On Capacity Dimensioning in Dynamic Scenarios: The Key Role of Peak Values

Florin Ciucu
University of Warwick

Felix Poloczek
University of Warwick / TU Berlin

Oliver Hohlfeld
RWTH Aachen University

Abstract—**This paper analyzes queueing behavior in queues with a *random number of parallel flows*, and not static as typically assumed. By deriving upper and lower bounds on the queue size distribution, the paper identifies extremal properties in such dynamic queues. The extremal *best-case* distribution (minimizing the queue) is simply the constant, whereas the *worst-case* distribution (maximizing the queue) has a bimodal structure. From a more practical point of view, this paper highlights an idiosyncrasy of dynamic queues: unlike in static queues whereby capacity dimensioning is dominated by *average-values* (subject to certain safety margins), in dynamic queues the capacity dimensioning is dominated instead by *peak-values*.**

I. Introduction

Resource allocation is an old problem which perpetually reincarnates itself in resource sharing systems such as the telephone network, the Internet, or data centers. The first influential related treatment was performed by Erlang who essentially looked at the problem of dimensioning the telephone network. One of Erlang's main results was a formula for the computation of the blocking probability that some shared resource is occupied [20]; remarkably, amongst many applications, this formula has been used for nearly a century to dimension telephone networks.

Erlang's seminal work triggered the development of queueing theory, which has become an indispensable mathematical framework for the performance analysis of resource sharing based systems. Over almost a century, the exact approach to queueing theory (a.k.a. the classical approach) has been generalized to cover a broad class of networks, largely known by the product-form property (Baskett *et al.* [5], Kelly [27]). Besides its large scope, the class of product-form queueing networks is numerically tractable using convolution (Buzen [13]) or mean value analysis algorithms (Reiser and Lavenberg [37]).

Several alternative theories to queueing have been developed to avoid the general limitation of Poisson arrivals of product-form networks. One is the theory of effective bandwidth (Kelly [28], Mazumdar [35]), which relies on large deviation techniques and provides a rather straightforward analysis of multiplexing regimes for a broad class of arrival processes. An extension of the effective bandwidth theory, which can additionally deal with many scheduling algorithms and especially multi-queue scenarios, is the stochastic network calculus (Chang [14], Jiang and Liu [26], Ciucu and Schmitt [18]). These conceivably attractive theories provide, however, queueing metric results in terms of either exact asymptotics or probabilistic bounds. While the relevance of asymptotics and (conceivably loose) bounds is often questioned (Abate *et al.* [1], Choudhury *et al.* [17], Shroff and Schwartz [41]), other advanced techniques yield much refined results (Duffield [20], Liu *et al.* [32], Chang [14], Mazumdar [35]) at the expense however of a more involved analysis.

The common challenge faced by queueing approaches, when modelling some unpredictable resource sharing based system, is capturing the system's inherent randomness. For instance, in the context of a network router, the high variability inherent to packet flows is captured by conventional queueing models with probability distributions (e.g., of the packets' inter-arrival times and sizes). Network calculus models use instead envelope functions, which enforce either deterministic or probabilistic bounds on the amounts of packets over time intervals. A very recent alternative to classical queueing theory uses deterministic models satisfying the implications of probability laws characteristic to the packets flows (e.g., the Law of the Iterated Logarithm, see Bertismas *et al.* [7]).

While capturing randomness is essential in modelling, different randomness models can lead to very different (and possibly bogus) insights on actual system behavior. Consider for instance a simple example of a router with capacity C which is being modelled by the classic M/M/1 queue: packets arrive as a Poisson process with rate λ, and their sizes are exponentially distributed with average $1/\mu$. Under the stability condition $\lambda/(\mu C) < 1$, the packets' average delay is

$$E\Big[delay\Big] = \frac{1}{\mu C - \lambda} \ . \tag{1}$$

Consider next the much simpler averaged-out D/D/1 model, in which the interarrival times are constant (i.e., equal to $1/\lambda$) and packet sizes are constant as well (i.e., equal to $1/\mu$). Under the same stability condition, the packets' average delay becomes

$$E\Big[delay\Big] = \frac{1}{\mu C} \ . \tag{2}$$

Note the different quantitative results predicted by the two models, with the observation that the 'more-random' one predicts higher delays. Such stochastic ordering properties, formalizing the manifestation of the folk principle that "*determinism minimizes the queue*", have been studied in the context of queueing systems (see the related work section) and even for risk management (see, e.g., Asmussen *et al.* [3]).

