),(q,l)) &= \alpha_h, \quad \text{for all} \quad q \geq 0, \\ r((q,l),(q,h)) &= \alpha_l, \quad \text{for all} \quad q \geq 0. \end{aligned}$$

This model with heterogeneous time-varying service rates was studied as early as 1971 by Yechiali and Naor [32]. However, the expression of the average delays is quite complex; it involves computing the root of a cubic equation (see also in Appendix). Even though it can be computed in principle, its explicit expression is very complex (it would take dozens of line to write it explicitly), and few works give analytical results that can be intuitively understood. Ross conjectured in 1978 that increased variability leads to increased averaged delays [29]. Variability was expressed by the sole parameter $\alpha = \alpha_h + \alpha_l$. A lower α means longer sojourn times in the "high" and "low" states: It leads to a higher heterogeneity of the service rates, i.e., a larger variability. In contrast, with very short sojourn times in each of the states (i.e., very frequent transitions and large α), the queue with heterogeneous service rates performs close to a homogeneous M/M/1 queue with an averaged service rate [13]. Ross' conjecture was proven in 1981 by Rolski [28] in the particular case $\mu_h = \mu_l$, with different arrival rates in high and low states: The average delay in this scenario is a decreasing function of α. The general case, with different service rates ($\mu_h \neq \mu_l$), was studied in 2006 by Gupta et al. [13]. They show that the average queue size is always monotonic in α, but not always decreasing. In our setting where the arrival rate is the same in both states, the average queue size, hence, the average delay is a decreasing function of α, as conjectured by Ross.

3 ANALYSIS OF A QUEUE WITH TIME-VARYING SERVICE RATES

We consider two queues that follow the model described in Section 2, with respective parameters $\mu_{l,i}$, $\mu_{h,i}$, $\alpha_{l,i}$, and $\alpha_{h,i}$ for $i = 1, 2$. In the remaining of the paper, we only study average delay, and sometimes refer to it simply as delay; delay of Queue i as a function of the arrival rate λ is denoted by $D_i(\lambda)$. We want to compare the two queues and to find out which one yields the lowest delay, when they have the same average service rate and arrival rate. Intuitively, we expect that a larger variability should yield larger delays. This is what happens for an M/G/1 queue, where variability is expressed by the variance of the service times: The well-known Pollaczek–Khinchine formula states indeed that when the two queues have the same average service rate $\hat{\mu}$ and the same arrival rate $\lambda \in [0, \hat{\mu})$, the queue with the largest delays is always the queue with the largest variance of the service times. Our model with heterogeneous time-varying service rates is different: In an M/G/1 queue, the service times are i.i.d., contrary to our model where they are drawn from two different exponential distributions, depending on the state (low or high) the system is in. We show in this section that, although this is often the case, larger variability does *not* always yield larger delays; we also show that for certain values of $\mu_{l,i}$, $\mu_{h,i}$, $\alpha_{l,i}$, and $\alpha_{h,i}$, the queue with the largest delays is not the same for all arrival rates λ.

Let R_i denote the random variable for the service rate, taking values in $\{\mu_{l,i}, \mu_{h,i}\}$ and distributed according to the stationary distribution of the process illustrated in Figure 1 (right). The average service rate of Queue i for $i \in \{1, 2\}$ is given by

$$\hat{\mu}_i = \mathbb{E}[R_i] = \frac{\mu_{h,i}/\alpha_{h,i} + \mu_{l,i}/\alpha_{l,i}}{1/\alpha_{h,i} + 1/\alpha_{l,i}} = \frac{\alpha_{l,i}\mu_{h,i} + \alpha_{h,i}\mu_{l,i}}{\alpha_{h,i} + \alpha_{l,i}},$$

For Queue i to be stable, we must have $\lambda < \hat{\mu}_i$ [32]. In this section, we are interested in studying how the variability of the service rates, rather than the average service rate itself, affects the average delays. Indeed, it is obvious that for a fixed level of variability, an increased average service rate yields lower delays. For this reason, we assume in this section that the two queues have the same average service rate, i.e., $\hat{\mu}_1 = \hat{\mu}_2 \doteq \hat{\mu}$. In particular, the two queues have the same stability region $\lambda \in [0, \hat{\mu})$. Previous work has studied the effect on the delays of the parameter $\alpha_i \doteq \alpha_{h,i} + \alpha_{l,i}$ when μ_h and μ_l are fixed [13, 16, 28]. For this reason, we assume in this section

that $\alpha_1 = \alpha_2 \doteq \alpha$ and our goal is to study the effect of the other parameters. We express variability by the variance V_i of the service rate of Queue i, which can be written as

$$V_i = \mathrm{Var}[R_i] = \frac{\alpha_{l,i}\mu_{h,i}^2 + \alpha_{h,i}\mu_{l,i}^2}{\alpha_{h,i} + \alpha_{l,i}} - \hat{\mu}^2. \quad (1)$$

Note that $V_i = 0$ **iff** Queue i is homogeneous ($\mu_{h,i} = \mu_{l,i}$).

The delay for Queue i for $i \in \{1, 2\}$ is determined by five parameters: $\mu_{h,i}$, $\mu_{l,i}$, $\alpha_{h,i}$, $\alpha_{l,i}$, and λ. When $\hat{\mu}$ and α are fixed, there are still three degrees of freedom, and the delay depends both on the service rates (μ's) and on the transition rates (α's). The first natural question that we want to answer is what happens when the variance V is fixed (i.e., $V_1 = V_2$). Theorem 3.1 shows that determining which queue has the largest delays now only depends on the service rates and not on the transition rates, and that the best queue is the same for all arrival rates. The proofs of the theorems of this section are given in Appendix.

Theorem 3.1. *Let us assume that $\hat{\mu}_1 = \hat{\mu}_2$ and $\alpha_1 = \alpha_2$. For $i \in \{1, 2\}$, let*

$$\pi_i = \mu_{h,i}\mu_{l,i}. \quad (2)$$

*If $V_1 = V_2$, then for all arrival rates $\lambda \in (0, \hat{\mu})$, $D_1(\lambda) \geq D_2(\lambda)$ **iff** $\pi_1 \leq \pi_2$, with equality **iff** $\pi_1 = \pi_2$. Conversely, if $\pi_1 = \pi_2$, then for all arrival rates $\lambda \in (0, \hat{\mu})$, $D_1(\lambda) \geq D_2(\lambda)$ **iff** $V_1 \geq V_2$, with equality **iff** $V_1 = V_2$.*

With $\hat{\mu}$ and α fixed, the delay for Queue i is fully determined by V_i (defined by (1)), π_i (defined by (2)), and λ. We now want to determine the queue with the lowest delays when neither the variance V nor the product of the two service rates π are fixed, i.e., when $V_1 \neq V_2$ and $\pi_1 \neq \pi_2$. Theorem 3.1 shows that for all arrival rates λ, the average delay is an increasing function of V_i when $\pi_1 = \pi_2$, and a decreasing function of π_i when $V_1 = V_2$. For this reason, we expect that the average delay increases when V_i increases and π_i decreases, and that it decreases when V_i decreases and π_i increases. Theorem 3.2 shows that this is indeed true.

Theorem 3.2. *Let us assume that $\hat{\mu}_1 = \hat{\mu}_2$ and $\alpha_1 = \alpha_2$. If $V_1 > V_2$ and $\pi_1 < \pi_2$, then for all arrival rates $\lambda \in (0, \hat{\mu})$, $D_1(\lambda) > D_2(\lambda)$. Conversely, if $V_1 < V_2$ and $\pi_1 > \pi_2$, then for all arrival rates $\lambda \in (0, \hat{\mu})$, $D_1(\lambda) < D_2(\lambda)$.*

When μ_h is fixed ($\mu_{h,1} = \mu_{h,2}$), V_i is a decreasing function of $\mu_{l,i}$ and π_i is an increasing function of $\mu_{l,i}$. (Note that because $\hat{\mu}$ and μ_h are fixed, increasing $\mu_{l,i}$ requires increasing the sojourn times in the low state.) A corollary of Theorem 3.2 is consequently that if $\mu_{h,1} = \mu_{h,2}$, the delay is an increasing function of V_i (and a decreasing function of $\mu_{l,i}$).

The situation becomes more complex when both V_i and π_i decrease, or both V_i and π_i increase. For certain values of V_i and π_i, one queue yields lower delays for all arrival rates, and in that case, we show that this must be the queue with the lowest variance V_i. However, we show that for certain values of V_i and π_i (precise conditions are given in Appendix), determining which queue yields lower delays depends on the arrival rate λ. In particular, for certain arrival rates, the queue with the lowest delays is the queue with the largest variance V_i (Corollary 3.4).

