Robust transient analysis of multi-server queueing systems and feed-forward networks

As Published: https://doi.org/10.1007/s11134-017-9566-6

Publisher: Springer US

Persistent URL: https://hdl.handle.net/1721.1/131913

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Noname manuscript No.
(will be inserted by the editor)

Robust Transient Analysis of Multi-server Queueing Systems and Feed-forward Networks

Chaithanya Bandi · Dimitris Bertsimas · Nataly Youssef

Received: date / Accepted: date

Abstract We propose an analytically tractable approach for studying the transient behavior of multi-server queueing systems and feed-forward networks. We model the queueing primitives via polyhedral uncertainty sets inspired by the limit laws of probability. These uncertainty sets are characterized by variability parameters that control the degree of conservatism of the model. Assuming the inter arrival and service times belong to such uncertainty sets, we obtain closed form expressions for the worst case transient system time in multi-server queues and feed-forward networks with deterministic routing. These analytic formulas offer rich qualitative insights on the dependence of the system times as a function of the variability parameters and the fundamental quantities in the queueing system. To approximate the average behavior, we treat the variability parameters as random variables and infer their density by using ideas from queues in heavy traffic under reflected Brownian motion. We then average the worst case values obtained with respect to the variability parameters. Our averaging approach yields approximations that match the diffusion approximations for a single queue with light-tailed primitives and allows to extend the framework to heavy-tailed feed-forward networks. Our methodology achieves significant computational tractability and provides accurate approximations for the expected system time relative to simulated values.

Keywords Transient Queueing Theory · Relaxation Time · Steady State · Robust Optimization · Heavy Tails · Feed-forward networks · Tandem Queues

1 Introduction

The origin of queueing theory dates back to the beginning of the twentieth century, when [26] published his fundamental paper on congestion in telephone traffic. Over the past century queueing theory has found many other applications, particularly in service, manufacturing and transportation industries. In recent years, new queueing applications have emerged, such as data centers and cloud computing, call centers and the Internet. These industries are experiencing surging growth rates, with call centers and cloud computing enjoying respective annual growth of 20% and 38%, according to the 2012 Gartner and Global Industry Analysts Survey.

Many applications operate under heavy-traffic conditions yielding a slow convergence to steady state, which may not be reached within the operation time window. Analyzing such queueing systems requires an understanding of **(a)** the evolution of the system time over time, and **(b)** the time it takes the queueing system to reach steady state. Furthermore, queueing systems that are characterized by heavy tailed arrivals and/or service times never reach steady state and therefore their behavior is essentially transient. For instance, heavy tailed arrivals and service times have been reported for the Internet by [43] and [24], for

Chaithanya Bandi
Assistant Professor of Economics and Decision Sciences
Kellogg School of Management
Northwestern University, Evanston, IL 60208 E-mail: c-bandi@kellogg.northwestern.edu

Dimitris Bertsimas
Boeing Professor of Operations Research
Co-director, Operations Research Center
Massachusetts Institute of Technology, Cambridge, MA 02139
E-mail: dbertsim@mit.edu

Nataly Youssef
Massachusetts Institute of Technology, Cambridge, MA 02139
E-mail: youssefn@mit.edu

call centers by [10], and for data centers by [45] and [12]. A steady state analysis in these situations is not relevant.

Despite the need for an understanding of the transient behavior, the probabilistic analysis of transient queues is by and large analytically intractable. For $M/M/1$ queues, the exact analysis of the queue length involves an infinite sum of Bessel functions and for $M/M/m$ queues, [36] obtained the transition probabilities of the Markov chain describing the queue length as functions of Poisson-Charlier polynomials. [6,7] used double transforms with respect to space and time to describe the transient behavior of an $M/M/1$ queue. This analysis was further extended in a series of papers, see [2], [3], [22], [23], [4], to obtain additional insights on the queue length process. These analyses also provide insights on the usefulness of reflected Brownian motion approximations for queues. [15] formulate the problem of finding the distribution of the transient waiting time as a two-dimensional Lindley process and then transform it to a Hilbert factorization problem. They obtain the solution for $GI/R/I$, $R/G/I$ queues, where R is the class of distributions with rational Laplace transforms. Extending these results, [16] use the "method of stages" to study $MGE_L/MGE_M/1$ queueing systems, where MGE is the class of mixed generalized Erlang distributions which can approximate an arbitrary distribution. [46,33] study the transient analysis problem for process sharing Markovian queues with time-varying rates using a technique known as "uniform acceleration". As discussed in [51], there are multiple approximations available but a tractable theory of transient analysis of G/G/m queues is lacking (see also [32], [35], and [37]). Further complicating the transient analysis is the effect of initial conditions, which gives rise to a significantly different behaviors as empirically investigated in [38] and [51]. Even numerically, the calculations involve complicated integrals which do not allow sensitivity analysis, an integral requirement for a system designer managing these systems.

Given these difficulties, a body of work has concentrated on developing approximate numerical solution techniques to investigate transient behavior (e.g., [40], [49], [47], [53], [30], [21], [41], [31], and [54]). [50], in his work on the diffusion approximation of $GI/G/1$ queueing systems under heavy traffic, obtains a closed-form expression and proposes an order of magnitude estimate of the time required for the transient effects to become negligible. [48], develops a numerical technique for estimating the transient behavior of the expected waiting time for $M/M/1$ and $M/D/1$ queueing systems on the basis of a recursive relationship involving waiting times of successive jobs. All of these approaches have focused on improving the efficiency and accuracy of numerical solution techniques, rather than on using their results to draw conclusions on general attributes of transient behavior. More recently, based on earlier work by [18], [52] use a semi-definite optimization approach to obtain qualitative insights on the transient behavior of queues. They derive upper bounds on the tail distribution of the transient waiting time, and use it to bound the expected waiting time, for $GI/GI/1$ queues starting with empty buffer for non-heavy-tailed distributions. [59] use an extension of the Stochastic Network Calculus framework to propose a temporal network calculus approach to obtain bounds on delays in internet networks. However, these approaches do not tackle heavy-tailed queues and the effect of initial buffer conditions.

Motivated by these challenges, we propose an analytically tractable approach for studying the transient behavior of multi-server queueing systems with heavy-tailed arrival and service processes. Building upon our earlier work in [9] for queues in steady state, we first model the queueing primitives via polyhedral uncertainty sets indexed by two parameters which control the degree of conservatism of the corresponding arrival and service processes. We then consider a robust optimization perspective which yields closed form formulas for the transient system time. These expressions offer new qualitative insights on the dependence of the system time as a function of fundamental quantities in the queueing system. We break new ground by treating the parameters characterizing the uncertainty sets as random variables and infer their density using ideas from queues in heavy traffic under reflected Brownian motion. We then approximate the expected behavior via averaging the worst case values over the variability parameters. This averaging approach achieves significant tractability by reducing the problem of transient analysis to a low dimensional integral. As a sanity check, we show that our results match the diffusion approximations for a single queue with light-tailed primitives. Furthermore, we also extend our approach to feed-forward networks with possibly heavy-tailed arrivals and/or service times.

The motivation behind our idea stems from the rich development of optimization as a scientific field during the second part of the twentieth century. From its early years ([25]), modern optimization has had the objective to solve multi-dimensional problems efficiently from a practical point of view. Today, many commercial codes are available which can solve truly large scale structured (linear, mixed integer and quadratic) optimization problems. In particular, Robust Optimization (RO), arguably one of the fastest growing areas in optimization in the last decade, provides, in our opinion, a natural modeling framework for stochastic systems. For a review of robust optimization, we refer the reader to [11], and [13]. The present paper is part of a broader investigation to analyze stochastic systems such as market design, information theory, finance, and other areas via robust optimization (see [8]).

1.1 Contributions and Structure of the paper

We make the following contributions in this paper:

1. We provide worst case and average case analysis of multi-server queueing systems in the presence of heavy tails, even when the queues are non-empty to begin with.
2. We extend our approach to tandem networks and feed-forward networks and present a tractable way to analyze the worst case and average case waiting time.

These contributions extend the robust optimization approach to analyzing queueing networks as introduced in [17] and [9], by focusing on the analysis of the transient regime rather than the steady state considered in these papers.

The structure of the paper is as follows. Section 2 provides an overview of our framework. In Section 3, we present our analysis for single multi-server queues with possibly heavy-tailed arrivals and service times. In Sections 4 and 5, we extend our approach to analyze tandem queueing systems and more complex feed-forward networks. Section 6 concludes the paper.

2 Proposed Framework

In this section, we present the main components of our framework and describe the main contributions. Let $\mathbf{T} = (T_1, \ldots, T_n)$ and $\mathbf{X} = (X_1, \ldots, X_n)$ denote the inter-arrival times and service times of n jobs, respectively. Note that in the traditional probabilistic study of queues, these primitives are modeled via renewal processes. In a first-come first-serve (FCFS) single-server queue, the waiting time $W_n = W_n(\mathbf{T}, \mathbf{X})$ and the system time $S_n = S_n(\mathbf{T}, \mathbf{X})$ are given by the Lindley recursion ([44]) as follows

$$S_n = W_n + X_n = \max\left(S_{n-1} + X_n - T_n, X_n\right) = \max_{1 \le k \le n} \left(\sum_{i=k}^{n} X_i - \sum_{i=k+1}^{n} T_i \right). \tag{1}$$

Analyzing the expected waiting and system times, given by

$$\overline{W}_n = \mathbb{E}_{\mathbf{T},\mathbf{X}}\left[W_n(\mathbf{T}, \mathbf{X})\right] \quad \text{and} \quad \overline{S}_n = \mathbb{E}_{\mathbf{T},\mathbf{X}}\left[S_n(\mathbf{T}, \mathbf{X})\right], \tag{2}$$

entails the understanding of the complex relationships between the random variables associated with the inter-arrival and service times. The high dimensional nature of the performance analysis problem makes the probabilistic analysis by and large intractable, especially in the transient domain. The study of multi-server queues is even more challenging. Instead, we propose an approximation of the expected system time by

- **(a)** using the modeling framework introduced in [9] to model the uncertainty in the arrival and service processes via parametrized polyhedral sets,
- **(b)** computing closed-form expressions for the worst case system time under our assumptions, and
- **(c)** taking advantage of the uncertainty dimensionality reduction and leveraging the worst case values to obtain analytical expressions that approximate the average-case system behavior.

In what follows, we present an overview of our approach in this section and illustrate our methodology through the case of a single-server queue with light-tailed arrivals and service times. We then extend our framework to analyze the average behavior of heavy-tailed multi-server queues (Section 3), tandem networks (Section 4), and feed-forward networks (Section 5).

2.1 Uncertainty Modeling

Given the structure of the Lindley recursion, [9] model the uncertainty around the partial sums of the inter-arrival and service times in Eq. (1) via uncertainty sets inspired by the Central Limit Theorem. In particular, [9] constrain the quantities T_i and X_i to take values while satisfying

$$\frac{\displaystyle\sum_{i=k+1}^{n} T_i - \frac{n-k}{\lambda}}{\sqrt{n-k}} \ge -\mathit{\Gamma}_a, \quad \text{and} \quad \frac{\displaystyle\sum_{i=k}^{n} X_i - \frac{n-k+1}{\mu}}{\sqrt{n-k+1}} \le \mathit{\Gamma}_s, \quad \forall k = 1, \ldots, n, \tag{3}$$

for some parameters $\mathit{\Gamma}_a$ and $\mathit{\Gamma}_s$ that we use to control the degree of conservatism.

Assumption 1.

We make the following assumptions on the inter-arrival and service times.

(a) The inter-arrival times $(T_1, \dots, T_n)$ belong to the parametrized uncertainty set

$$\mathcal{U}^a = \mathcal{U}^a\left(\varGamma_a\right) = \left\{ (T_1, \dots, T_n) \left| \sum_{i=k+1}^{n} T_i - \frac{n-k}{\lambda} \geq -\varGamma_a \sqrt{n-k}, \quad \forall\, 0 \leq k < n \right. \right\},$$

where $1/\lambda$ is the expected inter-arrival time and $\varGamma_a \in \mathbb{R}$ controls the degree of conservatism.
(b) For a single-server queue, the service times $(X_1, \dots, X_n)$ belong to the uncertainty set

$$\mathcal{U}^s = \mathcal{U}^s\left(\varGamma_s\right) = \left\{ (X_1, \dots, X_n) \left| \sum_{i=k+1}^{n} X_i - \frac{n-k}{\mu} \leq \varGamma_s \sqrt{n-k}, \ \forall\, 0 \leq k < n \right. \right\},$$

where $1/\mu$ is the expected service time, and $\varGamma_s \in \mathbb{R}$ controls the degree of conservatism.
Note that, in this paper, we allow $\varGamma_a$ and $\varGamma_s$ to take both negative and positive values. When these parameters are negative, the constraints on the inter arrival and service times imply

$$\sum_{i=k+1}^{n} T_i \geq \frac{n-k}{\lambda}, \ \forall k \leq n-1\ , \ \sum_{i=k+1}^{n} X_i - \frac{n-k}{\mu} \leq \varGamma_s \sqrt{n-k}, \ \forall\, k \leq n-1,$$

thus constraining the sums of the inter arrival times to exceed their mean and the sums of the service times to take values below the mean. This scenario constrains the analysis to realizations with generally longer inter arrival times and short service times, and therefore the jobs enter service without waiting in the queue. When these parameters are positive, the constraints on the partial sums of the inter arrival and service times allow realizations with shorter inter-arrival times and longer service times, and in these cases jobs may need to wait in the queue before entering service.

2.2 Worst Case Behavior

To characterize the worst case behavior, we formulate the related performance analysis question as a robust optimization problem. In particular, assuming inter-arrival and service times satisfy Assumption 1, we seek the worst case waiting and system times defined as

$$\widehat{W}_n = \max_{\mathcal{U}^a \times \mathcal{U}^s} W_n\left(\mathbf{T}, \mathbf{X}\right) \quad \text{and} \quad \widehat{S}_n = \max_{\mathcal{U}^a \times \mathcal{U}^s} S_n\left(\mathbf{T}, \mathbf{X}\right). \tag{4}$$

The maximization problems in Eq. (4) yield simple nonlinear optimization problems.

Unstable Queue: For a light-tailed queue with $\rho = \lambda/\mu > 1$, Eq. (4) gives rise to a closed form characterization of the worst case waiting and system times with

$$\widehat{S}_n\left(\varGamma\right) \leq \widehat{W}_n\left(\varGamma\right) + \left(\frac{1}{\mu} + \varGamma_s\right) \leq \left(\varGamma\sqrt{n} + \frac{\rho-1}{\lambda}n\right)^+ + \left(\frac{1}{\mu} + \varGamma_s\right) \tag{5}$$

where $\varGamma = \varGamma_a + \varGamma_s$ denotes the effective variability parameter and the notation $a^+ = \max\left(0, a\right)$. For the case where $\rho > 1$, the worst case waiting and system times increase linearly with the value of n.

Stable Queue: For a light-tailed queue with $\rho = \lambda/\mu < 1$, Eq. (4) gives rise to a closed form characterization of the worst case waiting and system times with

$$\begin{aligned} \widehat{S}_n\left(\varGamma\right) &\leq \widehat{W}_n\left(\varGamma\right) + \left(\frac{1}{\mu} + \varGamma_s\right) \\ &\leq \max \begin{cases} \varGamma\sqrt{n} - \dfrac{1-\rho}{\lambda}n + \left(\dfrac{1}{\mu} + \varGamma_s\right), & \text{if } \ n < \dfrac{\lambda^2 \left[\varGamma^+\right]^2}{4(1-\rho)^2}, \\ \dfrac{\lambda}{4} \cdot \dfrac{\left[\varGamma^+\right]^2}{1-\rho} + \left(\dfrac{1}{\mu} + \varGamma_s\right), & \text{otherwise}, \end{cases} \end{aligned} \tag{6}$$

where $\varGamma = \varGamma_a + \varGamma_s$ denotes the effective variability parameter and the notation $a^+ = \max\left(0, a\right)$. The evolution of the worst case behavior is characterized by two distinct states: **(a)** a *transient state* where the behavior is dependent on n with the system time in an initially empty queue increasing at an order of $\sqrt{n}$ when $\varGamma > 0$; and **(b)** a *steady state* where the behavior is independent of n. When $\varGamma < 0$, jobs do not experience any waiting time, and therefore the worst case system time is equal to the worst case service time. The characterization of the worst case behavior bears qualitative similarity to the bounds established by [52] and [39] for the transient and steady state expected waiting and system times in a $GI/GI/1$ queue, respectively,

$$\mathbb{E}\left[S_n\right] = \mathbb{E}\left[W_n\right] + \frac{1}{\mu} \leq \begin{cases} \dfrac{e}{2}\sqrt{\sigma_a^2+\sigma_s^2}\sqrt{n} + \dfrac{1}{\mu}, & \text{if } n < \dfrac{\lambda^2(\sigma_a^2+\sigma_s^2)}{e^2(1-\rho)^2}, \\ \dfrac{\lambda}{2}\dfrac{\left(\sigma_a^2+\sigma_s^2\right)}{1-\rho} + \dfrac{1}{\mu}, & \text{otherwise,} \end{cases}$$

where $e = \exp(1) = 2.718$. For ease of notation, we rewrite the worst case behavior in Eq. (6) as

$$\widehat{S}_n\left(\varGamma\right) \leq \widehat{S}_n^t\left(\varGamma\right) \cdot \mathbb{1}_n^t\left(\varGamma\right) + \widehat{S}^s\left(\varGamma\right) \cdot \mathbb{1}_n^s\left(\varGamma\right), \tag{7}$$

where the terms $\widehat{S}_n^t$ and $\widehat{S}^s$ respectively denote the quantities associated with the transient state and the steady state, i.e.,

$$\left\{\begin{aligned} \widehat{S}_n^t &= \varGamma\sqrt{n} - \frac{1-\rho}{\lambda}n + \frac{1}{\mu} + \varGamma_s \\ \widehat{S}^s &= \frac{\lambda}{4}\cdot\frac{\left[\varGamma^+\right]^2}{1-\rho} + \frac{1}{\mu} + \varGamma_s \end{aligned}\right\},$$

and the indicator functions $\mathbb{1}_n^t$ and $\mathbb{1}_n^s$ respectively reflect the condition for the system to be in the transient state and the steady state, with

$$\begin{cases} \mathbb{1}_n^t\left(\varGamma\right) = 1, & \text{if } \varGamma > \dfrac{2(1-\rho)}{\lambda}\cdot\sqrt{n}, \\ \mathbb{1}_n^s\left(\varGamma\right) = 1, & \text{otherwise.} \end{cases}$$

Note that the worst case values directly depend on the value of $\varGamma$. Larger values of $\varGamma$ yield increasingly more conservative estimates.

2.3 Average Case Behavior

We next propose to analyze the average case behavior leveraging our worst case analysis. Our approach is driven by the following two key observations:

1. The expected value of a random variable can be computed by "averaging" its quantiles with appropriate weights (density).
2. Our worst case analysis provides a way to calculate these quantiles.

We next elaborate on these observations and discuss the details of our approach.

For a given value of n, suppose that the waiting time $W_n = W_n\left(\mathbf{T}, \mathbf{X}\right)$ is governed by a distribution F_n, and assume that F_n is continuous. Note that, this can be derived from the joint distribution over the inter-arrival and service times by considering the dynamics of a queue. The expected waiting time is then given by

$$\overline{W}_n = \int w \mathrm{d}F_n(w).$$

The inverse of $F_n\left(\cdot\right)$ then corresponds to the quantile function $Q_n(\cdot)$ given by

$$Q_n(p) = F_n^{-1}\left(p\right) = \Big\{q : F_n(q) = p\Big\} = \Big\{q : \mathbb{P}\left(S_n \leq q\right) = p\Big\},$$

for some probability level $p \in (0,1)$. By a simple variable substitution, we can view the expected value as an "average" of quantiles, given by

$$\overline{W}_n = \int_0^1 Q_n(p)\mathrm{d}p. \tag{8}$$

Recall that we have obtained an analytic expression of the worst case waiting time as a function of the variability parameter $\varGamma$. We can map each quantile value $Q_n(p)$ to a corresponding worst case value $\widehat{W}_n\left(\varGamma\right)$. Let G_n denote the function that maps p to $\varGamma$ such that $Q_n(p) = \widehat{W}_n\left(\varGamma\right)$, i.e.,

$$p = \mathbb{P}\left(W_n \leq \widehat{W}_n\left(\varGamma\right)\right) = F_n\left(\widehat{W}_n\left(\varGamma\right)\right) = G_n\left(\varGamma\right). \tag{9}$$

In this context, the expected value of the waiting time in Eq. (8) can be written as an average over the worst case values, with

$$\overline{W}_n = \int \widehat{W}_n(\Gamma)\,\mathrm{d}G_n(\Gamma) = \mathbb{E}_\Gamma\left[\widehat{W}_n(\Gamma)\right]. \tag{10}$$

Philosophically, this approach distills all the probabilistic information contained in the random variables X_i's and T_i's into the parameter Γ, hence allowing a significant dimensionality reduction of the uncertainty. This in turn yields a tractable approximation of the expected transient waiting time by reducing the problem to solving a low-dimensional integral.