There is substantial prior work on randomness models at the flow level. What is much less known, and motivates this paper, concerns randomness models for the *number of parallel flows*. Understanding the impact of such randomness models is

978-1-4799-4894-9/14/$31.00 © 2014 IEEE

clearly important, e.g., to dimension a network router which is conceivably traversed by a highly fluctuating number of flows (for empirical evidence see Section III).

To this end, this paper derives closed-form upper and lower bounds on the queue size distribution in scenarios whereby the number of parallel flows $M(t)$ is a random process, and identifies several extremal properties characteristic to dynamic queues. The bounds are obtained using martingale-based techniques, which not only lead to tight but also concise results for broad distributions of $M(t)$. By leveraging the simplicity of the obtained bounds, it is first shown using convexity arguments that the best-case distribution from the perspective of the queue size is the intuitively obvious *constant distribution*, extending thus the folk principle that "determinism minimizes the queues" from static to dynamic queues.

The second extremal property concerns the corresponding worst-case distribution, i.e., which law of $M(t)$ maximizes the queue size? It is shown that this is a *bimodal distribution*, with mass on the extremes of $M(t)$'s range and therefore maximizing all the moments. This result also agrees with parallel results from static queues concerning extremal properties of bimodal distributions (see Section II-B).

In contrast with these apparent similarities between static and dynamic queues, there is a fundamental difference between the two. In static queues, capacity dimensioning rules roughly depend on the average-values of the flows' rates in multiplexing regimes, subject to some "safety margins" (see, e.g., van den Berg *et al.* [6]). For instance, a safety margin of 15% above the average suffices for guaranteeing end-to-end delays as low as 3 ms, according to a study on the Sprint network (see Fraleigh *et al.* [22]). When dimensioning dynamic queues, however, this paper uncovers that capacity dimensioning rules should depend on the peak-values *and not* the average-values of the number of parallel flows.

In conjunction with the previous two extremal properties, this paper indicates that approximating dynamic queues with static queues (by letting $M(t) = E[M(t)]$), as conventionally done, can severely underestimate queueing behavior and consequently lead to improper dimensioning rules.

The rest of the paper is structured as follows. First we overview related work. In Section III we provide some empirical evidence on the practical relevance for studying dynamic queues. In Section IV we derive upper and lower bounds on the queue distribution in a dynamic queue. In Section V we present the best and worst-case distributions in dynamic queues and illustrate numerical results. Section VI highlights the idiosyncrasy of dynamic queues that peak-values dominate capacity dimensioning rules. Section VII summarizes the paper.

II. Related Work

Here we overview previous work related to the main topics of this paper, i.e., 1) the relevance of studying dynamic queues and 2) extremal distributions for minimizing/maximizing queues.

A. Dynamic Queues and Analytical Approaches

The importance of accounting for the elastic nature of Internet traffic, determined by a dynamic or random number of parallel flows, has been recognized in the context of bandwidth sharing. Massoulié and Roberts showed that randomness in the number of parallel flows can have unpredictable consequences on the throughput of long-lived flows, irrespective of the assigned weights to the parallel flows [34]. In a similar setting, Bonald and Massoulié demonstrated that network stability is insensitive to a broad range of fair allocations [8], generalizing a result of de Veciana *et al.* for weighted max-min fairness [44]. A more recent study of Liu *et al.* showed that stability is actually sensitive to the settings of α-fairness, in networks with nonconvex and time-varying rate regions [31], generalizing an earlier result of Bonald and Proutière [9]. Another notable insensitivity result is that in dynamic scenarios with flows arriving as a Poisson process, the first moments of the number of flows and the flows' throughput do not depend on the flow size distribution or on the properties of the flows' arrivals (Fred *et al.* [23]).

A general way to model randomness in the number of flows is through a queue with bulk arrivals, i.e., the $G^{[M]}/G/1$ queue, whereby customers arrive in batches of random size M according to a renewal process, and customers have some service time distribution. In the case of Poisson renewals, exact solutions exist for various queueing metrics (e.g., Laplace transforms for waiting times) and various scheduling of the batches: FIFO (Burke [11]), with priorities (Takagi and Takahashi [43]), or PS (Bansal [4]); for more general renewals solutions are given numerically (Schleyer [39]) or in terms of bounds (Yao *et al.* [46]). For an excellent treatment of queues with bulk arrivals see Chaudhry and Templeton [15]. Our contribution herein is to analyze very general distributions (subject to a finite moment generating function (MGF)).