Theorem 3.3. *Let us assume that $\hat{\mu}_1 = \hat{\mu}_2$ and $\alpha_1 = \alpha_2$.*

- *There are values of V_i and π_i such that $D_1(\lambda) - D_2(\lambda)$ changes sign in $(0, \hat{\mu})$.*
- *If for all arrival rates $\lambda \in (0, \hat{\mu})$, $D_1(\lambda) \neq D_2(\lambda)$, then for all $\lambda \in (0, \hat{\mu})$, $D_1(\lambda) > D_2(\lambda)$ **iff** $V_1 > V_2$.*

Corollary 3.4. *For certain values of V_i and π_i, there is $\lambda_0 \in (0, \hat{\mu})$ such that for all arrival rates $\lambda \in (0, \lambda_0)$, $V_1 > V_2$ and $D_1(\lambda) < D_2(\lambda)$, or $V_1 < V_2$ and $D_1(\lambda) > D_2(\lambda)$.*

In Section 5, we present numerical and testbed evidences that show that by using the queue with the largest variance, the delay gain provided can be significant.

The next step is to understand why the queue with the lowest variance can sometimes yield larger delays. We make the following conjecture, that appears to hold numerically.

Conjecture 3.5. *Let us assume that $\hat{\mu}_1 = \hat{\mu}_2$ and $\alpha_1 = \alpha_2$. If $V_1 > V_2$ and $\mu_{l,1} < \mu_{l,2}$, then for all arrival rates $\lambda \in (0, \hat{\mu})$, $D_1(\lambda) > D_2(\lambda)$. Conversely, if $V_1 < V_2$ and $\mu_{l,1} > \mu_{l,2}$, then for all arrival rates $\lambda \in (0, \hat{\mu})$, $D_1(\lambda) < D_2(\lambda)$.*

This would mean that a necessary condition for the queue with the largest variance to yield the smallest delays would be to have its service rate in the low state ($\mu_{l,i}$) be larger than that of the queue with the lowest variance. (Note that this is *not* a sufficient condition.) Therefore, in low state, the size of the queue with the smallest variance would increase faster than the size of the queue with the largest variance, which can cause larger average delays despite a smaller variance. This conjecture has another direct consequence: If $\alpha_{h,1} = \alpha_{h,2} \doteq \alpha_h$ and $\alpha_{l,1} = \alpha_{l,2} \doteq \alpha_l$, then for all $\lambda \in (0, \hat{\mu})$, $D_1(\lambda) > D_2(\lambda)$ **iff** $V_1 > V_2$: If we fix the transition rates α_h and α_l, then the average delay is an increasing function of the variance. If α_h and α_l are fixed, we can already prove that if $\alpha_l\mu_{h,i} \geq \alpha_h\mu_{l,i}$, then $V_1 \geq V_2$ **iff** $\pi_1 \leq \pi_2$ (see Lemma A.7 in Appendix), and it follows therefore from Theorem 3.2 that for all $\lambda \in (0, \hat{\mu})$ $D_1(\lambda) > D_2(\lambda)$ **iff** $V_1 > V_2$. This is in particular the case if $\alpha_l \geq \alpha_h$, i.e., if the queues spend more time in the high state.

Note that if $\alpha_1 \neq \alpha_2$, it is possible to have $D_1(\lambda) < D_2(\lambda)$ with $V_1 > V_2$ and $\mu_{l,1} < \mu_{l,2}$ (e.g., when α_1 is large and α_2 is small).

4 TWO PATHS USED SIMULTANEOUSLY

We now move to the second setting, where two paths, potentially with different average service rates, can be used simultaneously. We first describe the model for two simultaneous paths, and then we analyze the average delay theoretically and numerically.

4.1 Model for Simultaneous Paths

The model for our scenario where two paths can be used simultaneously is illustrated in Figure 2. As described in Section 2, the

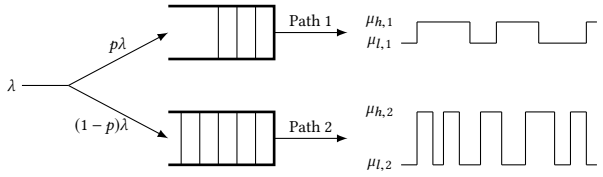

Figure 2: Model for two paths used simultaneously.

service rate of Queue i for $i \in \{1,2\}$ is modelled by low and high states, with respective service rates $\mu_{l,i}$ and $\mu_{h,i}$ ($\mu_{h,i} > \mu_{l,i} > 0$). Packets arrive in the system as a Poisson process with rate parameter λ, and are routed to the queues based on a Bernoulli trial: With probability p, a packet goes to Queue 1, with probability $1-p$, it goes to Queue 2. As opposed to most works that study queues in parallel [14, 17, 18], we assume that the routing decision is made without knowledge of the current size of the queues. Several reasons justify this choice. First, it might be impossible to obtain the size of the internal queue of a network interface (e.g., most PLC devices do not give such an information). Second, even when it is possible, it can be quite complex in practice. For example, in OpenWRT (the system that we use in our experiments of Section 5), the networking stack consists in three different queues [2]; accessing each of them on a packet-per-packet basis might cause significant overload. Third, to reduce the overhead due to MAC protocols, recent technologies employ frame aggregation. This is for example the case in IEEE 802.11n and 802.11ac, the most recent standards for WiFi, and in IEEE 1901, the most recent standard for PLC. This means that a queue might be non-empty while the channel is idle, which makes the model where the routing decision is based solely on queue size inadequate. Protocols that send feedback for (almost) every packet, such as MPTCP, offer the possibility to use indirect information on the queues, such as the round-trip time. However, this information is delayed, whereas classic models assume immediate information. In addition, feedback cannot be employed for a protocol such as UDP, whereas UDP might be preferred for delay-sensitive applications [20]. In Section 5, we compare experimentally the delays obtained with our model with UDP and those obtained with TCP/MPTCP.

One purpose of our analysis is to study the effect on delays of the parameter p, i.e., of the traffic splitting between the two paths. We want to find the value of p, denoted in the following by p^*, that yields the smallest average delay. When the packets are routed to the queues by using a Bernoulli trial, the two queues are independent, with respective arrival rates $p\lambda$ and $(1-p)\lambda$. With $N_i(\lambda_i)$ being the average size of Queue i as a function of the arrival rate λ_i at Queue i, the average total delay as a function of the splitting probability p is simply, using Little's law,

$$D(p) = N_t(p)/\lambda, \tag{3}$$

where $N_t(p)$ is the total average queue length as a function of the splitting probability $p \in [0,1]$ and reads

$$N_t(p) \doteq N_1(p\lambda) + N_2((1-p)\lambda). \tag{4}$$

For a given λ, minimizing the average total size of the queues and minimizing the average delay is thus equivalent.

Remember that $\hat{\mu}_i$ is the average service rate of Queue i (in this section, we can have $\hat{\mu}_1 \neq \hat{\mu}_2$). The natural *static* splitting probability (i.e., constant for all arrival rates λ) to use is

$$p_{\lim} = \frac{\hat{\mu}_1}{\hat{\mu}_1 + \hat{\mu}_2}. \tag{5}$$

It is easy to show that $p_{\lim}$ is the optimal static p, as it is the only static value of p that maintains the two queues stable for all arrival rates $\lambda < \hat{\mu}_1 + \hat{\mu}_2$. If $\lambda \geq \hat{\mu}_1 + \hat{\mu}_2$, no p maintains both queues stable. We now study the optimal splitting probability $p^*(\lambda)$, potentially different for each arrival rate $\lambda \in (0, \hat{\mu}_1 + \hat{\mu}_2)$.