Note: The knowledge of G_n allows us to compute the expected waiting time $\overline{W}_n$ exactly, however, this depends on the knowledge of the waiting time distribution function F_n. This is feasible for simple systems, e.g., analyzing the steady-state waiting time in an M/M/1 queue. For this particular example, it is well known that the conditional steady state waiting time $W_\infty\,|W_\infty > 0$ is exponentially distributed with rate $\mu(1-\rho)$. Therefore,

$$F_\infty(q) = 1 - \rho e^{-\mu(1-\rho)q}, \text{ for } q \geq 0, \quad \text{and} \quad Q(p) = -\frac{\ln\left((1-p)/\rho\right)}{\mu(1-\rho)}, \text{ for } p \in (0,1).$$

In this case, we can derive an exact characterization of the function G_∞ and obtain

$$p = F\left(\widehat{W}_\infty(\Gamma)\right) = G_\infty(\Gamma) = 1 - \rho \cdot \exp\left(-\frac{\lambda\mu}{4}\cdot\left(\Gamma^+\right)^2\right).$$

Note that the function G_∞ is a cumulative distribution function. Applying Eq. (10) yields

$$\int \widehat{W}_\infty(\Gamma)\,\mathrm{d}G_\infty(\Gamma) = \int_0^\infty \frac{\lambda}{4(1-\rho)}\cdot\Gamma^2\cdot\frac{\lambda\mu}{2}\cdot\Gamma\cdot\rho\cdot\exp\left(-\frac{\lambda\mu}{4}\cdot\Gamma^2\right)d\Gamma = \frac{\rho}{\mu(1-\rho)},$$

which matches the expression of the expected steady state waiting time $\overline{W}_\infty$ in an $M/M/1$ queue.

2.4 Robust Approximation

However, characterizing F_n (and therefore G_n) is challenging for more complex queueing systems, and depends directly on the distributions of the inter-arrival and service times. Instead, we propose an approximation to G_n, which we present next. We consider an initially empty GI/GI/1 queue and employ conclusions from the theory of *diffusion approximations* to obtain an approximation of the density G_n. From applying diffusion approximations to queueing theory, it is known that the waiting time of the n^{th} job arriving at the queue at time $t = n/\lambda$ is well approximated by a reflected Brownian motion

$$W_n \approx \frac{1}{\mu}\mathrm{RBM}\left(n/\lambda, \lambda-\mu, \lambda\left(\lambda^2\sigma_a^2 + \mu^2\sigma_s^2\right)\right), \tag{11}$$

where $\mathrm{RBM}\left(t,\theta,\sigma^2\right)$ denotes the state of the reflected Brownian motion with drift θ and variance σ^2 at time t, and (σ_a, σ_s) denote the standard deviations associated with the inter-arrival and service times, respectively (see [1]). Therefore, The distribution of the waiting time can be approximated by

$$\mathbb{P}\left(W_n \leq \omega\right) \approx \varPhi\left(\frac{\mu\omega - (\lambda-\mu)n/\lambda}{\sigma\sqrt{n/\lambda}}\right) - \varPhi\left(\frac{-\mu\omega - (\lambda-\mu)n/\lambda}{\sigma\sqrt{n/\lambda}}\right)\cdot \mathrm{e}^{2(\lambda-\mu)\mu\omega},$$

where $\varPhi(\cdot)$ denotes the distribution function of a standard normal and $\sigma^2 = \lambda\left(\lambda^2\sigma_a^2 + \mu^2\sigma_s^2\right)$. For heavy traffic systems, we have $\rho \to 1$, i.e., $\lambda \approx \mu$, thus yielding

$$\mathbb{P}\left(W_n \leq \omega\right) \approx \varPhi\left(\frac{\mu\omega}{\sigma\sqrt{n/\lambda}}\right) - \varPhi\left(\frac{-\mu\omega}{\sigma\sqrt{n/\lambda}}\right) \approx 2\cdot\varPhi\left(\frac{\omega}{\sqrt{\sigma_a^2+\sigma_s^2}\sqrt{n}}\right) - 1. \tag{12}$$

To derive an approximation of G_n, we assume $\rho < 1$ and focus on the worst case steady-state waiting time given by

$$\widehat{W}_n(\Gamma) = \frac{\lambda\left(\Gamma^+\right)^2}{4(1-\rho)}, \text{ for } n > \frac{\lambda^2\left(\Gamma^+\right)^2}{4(1-\rho)^2}.$$

Conditioned on Γ being positive, and applying Eq. (12), we obtain

$$\mathbb{P}\left(W_n \leq \widehat{W}_n(\Gamma)\,|\Gamma > 0\right) \approx 2\cdot\varPhi\left(\frac{\lambda\Gamma^2/4(1-\rho)}{\sqrt{\sigma_a^2+\sigma_s^2}\sqrt{n}}\right) - 1 \leq 2\cdot\varPhi\left(\frac{\Gamma}{2\sqrt{\sigma_a^2+\sigma_s^2}}\right) - 1.$$

By differentiating the right hand side of the above expression, we obtain an approximation to the conditional distribution of Γ, given $\Gamma > 0$ as follows

$$\frac{1}{\sqrt{\sigma_a^2+\sigma_s^2}} \cdot \phi\left(\frac{\Gamma}{2\sqrt{\sigma_a^2+\sigma_s^2}}\right),$$

which corresponds to the conditional distribution of a normal random variable Y with zero mean and standard deviation of $2\sqrt{\sigma_a^2+\sigma_s^2}$, given $Y > 0$.

This allows us to obtain an approximation of the expected waiting and system times as

$$\widetilde{W}_n \approx \mathbb{E}_\Gamma\left[\widehat{W}_n\left(\Gamma\right)\right] \text{ and } \widetilde{S}_n \approx \mathbb{E}_\Gamma\left[\widehat{S}_n\left(\Gamma\right)\right], \tag{13}$$

where we treat the effective variability parameter as a normally distributed random variable with

$$\Gamma \sim \mathcal{N}\left(0, 2\sqrt{\sigma_a^2+\sigma_s^2}\right). \tag{14}$$

Illustration of our Approach: Recovering Diffusion Approximations: We next show that by approximating the density of Γ using arguments borrowed from our worst case steady-state analysis, Eq. (13) yields values that match the standard approximation obtained via diffusion theory for light-tailed queues. The following approximations prove useful for our analysis (see [57])

$$\int_a^\infty x\phi(x)dx \approx \phi(a) \quad \text{and} \quad \int_a^\infty x^2\phi(x)dx \approx 1-\Phi(a)+a\phi(a), \tag{15}$$

where $\phi\left(\cdot\right)$ and $\Phi\left(\cdot\right)$ denote the standard normal density and distribution functions.

(a) Proposed Approach: Applying the approximation in Eq. (13) and given the expression of the worst case waiting time in Eq. (7), we obtain

$$\begin{aligned}
\widetilde{W}_n &\approx \mathbb{E}\left[\left(\Gamma\sqrt{n}-\frac{1-\rho}{\lambda}n\right)\cdot \mathbb{1}_{\Gamma>2\sqrt{n}(1-\rho)/\lambda}+\frac{\lambda}{4(1-\rho)}\Gamma^2\cdot \mathbb{1}_{0\le\Gamma\le 2\sqrt{n}(1-\rho)/\lambda}\right],\\
&= \int_\eta^\infty \left(2\sqrt{\sigma_a^2+\sigma_s^2}\cdot\sqrt{n}\cdot x-\frac{1-\rho}{\lambda}n\right)\phi(x)dx\\
&\quad +\int_0^\eta \frac{\lambda(\sigma_a^2+\sigma_s^2)}{1-\rho}\cdot x^2\phi(x)dx,
\end{aligned} \tag{16}$$

where $\phi\left(\cdot\right)$ and $\Phi\left(\cdot\right)$ denote the standard normal density and distribution functions, and

$$\eta = \frac{1-\rho}{\lambda}\sqrt{\frac{n}{\sigma_a^2+\sigma_s^2}} \quad \text{implying} \quad n = \frac{\lambda^2(\sigma_a^2+\sigma_s^2)}{(1-\rho)^2}\cdot\eta^2 = \frac{\lambda^2\sigma^2}{4(1-\rho)^2}\cdot\eta^2. \tag{17}$$

Using Eq. (17) and applying the approximations given in Eq. (15),

$$\begin{aligned}
\widetilde{W}_n &\approx \sqrt{\sigma_a^2+\sigma_s^2}\sqrt{n}\cdot\phi\left(\eta\right)-\frac{1-\rho}{\lambda}n\cdot[1-\Phi(\eta)]+\frac{\lambda\left(\sigma_a^2+\sigma_s^2\right)}{4(1-\rho)}\cdot\left[\Phi(\eta)-\eta\phi(\eta)-\frac{1}{2}\right],\\
&= \frac{\lambda(\sigma_a^2+\sigma_s^2)}{1-\rho}\left[\frac{1}{2}-\left(\eta^2+1\right)\cdot[1-\Phi(\eta)]+\eta\phi(\eta)\right].
\end{aligned} \tag{18}$$

(b) Diffusion Approximation: Given Eq. (11) and applying the results obtained by [1] to analyze the transient behavior of the reflected Brownian motion, [52] derive the diffusion approximation for $\overline{W}_n$ as

$$\widetilde{W}_n^{\text{diff}} = \frac{\lambda(\sigma_a^2+\sigma_s^2)}{1-\rho}\left[\frac{1}{2}-\left(\eta^2+1\right)\cdot[1-\Phi(\eta)]+\eta\phi(\eta)\right],$$

which matches our approximation given in Eq. (18).

Remark: For unstable queues ($\rho > 1$) and large n, we approximate the expected waiting time as

$$\widetilde{W}_n \approx \sqrt{\sigma_a^2+\sigma_s^2}\sqrt{n}\cdot\phi\left(\eta\right)-\frac{1-\rho}{\lambda}n\cdot[1-\Phi(\eta)] \approx -\frac{1-\rho}{\lambda}n$$

where η is defined in Eq. (17). It is known that, for single-server queues, the expected number of jobs in the queue is $(\lambda-\mu)t$ at any given time t. So on average, the n^{th} job will have to wait for $(\lambda-\mu)n/\lambda$ jobs to clear the queue, which yields

$$\overline{W}_n = (\lambda-\mu)\cdot\frac{n}{\lambda}\cdot\frac{1}{\mu} = -\frac{1-\rho}{\lambda}n,$$

which matches our approximation.

Our approach extends beyond the simple example of single-server queues with light-tailed arrivals and services. Sections 3-6 apply our proposed framework to study multi-server heavy-tailed queueing systems and feed-forward transient networks, for which standard approximations are not available to the best of our knowledge.

3 Extensions to Heavy-Tailed Queues

In this section, we extend our analysis of the worst and average case behavior to study the performance of a single multi-server queue with possibly heavy-tailed arrivals and services. We restrict our analysis to an FCFS scheduling policy and consider an m-server queueing system which begins its operation with n_0 initial jobs. We show that (a) the worst case approach yields closed form expressions for the worst case system time, and (b) averaging the worst case values yields a good approximation of the expected system time.

3.1 Uncertainty Modeling

To model uncertainty in the partial sums of the inter-arrival and service times, we invoke the generalized Central Limit Theorem reproduced below in Theorem 1.

Theorem 1 Generalized CLT ([55])
Let $\{Y_1, Y_2, \ldots\}$ be a sequence of independent and identically distributed random variables, with mean μ and undefined variance. Then, the normalized sum

$$\frac{\sum_{i=1}^{n} Y_i - n\mu}{C_\alpha n^{1/\alpha}} \sim Y, \tag{19}$$

where Y is a stable distribution with a tail coefficient $\alpha \in (1, 2]$ and C_α is a normalizing constant.

With the insight from Theorem 1, we adapt the uncertainty sets to handle possibly heavy-tailed arrivals and service times.
Assumption 2.:
We make the following assumptions on the inter-arrival and service times.
(a) The inter-arrival times $(T_{n_0+1}, \ldots, T_n)$ belong to the parametrized uncertainty set

$$\mathcal{U}^a\left(\varGamma_a\right) = \left\{ (T_{n_0+1}, \ldots, T_n) \,\middle|\, \sum_{i=k+1}^{n} T_i - \frac{n-k}{\lambda} \geq -\varGamma_a (n-k)^{1/\alpha_a}, \forall n_0 \leq k \leq n \right\},$$

where $1/\lambda$ is the expected inter-arrival time, n_0 is the initial buffer in the queue, $\varGamma_a \in \mathbb{R}$ controls the degree of conservatism, and $1 < \alpha_a \leq 2$ is a tail coefficient modeling possibly heavy-tailed inter-arrival times.
(b) For a single-server queue, the service times $(X_1, \ldots, X_n)$ belong to the uncertainty set

$$\mathcal{U}^s\left(\varGamma_s\right) = \left\{ (X_1, \ldots, X_n) \,\middle|\, \sum_{i=k+1}^{\ell} X_i - \frac{n-k}{\mu} \leq \varGamma_s \left(n-k\right)^{1/\alpha_s}, \quad \forall\ 0 \leq k \leq n \right\},$$

where $1/\mu$ is the expected service time, $\varGamma_s \in \mathbb{R}$ controls the degree of conservatism, and $1 < \alpha_s \leq 2$ is a tail coefficient modeling possibly heavy-tailed service times.
(c) For an m-server queue, $m \geq 2$, we let ν be a non-negative integer such that $\nu = \lfloor (n-1)/m \rfloor$, where n is the index corresponding to the n^{th} arriving job. We partition the job indices into sets $K_i = \{k \leq n : \lfloor (k-1)/m \rfloor = i\}$, for $i = 0, \ldots, \nu$, i.e.,

$$K_0 = \{1, \ldots, m\}, K_1 = \{m+1, \ldots, 2m\}, \ldots, K_\nu = \{\nu m + 1, \ldots, n, \ldots, (\nu+1)m\}. \tag{20}$$

Let $k_i \in K_i$ denote the index that selects a job from set K_i, for $i = 0, \ldots, \nu$. The service times for a multi-server queue belong to the parameterized uncertainty set

$$\mathcal{U}^m\left(\varGamma_m\right) = \left\{ \sum_{i \in \mathcal{I}} X_{k_i} - \frac{|\mathcal{I}|}{\mu} \leq \varGamma_m \left|\mathcal{I}\right|^{1/\alpha_s}, \ \forall\ k_i \in K_i, i \in \mathcal{I} \subseteq \{0, \ldots, \nu\}, \right\},$$

where $1/\mu$ is the expected service time, $\varGamma_m \in \mathbb{R}$ controls the degree of conservatism, and $1 < \alpha_s \leq 2$ is a tail coefficient modeling possibly heavy-tailed service times. Note that $\mathcal{U}^m \subset \mathcal{U}^s$ for the case of $m = 1$

We next study the worst case system time using the approach developed by [9].

3.2 Worst Case Behavior

Let C_n denote the completion time of the n^{th} job, i.e., the time the n^{th} job leaves the system (including service), and $C_{(n)}$ denote the time of the n^{th} departure from the system. In general, the following recursions describe the dynamics in a multi-server queue ([42])

$$\begin{aligned} C_n &= \max\left(A_n, C_{(n-m)}\right) + X_n, \\ S_n &= C_n - A_n = \max\left(C_{(n-m)} - A_n, 0\right) + X_n, \end{aligned} \tag{21}$$

where $A_n = \sum_{i=1}^{n} T_i$ denotes the time of arrival of the n^{th} job.

It is well known that the central difficulty in analyzing multi-server queues lies in the fact that overtaking may occur, i.e., the n^{th} departure may not correspond to the n^{th} job arriving to the queue. However, as noted in [9], taking a worst case approach allows us to overcome the challenges of multi-server queue dynamics and obtain an exact characterization of the worst case system time for the n^{th} job, for any $\mathbf{T}$. Proposition 1 presents an exact bound on the worst case system time in an m-server queue, for all possible realizations of the inter-arrival times.

Proposition 1 (Worst Case System Time in a Multi-Server Queue)
In an m-server queue under Assumption 2(c), the worst case system time for the n^{th} job for any realization of $\mathbf{T}$ is given by

$$\begin{aligned} \widehat{S}_n(\mathbf{T}) &= \max_{\mathcal{U}^m(\Gamma_m)} S_n(\mathbf{T}, \mathbf{X}) \\ &\le \max_{\mathcal{U}^m(\Gamma_m^+)} S_n(\mathbf{T}, \mathbf{X}) \\ &\le \max_{0 \le k \le \nu} \left(\max_{\mathcal{U}^m(\Gamma_m^+)} \sum_{i=k}^{\nu} X_{r(i)} - \sum_{i=r(k)+1}^{n} T_i \right), \end{aligned} \tag{22}$$

where $\nu = \lfloor (n-1)/m \rfloor$ and $r(i) = n - (\nu - i)m$ and $\Gamma_m^+ = \max(0, \Gamma_m)$.

The proof of Proposition 1 can be easily adapted from [9]. We next present our worst case analysis for initially empty and nonempty heavy-tailed queues.

Initially Empty Queues

Given Assumption 2, we bound Eq. (22) by the following one-dimensional optimization problem

$$\widehat{S}_n \le \max_{0 \le k \le \nu} \left\{ \frac{\nu - k + 1}{\mu} + \Gamma_m^+ (\nu - k + 1)^{1/\alpha_s} - \frac{m(\nu - k)}{\lambda} + \Gamma_a \left[m(\nu - k) \right]^{1/\alpha_a} \right\}. \tag{23}$$

This bound can be computed efficiently for the general case where $\alpha_s \neq \alpha_a$ by solving a simple constrained non-linear optimization problem. Furthermore, we can obtain a closed form expression for the upper bound on the worst case system time for the special case where the arrival and service tail coefficients are equal, i.e., $\alpha_a = \alpha_s$, as shown in Theorem 2.

Theorem 2 (Highest System Time in an Initially Empty Heavy-Tailed Queue)
In an initially empty m-server FCFS queue satisfying Assumptions 2, with $\alpha_a = \alpha_s = \alpha$ and $\rho < 1$, the worst-case system time is given by

$$\widehat{S}_n(\Gamma) \le \begin{cases} \Gamma \cdot \nu^{1/\alpha} - \dfrac{m(1-\rho)}{\lambda} \cdot \nu + \left(\dfrac{1}{\mu} + \Gamma_m^+ \right), & \text{if } \nu < \left(\dfrac{\lambda \Gamma / m}{\alpha(1-\rho)} \right)^{\alpha/(\alpha-1)}, \\ \dfrac{\alpha - 1}{\alpha^{\alpha/(\alpha-1)}} \cdot \dfrac{\lambda^{1/(\alpha-1)} \cdot \Gamma^{\alpha/(\alpha-1)}}{[m(1-\rho)]^{1/(\alpha-1)}} + \left(\dfrac{1}{\mu} + \Gamma_m^+ \right), & \text{otherwise,} \end{cases} \tag{24}$$

where $\nu = \lfloor (n-1)/m \rfloor$ and $\Gamma = m^{1/\alpha} \Gamma_a + \Gamma_m^+ > 0$.

Note that, for the case where $\Gamma \le 0$, the function in Eq. (24) is increasing in k over the interval $k \in [0, \nu]$, for $\rho = \lambda/(m\mu) < 1$. It is therefore maximized at $k = \nu$, which yields

$$\widehat{S}_n = \max_{\mathcal{U}^m} X_n \le \frac{1}{\mu} + \Gamma_m^+.$$

In this case, the n^{th} job does not experience a waiting time before entering service. This is due to the fact that the condition $\Gamma \le 0$ involves typically long inter arrival times and short service times.

Initially Nonempty Queues

We next analyze the case where $n_0 > 0$. For a single-server queue, and given that $T_i = 0$ for all $i = 1, \ldots, n_0$, the system time in Eq. (1) reduces to

$$\textbf{(a)} \text{ for } n \le n_0 : \quad S_n = \max_{1 \le k \le n_0} \sum_{i=k}^{n} X_i = \sum_{i=1}^{n} X_i \tag{25}$$

$$\textbf{(b)} \text{ for } n > n_0 : \quad S_n = \max \left\{ \sum_{i=1}^{n} X_i - \sum_{i=n_0+1}^{n} T_i, \; \max_{n_0+1 \le k \le n} \left(\sum_{i=k}^{n} X_i - \sum_{i=k+1}^{n} T_i \right) \right\}. \tag{26}$$

We note that Eqs. (25) and (26) involve the terms $\sum_{i=1}^{n} X_i$ and $\sum_{i=1}^{n} X_i - \sum_{i=n_0+1}^{n} T_i$, respectively. While the constraints in Assumption 1 allow us to obtain upper bounds on these terms, the resulting bound is not tight, since Γ_a and Γ_s bound all of the sums $\sum_{i=k+1}^{n} T_i$ and $\sum_{i=k+1}^{n} X_i$, for all values of k. To obtain tighter bounds, we introduce the parameters γ_a and γ_s which equal the sums

$$\frac{\sum_{i=n_0+1}^{n} T_i - \dfrac{n-n_0}{\lambda}}{(n-n_0)^{1/\alpha_a}} = -\gamma_a \quad \text{and} \quad \frac{\sum_{i=1}^{n} X_i - \dfrac{n}{\mu}}{n^{1/\alpha_s}} = \gamma_s, \tag{27}$$

where the parameters γ_a and γ_s are such that $\gamma_a \le \Gamma_a$ and $\gamma_s \le \Gamma_s$. Similarly, for an m-server queue, we introduce the parameter $\gamma_m \le \Gamma_m$ where

$$\frac{\sum_{i=0}^{\nu} X_{k_i} - \dfrac{\nu+1}{\mu}}{(\nu+1)^{1/\alpha_s}} \le \gamma_m, \; \forall \; k_i \in K_i, \tag{28}$$

where the set K_i is defined as $K_i = \{k \le n : \lfloor (k-1)/m \rfloor = i\}$, for $i = 0, \ldots, \nu$.