Other analytical approaches address queueing models with fluid arrivals. For instance, the classical Anick-Mitra-Sondhi model [2], with a fixed number of flows producing arrivals at some rates according to the states of Markov On-Off processes, can be regarded as a queue with a binomial number of flows. Queueing in related fluid models can be analyzed exactly in terms of spectral representations, at a cost of high computational complexity due to a combinatorial explosion in the number of states [42]. The advantage of our approach is that it provides *simple* (convex) upper and lower bounds on queueing metrics, which further permit the immediate analysis of extremal properties.

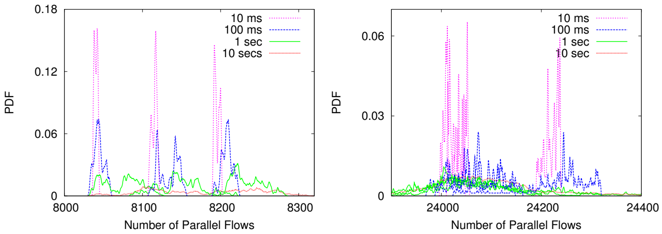
B. Extremal Distributions

A "folk theorem" in queueing theory states that, when the average inter-arrival (service) time is fixed, the constant inter-arrival (service) time distribution *minimizes* queueing metrics such as average waiting time. This result was proven for renewal processes (see Rogozin [38]) and also for more general arrival processes with exponential service times (see Hajek [24] and Humblet [25]). A related variant of the underlying intuitive principle that "determinism minimizes the waiting" is that round-robin server assignment outperforms random server assignment (see Makowski and Philips [33]).

In turn, bimodal distributions maximize queue lengths in GI/M/1 queues (Whitt [45]), in G/M/1 queues with bulk arrivals (Lee and Tsitsiklis [30]), and more recently in queues with bulk arrivals and finite buffers (Bušić *et al.* [12]). We will show that these extremal properties characteristic to static queues extend to dynamic queues as well.

III. Empirical Motivation

To confirm fluctuations in the number of parallel flows, and thus provide an empirical justification for addressing dynamic queues, we consider two packet-level traces referred to as Munich and MAWI. The first was captured in 2007 at the Munich Scientific Network, serving more than 50,000 hosts, on a 1 Gbps uplink to the German research backbone network (for more details see [19]). The second was collected over a trans-Pacific backbone line in July 2012 (MAWI repository, cf. [16]) and contains 15 minutes of traffic. For our study, we extracted connection-level information by analyzing packet headers using the tool BRO [36], and by setting inactivity timeouts for TCP flows to 1 min., and for UDP and ICMP flows to 10 secs.



(a) Munich trace (b) MAWI trace

Fig. 1. Densities of the number of parallel flows for two traces.

Figures 1.(a) and (b) illustrate the densities of the number of parallel flows over a randomly selected period of 10 secs, and three additional embedded periods of 10, 10^2, 10^3 ms in the 10 secs period. Both figures clearly show the existence of fluctuations in the number of parallel flows at relatively small time scales. While it is clearly of interest to investigate more closely the measured data (e.g., alike measurement studies showing subexponential flow interarrival times (Feldmann [21])) we do not seek to identify specific distributions but rather to motivate the need for investigating the effects of randomness in the number of parallel flows.

We remark that while the empirical justifications provided herein are based on realistic Internet traffic, the analytical models used in this paper are much simplified. In particular, they do not address closed-loop traffic (e.g., TCP), but rather open-loop traffic with i.i.d. assumptions at the flow level. All together, our analytical models provide an intuitively straightforward and yet effective convex optimization framework to investigate the effects of randomness in dynamic queues, and in particular to highlight extremal properties.

IV. Randomness in the Number of Parallel Flows

In this section we analyze the single-queue scenario from Figure 2. In addition to randomness at the flow level, the queueing model captures randomness in the number of parallel flows. The queue has an infinite sized buffer, whereas the server has a constant capacity C and serves the arrivals in a work-conserving manner.

After introducing the arrival model, we will present closed-form results on the queue size, and then discuss on the impact of randomness in the number of flows.

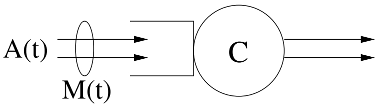


Fig. 2. A server with constant rate C serving a single queue with input $A(t)$ consisting of $M(t)$ flows.