4.2 Analysis of Optimal Splitting Probability

We start by noting that N_i is a strictly convex function of λ_i [23]. From (3), we know that, for a given λ, working with the average delays D and with the average total queue size N_t is equivalent, and that the delay is minimized when $N_t(p)$ given by (4) is minimized. When λ is given, N_t is minimized, either when $p=0$ or $p=1$, or when

$$N_t'(p) = \lambda N_1'(p\lambda) - \lambda N_2'((1-p)\lambda) = 0. \tag{6}$$

Let us assume first that $N_1'(0) < N_2'(0)$. Because $N_1(\lambda)$ has a vertical asymptote for $\lambda = \hat{\mu}_1$, $N_1'(\lambda) \to \infty$ when $\lambda \to \hat{\mu}_1$, i.e., there is a $\lambda_0 \in (0, \hat{\mu}_1)$ such that $N_1'(\lambda_0) = N_2'(0)$. Because N_1 is strictly convex and consequently N_1' is strictly increasing, λ_0 is unique. Then for all $\lambda \leq \lambda_0$, $N_t'(p) \leq 0$ for all $p \in [0,1]$, i.e., $N_t(p)$ is decreasing and $p^*(\lambda) = 1$. For $\lambda \in (\lambda_0, \hat{\mu}_1 + \hat{\mu}_2)$, (6) has a solution in (0,1) because $N_t'(0) = \lambda(N_1'(0) - N_2'(\lambda)) < \lambda(N_1'(0) - N_2'(0)) < 0$ and $N_t'(1) = \lambda(N_1'(\lambda) - N_2'(0)) > 0$, and because $N_t'(p_{\lim})$ is finite. This solution is unique because N_1 and N_2, and thus N_t, are strictly convex. For a given $\lambda \in [\lambda_0, \hat{\mu}_1 + \hat{\mu}_2)$, $p^*(\lambda)$ is then the unique solution of

$$N_1'(p^*\lambda) = N_2'((1-p^*)\lambda). \tag{7}$$

The optimal splitting probabiliy $p^*(\lambda)$ is therefore the function equal to 1 for each $\lambda \in (0, \lambda_0]$, and that associates to each $\lambda \in (\lambda_0, \hat{\mu}_1 + \hat{\mu}_2)$ the solution of (7). Note that $\lim_{\lambda \to \hat{\mu}_1 + \hat{\mu}_2} p^*(\lambda) = p_{\lim}$.

If $N_2'(0) < N_1'(0)$, everything is similar, except that $p^*(\lambda) = 0$ for $\lambda \in (0, \lambda_0]$. We have thus shown the following theorem.

Theorem 4.1. *If $N_1'(0) \neq N_2'(0)$, there is a $\lambda_0 \in (0, \hat{\mu}_1 + \hat{\mu}_2)$ such that for all $\lambda \in (0, \lambda_0]$, using a single queue yields smaller average delays than any traffic splitting between the two queues.*

For homogeneous service rates (i.e., when $\mu_{h,i} = \mu_{l,i} = \hat{\mu}_i$), $N_1'(0) = N_2'(0)$ **iff** the queues are identical ($\hat{\mu}_1 = \hat{\mu}_2$). For heterogeneous service rates, we can show with the notations of Section 3 that

$$N_i'(0) = \frac{\alpha_i \hat{\mu}_i + \pi_i + V_i}{\hat{\mu}_i(\alpha_i \hat{\mu}_i + \pi_i)}.$$

Although two non-identical queues can in theory have same value $N_i'(0)$, the set of parameters that meet $N_1'(0) = N_2'(0)$ has measure zero in the space of possible parameters for Queues 1 and 2. For this reason, two non-identical queues with parameters chosen at random are very unlikely to have $N_1'(0) = N_2'(0)$.

We first assume homogeneous service rates ($\mu_{h,i} = \mu_{l,i} = \hat{\mu}_i$) for both queues. Without loss of generality, we assume that Queue 1 is the queue with largest average service rate, i.e., $\hat{\mu}_1 > \hat{\mu}_2$. In the homogeneous case, we can obtain an explicit expression for $p^*(\lambda)$. We know that $N_i(\lambda) = \lambda/(\mu_i - \lambda)$, and a simple computation gives

$$\lambda_0 = \hat{\mu}_1 - \sqrt{\hat{\mu}_1 \hat{\mu}_2}. \tag{8}$$

Then, using (7), p^* is given for $\lambda \in (0, \hat{\mu}_1 + \hat{\mu}_2)$ by

$$p^*(\lambda) = 1 \textbf{ if } \lambda \in (0, \lambda_0],$$

$$p^*(\lambda) = \frac{\hat{\mu}_1(\lambda - 2\hat{\mu}_2) + (\hat{\mu}_1 + \hat{\mu}_2 - \lambda)\sqrt{\hat{\mu}_1 \hat{\mu}_2}}{\lambda(\hat{\mu}_1 - \hat{\mu}_2)} \textbf{ if } \lambda \in (\lambda_0, \hat{\mu}_1 + \hat{\mu}_2).$$

When the service rates are heterogeneous ($\mu_{h,i} \neq \mu_{l,i}$), the expression for N_i is too complex to provide an explicit expression of p^*. However, we have proven (Theorem 4.1) that there exists a $\lambda_0 > 0$ such that when $\lambda \leq \lambda_0$, $p^*(\lambda) = 0$ or $p^*(\lambda) = 1$: For low

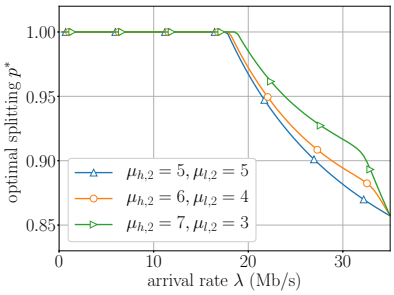

Figure 3: Optimal splitting p^* as a function of the arrival rate λ.

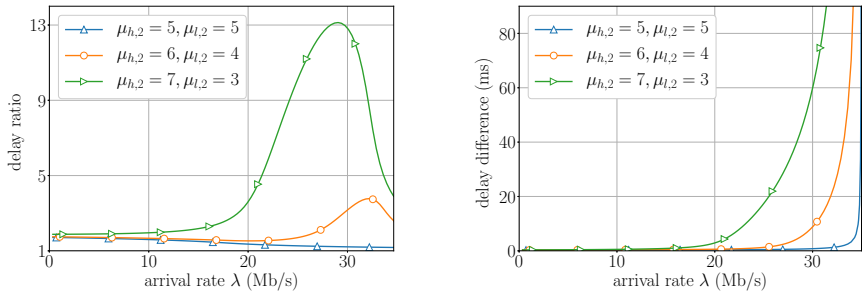

Figure 4: Ratio $D(p_{\text{lim}})/D(p^*)$ (left) and difference $D(p_{\text{lim}}) - D(p^*)$ (right) of the delays with static (p_{lim}) and optimal (p^*) splitting.

arrival rates, it is better in terms of delays to use only one path. In addition, p^* can be computed numerically using (7).

4.3 Numerical Results

In this section, we show numerically that in time-varying networks, the choice of p has a strong impact on the delays. In particular, we compare the delays obtained with $p^*(\lambda)$ and with p_{lim}. Queue 1 has an average service rate $\hat{\mu}_1 = 30$ Mb/s, and Queue 2 an average service rate $\hat{\mu}_2 = 5$ Mb/s. Packets have size 1 400 B, so that $\hat{\mu}_1 = 2\,678$ packets/s and $\hat{\mu}_2 = 446$ packets/s. Queue 1 has homogeneous service rates ($\mu_{h,1} = \mu_{l,1}$), and we study how the variability of Queue 2 affects delays. We set $\alpha_{h,2} = \alpha_{l,2} = 1$ transition/s, and we use three different sets of values for $\mu_{h,2}$ and $\mu_{l,2}$ with same average service rate $\hat{\mu}_2$: $\mu_{h,2} = \mu_{l,2} = 5$ Mb/s; $\mu_{h,2} = 6$ Mb/s and $\mu_{l,2} = 4$ Mb/s; $\mu_{h,2} = 7$ Mb/s and $\mu_{l,2} = 3$ Mb/s.

Figure 3 shows the optimal splitting p^* for the three sets of values for $\mu_{h,2}$ and $\mu_{l,2}$. When $\mu_{h,2} = \mu_{l,2} = 5$ Mb/s, both queues are homogeneous, and, as shown above, it is optimal to send all the traffic on Queue 1 when $\lambda < \lambda_0$ with $\lambda_0 \approx 17.8$ Mb/s given by (8). We see on Figure 3 that the effect of the variability ($\mu_{h,2} \neq \mu_{l,2}$) on λ_0 is quite small. However, the impact of the variability on the delays is very large. We show the ratio (Figure 4, left) and difference (Figure 4, right) of the packet delays when the traffic is split either with probability $p^*(\lambda)$ for an arrival rate λ, or with probability $p_{\text{lim}} \approx 0.86$ for all arrival rates. For example, when $\mu_{h,2} = 7$ Mb/s and $\mu_{l,2} = 3$ Mb/s and for an arrival rate $\lambda = 19$ Mb/s, $p^* = 1$, and the delay obtained by sending everything on Queue 1 is about 1 ms; when traffic is split with probability p_{lim}, the delay is 2.8 ms. For higher arrival rates, the delay ratio can be up to 13: For $\lambda = 29$ Mb/s, the delay with p^* is 3.9 ms, and the delay obtained with p_{lim} is 51 ms. Section 5.3 presents evidence of this behavior on a wireless testbed.