Now, for an m-server queue, let $\phi = \lfloor (n_0 - 1)/m \rfloor$. The first m jobs in the queue are routed immediately to the servers without any delays. For $n > m$, and given that $T_i = 0$ for all $i = 1, \ldots, n_0$, we rewrite Eq. (22) as

$$\textbf{(a)} \text{ for } n \le n_0 : \quad \widehat{S}_n(\mathbf{T}) \le \max_{\mathcal{U}^m} \left(\max_{0 \le k \le \nu \le \phi} \sum_{i=k}^{\nu} X_{r(i)} \right) = \max_{\mathcal{U}^m} \sum_{i=0}^{\nu} X_{r(i)} \tag{29}$$

$$\textbf{(b)} \text{ for } n > n_0 : \quad \widehat{S}_n(\mathbf{T}) \le \max \left\{ \begin{array}{l} \displaystyle \max_{\mathcal{U}^m} \sum_{i=0}^{\nu} X_{r(i)} - \sum_{i=n_0+1}^{n} T_i, \\ \displaystyle \max_{\phi < k \le \nu} \left(\max_{\mathcal{U}^m} \sum_{i=k}^{\nu} X_{r(i)} - \sum_{i=r(k)+1}^{n} T_i \right) \end{array} \right\}, \tag{30}$$

where $r = r(0) = n - \nu m$ and $\nu = \lfloor (n-1)/m \rfloor$. By applying Assumption 2 and the inequalities in Eqs. (27) and (28), we can bound Eqs. (29) and (30) and obtain an exact characterization of the worst case system time in an initially nonempty queue with heavy tails, where for $n \le n_0$

$$\widehat{S}_n \le \left(\frac{\nu+1}{\mu} + \gamma_m (\nu+1)^{1/\alpha_s} \right)^+, \tag{31}$$

and for $n > n_0$

$$\widehat{S}_n \le \max \left\{ \begin{array}{l} \displaystyle \left(\frac{\nu-k+1}{\mu} + \gamma_m (\nu-k+1)^{1/\alpha_s} \right)^+ - \frac{n-n_0}{\lambda} + \gamma_a (n-n_0)^{1/\alpha_a}, \\ \displaystyle \max_{\phi < k \le \nu} \left(\frac{\nu-k+1}{\mu} + \Gamma_m^+ (\nu-k+1)^{1/\alpha_s} - \frac{m(\nu-k)}{\lambda} + \Gamma_a [m(\nu-k)]^{1/\alpha_a} \right) \end{array} \right\}. \tag{32}$$

As for initially empty queues, the optimization problem in Eq. (32) can be computed efficiently for the general case where $\alpha_a \neq \alpha_s$. Theorem 3 provides a closed form expression for the upper bound on the worst case system time for the special case where $\alpha_a = \alpha_s$.

Theorem 3 (Highest System Time in an Initially Nonempty Heavy-Tailed Queue) *In an m-server FCFS queue under Assumption 2 with $n_0 \in \mathcal{K}_\phi$, where, $\phi = \lfloor (n_0-1)/m \rfloor$, $\alpha_a = \alpha_s = \alpha$ and $\rho < 1$, the worst case system time for $\Gamma = m^{1/\alpha}\Gamma_a + \Gamma_m^+ > 0$ is given by*

$$\widehat{S}_n(\Gamma) \le \max \left\{ \begin{array}{l} \left(\dfrac{\nu+1}{\mu} + \gamma_m (\nu+1)^{1/\alpha}\right)^+ - \dfrac{n-n_0}{\lambda} + \gamma_a (n-n_0)^{1/\alpha}, \\ \begin{cases} \Gamma(\nu-\phi)^{1/\alpha} - \dfrac{m(1-\rho)}{\lambda}(\nu-\phi) + \left(\dfrac{1}{\mu} + \Gamma_m^+\right), & \text{if } \nu - \phi < \left(\dfrac{\lambda\Gamma/m}{\alpha(1-\rho)}\right)^{\alpha/(\alpha-1)}, \\ \dfrac{\alpha-1}{\alpha^{\alpha/(\alpha-1)}} \dfrac{\lambda^{1/(\alpha-1)} \cdot \Gamma^{\alpha/(\alpha-1)}}{[m(1-\rho)]^{1/(\alpha-1)}} + \left(\dfrac{1}{\mu} + \Gamma_m^+\right), & \text{otherwise.} \end{cases} \end{array} \right\} \quad (33)$$

Note that, for the case where $\Gamma \le 0$, the worst case system time

$$\widehat{S}_n(\Gamma) \le \max \left\{ \left(\frac{\nu+1}{\mu} + \gamma_m(\nu+1)^{1/\alpha_s}\right)^+ - \frac{n-n_0}{\lambda} + \gamma_a(n-n_0)^{1/\alpha_a},\ \frac{1}{\mu} + \Gamma_m^+ \right\}.$$

In this case, the n^{th} job experiences a waiting time only due to the buildup effect left by the initial jobs. For big enough n, this effect becomes negligible and the system time eventually becomes equal to the service times, stabilizing at the value $1/\mu + \Gamma_m^+$.

Implications and Insights

In a multi-server queue, the worst case system time is characterized by two distinct states of behavior: **(a)** a *transient state* where the system time is dependent on n, and **(b)** a *steady state* where the system time is independent of n. Figure 3 shows a graphical representation of the evolution of the worst case system time under our modeling assumptions.

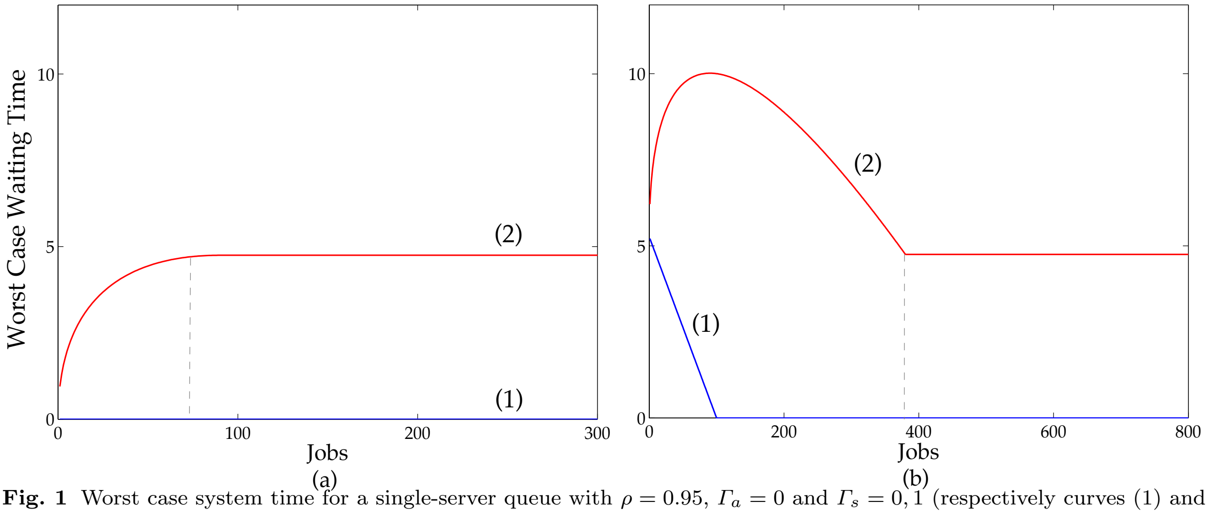


Fig. 1 Worst case system time for a single-server queue with $\rho = 0.95$, $\Gamma_a = 0$ and $\Gamma_s = 0, 1$ (respectively curves (1) and (2)), for (a) zero initial jobs, i.e., $n_0 = 0$, and (b) 5 initial jobs, i.e., $n_0 = 5$. The dotted lines indicate the phase change from transient to steady state.

In the queueing literature, the time it takes the system to reach steady state is referred to as *relaxation time*. We define the *robust relaxation time* as the number of jobs observed by the queue before reaching steady state in the worst case setting. Table 3 summarizes the effect of the traffic intensity on the steady-state system time and the robust relaxation time.

Remark: Under probabilistic assumptions, heavy-tailed queues are characterized by an infinitely long transient state as they never reach steady state (see [20]). However, in our robust framework, we attribute a steady state value, even for queues with heavy-tailed arrivals/services. The concept of a worst case steady state for systems with heavy tails stems from the assumptions of boundedness of the inter-arrival and service times implied by Assumption 2, which involve a truncation of the tails. Specifically, under the worst case paradigm, lower tail coefficients, and therefore heavier tails, yield an increase in both the relaxation and steady state system times as suggested by Table 1. To illustrate this, we consider an instance with $\rho = 0.95$, $m = 1$ and $\Gamma = 1$. By incrementally decreasing the tail coefficient from $\alpha = 2$ to $\alpha = 1.75$ and

Table 1 Effect of traffic intensity and heavy tails on worst case behavior of multi-server queues.

Worst Case Steady System Time*	Robust Relaxation Time*
$\mathcal{O}\left(\dfrac{(\Gamma^{+})^{\alpha/(\alpha-1)}}{m(1-\rho)^{1/(\alpha-1)}}\right)$	$\mathcal{O}\left(m\cdot\left[\dfrac{\Gamma^{+}}{m(1-\rho)}\right]^{\alpha/(\alpha-1)}\right)$

* $\Gamma = m^{1/\alpha}\Gamma_a + \Gamma_m$.

from $\alpha = 1.75$ to $\alpha = 1.5$, the steady state worst case system time experiences an respective increase by 115% and 420%, and the relaxation time increases by 190% and 680% respectively. Our averaging technique allows us to reconcile our approach with the conclusions from probabilistic queueing theory.

For ease of notation, we express the worst case system time in Eq. (33) as

$$\max\left\{\widehat{S}_n^{\,b}\left(\gamma_a,\gamma_m\right),\ \widehat{S}_n^{t}\left(\Gamma\right)\cdot\mathbb{1}_n^{\,t}\left(\Gamma\right)+\widehat{S}^{s}\left(\Gamma\right)\cdot\mathbb{1}_n^{\,s}\left(\Gamma\right)\right\},\tag{34}$$

where $\widehat{S}_n^{\,b}$, $\widehat{S}_n^{\,t}$, and $\widehat{S}^{s}$ denote the quantities associated with the system time effected by the initial buffer n_0, the transient state and the steady state, respectively, i.e.,

$$\left\{\begin{aligned}\widehat{S}_n^{\,b} &= \left(\frac{\nu+1}{\mu}+\gamma_m\left(\nu+1\right)^{1/\alpha}\right)^{+}-\frac{n-n_0}{\lambda}+\gamma_a\left(n-n_0\right)^{1/\alpha},\\ \widehat{S}_n^{\,t} &= \Gamma\left(\nu-\phi\right)^{1/\alpha}-\frac{m(1-\rho)}{\lambda}\left(\nu-\phi\right)+\left(\frac{1}{\mu}+\Gamma_m^{+}\right),\\ \widehat{S}^{s} &= \frac{(\alpha-1)}{\alpha^{\alpha/(\alpha-1)}}\frac{\lambda^{1/(\alpha-1)}\cdot\Gamma^{\alpha/(\alpha-1)}}{\left[m(1-\rho)\right]^{1/(\alpha-1)}}+\left(\frac{1}{\mu}+\Gamma_m^{+}\right),\end{aligned}\right\},$$

and the indicator functions $\mathbb{1}_n^{\,t}$ and $\mathbb{1}_n^{\,s}$ reflect the condition for the system to be in the transient state and the steady state, respectively. For $\alpha_a = \alpha_s = \alpha$, the indicator functions are such that

$$\begin{cases}\mathbb{1}_n^{\,t}\left(\Gamma\right)=1, & \text{if } \Gamma > \dfrac{\alpha m(1-\rho)}{\lambda}\cdot\Big[\lfloor n/m\rfloor-\lfloor n_0/m\rfloor\Big]^{(\alpha-1)/\alpha},\\ \mathbb{1}_n^{\,s}\left(\Gamma\right)=1, & \text{otherwise.}\end{cases}$$

3.3 Average Case Behavior

To analyze the average behavior of a multi-server queue, we treat the parameters (γ_a,Γ_a), and (γ_m,Γ_m) (correspondingly (γ_s,Γ_s) for a single-server queue) as random variables and compute the expected value of the worst case system time

$$\widetilde{S}_n = \mathbb{E}\left[\widehat{S}_n\right].$$

Similarly to the case of a single-server queue with light-tailed primitives, we propose to approximate the density of the variability parameters by invoking the limit laws of probability and leveraging the characterization of the effective variability in Eq. (14) to fit the analysis for multi-server queues with possibly heavy-tailed arrivals and services.

Choice of Variability Distribution

From Eq. (27), the parameters γ_a and γ_s can be viewed as normalized sums of the random variables $\{T_{n_0+1},\ldots,T_n\}$ and $\{X_1,\ldots,X_n\}$. Specifically,

$$\gamma_a = -\left[\frac{\displaystyle\sum_{i=n_0+1}^{n}T_i-\frac{n-n_0}{\lambda}}{\left(n-n_0\right)^{1/\alpha_a}}\right]\propto -Z_a \quad\text{and}\quad \gamma_s = \left[\frac{\displaystyle\sum_{i=1}^{n}X_i-\frac{n}{\mu}}{n^{1/\alpha_s}}\right]\propto Z_s.\tag{35}$$

By the limit laws of probability, γ_a and γ_s approximately behave as a random variable following a limiting distribution.

(a) Light Tails: For large enough n, γ_a and γ_s can be well approximated as normally distributed random variables by the central limit theorem. Specifically, $\gamma_a \sim \mathcal{N}(0, \sigma_a)$ and $\gamma_s \sim \mathcal{N}(0, \sigma_s)$, where σ_a and σ_s denote the standard deviations associated with the inter-arrival and service processes, respectively.

(b) Heavy Tails: By Theorem 1, the normalized sum of heavy-tailed random variables with tail coefficient α follows a stable distribution $S_\alpha(\psi, \xi, \phi)$ with a skewness parameter $\psi = 1$, a scale parameter $\xi = 1$ and a location parameter $\phi = 0$. Therefore, γ_a and γ_s as expressed in Eq. (35) are such that

$$\gamma_a \sim \mathcal{S}_{\alpha_a}(-1, C_{\alpha_a}, 0) \quad \text{and} \quad \gamma_s \sim \mathcal{S}_{\alpha_s}(1, C_{\alpha_s}, 0),$$

where C_α is a normalizing constant as introduced in Eq. (19). As a concrete example, for Pareto distributed inter-arrivals and service times,

$$C_\alpha = \left[\Gamma(1-\alpha)\cos(\pi\alpha/2)\right]^{1/\alpha},$$

where $\Gamma(\cdot)$ denotes the Gamma function. Note that, unlike the case of light tails, the distributions of γ_a and γ_s are asymmetrical. More specifically, the skewness of γ_a is negative since $\gamma_a = -Z_a$, where $Z_a = S_{\alpha_a}(1, C_{\alpha_a}, 0)$.

In a multi-server queue, and assuming without loss of generality that $n = (\nu+1)m$, we obtain

$$\gamma_s = \frac{\sum_{i=1}^{(\nu+1)m} X_i - \frac{(\nu+1)m}{\mu}}{\left[(\nu+1)m\right]^{1/\alpha}} = \frac{1}{m^{1/\alpha_s}} \cdot \sum_{j=1}^{m} \left[\frac{\sum_{i=0}^{\nu} X_{j+im} - \frac{\nu+1}{\mu}}{(\nu+1)^{1/\alpha_s}}\right] = \frac{1}{m^{1/\alpha_s}} \cdot \sum_{j=1}^{m} \gamma_m,$$

where the last inequality is due to Eq. (28). We can therefore express γ_m as

$$\gamma_m = \frac{1}{m^{(\alpha_s-1)/\alpha_s}} \cdot \gamma_s.$$

We next discuss how we choose the distribution of the effective parameter Γ. Since the exact characterization of the density of Γ is challenging, as we have observed in Section 2, we propose an approximation. Recall that for a single-server queue with light-tailed arrival and service times, we have proposed to treat Γ as

$$\Gamma \sim \mathcal{N}\left(0, 2\sqrt{\sigma_a^2 + \sigma_s^2}\right). \tag{36}$$

Put differently, we view $\Gamma = \Gamma_a + \Gamma_s$, where $\Gamma_a = \theta\gamma_a$ and $\Gamma_s = \theta\gamma_s$ with $\theta = 2$. We take a similar approach for multi-server queues and model the variability parameters as functions of γ_a, γ_s and γ_m as follows

$$\Gamma_a = \theta\gamma_a \quad \text{and} \quad \Gamma_m = \theta\gamma_m = \theta\frac{\gamma_s}{m^{(\alpha_s-1)/\alpha_s}},$$

and then inform the choice of the scaling parameter θ via known conclusions on the behavior of the system time (e.g., the bound on the steady-state behavior by [39]).

(a) Light Tails: We select θ so that the average worst case steady-state system time matches the bound provided by [39]. In other words, we ensure that

$$\frac{\lambda}{4(1-\rho)} \cdot \mathbb{E}\left[\left(\theta\gamma^+\right)^2\right] = \frac{\lambda}{2(1-\rho)} \cdot \left(\sigma_a^2 + \sigma_s^2/m^2\right), \tag{37}$$

where $\gamma = \gamma_a + \gamma_m^+/m^{1/2} = \gamma_a + \gamma_s^+/m$ and the expected value $\mathbb{E}\left[\left(\gamma^+\right)^2\right] \approx \mathbb{P}(\gamma \geq 0) \cdot \left(\sigma_a^2 + \sigma_s^2/m^2\right)$. By rearranging the terms in Eq. (37), we obtain

$$\theta = \left[\frac{2\left(\sigma_a^2 + \sigma_s^2/m^2\right)}{\mathbb{E}\left[(\gamma^+)^2\right]}\right]^{1/2} \approx \left(\frac{2}{\mathbb{P}(\gamma \geq 0)}\right)^{1/2}. \tag{38}$$

(b) Heavy Tails: The steady state in heavy-tailed queues does not exist. Instead, we propose to extend the formula in Eq. (38). For $\alpha_a = \alpha_s = \alpha$, we select the scaling parameter as

$$\theta \approx \left(\frac{\alpha}{\mathbb{P}(\gamma \geq 0)}\right)^{(\alpha-1)/\alpha}. \tag{39}$$

where the probability can be efficiently computed numerically. For asymmetric tails, we propose to model the variability parameters $\Gamma_a = \theta_a\gamma_a$ and $\Gamma_m = \theta_s\gamma_m$, with

$$\theta_a \approx \left(\frac{\alpha_a}{\mathbb{P}(\gamma \geq 0)}\right)^{(\alpha_a-1)/\alpha_a} \quad \text{and} \quad \theta_s \approx \left(\frac{\alpha_s}{\mathbb{P}(\gamma \geq 0)}\right)^{(\alpha_s-1)/\alpha_s}. \tag{40}$$

By expressing $\varGamma_a$ and $\varGamma_m$ in terms of γ_a and γ_s, we can approximate $\widetilde{S}_n$ by

$$\widetilde{S}_n \approx \mathbb{E}_{\gamma_a,\gamma_s}\left[\max\left\{\widehat{S}_n^{\,b}\left(\gamma_a,\gamma_s\right),\ \widehat{S}_n^{t}\left(\gamma_a,\gamma_s\right)\cdot\mathbb{1}_n^{\,t}\left(\gamma_a,\gamma_s\right)+\widehat{S}^{s}\left(\gamma_a,\gamma_s\right)\cdot\mathbb{1}_n^{s}\left(\gamma_a,\gamma_s\right)\right\}\right].$$

The above double integral can be efficiently computed using numerical integration. A key feature of our approximation approach is its computational tractability. Computing the average system time involves computing double integrals, which we compute by discretization the space of γ_a and γ_s. The average runtime to compute $\widetilde{S}_n$ for a given value of n is of the order of milli-seconds, irrespective of the system parameters: traffic ratio (ρ), number of servers (m), and light or heavy tailed nature (α). We contrast the computational requirement of our approach relative to simulations.

(a) **Computational Complexity:** When using simulation to calculate $\mathbb{E}[S_n]$, it is required to simulate all the jobs until n, requiring us to simulate an $\mathcal{O}(n)$–dimensional random vectors of inter-arrival times and service times. On the other hand, in our approach, we are required to perform only a double integration, which is significantly faster.

(b) **Effect of Heavy Tails and Heavy Traffic:** It is well known that the number of sample paths required grows for heavy traffic as well as heavy tailed systems (see [27,5,19]). In our approach, even for heavy tails and heavy traffic, we use the same level of discretization to calculate the double integrals.

(c) **Simulation of Multi-Server Systems:** A key step in simulating FCFS multi-server queues consists of sorting the workloads at each server to assign the next job to the first available server. This sorting process is required for each sample path. On the other hand, our approach provides a closed form expression for multi-server queues which does not involve sorting.

We next compare the performance of the proposed approximation with simulated values.