A. Arrival Model

The time model is discrete. The number of parallel flows active at time t is represented by a stationary stochastic process $M(t)$. The cumulative arrival process $A(t)$, counting the number of data units (e.g., packets) over the time interval $[0, t]$ is defined recursively as

$$A(t) = A(t-1) + \sum_{i=1}^{M(t)} a_i(t) \ , \tag{3}$$

with the initial condition $A(0) = 0$. The instantaneous arrival process at time t is represented by the random vector $\mathbf{a}(t) = (a_1(t), a_2(t), \dots)$. When clear from the context, we will refer to the elements of $M(t)$ by M, and to the elements of $\mathbf{a}(t)$ simply by a.

For some $\theta > 0$, we assume that the moment generating functions (MGFs)

$$\phi_a(\theta) := E\left[e^{\theta a}\right] \text{ and } \phi_M(\theta) := E\left[e^{\theta M}\right]$$

are finite. Moreover, for the sake of simplicity we assume that the elements of $\mathbf{a}(t)$ and $M(t)$ are each iid (independent and identically distributed), and jointly independently.

B. The Queue Distribution

Without loss of generality we assume that $A(t)$ is a reversible process such that the stationary queue length Q can be written as

$$Q = \sup_{t \geq 0} \left\{A(t) - Ct\right\} \ .$$

The next theorem provides upper and lower bounds for the distribution of Q for the case when the elements of $\mathbf{M}$ are statistically independent.

Theorem 1: (Q's Distribution) Consider the arrival process from Eq. (3) and assume that the elements of $\mathbf{A}$ are *i.i.d.* with MGF $\phi_a(\theta)$, and the elements of $\mathbf{M}$ are *i.i.d.* with MGF $\phi_M(\theta)$; also, $\mathbf{A}$ and $\mathbf{M}$ are independent. Consider a queue with service rate C and let

$$\theta = \sup\left\{\theta \geq 0 : \phi_M\left(\log \phi_a(\theta)\right) = \phi_C(\theta)\right\} \ . \tag{4}$$

Then we have the upper bound for all $x \geq 0$

$$\mathbb{P}\Big(Q \geq x\Big) \leq e^{-\theta x} \ . \tag{5}$$

If in addition there exists the constants $a_{\max}$ and $N_{\max}$ such that $a_1(1) \leq a_{\max}$ almost surely (a.s.), $M(1) \leq N_{\max}$ a.s., and $N_{\max} a_{\max} > C$, then we have the lower bound for all $x \geq 0$

$$\mathbb{P}\Big(Q \geq x\Big) \geq e^{-\theta(N_{\max} a_{\max} - C)} e^{-\theta x} \ .$$

The upper and lower bounds are asymptotically *exact* (i.e., the following limit $\lim_{x\to\infty} \frac{1}{x} \log \mathbb{P}\left(Q > x\right) = \theta$ holds) since the two exponential bounds have the same decay rate θ. We remark that the theorem immediately extends to the case of a queue with random instantaneous capacities $(C(1), C(2), \dots)$, if these are *i.i.d.*; the only modification is that $\phi_C(\theta)$ in Eq. (4) is to be replaced by $\phi_{C(1)}(\theta)$. In the theorem, we do not explicitly impose the stability condition $\phi'_a(0)\phi'_M(0) < C$. Unless this is true then $\theta = 0$ in Eq. (4). Also, for the lower bound, the condition $N_{\max}a_{\max} > C$ avoids the trivial situation of no queueing.

V. Extremal Distributions

In this section we first identify the best-case and worst-case distributions for $M(t)$ which minimize and maximize, respectively, the queue size in a scenario with a random number of parallel flows. Finally we discuss on conditions under which a particular distribution is 'better' or 'worse' than another, and present numerical results.

To formalize the underlying stochastic ordering, and thus the meaning of 'better' and 'worse', we say that a queue Q_1 is smaller than another queue Q_2 if the corresponding decay rates θ_1 and θ_2 (e.g., defined in Eq. (4)) satisfy

$$\theta_1 \leq \theta_2 \ .$$

A. Best-Case Distribution

First we briefly show the intuitive result that the best-case distribution of M is the constant one. What is more interesting is that neither of the distributions of M and a dominates the other, when jointly accounting for both.

In the simpler case when $M(t)$ is *i.i.d.*, Jensen's inequality applied to the exponential function (i.e., $e^{\theta E[X]} \leq E\left[e^{\theta X}\right]$ for some r.v. X) yields that

$$\phi_{E[M]}\left(\log \phi_a(\theta)\right) \leq \phi_M\left(\log \phi_a(\theta)\right) \ .$$

The left-hand side corresponds to the composition of MGFs from the definition of θ from Eq. (4) when there is no randomness in the number of parallel flows, i.e., when the elements of $M(t)$ are equal to a single constant. In turn, the right-hand side accounts for randomness in $M(t)$. Because of the inequality above, it follows that the value of θ from Eq. (4) decreases when accounting for randomness, which further means that the queue increases correspondingly. The best-distribution is thus the constant, which in particular minimizes all the moments.