In our analysis, we do not take reordering into account. A careful study of reordering is out of the scope of this paper and is left for future work, but reordering is likely to further increase the rate λ_0 before which using a single queue is preferable. Indeed, sending at a low rate on the second queue might harm delays (because an additional delay is required for packets to be ordered) more than the gains it offers.

5 EXPERIMENTAL RESULTS

We now show numerical and testbed results that support our analyses. We first illustrate the impact on delays of the service rate variability in a set of uncontrolled experiments. Next, in a set of controlled experiments, we show that, as unveiled by the analysis of Section 3, smaller variance sometimes yields larger delays. Finally, we show how the traffic splitting between two paths impacts delays and we show that, as proven in Section 4, using a single path is better for low arrival rates.

5.1 Variability Impacts Delays

We first illustrate the impact that variability has on delays in wireless networks. We carry an experiment between two nodes with two WiFi interfaces (Atheros AR9280) and one antenna per interface. The nodes are APU1D boards with an OpenWrt distribution patched for MPTCP [1] and ath9k wireless drivers. They are synchronized with the PTP protocol that offers a precision of a few microseconds (minimum delays are of the order of the millisecond). The two nodes are approximately 15 meters apart in two different offices, with two walls between them. Both WiFi interfaces use the 802.11n protocol; the first WiFi interface operates in the 2.4 GHz band, the second interface operates in the 5 GHz band, with a 20 MHz band for each. Lower frequencies are known to be less attenuated by walls than higher frequencies, and we observe that the maximum instantaneous rate between the two nodes is about 35 Mb/s in the 5 GHz band, whereas it is about 45 Mb/s in the 2.4 GHz band. However, in the building where the experiments take place, the 2.4 GHz band is also used by the WiFi network of the university, whereas no other node uses the 5 GHz band. When other nodes use the university WiFi network, the throughput on the 2.4 GHz link decreases. Consequently, the variability is larger in the 2.4 GHz band.

We run our experiments on a weekday, when the WiFi network of the university is more loaded. We send UDP traffic during 30 seconds at various rates with `iperf`, first in the 2.4 GHz band, then in the 5 GHz band, and we measure the one-way delay of the packets by using `tcpdump` on each interface. The experiment is repeated five times for each interface and each rate, and we present averaged results. Figure 5 (left) shows the receiving rate; the average throughput achieved in the 2.4 GHz band (about 38 Mb/s) is slightly higher than that achieved in the 5 GHz band, but it is below the maximum instantaneous throughput of 45 MB/s. This shows that the variability in the 2.4 GHz band is indeed quite large. In contrast, the throughput achieved in the 5 GHz is close to its maximum instantaneous throughput. Figure 5 (right) shows the average packet

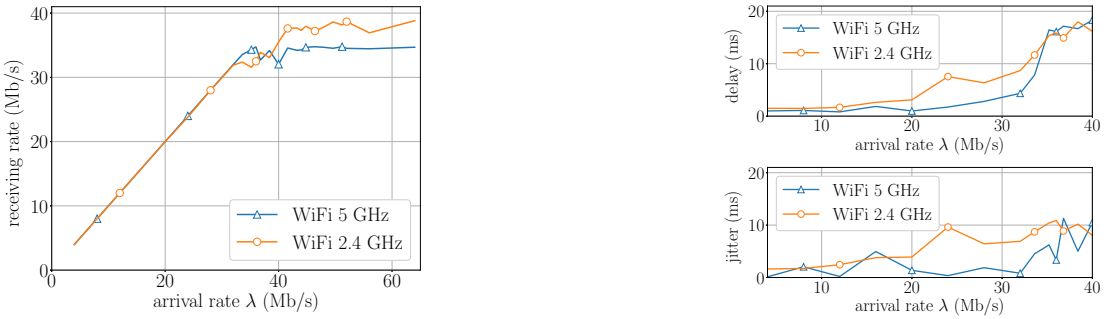

Figure 5: Throughput achieved (left), delays (top right), and jitter (bottom right) on a testbed with 2.4 GHz and 5 GHz WiFi.

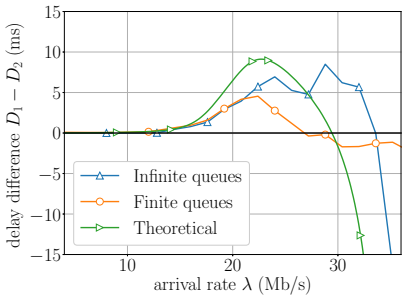

Figure 6: Delay difference $D_1 - D_2$ between Queue 1 (smaller variance) and Queue 2. Analytical and testbed results.

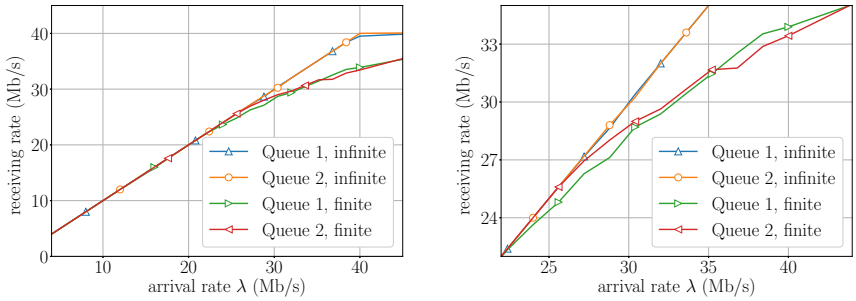

Figure 7: Experimental receiving rate for the two queues on a wireless testbed, with finite and infinite queues. Left: All arrival rates. Right: Zoom on some arrival rates.

delay (top) and jitter (bottom). Even though the average throughput is slightly higher in the 2.4 GHz band, delay and jitter both are smaller in the 5 GHz band (delay is up to 4.5x smaller when the arrival rate is 24 Mb/s). This illustrates that, as expected, variability has a high impact on the delays. Here, a larger variance yields larger delays. We now show that this is not always the case, as unveiled by the analysis of Section 3.

5.2 Larger Variance Can Imply Smaller Delays

In contrast to the experiments of Section 5.1, where we did not have control over the sources of variability, we want in the following to be able to control the parameters $\mu_{h,i}$, $\mu_{l,i}$, $\alpha_{h,i}$, and $\alpha_{l,i}$, in order to compare the experimental results with the analytical results. This is achieved with WiFi by setting the modulation and coding scheme (MCS) used by the interface. The ath9k driver enables us to set the MCS to the desired value. To avoid external sources of variability, we run all the experiments of this section in the 5 GHz band, not used by any other user. The analysis assumes infinite queues. In practice, the queues of the network interfaces are finite. We run our experiments in two modes: Either packets are queued in an infinite queue at the application level, by using the Click Modular Router [19]; or we use the default finite queues of the network interfaces.

We use the following values, with the service rates (μ) in packets/s, and the transition rates (α) in transitions/s: $\mu_{h,1} = 4\,196$ (MCS 5, about 46 Mb/s), $\mu_{l,1} = 1\,562$ (MCS 2, about 17.5 Mb/s), $\mu_{h,2} = 4\,687$ (MCS 6, about 52 Mb/s), $\mu_{l,2} = 2\,089$ (MCS 3, about 23 Mb/s), $\alpha_{h,1} = 2.92$, $\alpha_{l,1} = 7.01$, $\alpha_{h,2} = 4.84$, and $\alpha_{l,2} = 5.16$. These values are chosen randomly among some for which the smaller delays are achieved by the queue with larger variance for certain arrival rates, as shown by Corollary 3.4. More precisely, we choose values such that (18) in Appendix is verified. We have $\hat{\mu}_1 = \hat{\mu}_2 = 38.4$ Mb/s, $V_1 = 1.33 \times 10^6$, and $V_2 = 1.67 \times 10^6$: Queue 1 is the queue with smallest variance. The theoretical delay difference $D_1 - D_2$ obtained from (9) in Appendix is shown in Figure 6. For arrival rates lower than about 30 Mb/s, the queue with smallest delays is Queue 2, the queue with largest variance.