3.4 Computational Results

We investigate the performance of our approach relative to simulation and examine the effect of the system's parameters (traffic intensity, initial buffer and number of servers) on its accuracy. We run simulations for single and multi-server queues with $N = 5,000$ job arrivals and compute the expected system time for each job using 20,000 simulation replications. We pre-specify the arrival rate at the queue to be $\lambda = 0.1$ for all simulation instances, while varying the traffic intensity, the variances associated with the inter-arrival and service processes, the number of servers in the queue, and the number of initial jobs. We further consider a host of light-tailed distributions and simulate queues with normal, exponential, log-normal, and uniform inter-arrival and service times (including the service times for the initial jobs at the queue). To compare the simulated values $\overline{S}_n$ with our approximation $\widetilde{S}_n$, we report the average percent error defined as

$$\text{Average Percent Error} = \frac{1}{\widetilde{N}}\cdot\sum_{n=1}^{\widetilde{N}}\left|\frac{\overline{S}_n-\widetilde{S}_n}{\overline{S}_n}\right|\times 100\%,$$

where

$$\widetilde{N} = \min\left(N, \widetilde{n}_r\right), \tag{41}$$

$and\widetilde{n}_r$ denotes the number of jobs the queue observes until our approximation reaches steady state, i.e., $\widetilde{n}_r = \min\left(n : \widetilde{S}_n = \widetilde{S}_\infty\right)$.

We next present our results for multi-server queues with (a) light-tails ($\alpha_a = \alpha_s = 2$), (b) symmetric heavy tails ($\alpha_a = \alpha_s = \alpha$), and (c) asymmetric tails ($\alpha_a \neq \alpha_s$).

Light Tails: Table 2 reports the average percent error between simulation and our approximation for queues with normally distributed inter-arrival and service times. Note that the choice of the mean and standard deviations ensures that no more than 0.6% of values are negative. Whenever we obtain a negative value, we truncated at zero. Our approach generally yields percent errors within 10% relative to simulation. Figure 2 compares our approximation (dotted line) with simulation (solid line) for a single-server queue (top panels) and a 20-server queue (bottom panels) with normally distributed primitives.

As shown by simulations and empirical studies performed by [51] on light-tailed queueing systems, the expected transient system time has broadly four different behaviors depending on the initial jobs. Our averaging approach is capable of capturing these behaviors.

(a) The first behavior occurs when the system is initially empty. The average system time function is monotonic and concave in n. This behavior is detected in Figures 2(a),(d).

Table 2 Errors (%) relative to simulations for multi-server queues with normally distributed primitives.

		1 Server*			10 Servers†			20 Servers‡		
	ρ	$n_0 = 0$	$n_0 = 5$	$n_0 = 10$	$n_0 = 0$	$n_0 = 20$	$n_0 = 50$	$n_0 = 0$	$n_0 = 50$	$n_0 = 100$
$\sigma_a = 2.5$	.95	5.14	3.32	6.82	1.06	3.04	2.19	0.87	1.53	1.03
	.97	4.04	2.26	5.98	0.44	3.12	2.25	0.60	1.99	1.10
	.99	3.54	1.54	8.77	2.35	4.98	2.73	1.27	2.89	0.62
$\sigma_a = 4.0$	.95	2.23	2.57	6.44	0.64	3.28	3.59	1.21	2.60	2.11
	.97	1.75	2.16	7.65	1.49	4.14	4.85	0.59	3.33	3.39
	.99	5.05	4.09	8.51	4.47	7.70	5.31	2.83	5.08	1.50

* Instances with single-server queues with (a) $\sigma_a = \sigma_s = 2.5$ and (b) $\sigma_a = \sigma_s = 4.0$

† Instances with 10-server queues with (a) $\sigma_a = 2.5$ and $\sigma_s = 10$, and (b) $\sigma_a = 4.0$ and $\sigma_s = 20$

‡ Instances with 20-server queues with (a) $\sigma_a = 2.5$ and $\sigma_s = 20$, and (b) $\sigma_a = 4.0$ and $\sigma_s = 40$

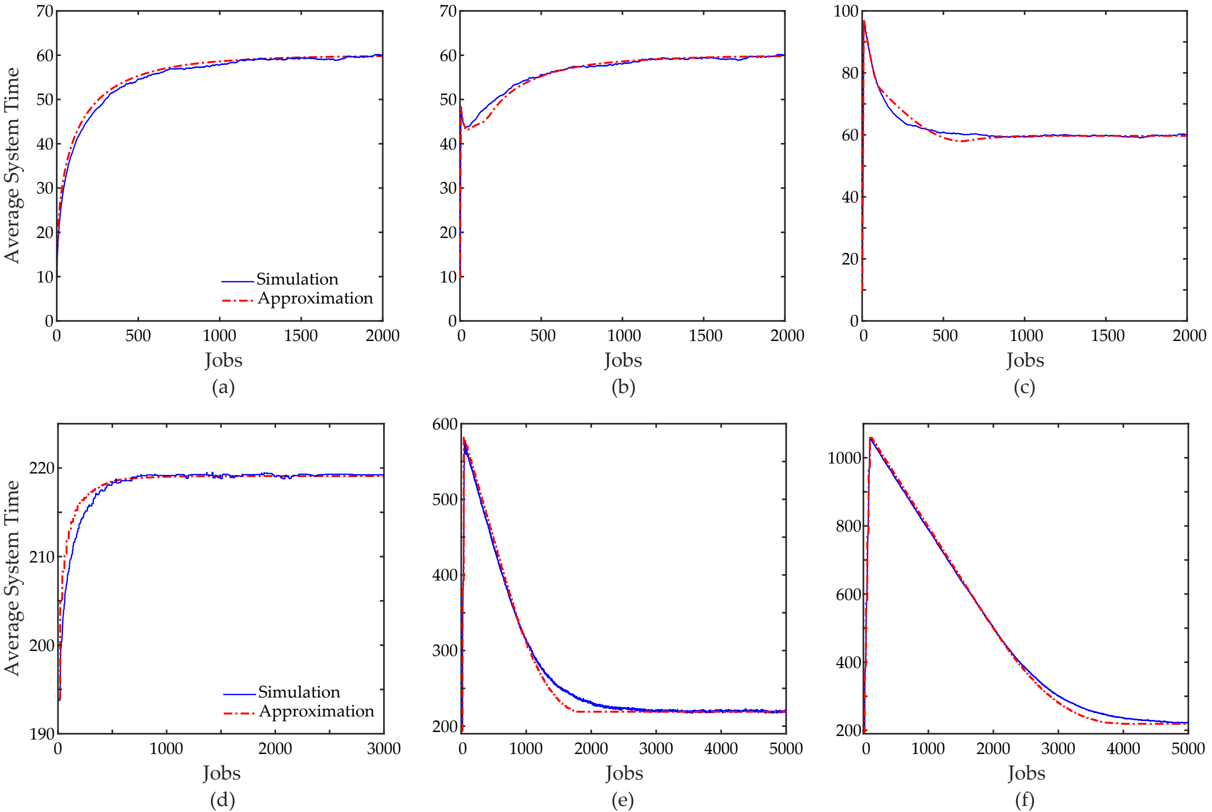


Fig. 2 Simulated (solid line) versus approximated values (dotted line) for a queue with normally distributed primitives with $\sigma_a = 4.0$ and $\rho = 0.97$. Panels (a)–(c) show a single-server queue with $\sigma_s = 4.0$ and $n_0 = 0, 5, 10$. Panels (d)–(f) show a 20-server queue with $\sigma_s = 40$ and $n_0 = 0, 50, 100$.

(b) The second behavior occurs when the number of initial jobs is small creating an initial system time $\widetilde{S}_{n_0}$ that is below the steady state value. The system time in this case initially decreases and subsequently increases until reaching steady state, as seen in Figure 2(b).

(c) The third behavior occurs when the number of initial jobs creates an initial system time $\widetilde{S}_{n_0}$ that is higher than the steady state value. In this case, the average system time is convex in n and decreases exponentially until reaching steady state, as detected in in Figure 2(c).

(d) The fourth behavior occurs when the initial buffer creates an initial system time $\widetilde{S}_{n_0}$ that is substantially larger than the steady state value. The initial decrease is approximately linear with jobs leaving the system at the rate of $\mu - \lambda$, as seen in Figures 2(e),(f).

Table 3 reports the average percent error between simulation and our approximation for queues with various combinations of light-tailed distributions (with $\lambda = 0.1$ and $\sigma_a = 10$). We consider in particular three pairs of distributions: (A) exponential arrivals and log-normal service times, (B) log-normal arrivals and service times, and (C) uniform arrivals and log-normal service times. We also vary the coefficients of variation associated with the inter-arrival times ($c_a = \lambda\sigma_a$) and the service times ($c_s = \mu\sigma_s$). Our approach yields errors within 10% relative to simulation. Figure 3 compares our approximation (dotted line) with simulation (solid lines) for an initially empty (a) single-server queue, (b) 10-server queue, and (c) 20-server queue for the various combination of distributions.

Table 3 Errors (%) relative to simulation for queues with light-tailed primitives.

		1 Server			10 Servers			20 Servers		
	Instance*	$\rho = .95$	$\rho = .97$	$\rho = .99$	$\rho = .95$	$\rho = .97$	$\rho = .99$	$\rho = .95$	$\rho = .97$	$\rho = .99$
$c_a = c_s$	A*	5.18	3.10	2.26	7.48	4.78	3.99	10.2	7.80	5.91
	B†	2.64	2.06	2.62	9.06	5.46	4.10	10.9	8.76	7.04
	C‡	3.75	2.52	1.50	6.97	4.37	3.55	9.45	7.58	6.05
$c_a = 2c_s$	A	8.14	4.66	2.82	3.39	2.23	2.98	5.37	2.71	2.03
	B	6.21	4.36	3.44	5.42	1.96	2.85	6.34	3.50	1.88
	C	4.70	3.14	1.17	2.11	2.52	2.97	4.25	1.72	1.87
$c_a = 5c_s$	A	4.17	3.63	1.71	5.81	2.51	2.09	6.18	3.77	1.48
	B	9.17	5.87	3.33	7.80	3.88	1.95	7.33	4.65	2.08
	C	0.71	0.82	1.43	3.76	1.34	1.89	4.88	2.67	1.63

* Instances with exponential arrivals and log-normal service times.
† Instances with log-normal arrivals and service times
‡ Instances with uniform arrivals and log-normal service times

Heavy Tails: Table 4 reports the average percent error between simulation and our approximation for queues with Pareto distributed inter-arrival and service times with $\alpha_a = \alpha_s = \alpha$. Our approach yields percent errors within 10% relative to simulation for single-server queues. While errors are higher for multi-server queues, our approximation still captures the heavy-tailed behavior. Figure 4 compares our approximation (dotted line) with simulation (solid line) for a single-server queue (top panels) and a 20-server queue (bottom panels) with Pareto distributed primitives ($\alpha_a = \alpha_s = 1.6$).

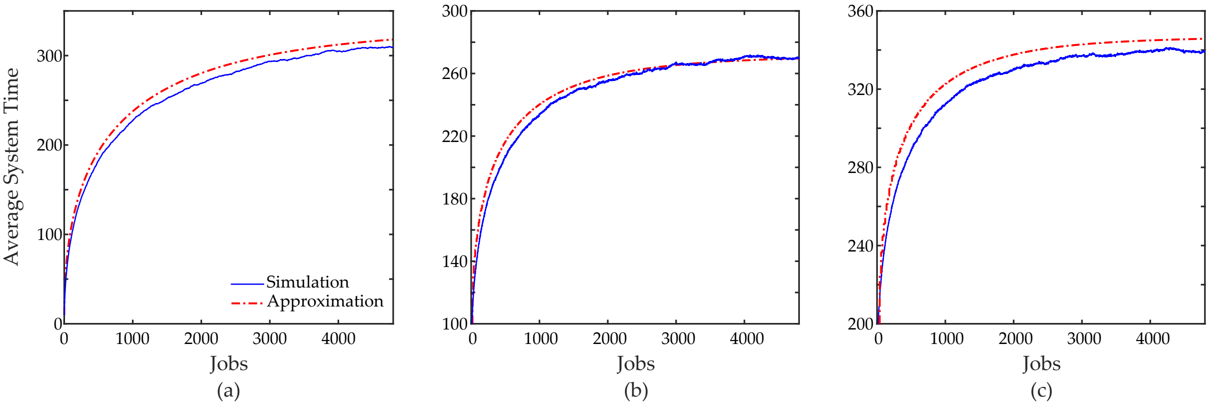
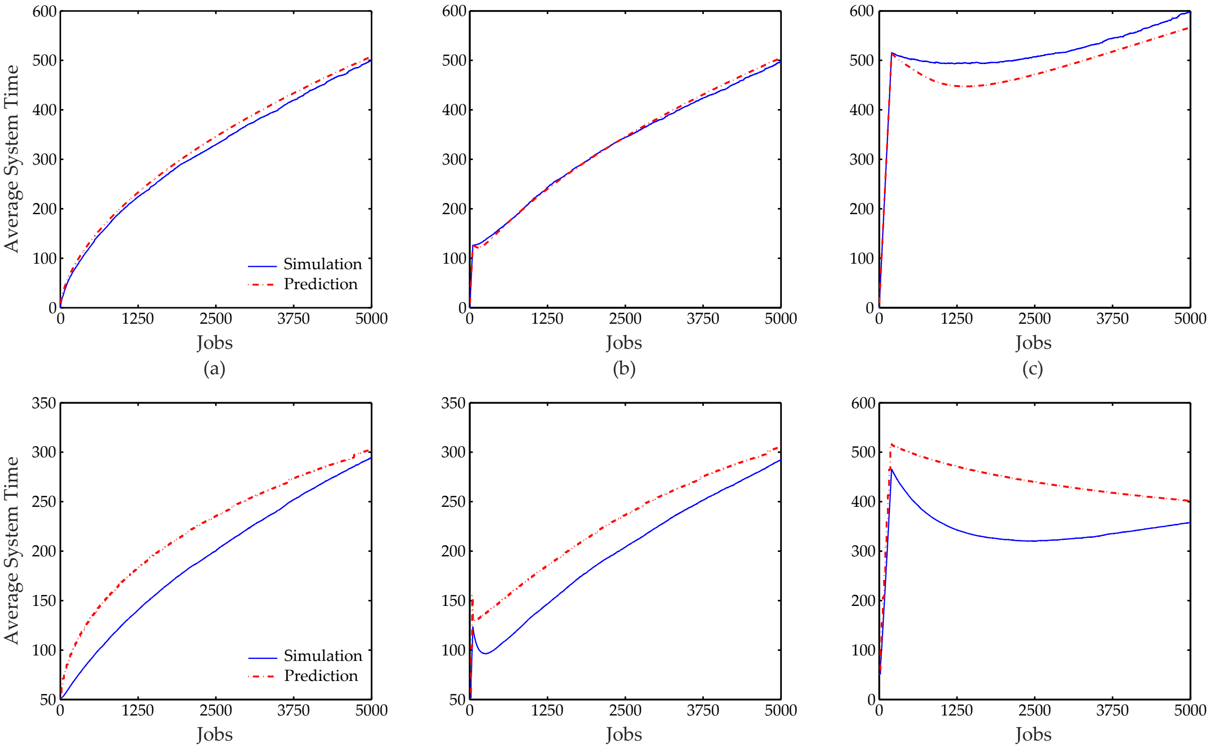


Fig. 3 Simulated (solid line) versus predicted values (dotted line) for a queue with $\rho = 0.97$. Panel (a) shows a single-server queue with exponential arrivals and log-normal service times with $c_a = c_s$. Panel (b) shows a 10-server queue with log-normal arrivals and service times with $c_a = 2c_s$. Panel (c) shows a 20-server queue with uniform arrivals and log-normal service times with $c_a = 5c_s$.

Note that our averaging technique allows us to reconcile our conclusions with probabilistic queueing theory for single server queues and for multi-server queues in heavy-traffic regime. In particular, it is well known ([20, 58]) that for single server queues under heavy tailed service distributions, the expected steady state waiting time is infinite. Additionally, [28, 29] have obtained a similar result for multi server queues under heavy traffic. We are able to match these results. In particular, from Table 1, the average system time is proportional to $\mathbb{E}\left[(\varGamma^{+})^{\alpha/(\alpha-1)}\right]$. For heavy-tailed primitives, the effective variability parameter $\varGamma$ is governed by a heavy-tailed distribution (concluded for the stable law). This implies that the moments of $\varGamma$



(a) (b) (c)

(d) (e) (f)

Fig. 4 Simulated (solid line) versus predicted values (dotted line) for a single queue with Pareto distributed primitives ($\alpha_a = \alpha_s = 1.6$) and $\rho = 0.97$. Panels (a)–(c) correspond to an instance with $m = 1$ and $n_0 = 0, 50, 200$. Panels (d)–(f) correspond to an instance with $m = 20$ and $n_0 = 0, 50, 200$.

higher than or equal to the second moment are infinite. As a result, $\mathbb{E}\left[(\varGamma^+)^{\alpha/(\alpha-1)}\right]$ is infinite for $\alpha < 2$. The average steady-state system time $\widetilde{S}_\infty$ and the relaxation time are therefore infinite.However, note that for multi-server queues under low ρs, we are only able to provide upper bounds.

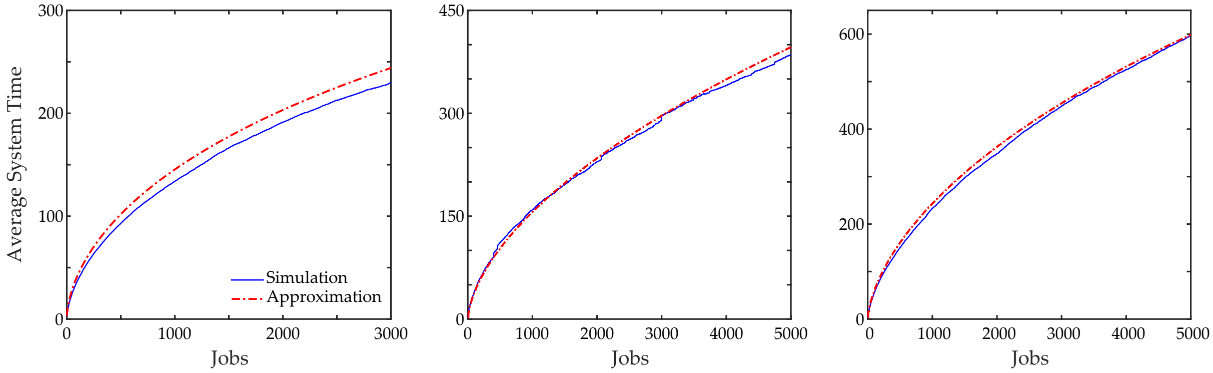
Table 4 Errors relative to simulations for multi-server queues with Pareto distributed primitives.

		1 Server			10 Servers			20 Servers		
	ρ	$n_0 = 0$	$n_0 = 50$	$n_0 = 200$	$n_0 = 0$	$n_0 = 50$	$n_0 = 200$	$n_0 = 0$	$n_0 = 50$	$n_0 = 200$
$\alpha = 1.6$	0.95	9.59	7.18	1.78	12.5	9.49	13.9	17.9	15.9	25.5
	0.97	4.86	1.49	5.98	12.1	9.56	13.7	19.6	17.8	28.6
	0.99	2.59	2.08	6.63	11.9	11.9	15.6	24.5	22.6	29.3
$\alpha = 1.7$	0.95	9.59	7.18	1.78	9.22	7.85	5.44	21.6	18.5	17.4
	0.97	8.75	3.14	2.92	12.7	9.63	9.76	21.7	17.7	19.8
	0.99	5.72	1.17	3.66	13.9	13.5	11.4	24.4	20.3	20.4

Asymmetric Tails: Figure 5 compares our approximation (dotted line) with simulation (solid lines) for a single-server queue with $\rho = 0.97$ and asymmetric tail coefficients. In particular, we consider three instances: (a) Pareto arrivals ($\alpha_a = 1.6$ and exponential service times, (b) exponential arrivals and Pareto service times ($\alpha_s = 1.6$), and (c) Pareto arrivals and services ($\alpha_a = 1.5$, $\alpha_s = 1.7$).

Remark

Note that the accuracy of our approach depends on the accuracy of limit laws, which depend on the value of n. However, in the heavy traffic transient regime, the relevant values of n are indeed high (at least in 100s) and in this regime CLT is a good approximation. In particular, as we observe in Table 4, the errors are indeed smaller in heavy traffic, but higher (still under 10%) for a single server queue.



(a) (b) (c)

Fig. 5 Simulated (solid line) versus predicted values (dotted line) for an initially empty single-server queue with $\rho = 0.97$ and (a) Pareto arrivals ($\alpha_a = 1.6$) and exponential service times, (b) exponential arrivals and Pareto service times ($\alpha_s = 1.6$), and (c) Pareto arrivals and services ($\alpha_a = 1.5$ and $\alpha_s = 1.7$). Percent errors with respect to simulation are 6.50%, 2.82%, and 3.23%, respectively.

4 Extensions to Tandem Networks

In this section, we extend our analysis of single queues to the analysis of tandem queues. We consider a network of J queues in series and study the expected overall system time $\overline{S}_n$ given by

$$\overline{S}_n = \mathbb{E}\left[S_n^{(1)} + \ldots + S_n^{(J)}\right] = \sum_{j=1}^{J} \mathbb{E}\left[S_n^{(j)}\right] = \sum_{j=1}^{J} \overline{S}_n^{(j)},$$

where $S_n^{(j)}$ is the system time of the n^{th} job in the j^{th} queue. Similarly to the analysis of a single queue, we assume the inter-arrival and service times belong to polyhedral sets which allow us to study the worst case system time. We then leverage the worst case values to perform an average case analysis.