Finally, we point out the interesting fact that none of the randomness in the number of parallel flows, or at the flow level, dominates the other. That is because there is no general ordering between the terms

$$\phi_{E[M]}\left(\log \phi_a(\theta)\right) \text{ and } \phi_M\left(\log \phi_{E[a]}(\theta)\right) \ .$$

Indeed, using Jensen's inequality, the left term is the smallest when a is non-random (i.e., $a = E[a]$) and M is random. In turn, the left term is the largest when M is non-random (i.e., $M = E[M]$) and a is random. This fundamental lack of monotonicity suggests that, even for the purpose of deriving bounds on the queue size distribution, both the randomness in the number of flows and at the flow level must be jointly accounted for. In other words, simplifying the queueing model by averaging-out either M or a can lend itself to bogus results.

B. Worst-Case Distribution

According to Theorem 1, the problem of determining the distribution of M which maximizes the queue reduces to solving for

$$\underset{M, E[M]=\overline{m}}{\text{argmax}} \ E\left[e^{\theta M}\right] \ , \tag{6}$$

for all $\theta > 0$. The next Lemma gives the solution.

Lemma 1: (Worst-case distribution) Assuming that M has the support $\{0, 1, \dots, m\}$, and $E[M] = \overline{m}$, then the solution of Eq. (6) is the bimodal distribution with

$$\pi_0 = 1 - \frac{\overline{m}}{m} \text{ and } \pi_m = \frac{\overline{m}}{m} \ .$$

Proof. Assume that there exists $0 < i < m$ such that $\pi_i := \mathbb{P}(M = i) > 0$. Denoting $x = \frac{m-i}{m}\pi_i$, let us observe that

$$\pi_0 + \pi_i e^{\theta i} + \pi_m e^{\theta m} \leq \pi_0 + x + \left(\pi_m + \pi_i - x\right) e^{\theta m} \ . \tag{7}$$

Indeed, showing this inequality reduces to showing that the function

$$f(i) := \frac{e^{\theta m} - e^{\theta i}}{m - i}$$

is monotonically increasing over $i \in \{0, 1, \dots, m-1\}$. This can be shown immediately by extending $f(\cdot)$ to continuous time, differentiating, and using the inequality $e^z \geq z + 1$ for $z \geq 0$.

Therefore, Eq. (7) shows that a 'worse' distribution can be obtained by appropriately spreading the distribution mass to the extremes. Note that the new distribution retains the average value $\overline{m}$ since

$$i\pi_i + m\pi_m = m\left(\pi_m + \pi_i - x\right) \ .$$

The proof is complete by repeatedly spreading the mass, as in Eq. (7), for all $0 < i < m$ for which $\pi_i > 0$. □

We note that the bimodal distribution was found to attain the maximum over a partial order set according to convex ordering (see Shaked and Shanthikumar [40], Theorem 3.A.24, p. 125); in our case, the ordering is restricted to MGFs only.

C. Numerical Results

We now provide numerical evidence on the discrepancy between static and dynamic queues, by varying the distribution of the number of parallel flows M.

To keep the analysis concise, we consider a homogenous scenario in which the elements of a are Bernoulli random variables taking the values 0 and 1 with probabilities $1-p$ and p, respectively. Figure 3 illustrates the queue size x, for a fixed violation probability $\varepsilon = 10^{-3}$, and as a function of the utilization factor; the other parameters are $E[M] = 10$, Peak$(M) = 20$, $C = 9$, and p is scaled accordingly for each utilization value. The worst-case distribution is the one from Lemma 1. The figure indicates that the impact of M's distribution on the queue size can be substantial (e.g., as large as many orders of magnitude). Moreover, simulation results (depicted with the '×' symbol, for each distribution) indicate that our analytical bounds are quite tight.

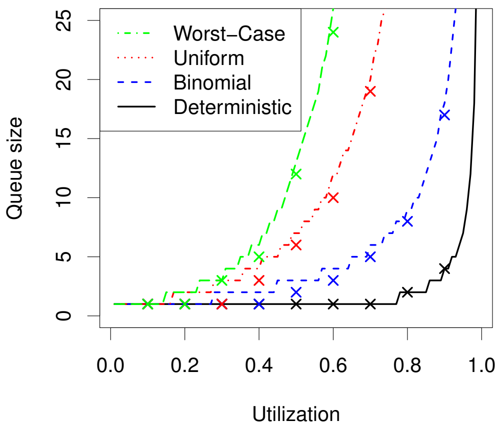


Fig. 3. Impact of several distributions for the number of parallel flows M on the queue size. Analytical bounds are depicted with lines, whereas corresponding simulation results are depicted with the '×' symbol.