We carry experiments with the two nodes described in Section 5.1, now in the same office. We send UDP traffic during 20 s on each queue and at various rates, and we measure the packet delays. Sending traffic on Queue i means that the WiFi interface switches between the MCS that corresponds to $\mu_{h,i}$ and $\mu_{l,i}$, with exponentially-distributed sojourn times with respective parameters $\alpha_{h,i}$ and $\alpha_{l,i}$. For all the experiments in Section 5, no assumption is made on the distribution of the service times of the packets, and only the average service rate is set; as opposed to the analytical part where they were assumed to be exponentially distributed, the service times depend on the wireless interface and their distribution is unknown. For the packet arrivals, we set the average arrival rate and we try exponential and deterministic distributions, but the choice has no effect on the delays. The results are shown with a deterministic distribution (i.e., the inter-arrival times between packets are constant). The experiments are repeated five times and we

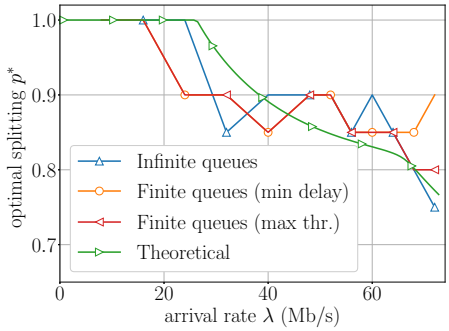

Figure 8: Optimal splitting probability $p^*(\lambda)$.

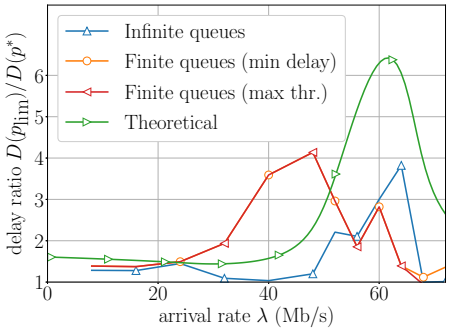

Figure 9: Ratio $D(p_{\text{lim}})/D(p^*)$ of the delays obtained with the static probability p_{lim} and with the optimal splitting p^*.

present averaged results. We first check that the average receiving rate is the same for both queues, which is the case (see Figure 7, left). As expected, it converges to approximately $\hat{\mu} = 38.4$ Mb/s for the infinite queues. The experimental average delay-difference is shown in Figure 6. It matches very well the analytical results, which shows that the queue with largest variance can indeed offer the smallest delays for certain arrival rates. This also shows that the best queue in terms of delays depends on the arrival rate λ. We note that the delay difference is, as expected, larger when the queues are infinite, and is significant for certain arrival rates: When $\lambda = 22.4$ Mb/s, the delay is 3x larger when using Queue 1, the queue with smallest variance (8.7 ms vs. 2.9 ms).

When the queues are finite, some packets are discarded when the queues are too long. In this case, variability has two consequences: larger delays and lower receiving rate. Delays can be observed in Figure 6: For example, when the arrival rate is around 23 Mb/s, Queue 1, the queue with the smallest variance, has a larger average delay (7.7 ms vs. 3.1 ms). For low arrival rates, Queue 1 not only has larger delays, but it also has lower receiving rates. This can be observed in Figure 7 (right), that shows a zoom on some arrival rates of Figure 7 (left). When the arrival rate is around 25 Mb/s, Queue 1 has a receiving rate of 24 Mb/s, vs. 25 Mb/s for Queue 2. When the arrival rate is around 35 Mb/s, the order gets reversed: Queue 1 has smaller delays and a higher receiving rate.

5.3 Simultaneous Paths in a Hybrid Network

We now study experimentally the impact of the parameter p described in Section 4. Our hybrid network consists in Queue 1, a WiFi interface in the 5 GHz band, and Queue 2, a WiFi interface in the 2.4 GHz band. In this section, to avoid external sources of variability, especially in the 2.4 GHz band used by the WiFi network of the university, we run all the experiments at night. We consider a simple and realistic case where the two queues have different average service rates. Queue 1 has homogeneous service rates: $\mu_{h,1} = \mu_{l,1} = 5\,000$ packets/s (MCS 7, about 56 Mb/s). Queue 2 has variable service rates: $\mu_{h,2} = 2\,053$ (MCS 3, about 23 Mb/s) and $\mu_{l,2} = 1\,071$ (MCS 1, about 12 Mb/s). The transition rates for Queue 2 are $\alpha_{h,2} = \alpha_{l,2} = 3$ transitions/s. We send UDP traffic during 20 seconds for various arrival rates λ and for all probabilities p between 0 and 1, with 0.05 increments. The experiments are repeated three times for each λ and p. Figure 8 shows the theoretical and experimental values for p^*. For the infinite queues, the experimental p^* is found for each arrival rate as the p that yields the smallest average delays. For the finite queues, there might be packet losses as shown in Section 5.2, and we measure both the splitting probability p that yields the smallest average delays and the splitting probability p that yields the maximal received throughput (we note that the two are equal, except when the arrival rate is very close to the saturation rate). The experimental values are close to the theoretical ones. The optimal static p, as defined by (5), is $p_{\text{lim}} \approx 0.75$. Figure 9 shows the ratio of the delays obtained with p_{lim} over the delays obtained with $p^*(\lambda)$. For certain arrival rates, the negative impact of a static splitting probability is quite strong: With infinite queues, for $\lambda = 64$ Mb/s, the delay is 6 ms with $p = p^*$, whereas it is 25 ms with $p = p_{\text{lim}}$. The impact is strong with finite queues as well: For $\lambda = 48$ Mb/s, the delay is 2 ms with $p = p^*$, whereas it is 9 ms with $p = p_{\text{lim}}$.

In contrast, when we set $\mu_{h,2} = \mu_{l,2} = 1\,562$ (MCS 2, about 17.5 Mb/s), i.e., when the two queues have homogeneous service rates, using p_{lim} instead of p^* has a small effect on delays: The maximum delay-ratio is around 1.5x for analytical and experimental results, and the difference is always less than a millisecond (the results are not shown due to lack of space).

Finally, we compare the results obtained with our model with the results obtained with single-path TCP (referred as TCP) and multipath TCP (referred as MPTCP). Because TCP and MPTCP are by default designed for favoring throughput, we do some modifications in order to favor delays, which reduces the delay at low throughput from about 2 ms to about 0.7 ms. This is the same value as with UDP traffic. We keep using $\mu_{h,1} = \mu_{l,1} = 5\,000$, $\mu_{h,2} = 2\,053$, $\mu_{l,2} = 1\,071$, and $\alpha_{h,2} = \alpha_{l,2} = 3$. TCP is used on Queue 1 only, the queue with higher average service rate (about 50 Mb/s with TCP). For MPTCP, the default scheduler is used. For each rate, the experiment is repeated five times and we present averaged results. Figure 10 shows the probability that traffic with MPTCP is sent on Queue 1, along with p^* found experimentally with UDP traffic. For low arrival rates, MPTCP sends more traffic on the second queue when compared to p^*, the optimal splitting in our model. Figure 11 shows the delays obtained with the different protocols. UDP is shown for reference, but a fair comparison with TCP or MPTCP is difficult, because the performance of TCP/MPTCP in terms of delays is very dependent on the configuration (scheduler, congestion window, slow start, etc.). It is more interesting to compare TCP and

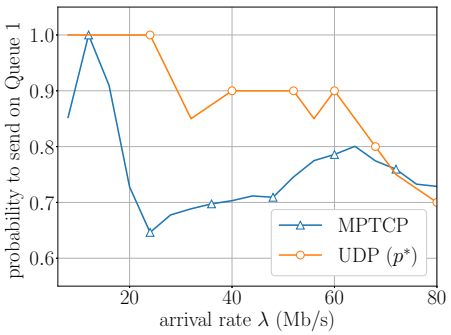

Figure 10: Probability to send on Queue 1.

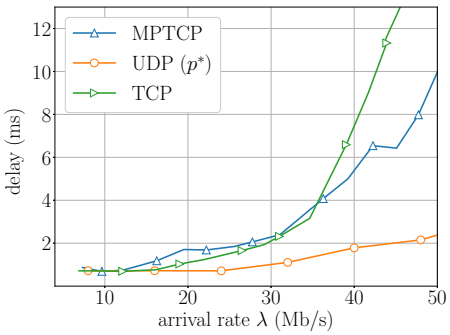

Figure 11: Delays obtained with TCP, MPTCP, and UDP (p^*).

MPTCP, because they use the same configurations. For low arrival rates, TCP (i.e., single-path) yields slightly smaller delays, whereas MPTCP reduces the delays for higher arrival rates. This confirms our analysis and previous experiments: For low arrival rates, it is better in terms of delays to use a single path.