We assume that the inter arrival times $\mathbf{T} = (T_1, \ldots, T_n)$ to the tandem network belong to the uncertainty set $\mathcal{U}^a$, and the service times $\mathbf{X}^{(j)} = \left\{X_1^{(j)}, \ldots, X_n^{(j)}\right\}$ at each queue j, for $j = 1, \ldots, J$ satisfy the uncertainty sets as described in Assumption 2. We summarize the assumptions on the service times as follows.

Assumption 3.:
We make the following assumptions on the service times in a tandem queue.

(a) For a single-server queue j, the service times belong to the uncertainty set

$$\mathcal{U}_j^s = \left\{ \left(X_1^{(j)}, \ldots, X_n^{(j)}\right) \;\middle|\; \begin{array}{l} \displaystyle\sum_{i=1}^{n} X_i^{(j)} - n/\mu_j \leq \gamma_s^{(j)}\, n^{1/\alpha_s^{(j)}}, \\ \displaystyle\sum_{i=k+1}^{\ell} X_i^{(j)} - \frac{\ell - k}{\mu_j} \leq \Gamma_s^{(j)}\, (\ell - k)^{1/\alpha_s^{(j)}}, \quad \forall\; 0 \leq k < \ell \leq n \end{array} \right\},$$

where the parameters $\gamma_s^{(j)}, \Gamma_s^{(j)} \in \mathbb{R}$ control the degree of conservatism, and $1 < \alpha_s^{(j)} \leq 2$ is a tail coefficient modeling possibly heavy tailed service times.

(b) For an m-server queue j, the service times belong to the uncertainty set

$$\mathcal{U}_j^m = \left\{ \left(X_1^{(j)}, \ldots, X_n^{(j)}\right) \;\middle|\; \begin{array}{l} \displaystyle\sum_{i=0}^{\nu} X_{k_i}^{(j)} - \frac{\nu + 1}{\mu_j} \leq \gamma_m^{(j)}\, (\nu + 1)^{1/\alpha_s^{(j)}}, \; \forall\; k_i \in K_i \\ \displaystyle\sum_{i \in \mathcal{I}} X_{k_i}^{(j)} - \frac{|\mathcal{I}|}{\mu_j} \leq \Gamma_m^{(j)}\, |\mathcal{I}|^{1/\alpha_s^{(j)}}, \quad \forall\; k_i \in K_i, \text{ and } \; i \in \mathcal{I} \subseteq \{0, \ldots, \nu\}, \end{array} \right\}.$$

where $\nu = \lfloor (n-1)/m \rfloor$, the set $K_i = \left\{im + 1, \ldots, (i+1)m\right\}$, the parameters $\gamma_m^{(j)}, \Gamma_m^{(j)} \in \mathbb{R}$ control the degree of conservatism, and $1 < \alpha_s^{(j)} \leq 2$ is a tail coefficient modeling possibly heavy tailed service times.

In a single-server tandem network, the system time of the n^{th} job at the j^{th} queue is given by

$$S_n^{(j)} = \max_{0 \leq k_j \leq n} \left(\sum_{i=k_j}^{n} X_i^{(j)} - \sum_{i=k_j+1}^{n} T_i^{(j)} \right),$$

where $\mathbf{T}^{(j)} = \left(T_1^{(j)}, \ldots, T_n^{(j)}\right)$ denote the inter arrival times to queue j. Note that $\mathbf{T}^{(j)}$ is exactly the vector of inter departure times $\mathbf{D}^{(j-1)}$ from queue $j-1$, which are given by

$$\sum_{i=k_j+1}^{n} T_i^{(j)} = \sum_{i=k_j+1}^{n} D_i^{(j-1)} = \sum_{i=k_j+1}^{n} T_i^{(j-1)} + S_n^{(j-1)} - S_{k_j}^{(j-1)}.$$

Recursively, the inter arrival times to queue j can be expressed as a function of the inter arrival times $\mathbf{T}$ to the network and the service times $\mathbf{X}^{(1)}$ through $\mathbf{X}^{(j-1)}$.

[9] show that the inter-departure times belong to the inter-arrival uncertainty set $\mathcal{U}^a$, under the assumption of adversarial servers (see Theorem 4). Specifically, [9] view each queue j from an adversarial perspective, where the servers act so as to maximize the system time of the n^{th} job, for all possible sequences of inter-arrival times. In other words, the servers choose their adversarial service times $\widehat{\mathbf{X}}^{(j)} = \left(\widehat{X}_1^{(j)}, \ldots, \widehat{X}_n^{(j)}\right)$ to achieve $\widehat{S}_n^{(j)}(\mathbf{T})$, for all $\mathbf{T}$.

Theorem 4 (Passing through a Queue With Adversarial Servers)
For a multi-server queue j with inter-arrival times $\mathbf{T}^{(j)} \in \mathcal{U}^a$, adversarial service times $\widehat{\mathbf{X}}^{(j)}$, and $\rho < 1$, the inter-departure times $\mathbf{D} = (D_1, \ldots, D_n)$ belongs to the set $\mathcal{U}^d$ satisfying

$$\mathcal{U}^d \subseteq \mathcal{U}^a = \left\{ (D_1, D_2, \ldots, D_n) \left| \frac{\displaystyle\sum_{i=k+1}^{n} D_i - \frac{n-k}{\lambda}}{(n-k)^{1/\alpha_a}} \geq -\Gamma_a, \ \forall\ 0 \leq k \leq n-1 \right. \right\}. \tag{42}$$

The characterization $\mathcal{U}^d \subseteq \mathcal{U}^a$ is true for all values of n, though its tightness improves for increasing values of n. Consequently, Theorem 4 is only tight under steady-state conditions and is therefore akin to Burke's theorem. We next discuss the implications of this result on our steady-state and transient analysis of tandem networks and illustrate our points using a simple example of single-server queues in tandem with $\alpha_a = \alpha_s^{(j)} = \alpha$, for all $j = 1, \ldots, J$.

Steady-State Analysis: To compute the overall system time under steady-state, [9] decomposed the queueing networks and obtained formulas to compute the effective arrival rate λ_j and the effective parameter $\Gamma_a^{(j)}$ observed at each queue j in the network.

For a tandem queueing network, $\lambda_j = \lambda$ and $\Gamma_a^{(j)} = \Gamma_a$ for all $j = 1, \ldots, J$. By Theorem 2, the worst case steady-state system time at queue j can then be expressed as

$$\widehat{S}_\infty^{(j)} = \frac{(\alpha-1)}{\alpha^{\alpha/(\alpha-1)}} \frac{\lambda_j^{1/(\alpha-1)} \cdot \left(\Gamma^{(j)+}\right)^{\alpha/(\alpha-1)}}{(1-\rho)^{1/(\alpha-1)}} + \left(\frac{1}{\mu} + \Gamma_s^{(j)}\right), \tag{43}$$

where $\Gamma^{(j)} = \Gamma_a + \Gamma_s^{(j)}$, for all $j = 1, \ldots, J$. For light-tailed queues, we compute $\widetilde{S}_\infty^{(j)}$ as in Section 3.2, and approximate the overall expected steady-state system time value by

$$\overline{S}_\infty \approx \widetilde{S}_\infty = \sum_{j=1}^{J} \mathbb{E}\left[\widehat{S}_\infty^{(j)}\right] = \sum_{j=1}^{J} \widetilde{S}_\infty^{(j)} = \sum_{j=1}^{J} \frac{\lambda \left[\sigma_a^2 + \left(\sigma_s^{(j)}\right)^2\right]}{2(1-\rho)} + \frac{1}{\mu_j}. \tag{44}$$

In particular, when $\mu_j = \mu$ and $\sigma_s^{(j)} = \sigma_s$ for all $j = 1, \ldots, J$, the steady-state system time becomes

$$\overline{S}_\infty \approx J \cdot \left[\frac{\lambda\left(\sigma_a^2 + \sigma_s^2\right)}{2(1-\rho)} + \frac{1}{\mu}\right]. \tag{45}$$

Note that this case is a special case of a feed-forward with equal coefficient of variation for all service times. [34] have shown that approximating the behavior of such systems under heavy traffic assumptions can be done through a reflected Brownian motion with a product-form stationary distribution. This implies a decoupling of the queues in steady-state, which is in agreement with our findings. Given that our approximations at each station match those obtained by diffusion theory, our approach yields the same conclusions of [34], and further apply to more complex Jackson networks in steady-state as shown in [9].

Transient Analysis: As noted earlier, the characterization of the inter-departure times in [9] holds for transient regimes, however, it generates loose upper bounds for smaller values of n. Consequently, decoupling the queues and taking a similar approach to the one we took for the steady-state analysis does not generate approximations that are close to simulated values. Figure 6 illustrates our point.

Instead of decomposing the network, we propose to use the recursive formulas that define the dynamics in a network of queues in series to study the overall system time. [14] obtain an exact characterization of the system time for single-server queues in series, with

$$S_n = S_n^{(1)} + \ldots + S_n^{(J)} = \max_{1 \le k_1 \le \ldots \le k_J \le n} \left(\sum_{i=k_1}^{k_2} X_i^{(1)} + \sum_{i=k_2}^{k_3} X_i^{(2)} + \ldots + \sum_{i=k_J}^{n} X_i^{(J)} - \sum_{i=k_1+1}^{n} T_i \right). \quad (46)$$

Given Eq. (46), we analyze the worst case system time and leverage these values to approximate the average behavior. Our approximations are comparable with simulations (see Figure 7).

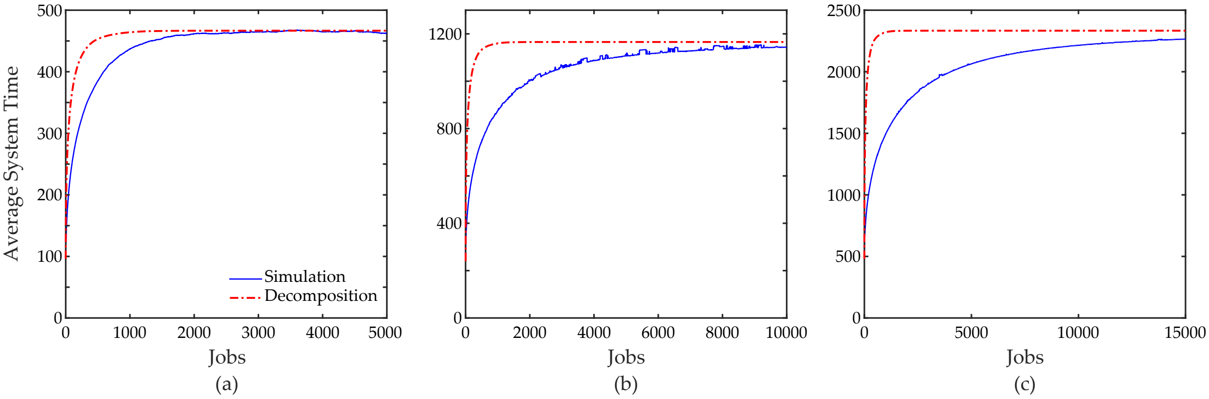


Fig. 6 Simulated (solid line) versus approximation via network decomposition (dotted line) for initially empty tandem networks with normally distributed primitives, $\rho = \rho_j = 0.96$ and $\sigma_a = \sigma_s^{(j)} = 4.0$ for all $j = 1, \ldots, J$, where (a) $J = 10$, (b) $J = 25$, and (c) $J = 50$.

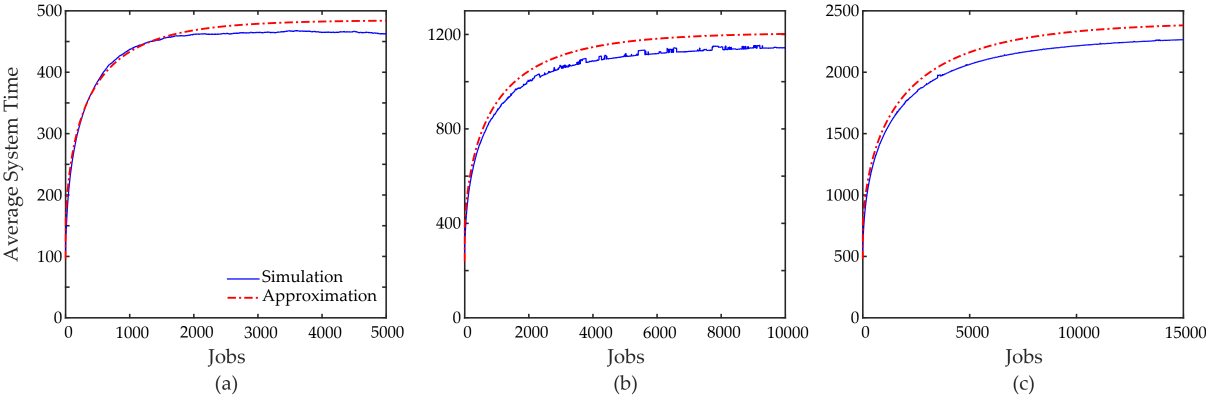


Fig. 7 Simulated (solid line) versus our approximation (dotted line) for initially empty tandem networks with normally distributed primitives, $\rho = \rho_j = 0.96$ and $\sigma_a = \sigma_s^{(j)} = 4.0$ for all $j = 1, \ldots, J$, where (a) $J = 10$, (b) $J = 25$, and (c) $J = 50$. The average percent errors between simulation and our approximation are (a) 2.49% ($\widetilde{N} = 5,000$), (b) 5.02% ($\widetilde{N} = 10,000$), and (c) 5.01% ($\widetilde{N} = 15,000$).

4.1 Worst Case Performance

Under the worst case approach, and applying the adversarial service times at each queue, the worst case system time of the n^{th} job for any realization of $\mathbf{T}$ is given by

$$\widehat{S}_n(\mathbf{T}) = \max_{1 \le k_1 \le \ldots \le k_J \le n} \left(\max_{\mathcal{U}_1^s} \sum_{i=k_1}^{k_2} X_i^{(1)} + \max_{\mathcal{U}_2^s} \sum_{i=k_2}^{k_3} X_i^{(2)} + \ldots + \max_{\mathcal{U}_J^s} \sum_{i=k_J}^{n} X_i^{(J)} - \sum_{i=k_1+1}^{n} T_i \right). \quad (47)$$

Proposition 2 provides a similar result for multi-server queues in series, under the assumption that each queue acts in an adversarial manner to maximize its system time, for all $\mathbf{T}$.

Proposition 2 (Worst Case System Time in a Tandem Queue with Multiple Servers) *In a network of J multi-server queues in series satisfying Assumption 3(b), the overall system time of the n^{th} job for all $\mathbf{T}$ is given by*

$$\widehat{S}_n\left(\mathbf{T}\right) = \max_{0\le k_1\le\ldots\le k_J\le\nu} \left(\max_{\mathcal{U}_1^m} \sum_{i=k_1}^{k_2} X_{r(i)}^{(1)} + \max_{\mathcal{U}_2^m} \sum_{i=k_2}^{k_3} X_{r(i)}^{(2)} + \ldots + \max_{\mathcal{U}_J^m} \sum_{i=k_J}^{n} X_{r(i)}^{(J)} - \sum_{i=r(k_1)+1}^{n} T_i \right), \quad (48)$$

where $r(i) = n-(\nu-i)m$.

The proof is presented in Appendix 1. By minimizing the partial sum of the inter-arrival times, we obtain an exact characterization of the worst case system time in a tandem queue as

$$\widehat{S}_n = \max_{0\le k_1\le\ldots\le k_J\le\nu} \left(\max_{\mathcal{U}_1^m} \sum_{i=k_1}^{k_2} X_{r(i)}^{(1)} + \ldots + \max_{\mathcal{U}_J^m} \sum_{i=k_J}^{n} X_{r(i)}^{(J)} - \min_{\mathcal{U}^a} \sum_{i=r(k_1)+1}^{n} T_i \right). \quad (49)$$

Initially Empty Queues in Tandem

By Assumption 1, the worst case system time is bounded by

$$\widehat{S}_n \le \max_{0\le k_1\le\ldots\le k_J\le\nu} \left\{ \sum_{j=1}^{J} \frac{k_{j+1}-k_j+1}{\mu_j} + \Gamma_m^{(j)+}\left(k_{j+1}-k_j+1\right)^{1/\alpha_s^{(j)}} - \frac{m(\nu-k_1)}{\lambda} + \Gamma_a\left[m\left(\nu-k_1\right)\right]^{1/\alpha_a} \right\} \quad (50)$$

which involves a J-dimensional nonlinear optimization problem. Theorem 5 provides a closed form upper bound on the worst case system time in an initially empty network of J identical queues in tandem, with $\mu_1 = \ldots = \mu_J$ and $\alpha_a = \alpha_s^{(1)} = \ldots = \alpha_s^{(J)} = \alpha$.

Theorem 5 (Highest System Time in an Initially Empty Tandem Queue) *In an initially empty network of J multi-server queues in series satisfying Assumptions 1(a) and 3(b), with $\alpha_a = \alpha_s^{(1)} = \ldots = \alpha_s^{(J)} = \alpha$, $\mu_1 = \ldots = \mu_J$, $\rho < 1$, and $\Gamma = m^{1/\alpha}\Gamma_a + \Gamma_m > 0$, where*

$$\Gamma_m = \left(\sum_{j=1}^{J} (\Gamma_m^{(j)+})^{\alpha/(\alpha-1)} \right)^{(\alpha-1)/\alpha}, \quad (51)$$

the worst-case system time of the n^{th} job with $\nu = \lfloor (n-1)/m \rfloor$ is given by

$$\widehat{S}_n \le \begin{cases} \Gamma\cdot\nu^{1/\alpha} - \dfrac{m(1-\rho)}{\lambda}\nu + \left(\dfrac{J}{\mu} + \displaystyle\sum_{i=1}^{J}\Gamma_m^{(i)+}\right), & \text{if } \nu \le \left[\dfrac{\lambda\Gamma}{\alpha m(1-\rho)}\right]^{\alpha/(\alpha-1)} \\ \dfrac{(\alpha-1)}{\alpha^{\alpha/(\alpha-1)}}\cdot\dfrac{\lambda^{1/(\alpha-1)}\cdot\Gamma^{\alpha/(\alpha-1)}}{[m(1-\rho)]^{1/(\alpha-1)}} + \left(\dfrac{J}{\mu} + \displaystyle\sum_{i=1}^{J}\Gamma_m^{(i)+}\right), & \text{otherwise.} \end{cases} \quad (52)$$

The case where $\Gamma = m^{1/\alpha}\Gamma_a + \Gamma_m \le 0$ arises when $\Gamma_a < 0$, since $\Gamma_m > 0$ as defined in Eq. (51). This scenario is characterized by long inter-arrival times yielding zero waiting times. The worst case system time therefore reduces to

$$\widehat{S}_n = \sum_{j=1}^{J} \widehat{X}_n^{(j)} \le \frac{J}{\mu} + \sum_{j=1}^{J} \Gamma_m^{(j)+}.$$

Note that this scenario becomes less likely with an increased number of queues in series.

Initially Nonempty Queues in Tandem

We next analyze the case where $n_0 > 0$ and let $\phi = \lfloor (n_0 - 1)/m \rfloor$. The first m jobs in the queue are routed immediately to the servers of the first queue without any delays. We are interested in the behavior for $n_0 > m$. Since $T_i = 0$ for all $i = 1, \ldots, n_0$, we can rewrite Eq. (49) as

(a) for $n \leq n_0$:

$$\widehat{S}_n = \max_{0 \leq k_1 \leq \ldots \leq k_J \leq \nu \leq \phi} \left(\max_{\mathcal{U}_1^m} \sum_{i=k_1}^{k_2} X_{r(i)}^{(1)} + \ldots + \max_{\mathcal{U}_J^m} \sum_{i=k_J}^{n} X_{r(i)}^{(J)} \right) \tag{53}$$

(b) for $n > n_0$:

$$\widehat{S}_n = \max \left\{ \begin{array}{l} \displaystyle\max_{\substack{0 \leq k_1 \leq \ldots \leq k_J \leq \nu \\ k_1 \leq \phi}} \left(\max_{\mathcal{U}_1^m} \sum_{i=k_1}^{k_2} X_{r(i)}^{(1)} + \ldots + \max_{\mathcal{U}_J^m} \sum_{i=k_J}^{n} X_{r(i)}^{(J)} \right) - \min_{\mathcal{U}^a} \sum_{i=n_0+1}^{n} T_i, \\ \displaystyle\max_{\phi < k_1 \leq \ldots \leq k_J \leq \nu} \left(\max_{\mathcal{U}_1^m} \sum_{i=k_1}^{k_2} X_{r(i)}^{(1)} + \ldots + \max_{\mathcal{U}_J^m} \sum_{i=k_J}^{n} X_{r(i)}^{(J)} - \min_{\mathcal{U}^a} \sum_{i=r(k_1)+1}^{n} T_i \right) \end{array} \right\}. \tag{54}$$

By Assumption 1, the worst case system time involves solving J-dimensional nonlinear optimization problems. Theorem 6 provides a closed form bound on the worst case system time in an initially nonempty network of J queues in tandem, with $\alpha_a = \alpha_s^{(1)} = \ldots = \alpha_s^{(J)} = \alpha$ and $\mu_1 = \ldots = \mu_J$.