VI. The Peak-Value Matters

In this section we show the interesting fact that randomness in the number of parallel flows M is fundamentally different, from a practical point of view, than randomness in the elements of the flows a. We expose this discrepancy by analyzing the dominant roles of average and peak-values in capacity dimensioning.

To this end, we consider a simple capacity dimensioning problem: given M parallel flows and the overflow probability constraint

$$\mathbb{P}\Big(Q > x_0\Big) \le \varepsilon_0 \ ,$$

for some *fixed* values x_0 and ε_0, one has to determine the required capacity C_0.

We first give the solution when M is non-random, i.e., $M = E[M]$. According to Theorem 1, for instance, the required capacity must satisfy

$$C_0 \ge E[M]\alpha_a(\theta_0) \ ,$$

where $\alpha(\theta_0)$ is the so-called effective bandwidth (see, e.g., Kelly [28]) of a single source, which is defined here as $\alpha_a(\theta_0) = \frac{\log \phi_a(\theta_0)}{\theta_0}$. It is important to point out now the range of the effective bandwidth, i.e.,

$$E[a] \le \alpha_a(\theta_0) \le \text{Peak}(a) \ ,$$

where $\text{Peak}(a)$ denotes the peak-value of a; also, $\alpha_a(\theta_0)$ is non-decreasing in θ_0. Accordingly, the above dimensioning rule says that the required capacity scales linearly with a rate $\alpha_a(\theta_0)$, which lies between the average and peak-value of a. In particular, in the case of a weak constraint of the dimensioning problem (e.g., for large values of x_0 or of the violation probability ε_0), the required capacity C_0 scales roughly linearly with $E[a]$.

In contrast, we next show that in the case when M is random, it is generally the peak-value of M, and *not* its average-value, which determines the required capacity. Let us assume that M has some distribution with support $\{0, 1, \ldots, m\}$, where $m = \text{Peak}(M)$ is the peak-value having some positive mass π_m. According to Theorem 1, the required capacity for the above dimensioning problem satisfies

$$\begin{aligned} C_0 &\ge \frac{1}{\theta_0} \log E\left[\phi_a(\theta_0)^M\right] \\ &\ge \text{Peak}(M)\alpha_a(\theta_0) + \frac{\log \pi_m}{\theta_0} \ . \end{aligned}$$

This shows that the required capacity C_0 grows linearly in the peak-value of M as long as $\pi_m = \omega\left(\left(\frac{1}{\phi_a(\theta_0)}\right)^m\right)$, i.e., the mass of the peak-value m does not decay (at least) exponentially fast with a certain base depending on the overflow constraint.

To summarize, we have showed (first the known fact) that when M is non-random then the following capacity dimensioning rule applies

$$\text{Capacity} \approx \text{AvgValue (\# of flows)} * (\text{AvgRate (flows)} + \text{SM}) \ ,$$

where SM is some safety margin (e.g., $\text{SM} = \alpha_a(\theta_0) - E[a]$) (van den Berg *et al.* [6]) . This rule captures the manifestation of *statistical multiplexing* or *economies of scale*; see Knightly and Shroff [29], or Boorstyn *et al.* [10] for numerical evidence. In turn, when M is random, then a fundamentally different capacity dimensioning rule applies, i.e.,

$$\text{Capacity} \approx \text{PeakValue (\# of flows)} * \text{AvgRate (flows)} \ .$$

As a side remark, the obtained results so far uncover several fundamental similarities and differences amongst the concepts of capacity when defined in 1) information theory (e.g., as the channel capacity), 2) static, and 3) dynamic queues (e.g., as the required capacity to guarantee some queueing constraints). All three corresponding maximal capacities are attained by the intuitively obvious constant distribution, which in particular has zero entropy. In turn, while the minimal channel capacity is attained by the uniform distribution (which maximizes the entropy), the two queueing minimal capacities are attained by bimodal distributions; this conceptual difference stems from the different scalar measures of a distribution used in information theory (i.e., the entropy) and queues (i.e., moments accounting for actual values). A fundamental distinction between 2) and 3) is that the corresponding capacities are dominated in 2) by average-values (or more precisely by effective bandwidths), and in 3) by peak-values, which are themselves fundamentally different scalar metrics of a distribution.