6 RELATED WORK

Queueing Models. Time-varying queueing models were first studied in 1956 by Clarke [10]. The model described in Section 2 was introduced and solved for the first time by Yechiali and Naor [32]. It was then studied with more general assumptions [15] or solved with different techniques [24, 26]. These works give powerful tools to derive numerical results, but only a few give intuitive insights and fundamental properties of the delays generated by this queue model. The works that do so study the effect on delays of the transition rates α [13, 16, 28, 29]. Other works have studied M/G/1 or MAP/G/1 queues with correlated service times [5, 21, 25]; they find recursive equations for the delays and solve them numerically, but they do not give intuitive insights and fundamental properties of the average delays. To the best of our knowledge, our work is the first that shows that with the same average service rate, a largest variance can yield lower average delays.

A model with two queues in parallel was introduced in 1958 by Haight [14]. A large number of works study models that assume identical servers with homogeneous service rates ($\mu_h = \mu_l$), when the packets are routed based on the current queue sizes [6, 31]. These models were extended to support non-identical servers, still with homogeneous service rates [17, 22]. They show that the optimal decision is threshold-based, i.e., packets are routed to the fastest queue, unless the queue-size difference is above some threshold. To the best of our knowledge, this paper is the first work that studies queues in parallel with heterogeneous time-varying services rates. Here, we assume that packets are routed based on a Bernoulli trial. Only a few models study Bernoulli routing, either with identical servers and homogeneous service rates [7], or by only looking at the case where the parameter p is the same for all arrival rates [17].

Delays in Hybrid Networks. With the recent development of hybrid networks, much attention has been recently given on knowing if multipath, in particular MPTCP, could help reduce delays. When the characteristics of the two paths are very different, in particular in terms of RTT (e.g., with LTE and WiFi), MPTCP is found to increase slightly the delays [8, 9]. When the characteristics of the two paths are close, MPTCP can reduce delays significantly [33]. In this paper, we find in addition that it depends on the arrival rates: For low arrival rates, it is usually better to use a single path, whereas it is better to use two paths for larger arrival rates. Finally, MPTCP can also be used to improve performance by sending redundant messages on the two paths [11]. This reduces the delays, but at the cost of higher utilization. This model with redundant messages is out of the scope of our paper.

7 CONCLUSION

In this paper, we have studied delays in time-varying networks such as wireless networks with varying signals and concurrent users, data centers with high-rate and low-rate periods, etc. The variability of the service rate severely impacts packet delays, as illustrated by the experiments we carried on a wireless testbed. Based on a queue model with heterogeneous time-varying service rates, we have shown that for a given average service rate, the queue that offers the smallest delays is not necessarily the same for all arrival rates, and that the queue with the largest variance sometimes yields the smallest delays. These results obtained with the time-varying queuing model of Section 2 are supported by experiments carried out on a wireless testbed, which have shown that the delay gain provided by using the queue with the largest variance can be significant. We have then studied the conditions under which using simultaneously two independent paths (for example, two different technologies in hybrid networks, e.g., WiFi and LTE or WiFi and PLC) reduces the delays, compared to using a single path. We have shown through analysis and testbed experiments that the optimal splitting between the two paths depends on the arrival rate, and that for low arrival rates, using a single path is usually better than using two. Numerically and experimentally, we have shown that the larger the variability, the higher the impact of the splitting decision on the delays.

REFERENCES

[1] https://multipath-tcp.org/pmwiki.php/Users/OpenWRT.
[2] https://wiki.openwrt.org/doc/networking/praxis.
[3] IEEE 1905.1-2013 Standard for Heterogeneous Technologies. 2013.
[4] IMT-Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond. *Recommendation ITU-R*, 2015.
[5] I. J.-B. F. Adan and V. G. Kulkarni. Single-Server Queue with Markov-Dependent Inter-Arrival and Service Times. *Queueing Systems*, 2003.

[6] O. T. Akgun, R. Righter, and R. Wolff. Multiple-server System with Flexible Arrivals. *Advances in Applied Probability*, 2011.
[7] C.-S. Chang, X. Chao, and M. Pinedo. A Note on Queues with Bernoulli Routing. In *IEEE CDC*, 1990.
[8] Y.-C. Chen, Y.-s. Lim, R. J. Gibbens, E. M. Nahum, R. Khalili, and D. Towsley. A Measurement-based Study of Multipath TCP Performance over Wireless Networks. In *ACM IMC*, 2013.
[9] Y.-C. Chen and D. Towsley. On Bufferbloat and Delay Analysis of Multipath TCP in Wireless Networks. In *IFIP Networking*, 2014.
[10] A. B. Clarke. A Waiting Line Process of Markov Type. *The Annals of Mathematical Statistics*, 1956.
[11] A. Frömmgen, T. Erbshäußer, Buchmann, T. Zimmermann, and K. Wehrle. ReMP TCP: Low Latency Multipath TCP. In *IEEE ICC*, 2016.
[12] K. Gardner, M. Harchol-Balter, A. Scheller-Wolf, M. Velednitsky, and S. Zbarsky. Redundancy-d: The Power of d Choices for Redundancy. *Operations Research*, 2017.
[13] V. Gupta, M. Harchol-Balter, A. Wolf, and U. Yechiali. Fundamental Characteristics of Queues with Fluctuating Load. In *ACM Sigmetrics*, 2006.
[14] F. A. Haight. Two Queues in Parallel. *Biometrika*, 1958.
[15] P. G. Harrison and H. Zatschler. Sojourn time distributions in modulated g-queues with batch processing. In *IEEE QEST*, 2004.
[16] D. Heyman. On Ross's Conjectures about Queues with Non-stationary Poisson Arrivals. *Journal of Applied Probability*, 1982.
[17] E. Hyytiä. Optimal Routing of Fixed Size Jobs to Two Parallel Servers. *INFOR: Information Systems and Operational Research*, 2013.
[18] A. Izagirre and A. M. Makowski. Light traffic performance under the power of two load balancing strategy: The case of server heterogeneity. *ACM Sigmetrics Performance Evaluation Review*, 2014.
[19] E. Kohler, R. Morris, B. Chen, J. Jannotti, and M. F. Kaashoek. The Click Modular Router. *ACM TOCS*, 2000.
[20] J. F. Kurose and K. W. Ross. *Computer Networking: A Top-down Approach.* 2009.
[21] J. Lambert, B. Van Houdt, and C. Blondia. Queues with Correlated Service and Inter-Arrival Times and Their Application to Optical Buffers. *Stochastic Models*, 2006.
[22] R. L. Larsen. *Control of Multiple Exponential Servers with Application to Computer Systems.* PhD thesis, University of Maryland, 1981.
[23] L. Liyanage and J. Shanthikumar. Second-order Properties of Single-stage Queueing Systems. *Oxford Statistical Science Series*, 1992.
[24] S. R. Mahabhashyam and N. Gautam. On Queues with Markov Modulated Service Rates. *Queueing Systems*, 2005.
[25] C. Mitchell, A. Paulson, and C. Beswick. The Effect of Correlated Exponential Service Times on Single Server Tandem Queues. *Naval Research Logistics*, 1977.
[26] M. F. Neuts. *The M/M/1 Queue with Randomly Varying Arrival and Service Rates.* PhD thesis, Delaware University, 1977.
[27] C. Raiciu, C. Paasch, S. Barre, A. Ford, M. Honda, F. Duchene, O. Bonaventure, and M. Handley. How Hard Can It Be? Designing and Implementing a Deployable Multipath TCP. In *USENIX NSDI*, 2012.
[28] T. Rolski. Queues with Non-stationary Input Stream: Ross's Conjecture. *Advances in Applied Probability*, 1981.
[29] S. M. Ross. Average Delay in Queues with Non-stationary Poisson Arrivals. *Journal of Applied Probability*, 1978.
[30] C. Vlachou, S. Henri, and P. Thiran. Electri-Fi Your Data: Measuring and Combining PLC with WiFi. In *ACM IMC*, 2015.
[31] W. Winston. Optimality of the Shortest Line Discipline. *Journal of Applied Probability*, 1977.
[32] U. Yechiali and P. Naor. Queuing Problems with Heterogeneous Arrivals and Service. *Operations Research*, 1971.
[33] K. Yedugundla, S. Ferlin, T. Dreibholz, Ö. Alay, N. Kuhn, P. Hurtig, and A. Brunstrom. Is Multi-path Transport Suitable for Latency Sensitive Traffic? *Computer Networks*, 2016.