Theorem 6 (Highest System Time in an Initially Nonempty Tandem Queue) *In an initially nonempty network of J multi-server queues in series satisfying Assumptions 1(a) and 3(b), with $n_0 > m$, $\mu_1 = \ldots = \mu_J$, $\alpha_a = \alpha_s^{(1)} = \ldots = \alpha_s^{(J)} = \alpha$, $\rho < 1$, and $\Gamma = m^{1/\alpha}\Gamma_a + \Gamma_m > 0$, where Γ_m is defined in Eq. (51), the worst-case system time for $n > n_0$ is given by*

$$\widehat{S}_n \leq \max \left\{ \begin{array}{l} \displaystyle\frac{\nu + J}{\mu} + \sum_{j=1}^{J} \Gamma_m^{(j)+} + \Gamma_m \cdot \nu^{1/\alpha} - \frac{n - n_0}{\lambda} + \gamma_a \left(n - n_0\right)^{1/\alpha}, \\ \begin{cases} \displaystyle\Gamma \left(\nu - \phi\right)^{1/\alpha} - \frac{m(1-\rho)}{\lambda}\left(\nu - \phi\right) + \left(\frac{J}{\mu} + \sum_{i=1}^{J} \Gamma_m^{(i)+}\right), \; \textit{if} \;\; (\nu - \phi) < \left[\frac{\lambda \Gamma / m}{\alpha(1-\rho)}\right]^{\alpha/(\alpha-1)}, \\ \displaystyle\frac{(\alpha-1)}{\alpha^{\alpha/(\alpha-1)}} \frac{\lambda^{1/(\alpha-1)} \cdot \Gamma^{\alpha/(\alpha-1)}}{\left[m(1-\rho)\right]^{1/(\alpha-1)}} + \left(\frac{J}{\mu} + \sum_{i=1}^{J} \Gamma_m^{(i)+}\right), \quad \textit{otherwise.} \end{cases} \end{array} \right\} \tag{55}$$

Note that, for the case where $\Gamma = m^{1/\alpha}\Gamma_a + \Gamma_m \leq 0$, the worst case system time is given by

$$\widehat{S}_n \leq \max \left\{ \frac{\nu + J}{\mu} + \Gamma_m \cdot \nu^{1/\alpha} + \sum_{j=1}^{J} \Gamma_m^{(j)+} - \frac{n - n_0}{\lambda} + \gamma_a \left(n - n_0\right)^{1/\alpha}, \; \frac{J}{\mu} + \sum_{j=1}^{J} \Gamma_m^{(j)+} \right\}.$$

In this case, the n^{th} job experiences a waiting time only due to the buildup effect left by the initial jobs. For big enough n, this effect becomes negligible and the system time eventually becomes equal to the sum of the service times.

For ease of notation, we express the worst case system time in Eq. (55) as

$$\max \left\{ \widehat{S}_n^b \left(\gamma_a, \Gamma_m\right), \; \widehat{S}_n^t \left(\Gamma\right) \cdot \mathbb{1}_n^t \left(\Gamma\right) + \widehat{S}^s \left(\Gamma\right) \cdot \mathbb{1}_n^s \left(\Gamma\right) \right\}, \tag{56}$$

where $\widehat{S}_n^b$, $\widehat{S}_n^t$, and $\widehat{S}^s$ denote the quantities associated with the system time effected by the initial buffer n_0, the transient state and the steady state, respectively, and the indicator functions $\mathbb{1}_n^t$ and $\mathbb{1}_n^s$ reflect the condition for the system to be in the transient state and the steady state, respectively. For $\alpha_a = \alpha_s = \alpha$, the indicator functions are such that

$$\begin{cases} \mathbb{1}_n^t \left(\Gamma\right) = 1, & \text{if } \; \Gamma > \dfrac{\alpha m (1-\rho)}{\lambda} \cdot \Big[\lfloor n/m \rfloor - \lfloor n_0/m \rfloor \Big]^{(\alpha-1)/\alpha}, \\ \mathbb{1}_n^s \left(\Gamma\right) = 1, & \text{otherwise.} \end{cases}$$

4.2 Average Case Behavior

To analyze the average behavior of a multi-server queue, we treat the variability parameters as random variables and compute the expected value of the worst case system time

$$\widetilde{S}_n = \mathbb{E}\left[\widehat{S}_n\right].$$

Similarly to the case of a single-server queue with light-tailed primitives, we propose to approximate the density of the variability parameters by invoking the limit laws of probability and leveraging the characterization of the effective variability in Eq. (14) to fit the analysis for tandem queueing networks with possibly heavy-tailed arrivals and services.

Choice of Variability Distributions

For a network of J queues in series, we express the parameters

$$\Gamma_a = \theta_a \gamma_a \ , \ \Gamma_s^{(j)} = \theta_s \gamma_s^{(j)} \ \text{ and } \Gamma_m^{(j)} = \theta_s \gamma_m^{(j)} = \theta_s \frac{\gamma_s^{(j)}}{m^{(\alpha-1)/\alpha}},$$

where γ_a and $\gamma_s^{(j)}$ follow limiting distributions as defined in the case of a single queue, for $j = 1, \dots, J$. More specifically, $\gamma_a \sim \mathcal{N}(0, \sigma_a)$ and $\gamma_s^{(j)} \sim \mathcal{N}\left(0, \sigma_s^{(j)}\right)$ for light-tailed primitives, $\gamma_a \sim S_\alpha(-1, C_\alpha, 0)$ and $\gamma_s^{(j)} \sim S(1, C_\alpha, 0)$ for heavy-tailed primitives. Note that the effective parameter Γ_m is a function of $\Gamma_m^{(j)}$s, for $j = 1, \dots, J$. Specifically, by Eq. (51),

$$\Gamma_m = \left(\sum_{j=1}^{J} (\Gamma_m^{(j)+})^{\alpha/(\alpha-1)} \right)^{(\alpha-1)/\alpha} = \theta_s \cdot \frac{\gamma_s^{+}}{m^{(\alpha-1)/\alpha}}, \quad \text{where} \quad \gamma_s^{+} = \left(\sum_{j=1}^{J} (\gamma_s^{(j)+})^{\alpha/(\alpha-1)} \right)^{(\alpha-1)/\alpha}. \tag{57}$$

We next propose to approximate the distribution of γ_s^{+} by fitting a generalized extreme value distribution to the sampled distribution with a shape parameter ψ_s, scale parameter ξ_s and a location parameter ϕ_s. This approximation is motivated by observing that γ_s^{+} is the $\frac{\alpha}{\alpha-1}$th norm of the vector of random variables $\{\gamma_s^{(j)+}\}_{j=1}^{J}$ and by invoking Theorems 2.1 to 2.4 in [56]. In [56], the authors show that the norms of vectors of random variables X_i distributed according to F, are also approximately distributed according to F; for distributions such as Normal, Weibull, Frechet etc. This step, although an approximation, allows us to reduce the computational effort to obtain $\widetilde{S}_n$ from solving a $(J+1)$-dimensional integral with respect to γ_a and $\gamma_s^{(j)}$ to a double integral with respect to γ_a and γ_s^{+}.

Table 5 summarizes the parameters defining the generalized extreme value distribution for light-tailed service times with $\sigma_s^{(1)} = \ldots = \sigma_s^{(J)} = 1$ and heavy-tailed queues for $J = 10, 25$ and 50. Figure 8 shows that this fit provides a good approximation of the sampled distribution for $J = 25$.

Table 5 Generalized extreme value distributions for γ_s^{+} for light ($\sigma_s = 1$) and heavy-tailed services.

	10 Queues			25 Queues			50 Queues		
Parameters	$\alpha = 2$	$\alpha = 1.6$	$\alpha = 1.7$	$\alpha = 2$	$\alpha = 1.6$	$\alpha = 1.7$	$\alpha = 2$	$\alpha = 1.6$	$\alpha = 1.7$
ψ_s	-0.20	0.32	0.42	-0.21	0.36	0.44	-0.22	0.42	0.50
ξ_s	0.76	1.70	1.95	0.77	2.34	2.94	0.78	3.10	4.10
ϕ_s	1.78	2.36	2.37	3.13	4.63	4.92	4.65	7.89	7.89

We next inform the choice of the scaling parameters (θ_a, θ_s) via known conclusions on the behavior of the system time in tandem queueing networks.

(a) Light Tails: We select the value of the scaling parameter θ so that the average worst case steady-state system time matches the steady-state bound obtained in Eq. (45). We ensure that

$$\frac{\lambda}{4(1-\rho)} \cdot \mathbb{E}\left[\left(\gamma^{+}\right)^2\right] = \frac{\lambda}{2(1-\rho)} \cdot \sum_{j=1}^{J} \left[\sigma_a^2 + \left(\sigma_s^{(j)}\right)^2 / m^2\right], \tag{58}$$

where $\gamma = \theta_a \gamma_a + \theta_s \gamma_s^{+}/m$ and γ_s^{+} is defined in Eq. (57). We approximate the expected value

$$\mathbb{E}\left[\left(\gamma^{+}\right)^2\right] \approx \mathbb{P}(\gamma \geq 0) \cdot \left(\theta_a^2 \sigma_a^2 + \theta_s^2 \sum_{j=1}^{J} \left(\sigma_s^{(j)}\right)^2 / m^2 \right).$$

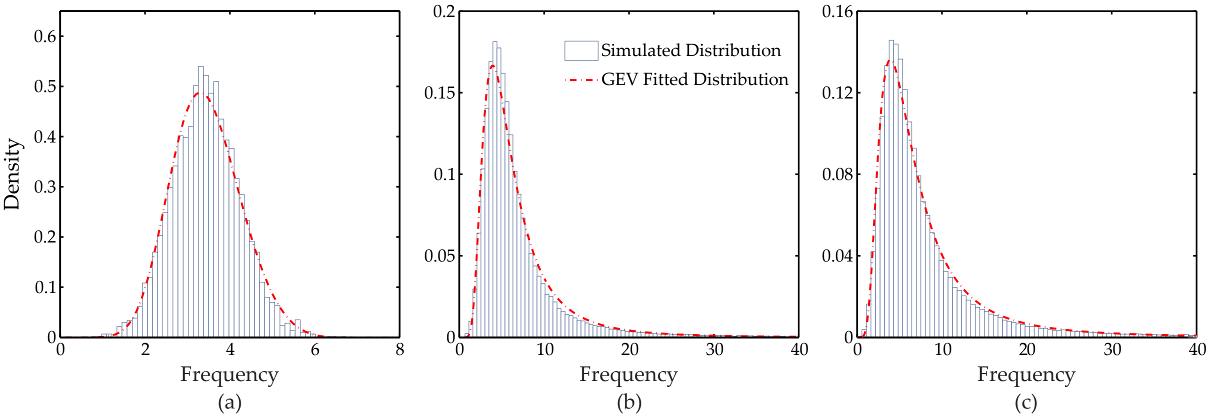


Fig. 8 Sampled distribution and fitted generalized extreme value distribution for the effective service parameter γ_s^+ for the case of $J = 25$ queues in series with (a) $\alpha = 2$, (b) $\alpha = 1.7$, and (c) $\alpha = 1.6$.

By rearranging the terms in Eq. (58), we obtain

$$\theta_a \approx \left(\frac{2J}{\mathbb{P}(\gamma \geq 0)}\right)^{1/2} \quad \text{and} \quad \theta_m \approx \left(\frac{2}{\mathbb{P}(\gamma \geq 0)}\right)^{1/2}, \tag{59}$$

where the probability $\mathbb{P}(\gamma \geq 0) = \mathbb{P}\left(J^{1/2} \cdot \gamma_a + \gamma_s^+/m \geq 0\right)$ can be efficiently computed numerically.

(b) Heavy Tails: The steady state in heavy-tailed queues does not exist. Instead, we propose to extend the formula in Eq. (59). For $\alpha_a = \alpha_s = \alpha$, we select the scaling parameter as

$$\theta_a \approx \left(\frac{\alpha J}{\mathbb{P}(\gamma \geq 0)}\right)^{(\alpha-1)/\alpha} \quad \text{and} \quad \theta_s \approx \left(\frac{\alpha}{\mathbb{P}(\gamma \geq 0)}\right)^{(\alpha-1)/\alpha}, \tag{60}$$

where the probability $\mathbb{P}(\gamma \geq 0) = \mathbb{P}\left(J^{(\alpha-1)/\alpha} \cdot \gamma_a + \gamma_s^+/m \geq 0\right)$ can be efficiently computed numerically given the distributions of γ_a and γ_s^+.

4.3 Computational Results

We investigate the performance of our approach relative to simulation and examine the effect of the system's parameters on its accuracy. We run simulations for tandem queueing networks with $N = 20,000$ job arrivals and compute the expected system time for each job using 20,000 simulation replications. We pre-specify the arrival rate at the queue to be $\lambda = 0.1$ for all simulation instances, while varying the traffic intensity, the variances associated with the inter-arrival and service processes, the number of servers in the queue, and the number of initial jobs. To compare the simulated values $\overline{S}_n$ with our approximation $\widetilde{S}_n$, we report the average percent error

$$\text{Average Percent Error} = \frac{1}{\widetilde{N}} \cdot \sum_{n=1}^{\widetilde{N}} \left| \frac{\overline{S}_n - \widetilde{S}_n}{\overline{S}_n} \right| \times 100\%,$$

where $\widetilde{N} = \min(N, \widetilde{n}_r)$ and $\widetilde{n}_r$ denotes the number of jobs the queue observes until our approximation reaches steady state, i.e., $\widetilde{n}_r = \min\left(n : \widetilde{S}_n = \widetilde{S}_\infty\right)$. We next present our results for tandem networks with (a) light-tails ($\alpha_a = \alpha_s = 2$), and (b) symmetric heavy tails ($\alpha_a = \alpha_s = \alpha$).

Light Tails: Table 6 reports the average percent error between simulation and our approximation for tandem queues with normally distributed inter-arrival and service times. Our approach generally yields percent errors within 10% relative to simulation. Figure 9(a)-(d) compare our approximation (dotted line) with simulation (solid line) for tandem networks of queues with normally distributed primitives. Note that, for $n_0 > 0$, the system exhibits slower recovery from the initial perturbation than for a single queue.

Heavy Tails: Table 7 reports the average percent error between simulation and our approximation for tandem queues with Pareto distributed inter-arrival and service times. Our approach generally yields percent errors within 10% relative to simulation, with occasional outliers. Figure 9(e)-(f) compare our approximation (dotted line) with simulation (solid line) for tandem networks of queues with Pareto distributed primitives. Note that, since the effective variability parameter Γ is heavy-tailed distributed, $\mathbb{E}\left[(\Gamma^+)^{\alpha/(\alpha-1)}\right]$ is infinite for $\alpha < 2$, suggesting that heavy-tailed tandem queueing systems never reach steady state.

Table 6 Errors for multi-server tandem queues with normally distributed primitives.

		10 Queues*			25 Queues†		50 Queues‡	
	ρ	$n_0 = 0$	$n_0 = 20$	$n_0 = 50$	$n_0 = 0$	$n_0 = 50$	$n_0 = 0$	$n_0 = 100$
$\sigma_a = \sigma_s = 2.5$	0.90	4.44	2.85	5.61	0.76	1.61	0.85	2.39
	0.92	4.85	2.82	5.58	0.81	1.96	0.82	2.41
	0.94	4.67	3.07	5.77	1.05	2.02	0.81	2.33
	0.96	5.04	3.42	4.59	1.41	3.20	0.77	2.26
$\sigma_a = \sigma_s = 4.0$	0.90	1.23	2.38	7.65	1.74	2.64	1.77	2.62
	0.92	2.02	1.65	5.91	2.28	3.14	1.73	2.32
	0.94	2.95	2.86	3.93	2.45	4.37	1.80	2.23
	0.96	3.12	3.81	3.07	2.46	4.74	4.39	5.74

$^\star m = 1$ for 10 tandem queues, $^\dagger m = 10$ for 25 tandem queues, $^\ddagger m = 20$ for 50 tandem queues.

Note: Simulating the expected overall system time of the n^{th} job in a tandem queue requires simulating each queue in the system for all n jobs, yielding run-times which highly depend on the number of queues J in the system. Our approach, on the other hand, involves (a) running a simulation to fit a generalized extreme value distribution to γ_s^+ as defined in Eq. (57) for a given α, and (b) computing double integrals with respect to γ_a and γ_s^+. Both steps can be computed efficiently for both single and multi-server tandem queues irrespective of the magnitude of J, with similar run-times to those observed for a single queue.

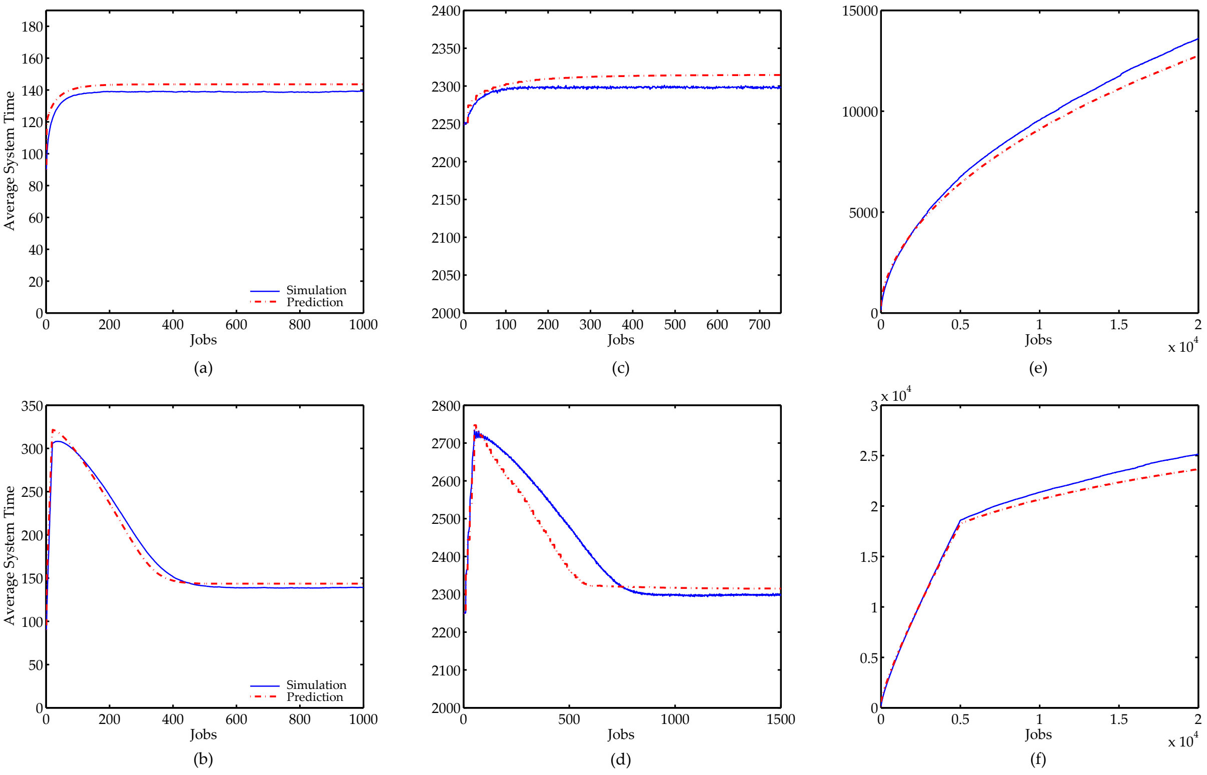


Fig. 9 Simulated (solid line) versus predicted values (dotted line). Panels (a)-(d) correspond to normally distributed queues in series with $\sigma_a = 2.5$ and $\rho = 0.90$ with $J = 10$, $m = 1$, and $n_0 = 0, 20$ (panels (a) and (b), respectively) and $J = 25$, $m = 10$, and $n_0 = 0, 50$ (panels (c) and (d), respectively). Panels (e) and (f) correspond to a tandem network with $J = 50$ single-server queues with Pareto distributed primitives ($\alpha_a = \alpha_s = 1.7$), $\rho = 0.90$, and $n_0 = 0$ and $n_0 = 5000$, respectively.

5 Extensions to Feed-forward Networks

In this section, we extend our approach to analyze open feed-forward queueing networks with no feedback. In feed-forward queueing networks, a job can visit a queue at most once before exiting the network. We consider a feed-forward network with a set of queueing nodes $\mathcal{J}$ with

(a) external arrival processes with parameters (λ_j, α_a) that arrive at queue $j \in \mathcal{J}$,

Table 7 Errors for single-server tandem queues with Pareto distributed primitives.

		10 Queues		25 Queues		50 Queues	
	ρ	$n_0 = 0$	$n_0 = 2000$	$n_0 = 0$	$n_0 = 3500$	$n_0 = 0$	$n_0 = 5000$
$\alpha_a = 1.6$	0.90	9.80	5.11	2.89	2.31	4.88	4.77
	0.92	4.30	3.52	7.88	1.82	3.13	1.81
	0.94	2.40	2.10	7.94	2.95	16.6	7.84
	0.96	2.82	2.54	14.7	5.22	16.5	6.71
$\alpha_a = \alpha_s = 1.7$	0.90	24.3	7.79	5.61	2.17	5.31	3.93
	0.92	15.8	6.69	2.85	1.04	10.0	2.82
	0.94	11.6	4.72	3.45	2.77	12.6	5.91
	0.96	6.34	3.92	5.67	3.55	11.6	5.92

(b) service processes with parameters $(\mu_j, \alpha_s^{(j)})$ with the number of servers m_j at queue $j \in \mathcal{J}$,

(c) a routing matrix $\mathbf{F} = [f_{ij}]$, $i, j \in \mathcal{J}$, where f_{ij} denotes the fraction of the jobs passing through queue i which are routed to queue j. The fraction of jobs leaving queue i is $1 - \sum_j f_{ij}$.