VII. Conclusions

In this paper we have investigated queueing behavior in typically neglected but highly relevant dynamic queues characterized by a *random number of parallel flows*. We have showed that dynamic queues retain some extremal properties from static queues, i.e., capacities are maximized by constant distributions and are minimized by bimodal distributions. The underlying apparent high sensitivity of dynamic queues to randomness is further evidenced by the observation that capacities are dominated by peak *and not* by average-values; the opposite holds instead in static queues. These results jointly highlight the pitfall of approximating dynamic by static queues, by averaging-out the number of parallel flows, which can lead to very misleading results.

References

[1] J. Abate, G. L. Choudhury, and W. Whitt. Waiting-time tail probabilities in queues with long-tail service-time distributions. *Queueing Systems*, 16(3-4):311–338, Sept. 1994.

[2] D. Anick, D. Mitra, and M. Sondhi. Stochastic theory of a data-handling system with multiple sources. *Bell Systems Technical Journal*, 61(8):1871–1894, Oct. 1982.

[3] S. Asmussen, A. Frey, T. Rolski, and V. Schmidt. Does Markov-modulation increase the risk? *ASTIN Bulletin*, 25(1):49–66, May 1995.

[4] N. Bansal. Analysis of the M/G/1 processor-sharing queue with bulk arrivals. *Operations Research Letters*, 31(5):401 – 405, Sept. 2003.

[5] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios. Open, closed and mixed networks of queues with different classes of customers. *Journal of the ACM*, 22(2):248–260, Apr. 1975.

[6] H. van den Berg, M. Mandjes, R. van de Meent, A. Pras, F. Roijers, and P. Venemans. Qos-aware bandwidth provisioning for IP network links. *Computer Networks*, 50(5):631–647, Apr. 2006.

[7] D. Bertsimas, D. Gamarnik, and A. A. Rikun. Performance analysis of queueing networks via robust optimization. *Operations Research*, 59(2):455–466, Mar. 2011.

[8] T. Bonald and L. Massoulié. Impact of fairness on internet performance. In *ACM Sigmetrics*, pages 82–91, 2001.

[9] T. Bonald and A. Proutière. Flow-level stability of utility-based allocations for non-convex rate regions. In *40th Annual Conference on Information Sciences and Systems*, pages 327–332, 2006.

[10] R. Boorstyn, A. Burchard, J. Liebeherr, and C. Oottamakorn. Statistical service assurances for traffic scheduling algorithms. *IEEE Journal on Selected Areas in Communications. Special Issue on Internet QoS*, 18(12):2651–2664, Dec. 2000.

[11] P. J. Burke. Delays in single-server queues with batch input. *INFORMS-Operations Research*, 23(4):830–833, July-August 1975.

[12] A. Bušić, J.-M. Fourneau, and N. Pekergin. Worst case analysis of batch arrivals with the increasing convex ordering. In A. Horváth and M. Telek, editors, *Formal Methods and Stochastic Models for Performance Evaluation*, volume 4054 of *Lecture Notes in Computer Science*, pages 196–210. Springer, 2006.

[13] J. P. Buzen. Computational algorithms for closed queueing networks with exponential servers. *Communications of the ACM*, 16(9):527–531, Sept. 1973.

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

[15] M. L. Chaudhry and J. G. C. Templeton. *A First Course in Bulk Queues*. John Wiley and Sons, 1983.

[16] K. Cho, K. Mitsuya, and A. Kato. Traffic data repository at the wide project. In *USENIX Annual Technical Conference*, pages 263–270, 2000.

[17] G. Choudhury, D. Lucantoni, and W. Whitt. Squeezing the most out of ATM. *IEEE Transactions on Communications*, 44(2):203–217, Feb. 1996.

[18] F. Ciucu and J. Schmitt. Perspectives on network calculus - No free lunch but still good value. In *ACM Sigcomm*, 2012.

[19] H. Dreger, A. Feldmann, M. Mai, V. Paxson, and R. Sommer. Dynamic application-layer protocol analysis for network intrusion detection. In *15th USENIX Security Symposium*, pages 257–272, 2006.

[20] N. G. Duffield. Exponential bounds for queues with markovian arrivals. *Queueing Systems*, 17(3-4):413–430, Sept. 1994.

[21] A. Feldmann. Characteristics of TCP connection arrivals. In *Self-Similar Network Traffic and Performance Evaluation. (Editors: K. Park and W. Willinger)*, pages 367–299. Wiley, 1999.