A APPENDIX

A.1 Proofs of Section 3

We study which queue has the largest average delays, i.e., the sign of $D_1 - D_2$. From Little's law, we know that working with the average delays D_i and with the average queue size N_i is equivalent, because $N_i = \lambda D_i$, hence, $D_1 - D_2$ and $N_1 - N_2$ have same sign. The average queue size of a queue with heterogeneous service rates is given by Yechiali and Naor [32]. After some manipulations, it can be rewritten for Queue i for $i \in \{1, 2\}$ and for any $\lambda < \hat{\mu}_i$ as

$$N_i(\lambda) = \frac{\lambda}{\hat{\mu}_i - \lambda} + \frac{\lambda}{\alpha_i(\hat{\mu}_i - \lambda)} f_i(\lambda) V_i, \tag{9}$$

where we use the notations of Section 3 and where

$$f_i(\lambda) = \frac{1 - z_i(\lambda)}{\hat{\mu}_i - \lambda z_i(\lambda)}$$

with $z_i(\lambda)$ defined as the only root in (0,1) of

$$g_i(z) = \alpha_i z(\hat{\mu}_i - \lambda z) - (1 - z)(\mu_{h,i} - \lambda z)(\mu_{l,i} - \lambda z). \tag{10}$$

We also define $\sigma_i = \mu_{h,i} + \mu_{l,i}$ and

$$M = \frac{\pi_2 - \pi_1}{\sigma_2 - \sigma_1}. \tag{11}$$

We start with useful lemmas. Remember that $\alpha_1 = \alpha_2 = \alpha$ and $\hat{\mu}_1 = \hat{\mu}_2 = \hat{\mu}$.

Lemma A.1. *$z_i(\lambda)$ is a decreasing function of λ in $(0, \hat{\mu})$.*

Proof. z_i is the only root of $g_i(z)$ in (0,1). We write $g_i(z, \lambda)$ for the function g_i, making its dependency on λ explicit. Let $\epsilon > 0$. After some manipulations using $g_i(z_i, \lambda) = 0$, we have $g_i(z_i, \lambda + \epsilon) = \epsilon z_i k_i(z, \epsilon)$, where

$$k_i(z, \epsilon) = \sigma_i - (\alpha + \epsilon + 2\lambda + \sigma_i) z_i + (\epsilon + 2\lambda) z_i^2.$$

If $k_i(z_i, 0) > 0$, then for $\epsilon > 0$ small enough, $k_i(z_i, \epsilon) > 0$, and $g_i(z_i, \lambda + \epsilon) = \epsilon z_i k(z_i, \epsilon) > 0$, i.e., increasing λ increases g_i around z_i, and thus increasing λ decreases the root z_i, because g_i is increasing near z_i as $g_i(0) = -\pi_i < 0$ and $g_i(1) = \alpha(\hat{\mu} - \lambda) > 0$. Now, $k_i(z_i, 0) = (1 - z_i)(\sigma_i - 2\lambda z_i) - \alpha z_i$, which for $z_i \in (0, 1)$ has same sign as

$$\begin{aligned} Q_i \doteq \frac{k(z_i, 0)}{1 - z_i} &= (\sigma_i - 2\lambda z_i) - \frac{\alpha z_i}{1 - z_i} \\ &= (\sigma_i - 2\lambda z_i) - \frac{(\mu_{h,i} - \lambda z_i)(\mu_{l,i} - \lambda z_i)}{\hat{\mu} - \lambda z_i}, \end{aligned}$$

where for the last equality we used (10) and $g_i(z_i, \lambda) = 0$. If $\mu_{h,i} = \mu_{l,i}$, then $\mu_{h,i} = \hat{\mu}$ and we have $Q_i = \hat{\mu} - \lambda z_i > 0$. Otherwise, necessarily $\mu_{h,i} > \hat{\mu}$ and $\mu_{l,i} < \hat{\mu}$. Then

$$\begin{aligned} Q_i &= (\mu_{h,i} - \lambda z_i) + (\mu_{l,i} - \lambda z_i) - \frac{(\mu_{h,i} - \lambda z_i)(\mu_{l,i} - \lambda z_i)}{\hat{\mu} - \lambda z_i} \\ &= (\mu_{h,i} - \lambda z_i) \left(1 - \frac{\mu_{l,i} - \lambda z_i}{\hat{\mu} - \lambda z_i}\right) + (\mu_{l,i} - \lambda z_i). \end{aligned}$$

Using (10) again, we know that $\mu_{l,i} - \lambda z_i > 0$. Also, $\mu_{l,i} < \hat{\mu}$, consequently $k(z_i, 0) > 0$, which concludes the proof. □

Lemma A.2. *If $\sigma_1 = \sigma_2$ and $\pi_1 = \pi_2$, then $g_1 = g_2$. If $\sigma_1 = \sigma_2$ and $\pi_1 \neq \pi_2$, the only root of $g_1(z) - g_2(z)$ is 1. If $\sigma_1 \neq \sigma_2$, the only roots of $g_1(z) - g_2(z)$ are 1 and M/λ.*

Proof. This follows directly from (10). □

Lemma A.3. *If there is no root of $g_1(z) - g_2(z)$ in (0,1), then $z_1 > z_2$ **iff** either $\pi_1 > \pi_2$, or $\pi_1 = \pi_2$ and $\sigma_1 < \sigma_2$.*

Proof. If $g_1(z) - g_2(z)$ has no root in (0,1) and because, g_i is increasing near z_i, we have $z_1 > z_2$ **iff** $\forall z \in (0, 1)$, $g_1(z) < g_2(z)$. If $\pi_1 \neq \pi_2$, $g_1(z) < g_2(z)$ **iff** $\pi_1 > \pi_2$, because $g_i(0) = -\pi_i$. If $\pi_1 = \pi_2$, we have $g_1 - g_2 = \lambda z(1 - z)(\sigma_1 - \sigma_2)$, i.e., $g_1(z) < g_2(z)$ **iff** $\sigma_1 < \sigma_2$. □

Lemma A.4. *For all $\lambda \in (0, \hat{\mu})$, $f_1(\lambda) < f_2(\lambda)$ **iff** $z_1 > z_2$.*

Proof. Easy computation with $\lambda < \hat{\mu}$ and $0 < z_i < 1$. □

Lemma A.5. $(V_1-V_2)(\pi_1-\pi_2) < 0$ ***iff*** *either* $\sigma_1 = \sigma_2$, *or* $\hat{\mu}/M < 1$. *Also,* $V_1 = V_2$ ***iff*** *either* $M = \hat{\mu}$, *or* $\sigma_1 = \sigma_2$ *and* $\pi_1 = \pi_2$.

Proof. After some manipulations, we can rewrite

$$V_i = \hat{\mu}\sigma_i - \pi_i - \hat{\mu}^2, \tag{12}$$

and the cases of equality become clear. It is also clear with simple manipulations that if $V_1 < V_2$ and $\pi_1 > \pi_2$, or $V_1 > V_2$ and $\pi_1 < \pi_2$, then $\hat{\mu}/M < 1$.

Reciprocally, if $\hat{\mu}/M < 1$, then either $\hat{\mu} < M$ and $M > 0$, or $\hat{\mu} > M$ and $M < 0$. In the first case, we have

$$\hat{\mu} < \frac{\pi_2 - \pi_1}{\sigma_2 - \sigma_1}. \tag{13}$$

Either $\sigma_1 > \sigma_2$ and we have $\pi_1 > \pi_2$ (because $M > 0$) and $V_1 < V_2$ (because of (13)) and thus $(V_1 - V_2)(\pi_1 - \pi_2) < 0$; or $\sigma_1 < \sigma_2$ and we have similarly $\pi_1 < \pi_2$ and $V_1 > V_2$. Similar manipulations show the result in the second case. □

Lemma A.6. *If* $V_1 \neq V_2$, *there exists a* $\lambda_m \in [0, \hat{\mu})$ *such that for all* $\lambda \in (\lambda_m, \hat{\mu})$, $N_1(\lambda) > N_2(\lambda)$ ***iff*** $V_1 > V_2$.

Proof. We have

$$\frac{f_1(\hat{\mu})}{f_2(\hat{\mu})} = 1, \tag{14}$$

consequently there is a $\lambda_m \in [0, \hat{\mu})$ such that for all $\lambda \in (\lambda_m, \hat{\mu})$, $\left|1 - \frac{f_1(\lambda)}{f_2(\lambda)}\right| < \left|\frac{V_1 - V_2}{2}\right|$, which shows that for $\lambda \in (\lambda_m, \hat{\mu})$, $N_1(\lambda) - N_2(\lambda)$ has same sign as $V_1 - V_2$. □

Lemma A.7. *Let* $\beta_i = \alpha_{l,i}/\alpha$. *If* $\beta_1 = \beta_2 \doteq \beta$ *and if* $\beta\alpha_{h,i} \geq (1-\beta)\alpha_{l,i}$, *then* $V_1 \geq V_2$ ***iff*** $\pi_1 \leq \pi_2$.