We study the expected overall system time of the n^{th} job passing through the network. Let $\mathcal{P}$ be the set of all possible paths that job n may take and f_P denote the probability that a job n takes a particular path $P \in \mathcal{P}$. The expected overall system time can the be expressed as

$$\overline{S}_n = \sum_{P \in \mathcal{P}} f_P \cdot \mathbb{E}\left[S_n^P\right] = \sum_{P \in \mathcal{P}} f_P \cdot \overline{S}_n^P,$$

where S_n^P denote the system time of the n^{th} job when traversing the network through path P. Since it is challenging to analyze the expected system time using traditional probabilistic approaches, we propose a similar approach to the one undertaken for single and tandem queues.

To make the exposition clear, we assume that the network starts operation without any initial jobs, i.e., $n_0 = 0$ at all queues. We let $\mathcal{L}_i$ denote the set of jobs departing from queue i, and $\mathcal{E}_{ij}$ the set of jobs routed from queue i to queue j (see Figure 6 for an illustration). Under a probabilistic routing scheme, these sets are not known until after an instance of the network is realized. For the purpose of our analysis, we propose to approximate the dynamics of a probabilistic feed-forward network as follows.

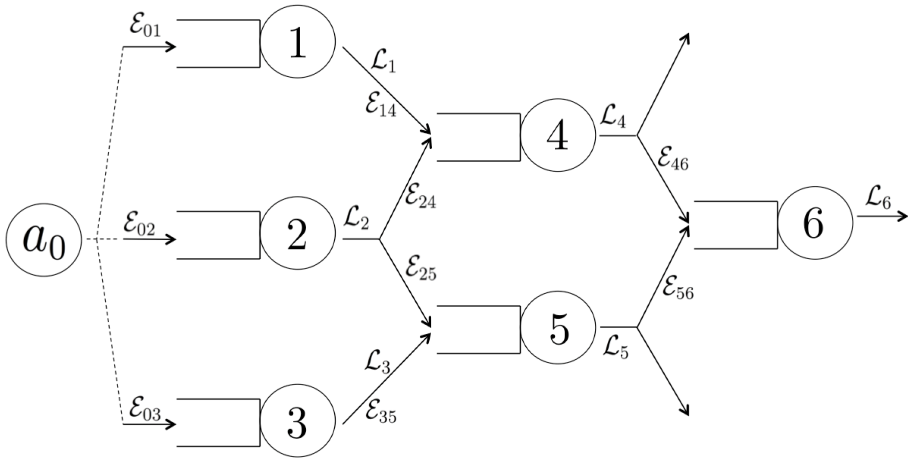


Fig. 10 Feed-forward network with deterministic routing.

(a) **Deterministic Routing:** We consider a deterministic approximation of probabilistic routing. Suppose that f_{ij} and f_{ik} denote the fraction of the jobs leaving from queue i that are routed to queues j and k, respectively, while the remaining jobs exit the system. We assume that the fractions f_{ij} and f_{ik} are rational and given by

$$f_{ij} = \frac{p_{ij}}{q_i} \quad \text{and} \quad f_{ik} = \frac{p_{ik}}{q_i},$$

where $p_{ij}, p_{ik} \geq 0$ and $q_i > 0$ are integers, with $p_{ij} + p_{ik} \leq q_i$. This assumption of rationality is not restrictive, since any irrational number can be arbitrarily closely approximated by a rational number. Under deterministic routing, the jobs are routed as follows. We divide the set of jobs $\mathcal{L}_i$ departing queue i into q_i sets of jobs

$$\mathcal{B}_t^i = \{t, t + q_i, t + 2q_i, \ldots\}, \quad \forall t = 1, \ldots, q_i,$$

and then route jobs from the jobs in sets $\mathcal{E}_{ij}$ and $\mathcal{E}_{ik}$ to queues j and k, respectively, where

$$\mathcal{E}_{ij} = \mathcal{B}_1^i \cup \ldots \cup \mathcal{B}_{p_{ij}}^i \quad \text{and} \quad \mathcal{E}_{ik} = \mathcal{B}_{p_{ij}+1}^i \cup \ldots \cup \mathcal{B}_{p_{ik}}^i .$$

Note that, with this deterministic routing scheme, for a large number of jobs, approximately a fraction f_{ij} and f_{ik} of jobs are routed to queues j and k, respectively. To illustrate, consider queue 2 in Figure 6, and suppose $\mathcal{L}_2 = \{2, 3, 5, 7, 10, 11, 14, 15\}$, $f_{24} = 1/3$ and $f_{25} = 2/3$. Then, by our routing scheme,

$$\mathcal{E}_{24} = \{2, 7, 14\} \quad \text{and} \quad \mathcal{E}_{25} = \{3, 5, 10, 11, 15\} .$$

(b) External Arrivals: We assume that the external arrivals emanate from a single node a_0. In other words, we assume jobs enter the network at node a_0 with rate $\lambda = \sum_{j \in \mathcal{J}} \lambda_j$ and tail coefficient α_a. The arrivals are then routed to the nodes $j \in \mathcal{J}$ such that

$$f_{0j} = \frac{\lambda_j}{\lambda}, \ \forall \ j \in \mathcal{J}.$$

Note: The number of jobs passing through some queue $j \in \mathcal{J}$ is a subset of all the jobs that are routed through the network. We let ϕ_j denote the fraction of jobs passing through queue j, which is computed recursively using the routing matrix F as

$$\phi_j = \sum_{i \in \mathcal{J}} \phi_i \cdot f_{ij}. \tag{61}$$

Furthermore, under steady-state, the traffic intensity observed by queue j is equal to the ratio of the arrival rate it experiences and its service rate. Given the fraction of jobs ϕ_j that pass by queue j, the traffic intensity observed is

$$\rho_j = \frac{\lambda_j}{\mu_j} = \frac{\lambda \cdot \phi_j}{\mu_j}. \tag{62}$$

We further assume that the inter arrival times $\mathbf{T}$ to node a_0 satisfy the uncertainty set $\mathcal{U}^a$ as defined in Assumption 1(a) and that the service times $\mathbf{X}^{(j)}$ at node j satisfy $\mathcal{U}_j^s$ in case of a single server ($\mathcal{U}_j^m$ in case of multiple servers) as defined in Assumption 1, for all $j \in \mathcal{J}$.

Steady-State Analysis: [9] have studied this network's steady-state behavior using the robust framework. In particular, [9] show that the inter-departure times belong to the inter-arrival uncertainty set $\mathcal{U}^a$. This characterization is akin to Burke's theorem and is particularly tight under steady-state conditions. This allows [9] to study the phenomena of merging and splitting with a queueing network. Specifically, the effective inter-arrival times $\mathbf{T}^{(j)}$ to some queue j satisfy the uncertainty set

$$\mathcal{U}_j^a = \left\{ \left(T_1^{(j)}, \ldots, T_n^{(j)}\right) \middle| \sum_{i=k+1}^{n} T_i^{(j)} - \frac{n-k}{\lambda_j} \geq -\Gamma_a^{(j)} (n-k)^{1/\alpha_a}, \quad \forall\ 0 \leq k \leq n \right\},$$

where $\lambda_j = \lambda \cdot \phi_j$ and $\Gamma_a^{(j)} = \Gamma_a / \phi_j^{1/\alpha_a}$, for all $j \in \mathcal{J}$. By this network decomposition, the worst case steady-state system time of a job passing by queue j can be expressed as

$$\widehat{S}_\infty^{(j)} = \frac{(\alpha-1)}{\alpha^{\alpha/(\alpha-1)}} \frac{\lambda_j^{1/(\alpha-1)} \cdot \left(\Gamma^{(j)+}\right)^{\alpha/(\alpha-1)}}{(1-\rho_j)^{1/(\alpha-1)}} + \left(\frac{1}{\mu_j} + \Gamma_s^{(j)+}\right), \tag{63}$$

where $\alpha_a = \alpha_s^{(j)} = \alpha$ and $\Gamma^{(j)} = \Gamma_a / \phi_j^{1/\alpha} + \Gamma_m^{(j)}$, for all $j \in \mathcal{J}$. For light-tailed queues, obtaining $\widetilde{S}_\infty^{(j)}$ as in Section 3.2., we approximate the overall expected steady-state system time value by

$$\begin{aligned} \overline{S}_\infty \approx \widetilde{S}_\infty &= \sum_{P \in \mathcal{P}} f_P \sum_{j \in P} \widetilde{S}_\infty^{(j)} \\ &= \sum_{P \in \mathcal{P}} f_P \sum_{j \in P} \left[\frac{\lambda \phi_j}{2(1-\rho_j)} \mathbb{E}\left[\sigma_a^2/\phi_j + \left(\sigma_s^{(j)}\right)^2 /m^2 \right] + \frac{1}{\mu_j} + \mathbb{E}\left[\Gamma_m^{(j)+}\right] \right]. \end{aligned} \tag{64}$$

Transient Analysis: While the characterization of the inter-departure times in [9] holds for transient regimes, it however provides loose bounds. Obtaining an exact transient characterization of the inter-departure process is challenging. Instead of decomposing the network, we propose to obtain a recursive formula that defines the dynamics in a feed-forward network similarly to the one obtained for tandem queues in Eq. (46). To make the exposition clear, we consider the case of a feed-forward network with single-server queues. To illustrate how we derive a characterization of the system time for the n^{th} job in a feed-forward network with deterministic routing, we consider the network instance depicted in Figure 6.

Suppose that job n exits the system at node 6 after passing through queue 1 and queue 4, i.e., $n \in \mathcal{E}_{46}$ and $n \in \mathcal{E}_{14}$. The overall system time of the n^{th} job is given by

$$S_n = S_n^{(1)} + S_n^{(4)} + S_n^{(6)}.$$

The system time of the n^{th} job at queue 6 is given by

$$S_n^{(6)} = \max_{1 \le k_6 \le n} \left(\sum_{\substack{i=k_6 \\ i \in \mathcal{L}_6}}^{n} X_i^{(6)} - \sum_{\substack{i=k_6+1 \\ i \in \mathcal{L}_6}}^{n} T_i^{(6)} \right),$$

where $\mathbf{T}^{(6)}$ denotes the inter arrival times of jobs entering queue 6. Job k_6 could have either come from queue 4, i.e., $k_6 \in \mathcal{E}_{46}$, or from queue 5, i.e., $k_6 \in \mathcal{E}_{56}$.

(a) If $k_6 \in \mathcal{E}_{46}$, and given that $n \in \mathcal{E}_{46}$, the time between the arrivals of jobs k_6 and n to queue 6 is the same as the time between the departures of jobs k_6 and n from queue 4, i.e.,

$$\sum_{\substack{i=k_6+1 \\ i \in \mathcal{L}_6}}^{n} T_i^{(6)} = \sum_{\substack{i=k_6+1 \\ i \in \mathcal{L}_4}}^{n} D_i^{(4)} = \sum_{\substack{i=k_6+1 \\ i \in \mathcal{L}_4}}^{n} T_i^{(4)} + S_n^{(4)} - S_{k_6}^{(4)},$$

where $\mathbf{D}^{(4)}$ denotes the inter departure times from queue 4. Similarly to a tandem queue, the system time spent by the n^{th} job at queues 4 and 6 is given by

$$S_n^{(4)} + S_n^{(6)} = \max_{1 \le k_4 \le k_6 \le n} \left(\sum_{\substack{i=k_4 \\ i \in \mathcal{L}_4}}^{k_6} X_i^{(4)} + \sum_{\substack{i=k_6 \\ i \in \mathcal{L}_6}}^{n} X_i^{(6)} - \sum_{\substack{i=k_4+1 \\ i \in \mathcal{L}_4}}^{n} T_i^{(4)} \right).$$

(1) If $k_4 \in \mathcal{E}_{14}$, and since $n \in \mathcal{E}_{14}$, the overall system time is given by

$$S_n = \max_{1 \le k_1 \le k_4 \le k_6 \le n} \left(\sum_{\substack{i=k_1 \\ i \in \mathcal{L}_1}}^{k_4} X_i^{(1)} + \sum_{\substack{i=k_4 \\ i \in \mathcal{L}_4}}^{k_6} X_i^{(4)} + \sum_{\substack{i=k_6 \\ i \in \mathcal{L}_6}}^{n} X_i^{(6)} - \sum_{\substack{i=k_1+1 \\ i \in \mathcal{L}_1}}^{n} T_i^{(1)} \right).$$

(2) If $k_4 \in \mathcal{E}_{24}$, then the time between the arrivals of jobs k_4 and n to queue 4 is equal to the time between the departures of jobs k_4 and n from queues 2 and 1, respectively, i.e.,

$$\sum_{\substack{i=k_4+1 \\ i \in \mathcal{L}_4}}^{n} T_i^{(4)} = \sum_{\substack{i=1 \\ i \in \mathcal{L}_1}}^{n} D_i^{(1)} - \sum_{\substack{i=1 \\ i \in \mathcal{L}_2}}^{k_4} D_i^{(2)} = \left(\sum_{\substack{i=1 \\ i \in \mathcal{L}_1}}^{n} T_i^{,(1)} + S_n^{(1)} \right) - \left(\sum_{\substack{i=1 \\ in \mathcal{L}_2}}^{k_4} T_i^{,(2)} + S_{k_4}^{(2)} \right).$$

Under this scenario, the overall system time of the n^{th} job becomes

$$S_n = \max_{1 \le k_2 \le k_4 \le k_6 \le n} \left(\sum_{\substack{i=k_2 \\ i \in \mathcal{L}_2}}^{k_4} X_i^{(2)} + \sum_{\substack{i=k_4 \\ i \in \mathcal{L}_4}}^{k_6} X_i^{(4)} + \sum_{\substack{i=k_6 \\ i \in \mathcal{L}_6}}^{n} X_i^{(6)} - \sum_{\substack{i=1 \\ i \in \mathcal{L}_1}}^{n} T_i^{(1)} + \sum_{\substack{i=1 \\ i \in \mathcal{L}_2}}^{k_2} T_i^{(2)} \right).$$

(b) If $k_6 \in \mathcal{E}_{56}$, and by similar arguments to those presented in part (a),

(1) If $k_5 \in \mathcal{E}_{25}$, then $S_n = \max_{1 \le k_2 \le k_5 \le k_6 \le n} \left(\sum_{\substack{i=k_2 \\ i \in \mathcal{L}_2}}^{k_5} X_i^{(2)} + \sum_{\substack{i=k_5 \\ i \in \mathcal{L}_5}}^{k_6} X_i^{(5)} + \sum_{\substack{i=k_6 \\ i \in \mathcal{L}_6}}^{n} X_i^{(6)} - \sum_{\substack{i=1 \\ i \in \mathcal{L}_1}}^{n} T_i^{(1)} + \sum_{\substack{i=1 \\ i \in \mathcal{L}_2}}^{k_2} T_i^{(2)} \right),$

(2) If $k_5 \in \mathcal{E}_{35}$, then $S_n = \max_{1 \le k_3 \le k_5 \le k_6 \le n} \left(\sum_{\substack{i=k_3 \\ i \in \mathcal{L}_3}}^{k_5} X_i^{(3)} + \sum_{\substack{i=k_5 \\ i \in \mathcal{L}_5}}^{k_6} X_i^{(5)} + \sum_{\substack{i=k_6 \\ i \in \mathcal{L}_6}}^{n} X_i^{(6)} - \sum_{\substack{i=1 \\ i \in \mathcal{L}_1}}^{n} T_i^{(1)} + \sum_{\substack{i=1 \\ i \in \mathcal{L}_3}}^{k_3} T_i^{(3)} \right).$

Note that the arrival times of jobs to queues 1, 2 and 3 is equal to the time of arrival at node a_0, since there is no service delay at node a_0, which yields

$$\sum_{\substack{i=1\\ i\in\mathcal{L}_\ell}}^{k_\ell} T_i^{(\ell)} = \sum_{i=1}^{k_\ell} T_i, \text{ for all jobs } k_\ell \text{ arriving at queue } \ell = 1,2,3.$$

Consequently, for job $n \in \mathcal{L}_6$ leaving the system at queue 6, combining parts (a) and (b) gives us the following characterization of the overall system time

$$S_n\left(\mathcal{P}_6\right) = \max_{P\in\mathcal{P}_6}\left\{\max_{\substack{1\leq k_{a_1}\leq\ldots\leq k_6\leq n\\ k_{a_{j+1}}\in\mathcal{E}_{a_j a_{j+1}}}}\left(\sum_{\substack{i=k_{a_1}\\ i\in\mathcal{L}_{a_1}}}^{k_{a_2}} X_i^{(a_1)} + \ldots + \sum_{\substack{i=k_6\\ i\in\mathcal{L}_6}}^{n} X_i^{(6)} - \sum_{i=k_{a_1}}^{n} T_i\right)\right\}, \tag{65}$$

where $\mathcal{P}_6 = \{(1,4,6), (2,4,6), (2,5,6), (3,5,6)\}$ is the set of all the paths $P = (a_0, a_1, a_2, \ldots, \ell)$ that leave the network at queue 6. Proposition 3 presents the characterization of the overall system time of the n^{th} job in a generalized feed-forward network with deterministic routing.

Proposition 3 (System Time in Feed-Forward Networks with Deterministic Routing) *In a feed-forward network composed of single-server queues with service times $\mathbf{X}^{(j)}$, $j \in \mathcal{J}$ and external inter-arrivals $\mathbf{T}$, the overall system time of the n^{th} job exiting at node ℓ is given by*

$$S_n\left(\mathcal{P}_\ell\right) = \max_{P\,\in\,\mathcal{P}_\ell}\left\{\max_{\substack{1\leq k_{a_1}\leq k_{a_2}\leq\ldots\leq k_\ell\leq n\\ k_{a_{j+1}}\in\mathcal{E}_{a_j a_{j+1}}}}\left(\sum_{\substack{i=k_{a_1}\\ i\in\mathcal{L}_{a_1}}}^{k_{a_2}} X_i^{(a_1)} + \sum_{\substack{i=k_{a_2}\\ i\in\mathcal{L}_{a_2}}}^{k_{a_3}} X_i^{(a_2)} + \ldots + \sum_{\substack{i=k_\ell\\ i\in\mathcal{L}_\ell}}^{n} X_i^{(\ell)} - \sum_{i=k_{a_1}+1}^{n} T_i\right)\right\}, \tag{66}$$

where $\mathcal{P}_\ell$ denotes the set of all paths $P = (a_0, a_1, a_2, \ldots, \ell)$ that leave the network at node ℓ.

A detailed proof of Proposition 3 is provided in Appendix 2. Similarly to the analysis of a single and tandem queue, we propose an analysis of the worst case overall system time in a feed-forward network. We then leverage the analytic expressions of the worst case system time to understand the behavior of feed-forward networks with deterministic routing.

5.1 Worst Case Behavior

To analyze the worst case behavior of the system time in the feed-forward network, we apply the bounds on the inter-arrival and service times presented in Assumptions 1(a) and 3(a) and obtain

$$\widehat{S}_n\left(\mathcal{P}_\ell\right) \;=\; \max_{P\,\in\,\mathcal{P}_\ell}\left\{\max_{\substack{1\leq k_{a_1}\leq\ldots\leq k_\ell\leq n\\ k_{a_{j+1}}\in\mathcal{E}_{a_j a_{j+1}}\subseteq\mathcal{L}_{a_{j+1}}}}\left(\max_{\mathcal{U}^s_{a_1}}\sum_{\substack{i=k_{a_1}\\ i\in\mathcal{L}_{a_1}}}^{k_{a_2}} X_i^{(a_1)} + \ldots + \max_{\mathcal{U}^s_\ell}\sum_{\substack{i=k_\ell\\ i\in\mathcal{L}_\ell}}^{n} X_i^{(\ell)} - \min_{\mathcal{U}^a}\sum_{i=k_{a_1}+1}^{n} T_i\right)\right\}, \tag{67}$$

where $\mathcal{P}_\ell$ denotes the set of all paths $P = (a_0, a_1, a_2, \ldots, \ell)$ that leave the network at node ℓ. By Assumptions 1, Eq. (67) involves solving a $|P|$-dimensional optimization problem for every path $P \in \mathcal{P}_\ell$, which can be computed efficiently. Theorem 7 provides a closed form upper bound for the worst case system time of the n^{th} job exiting the network at node ℓ in a feed-forward network with $\alpha_a = \alpha_s^{(j)} = \alpha$, for all $j \in \mathcal{J}$.