[22] C. Fraleigh, F. Tobagi, and C. Diot. Provisioning ip backbone networks to support latency sensitive traffic. In *IEEE Infocom*, pages 375–385, 2003.

[23] S. B. Fred, T. Bonald, A. Proutiere, G. Régnié, and J. W. Roberts. Statistical bandwidth sharing: a study of congestion at flow level. In *ACM Sigcomm*, pages 111–122, 2001.

[24] B. Hajek. The proof of a folk theorem on queuing delay with applications to routing in networks. *Journal of the ACM*, 30(4):834–851, Oct. 1983.

[25] P. A. Humblet. Determinism minimizes waiting time in queues. Technical report, MIT Laboratory for Information and Decision Systems, LIDS-P-1207, 1982.

[26] Y. Jiang and Y. Liu. *Stochastic Network Calculus*. Springer, 2008.

[27] F. P. Kelly. Networks of queues with customers of different types. *Journal of Applied Probability*, 3(12):542–554, Sept. 1975.

[28] F. P. Kelly. Notes on effective bandwidths. In *Stochastic Networks: Theory and Applications. (Editors: F.P. Kelly, S. Zachary and I.B. Ziedins) Royal Statistical Society Lecture Notes Series, 4*, pages 141–168. Oxford University Press, 1996.

[29] E. W. Knightly and N. B. Shroff. Admission control for statistical QoS: Theory and practice. *IEEE Network*, 13(2):20–29, Mar./Apr. 1999.

[30] D. C. Lee and J. N. Tsitsiklis. The worst bulk arrival process to a queue. Technical report, MIT Laboratory for Information and Decision Systems, LIDS-P-2116, 1992.

[31] J. Liu, A. Proutière, Y. Yi, M. Chiang, and H. Poor. Stability, fairness, and performance: A flow-level study on nonconvex and time-varying rate regions. *IEEE Transactions on Information Theory*, 55(8):3437–3456, Aug. 2009.

[32] Z. Liu, P. Nain, and D. Towsley. Exponential bounds with applications to call admission. *Journal of the ACM*, 44(3):366–394, May 1997.

[33] A. M. Makowski and T. Philips. Simple proofs of some folk theorems for parallel queues. Technical report, Institute for Systems Research, ISR-TR-1989-37, 1989.

[34] L. Massoulié and J. Roberts. Bandwidth sharing and admission control for elastic traffic. *Telecommunication Systems*, 15(1-2):185–201, Nov. 2000.

[35] R. R. Mazumdar. *Performance Modeling, Loss Networks, and Statistical Multiplexing*. Synthesis Lectures on Communication Networks. Morgan & Claypool Publishers, 2009.

[36] V. Paxson. End-to-end Internet packet dynamics. *IEEE/ACM Transactions on Networking*, 7(3):277–292, June 1999.

[37] M. Reiser and S. S. Lavenberg. Mean-value analysis of closed multichain queuing networks. *Journal of the ACM*, 27(2):313–322, Apr. 1980.

[38] B. Rogozin. Some extremal problems in the theory of mass service. *Theory of Probability & Its Applications*, 11(1):144–151, 1966.

[39] M. Schleyer. An analytical method for the calculation of the waiting time distribution of a discrete time G/G/1-queueing system with batch arrivals. *OR Spectrum*, 29(4):745–763, Oct. 2007.

[40] M. Shaked and J. G. Shanthikumar. *Stochastic Orders*. Springer, 2007.

[41] N. B. Shroff and M. Schwartz. Improved loss calculations at an ATM multiplexer. *IEEE/ACM Transactions on Networking*, 6(4):411–421, Aug. 1998.

[42] T. E. Stern and A. I. Elwalid. Analysis of separable Markov-modulated rate models for information-handling systems. *Advances in Applied Probability*, 23(1):105–139, Mar. 1991.

[43] H. Takagi and Y. Takahashi. Priority queues with batch Poisson arrivals. *Operations Research Letters*, 10(4):225–232, June 1991.

[44] G. de Veciana, T.-J. Lee, and T. Konstantopoulos. Stability and performance analysis of networks supporting services with rate control - could the internet be unstable? In *IEEE Infocom*, pages 802–810, 1999.

[45] W. Whitt. On approximations for queues, I: Extremal distributions. Technical report, Institute for Systems Research, ISR-TR-1989-37, 1989.

[46] D. D. W. Yao, M. L. Chaudhry, and J. G. C. Templeton. On bounds for bulk arrival queues. *European Journal of Operational Research*, 15(2):237 – 243, Feb. 1984.