Proof. Let $v_i = \mu_{h,i} - \mu_{l,i}$. We have after easy computations $V_i = \beta(1-\beta)v_i^2$, thus $V_1 > V_2$ **iff** $v_1 > v_2$. After computations, we get $\pi_i = \hat{\mu}^2 + (1-2\beta)\hat{\mu}v_i - \beta(1-\beta)v_i^2$, hence

$$\pi_1 - \pi_2 = -(v_1 - v_2)(\beta(1-\beta)(v_1 + v_2) - (1-2\beta)\hat{\mu}).$$

If

$$\beta(1-\beta)(v_1 + v_2) \geq (1-2\beta)\hat{\mu}, \tag{15}$$

then $V_1 \geq V_2$ **iff** $\pi_1 \leq \pi_2$. After simplifications, we see that (15) is equivalent to $\beta(\mu_{h,1} + \mu_{h,2}) > \hat{\mu}$. This is true if, for $i \in \{1, 2\}$, $\beta\alpha_{h,i} \geq (1-\beta)\alpha_{l,i}$, which concludes the proof. □

We can now prove the theorems.

Proof of Theorem 3.1. From (9), we write

$$N_1(\lambda) - N_2(\lambda) = \frac{\lambda}{\alpha(\hat{\mu} - \lambda)}\left(f_1(\lambda)V_1 - f_2(\lambda)V_2\right), \tag{16}$$

If $V_1 = V_2$ then from Lemma A.5, $M = \hat{\mu}$, or $\sigma_1 = \sigma_2$ and $\pi_1 = \pi_2$. If $\pi_1 = \pi_2$, then $\sigma_1 = \sigma_2$ and thus $g_1 = g_2$ and $N_1 = N_2$. Otherwise, $M = \hat{\mu}$ and because $\lambda < \hat{\mu}$, Lemma A.2 shows that there is no root of $g_1 - g_2$ in (0,1), and thus Lemma A.3 and A.4 along with (16) show that $N_1 > N_2$ **iff** $\pi_1 < \pi_2$.

If $\pi_1 = \pi_2$, then $V_1 = V_2$ is equivalent to $\sigma_1 = \sigma_2$ because of (12). If $\sigma_1 = \sigma_2$, we have proved $N_1 = N_2$. If $V_1 \neq V_2$, then $\sigma_1 \neq \sigma_2$ and $M = 0$, consequently there is no root of $g_1 - g_2$ in (0,1) (Lemma A.2), and thus from Lemmas A.3 and A.4, $f_1(\lambda) > f_2(\lambda)$ for all $\lambda \in (0, \hat{\mu})$ **iff** $\sigma_1 > \sigma_2$, which happens **iff** $V_1 > V_2$. Thus $N_1 > N_2$ **iff** $V_1 > V_2$. □

Proof of Theorem 3.2. From Lemma A.5, $(V_1 - V_2)(\pi_1 - \pi_2) < 0$ **iff** $\sigma_1 = \sigma_2$ or $\hat{\mu}/M < 1$. In both cases, Lemma A.2 shows that there is no root of $g_1 - g_2$ in (0,1), and consequently, from Lemmas A.3 and A.4, $f_1(\lambda) > f_2(\lambda)$ for all $\lambda \in (0, \hat{\mu})$ **iff** $\pi_1 < \pi_2$, which happens **iff** $V_1 > V_2$. Thus for all $\lambda \in (0, \hat{\mu})$, $N_1(\lambda) > N_2(\lambda)$ **iff** $V_1 > V_2$. □

Theorem 3.3 and Corollary 3.4 are direct consequences of the following theorem.

Theorem A.8. *(1) If for all* $\lambda \in (0, \hat{\mu})$, $N_1(\lambda) \neq N_2(\lambda)$, *then for all* $\lambda \in (0, \hat{\mu})$, $N_1(\lambda) > N_2(\lambda)$ ***iff*** $V_1 > V_2$.
(2) If

$$(V_1 - V_2)(\pi_1 - \pi_2) > 0. \tag{17}$$

and

$$\left|1 - \frac{V_2}{V_1}\right| < \left|1 - \frac{\pi_2 + \alpha\hat{\mu}}{\pi_1 + \alpha\hat{\mu}}\right| \doteq B_0, \tag{18}$$

then $N_1(\lambda) - N_2(\lambda)$ *changes sign in* $(0, \hat{\mu})$, *and there is a* $\lambda_0 \in (0, \hat{\mu})$ *such that for all* $\lambda < \lambda_0$, $V_1 > V_2$ *and* $N_1(\lambda) < N_2(\lambda)$, *or* $V_1 < V_2$ *and* $N_1(\lambda) > N_2(\lambda)$.
(3) If (17) is verified and

$$\frac{\max(V_1, V_2)}{\min(V_1, V_2)} > 1 + \frac{\max(\pi_1, \pi_2)}{\alpha\hat{\mu}}, \tag{19}$$

then for all $\lambda \in (0, \hat{\mu})$, $N_1(\lambda) \neq N_2(\lambda)$.

Proof. (1) This follows directly from Lemma A.6.
(2) A simple computation using (10) with $\lambda = 0$ shows that

$$z_i(0) = \frac{\pi_i}{\pi_i + \alpha\hat{\mu}} \text{ and } \frac{f_1(0)}{f_2(0)} = \frac{\pi_2 + \alpha\hat{\mu}}{\pi_1 + \alpha\hat{\mu}}.$$

So, when λ is close to 0, the difference $N_1(\lambda) - N_2(\lambda)$ has same sign as $f_1(0)V_1 - f_2(0)V_2$, i.e., it has same sign as

$$Q \doteq \frac{f_1(0)}{f_2(0)} - \frac{V_2}{V_1} = \frac{\pi_2 + \alpha\hat{\mu}}{\pi_1 + \alpha\hat{\mu}} - \frac{V_2}{V_1}.$$

If (17) and (18) hold, $Q > 0$ **iff** $V_1 < V_2$: When λ is close to 0, $N_1(\lambda) > N_2(\lambda)$ **iff** $V_1 < V_2$. From Lemma A.6, we know that when λ is close to $\hat{\mu}$, $N_1(\lambda) > N_2(\lambda)$ **iff** $V_1 > V_2$, which means that $N_1 - N_2$ changes sign in $(0, \hat{\mu})$. In addition, if λ_0 is the first $\lambda \in (0, \hat{\mu})$ where $N_1 - N_2$ changes sign, then for all $\lambda \in (0, \lambda_0)$, $V_1 > V_2$ and $N_1(\lambda) < N_2(\lambda)$, or $V_1 < V_2$ and $N_1(\lambda) > N_2(\lambda)$.
(3) With simple manipulations, it is easy to see that (19) is equivalent to the inequality $\frac{f_1(0)}{f_2(\hat{\mu})} \leq \frac{V_2}{V_1} \leq \frac{f_1(\hat{\mu})}{f_2(0)}$ being false. Lemmas A.1 and A.4 prove that f_i is an increasing function of λ, (19) consequently implies that for all $\lambda \in (0, \hat{\mu})$, $\left|1 - \frac{V_2}{V_1}\right| > \left|1 - \frac{f_1(\lambda)}{f_2(\lambda)}\right|$, i.e., $N_1(\lambda) - N_2(\lambda)$ always has same sign as $V_1 - V_2$. □

A.2 Additional Remarks on the Bounds

When (17) is verified but neither (18) nor (19) are, then numerical results show that both cases can happen: Either for all $\lambda \in (0, \hat{\mu})$, $N_1(\lambda) \neq N_2(\lambda)$ (and in that case, $N_1(\lambda) > N_2(\lambda)$ **iff** $V_1 \geq V_2$); or $N_1(\lambda) - N_2(\lambda)$ changes sign. In that case, $N_1 - N_2$ changes sign at least twice in $(0, \hat{\mu})$. One open question is to determine the precise bound B as a function of π_i, α, and $\hat{\mu}$, such that $N_1(\lambda) \neq N_2(\lambda)$ for all $\lambda \in (0, \hat{\mu})$ **iff** $\left|1 - \frac{V_2}{V_1}\right| > B$. For that, one needs to find the extrema of $f_1(\lambda)/f_2(\lambda)$ in $(0, \hat{\mu})$. We conjecture that B is close to B_0 defined in (18). In fact, numerically, it seems that $\left|1 - \frac{V_2}{V_1}\right| > 2B_0$ already ensures that for all $\lambda \in (0, \hat{\mu})$, $N_1(\lambda) \neq N_2(\lambda)$.