Theorem 7 (Highest System Time in a Feed-Forward Network) *In a feed-forward network composed of single-server queues satisfying Assumptions 1(a) and 3(a) with $\alpha_a = \alpha_s^{(j)} = \alpha$, for all $j \in \mathcal{J}$, the set $\mathcal{P}_\ell$ containing all paths $P = (a_0, a_1, a_2, \ldots, \ell)$ that leave from node ℓ, and*

$$\rho_P = \frac{\lambda}{\min\limits_{j\in P}\mu_j/\phi_j} \quad \text{and} \quad \varGamma_P = \varGamma_a + \left[\sum_{j\in P}\left(\varGamma_s^{(j)+}\cdot\phi_j^{1/\alpha}\right)^{\alpha/(\alpha-1)}\right]^{(\alpha-1)/\alpha} > 0, \tag{68}$$

the overall system time of the n^{th} job exiting the network at node ℓ is bounded by

$$\widehat{S}_n\left(\mathcal{P}_\ell\right) \leq \max_{P \in \mathcal{P}_\ell} \begin{cases} \Gamma_P \cdot n^{1/\alpha} - \dfrac{1-\rho_P}{\lambda} n + \displaystyle\sum_{j \in P} \left(\dfrac{1}{\mu_j} + \Gamma_s^{(j)+}\right), & \text{if } n \leq \left[\dfrac{\lambda \Gamma_P}{\alpha(1-\rho_P)}\right]^{\alpha/(\alpha-1)}, \\ \dfrac{(\alpha-1)}{\alpha^{\alpha/(\alpha-1)}} \cdot \dfrac{\lambda^{1/(\alpha-1)} \cdot \Gamma_P^{\alpha/(\alpha-1)}}{(1-\rho_p)^{1/(\alpha-1)}} + \displaystyle\sum_{j \in P} \left(\dfrac{1}{\mu_j} + \Gamma_s^{(j)+}\right), & \textit{otherwise.} \end{cases} \tag{69}$$

The bound presented in Theorem 7 is particularly tight for the special case where $\rho_j = \rho$ (i.e., $\mu_j = \lambda \cdot \phi_j / \rho$) for all $j \in \mathcal{J}$ for some value ρ. This yields $\rho_P = \rho$ for all $P \in \mathcal{P}_\ell$. For this case, a higher value of the effective parameter Γ_P results in a higher system and relaxation times, as suggested by Eq. (69). The worst case system time $\widehat{S}_n\left(\mathcal{P}_\ell\right)$ therefore corresponds to

$$\Gamma\left(\mathcal{P}_\ell\right) = \max_{P \in \mathcal{P}_\ell} \Gamma_P.$$

Theorem 8 provides the analytic expression of the worst case system time of the n^{th} job exiting the network at node ℓ in a feed-forward network with $\alpha_a = \alpha_s^{(j)} = \alpha$ and $\rho_j = \rho$ for all $j \in \mathcal{J}$.

Theorem 8 (Highest System Time in a Feed-Forward Network for Fixed Traffic Rate) *In a feed-forward network composed of single-server queues satisfying Assumptions 1(a) and 3(a) with $\alpha_a = \alpha_s^{(j)} = \alpha$, and $\rho_j = \rho$ (i.e., $\mu_j = \lambda \cdot \phi_j / \rho$) for all $j \in \mathcal{J}$, and given the set $\mathcal{P}_\ell$ containing all paths $P = (a_0, a_1, \ldots, \ell)$ that leave the network at node ℓ, and*

$$\Gamma\left(\mathcal{P}_\ell\right) = \Gamma_a + \Gamma_s\left(\mathcal{P}_\ell\right) = \Gamma_a + \max_{P \in \mathcal{P}_\ell} \left[\sum_{j \in P} \left(\Gamma_s^{(j)+} \cdot \phi_j^{1/\alpha}\right)^{\alpha/(\alpha-1)}\right]^{(\alpha-1)/\alpha} > 0, \tag{70}$$

the overall system time of the n^{th} job exiting the network at node ℓ is given by

$$\widehat{S}_n\left(\mathcal{P}_\ell\right) \leq \begin{cases} \Gamma\left(\mathcal{P}_\ell\right) \cdot n^{1/\alpha} - \dfrac{1-\rho}{\lambda} n + \displaystyle\sum_{P \in \mathcal{P}_\ell} \sum_{j \in P} \left(\dfrac{1}{\mu_j} + \Gamma_s^{(j)+}\right), & \text{if } n \leq \left[\dfrac{\lambda \Gamma\left(\mathcal{P}_\ell\right)}{\alpha(1-\rho)}\right]^{\alpha/(\alpha-1)}, \\ \dfrac{(\alpha-1)}{\alpha^{\alpha/(\alpha-1)}} \cdot \dfrac{\lambda^{1/(\alpha-1)} \cdot \Gamma\left(\mathcal{P}_\ell\right)^{\alpha/(\alpha-1)}}{(1-\rho)^{1/(\alpha-1)}} + \displaystyle\sum_{P \in \mathcal{P}_\ell} \sum_{j \in P} \left(\dfrac{1}{\mu_j} + \Gamma_s^{(j)+}\right), & \textit{otherwise.} \end{cases} \tag{71}$$

The case where $\Gamma\left(\mathcal{P}_\ell\right) \leq 0$ arises when $\Gamma_a < 0$. This scenario is characterized by long inter-arrival times yielding zero waiting times. The worst case system time therefore reduces to

$$\widehat{S}_n\left(\mathcal{P}_\ell\right) \leq \max_{P \in \mathcal{P}_\ell} \sum_{j \in P} \left(\frac{1}{\mu_j} + \Gamma_s^{(j)+}\right) \leq \sum_{P \in \mathcal{P}_\ell} \sum_{j \in P} \left(\frac{1}{\mu_j} + \Gamma_s^{(j)+}\right).$$

We next extend our averaging approach to analyze feed-forward queueing networks with $\alpha_a = \alpha_s^{(j)} = \alpha$ and $\rho_j = \rho$ (i.e., $\mu_j = \lambda \cdot \phi_j / \rho$) for all $j \in \mathcal{J}$.

5.2 Average Case Behavior

The expected system time spent by the n^{th} job in the feed-forward network can be computed as

$$\overline{S}_n = \sum_{P \in \mathcal{P}} f_P \cdot \overline{S}_n^P = \sum_{\ell \in \mathcal{J}} p_\ell \cdot \overline{S}_n\left(\mathcal{P}_\ell\right), \tag{72}$$

where $\mathcal{P}$ denotes the set of all possible paths that can be taken by jobs passing through the network, f_P denotes the probability of taking a certain path P, $\overline{S}_n^P$ denotes the expected system time of the n^{th} job that is routed through the network via path P, $\overline{S}_n\left(\mathcal{P}_\ell\right)$ denotes the expected system time of the n^{th} job that leaves from node ℓ (i.e., job n takes any path $P \in \mathcal{P}_\ell$), and p_ℓ denotes the probability of a job exiting the network at node ℓ, i.e.,

$$p_\ell = \phi_\ell \cdot \left(1 - \sum_{j \in \mathcal{J}} f_{\ell j}\right).$$

Instead of taking the expectation of the system time over the random variables $\mathbf{T}$ and $\mathbf{X}$ to obtain $\overline{S}_n(P)$, for all paths $P \in \mathcal{P}$ or $\overline{S}_n\left(\mathcal{P}_\ell\right)$, for all $\ell \in \mathcal{J}$, we propose to compute the expected value of the worst case system time with respect to the parameters $\mathit{\Gamma}_a$ and $\mathit{\Gamma}_s(\mathcal{P}_\ell)$ which we treat as random variables. Mathematically, we compute

$$\widetilde{S}_n = \sum_{\ell \in \mathcal{J}} p_\ell \cdot \widetilde{S}_n(\mathcal{P}_\ell) = \sum_{\ell \in \mathcal{J}} p_\ell \cdot \mathbb{E}\left[\widehat{S}_n(\mathcal{P}_\ell)\right].$$

Given Theorem 8, we can express $\widehat{S}_n(\mathcal{P}_\ell)$ as a function of $\mathit{\Gamma}_a$ and $\mathit{\Gamma}_s(\mathcal{P}_\ell)$ as follows

$$\widehat{S}_n \leq \begin{cases} \widehat{S}_n^{\,t}\left(\mathit{\Gamma}_a, \mathit{\Gamma}_s(\mathcal{P}_\ell)\right), & \text{if } \; n < \left[\dfrac{\lambda\left(\mathit{\Gamma}_a + \mathit{\Gamma}_s(\mathcal{P}_\ell)\right)^+}{\alpha(1-\rho)}\right]^{\alpha/(\alpha-1)}, \\ \widehat{S}^s\left(\mathit{\Gamma}_a, \mathit{\Gamma}_s(\mathcal{P}_\ell)\right), & \text{otherwise,} \end{cases} \tag{73}$$

where $\mathit{\Gamma}_s(\mathcal{P}_\ell)$ is defined in Eq. (70) in terms of $\mathit{\Gamma}_m^{(j)}$, for $j \in \mathcal{J}$, and $\widehat{S}_n^{\,t}$, and $\widehat{S}^s$ denote the quantities associated with the transient state and the steady state, respectively. We rewrite Eq. (73) as

$$\widehat{S}_n^t\left(\mathit{\Gamma}_a, \mathit{\Gamma}_s(\mathcal{P}_\ell)\right) \cdot \mathbb{1}_n^{\,t}\left(\mathit{\Gamma}_a, \mathit{\Gamma}_s(\mathcal{P}_\ell)\right) + \widehat{S}^s\left(\mathit{\Gamma}_a, \mathit{\Gamma}_s(\mathcal{P}_\ell)\right) \cdot \mathbb{1}_n^s\left(\mathit{\Gamma}_a, \mathit{\Gamma}_s(\mathcal{P}_\ell)\right),$$

where the indicator functions $\mathbb{1}_n^t$ and $\mathbb{1}_n^s$ reflect the condition for the system to be in the transient state and the steady state, respectively, with

$$\begin{cases} \mathbb{1}_n^{\,t}\left(\mathit{\Gamma}_a, \mathit{\Gamma}_s(\mathcal{P}_\ell)\right) = 1, & \text{if } \; \mathit{\Gamma}_a + \mathit{\Gamma}_s(\mathcal{P}_\ell) > \dfrac{\alpha(1-\rho)}{\lambda} \cdot n^{(\alpha-1)/\alpha}, \\ \mathbb{1}_n^s\left(\mathit{\Gamma}_a, \mathit{\Gamma}_s(\mathcal{P}_\ell)\right) = 1, & \text{otherwise.} \end{cases}$$

By positing some assumptions on the distributions of $\mathit{\Gamma}_a$ and $\mathit{\Gamma}_s(\mathcal{P}_\ell)$, we express $\widetilde{S}_n$ as

$$\widetilde{S}_n = \mathbb{E}\left[\widehat{S}_n^t\left(\mathit{\Gamma}_a, \mathit{\Gamma}_s(\mathcal{P}_\ell)\right) \cdot \mathbb{1}_n^{\,t}\left(\mathit{\Gamma}_a, \mathit{\Gamma}_s(\mathcal{P}_\ell)\right) + \widehat{S}^s\left(\mathit{\Gamma}_a, \mathit{\Gamma}_s(\mathcal{P}_\ell)\right) \cdot \mathbb{1}_n^s\left(\mathit{\Gamma}_a, \mathit{\Gamma}_s(\mathcal{P}_\ell)\right)\right],$$

which can be efficiently computed via numerical integration. We next discuss our choice of the parameter distributions.

Choice of Variability Distributions

We propose to express the parameters $\mathit{\Gamma}_a = \theta_a \gamma_a$ and $\mathit{\Gamma}_s^{(j)} = \theta_s \gamma_s^{(j)}$, where γ_a and $\gamma_s^{(j)}$ follow limiting distributions for all $j \in \mathcal{J}$. More specifically, $\gamma_a \sim \mathcal{N}(0, \sigma_a)$ and $\gamma_s^{(j)} \sim \mathcal{N}\left(0, \sigma_s^{(j)}\right)$ for light-tailed primitives, $\gamma_a \sim S_\alpha(-1, C_\alpha, 0)$ and $\gamma_s^{(j)} \sim S(1, C_\alpha, 0)$ for heavy-tailed primitives. Note that the effective service parameter $\mathit{\Gamma}_s(\mathcal{P}_\ell)$ is a function of $\mathit{\Gamma}_s^{(j)}$s, for $j \in \mathcal{J}$. Specifically, by Eq. (70),

$$\mathit{\Gamma}_s(\mathcal{P}_\ell) = \theta_s \gamma_s^+(\mathcal{P}_\ell) \quad \text{where} \quad \gamma_s^+(\mathcal{P}_\ell) = \max_{P \in \mathcal{P}_\ell} \left[\sum_{j \in P} \left(\gamma_s^{(j)+} \cdot \phi_j^{1/\alpha}\right)^{\alpha/(\alpha-1)}\right]^{(\alpha-1)/\alpha}. \tag{74}$$

Similarly to our approach for tandem queues, we propose an approximation of the distribution of $\gamma_s^{\ell+}$ by fitting generalized extreme value distribution to the sampled distribution.

For light-tailed queues, by Theorem 8, the expected value of the overall worst case steady-state system time for a feed-forward network is given by

$$\begin{aligned} \widetilde{S}_\infty = \sum_{\ell \in \mathcal{J}} p_\ell \widetilde{S}_\infty(\mathcal{P}_\ell) &= \sum_{\ell \in \mathcal{J}} p_\ell \cdot \frac{\lambda}{4(1-\rho)} \cdot \mathbb{E}\left[\left(\gamma(\mathcal{P}_\ell)^+\right)^2\right] + \sum_{\ell \in \mathcal{J}} p_\ell \sum_{P \in \mathcal{P}_\ell} \sum_{j \in P} \left(\frac{1}{\mu_j} + \mathbb{E}\left[\mathit{\Gamma}_m^{(j)+}\right]\right), \\ &= \sum_{\ell \in \mathcal{J}} p_\ell \cdot \frac{\lambda}{4(1-\rho)} \cdot \mathbb{E}\left[\left(\gamma(\mathcal{P}_\ell)^+\right)^2\right] + \sum_{P \in \mathcal{P}} f_P \sum_{j \in P} \left(\frac{1}{\mu_j} + \mathbb{E}\left[\mathit{\Gamma}_m^{(j)+}\right]\right), \end{aligned} \tag{75}$$

where $\gamma(\mathcal{P}_\ell) = \theta_a \gamma_a + \theta_s \gamma_s^+(\mathcal{P}_\ell)$ and $\gamma_s^+(\mathcal{P}_\ell)$ is defined in Eq. (74). The expected value in Eq. (75)

$$\mathbb{E}\left[\left(\gamma(\mathcal{P}_\ell)^+\right)^2\right] \approx \mathbb{P}\left(\gamma(\mathcal{P}_\ell) \geq 0\right) \cdot \mathbb{E}\left[\gamma(\mathcal{P}_\ell)^2\right] = \mathbb{P}\left(\gamma(\mathcal{P}_\ell) \geq 0\right) \cdot \left(\theta_a^2 \sigma_a^2 + \theta_s^2\, \mathbb{E}\left[\gamma_s^+(\mathcal{P}_\ell)^2\right]\right).$$

Similarly to the case of a single light-tailed queue, we select the parameters θ_a and θ_m to ensure $\widetilde{S}_\infty = \overline{S}_\infty$. Finding $\overline{S}_\infty$ in a general feed-forward network is however challenging. Instead, we ensure that the expression in Eq. (75) matches the approximation of the expected steady-state system time obtained via network decomposition, presented in Eq. (64). We then choose θ_a and θ_s as

$$\theta_a \approx \left[\frac{2\sum\limits_{P\in\mathcal{P}} f_P\cdot|P|}{\sum\limits_{\ell\in\mathcal{J}}\mathbb{P}\left(\gamma\left(\mathcal{P}_\ell\right)\geq 0\right)}\right]^{1/2} \quad \text{and} \quad \theta_s \approx \left[\frac{2\sum\limits_{P\in\mathcal{P}} f_P\sum\limits_{j\in P}\phi_j\left(\sigma_s^{(j)}\right)^2}{\sum\limits_{\ell\in\mathcal{J}}\mathbb{P}\left(\gamma\left(\mathcal{P}_\ell\right)\geq 0\right)\cdot\mathbb{E}\left[\gamma_s^+\left(\mathcal{P}_\ell\right)^2\right]}\right]^{1/2}. \tag{76}$$

Note: We introduce the parameter $\Gamma^\ell = \theta_a\gamma_a + \theta_s\gamma_s^{\ell+}$, where

$$\gamma_s^{\ell+} = \left[\sum_{P\in\mathcal{P}_\ell}\sum_{j\in P}\left(\gamma_s^{(j)+}\cdot\phi_j^{1/\alpha}\right)^{\alpha/(\alpha-1)}\right]^{(\alpha-1)/\alpha}. \tag{77}$$

Notice that $\gamma_s^{\ell+}\geq\gamma_s^+\left(\mathcal{P}_\ell\right)$, and therefore the parameter $\Gamma^\ell\geq\Gamma\left(\mathcal{P}_\ell\right)$, for all $\ell\in\mathcal{J}$. Since a higher parameter value yields higher system and relaxation time, we can bound $\widehat{S}_n\left(\mathcal{P}_\ell\right)=\widehat{S}_n\left(\Gamma\left(\mathcal{P}_\ell\right)\right)$ by $\widehat{S}_n\left(\Gamma^\ell\right)$, and hence we can bound $\widetilde{S}_n$ by

$$\widetilde{S}_n = \sum_{\ell\in\mathcal{J}} p_\ell\cdot\widetilde{S}_n\left(\mathcal{P}_\ell\right) \leq \sum_{\ell\in\mathcal{J}} p_\ell\cdot\widetilde{S}_n\left(\Gamma^\ell\right) = \sum_{\ell\in\mathcal{J}} p_\ell\cdot\mathbb{E}\left[\widehat{S}_n\left(\Gamma^\ell\right)\right].$$

We next show that the choice of the parameters θ_a and θ_s for the above approximation allows for simpler computations.

(a) *Light-Tailed Primitives:* By using the upper bound $\widetilde{S}_n\left(\Gamma^\ell\right)$ introduced above, the expected value of the overall worst case steady-state system time in Eq. (75) can be bounded by

$$\widetilde{S}_\infty \leq \sum_{\ell\in\mathcal{J}} p_\ell\cdot\frac{\lambda}{4(1-\rho)}\cdot\mathbb{E}\left[\left(\gamma_\ell^+\right)^2\right] + \sum_{P\in\mathcal{P}} f_P\sum_{j\in P}\left(\frac{1}{\mu_j}+\mathbb{E}\left[\Gamma_m^{(j)+}\right]\right), \tag{78}$$

where $\gamma_\ell = \theta_a\gamma_a+\theta_s\gamma_s^{\ell+}$ and $\gamma_s^{\ell+}$ is defined in Eq. (77). The expected value in Eq. (78)

$$\mathbb{E}\left[\left(\gamma_\ell^+\right)^2\right] \approx \mathbb{P}\left(\gamma_\ell\geq 0\right)\cdot\mathbb{E}\left[\gamma_\ell^2\right] = \mathbb{P}\left(\gamma_\ell\geq 0\right)\cdot\left(\theta_a^2\sigma_a^2+\theta_s^2\,\mathbb{E}\left[\left(\gamma_s^{\ell+}\right)^2\right]\right),$$

where, the second moment of $\gamma_s^{\ell+}$ can be expressed as

$$\mathbb{E}\left[\left(\gamma_s^{\ell+}\right)^2\right] = \sum_{P\in\mathcal{P}_\ell}\sum_{j\in P}\phi_j\cdot\mathbb{E}\left[\left(\gamma_s^{(j)+}\right)^2\right] = \mathbb{P}\left(\gamma_s^{(1)}\geq 0\right)\cdot\sum_{P\in\mathcal{P}_\ell}\sum_{j\in P}\phi_j\cdot\left(\sigma_s^{(j)}\right)^2.$$

We proceed by performing an additional bounding procedure to help simplify the computations. Specifically, we propose to bound the expression

$$\begin{aligned}\sum_{\ell\in\mathcal{J}}\mathbb{P}\left(\gamma_\ell\geq 0\right)\cdot\sum_{P\in\mathcal{P}_\ell}\sum_{j\in P}\phi_j\cdot\left(\sigma_s^{(j)}\right)^2 &\leq \sum_{\ell\in\mathcal{J}}\mathbb{P}\left(\gamma_\ell\geq 0\right)\cdot\sum_{\ell\in\mathcal{J}}\sum_{P\in\mathcal{P}_\ell}\sum_{j\in P}\phi_j\cdot\left(\sigma_s^{(j)}\right)^2,\\ &= \sum_{\ell\in\mathcal{J}}\mathbb{P}\left(\gamma_\ell\geq 0\right)\cdot\sum_{P\in\mathcal{P}}\sum_{j\in P}\phi_j\cdot\left(\sigma_s^{(j)}\right)^2. \end{aligned} \tag{79}$$

To match the approximation of the expected steady-state system time obtained via network decomposition presented in Eq. (64) and the resulting upper bound on $\widetilde{S}_\infty$ from combining Eqs. (78) and (79), we choose θ_a and θ_s as

$$\theta_a \approx \left(\frac{2\sum\limits_{P\in\mathcal{P}} f_P\cdot|P|}{\sum\limits_{\ell\in\mathcal{J}}\mathbb{P}\left(\gamma_\ell\geq 0\right)}\right)^{1/2} \quad \text{and} \quad \theta_s \approx \left(\frac{2}{\sum\limits_{\ell\in\mathcal{J}}\mathbb{P}\left(\gamma_\ell\geq 0\right)\cdot\mathbb{P}\left(\gamma_s^{(1)}\geq 0\right)}\right)^{1/2}. \tag{80}$$

The above expressions reduce to Eq. (59) for the case of a tandem queue, where $\mathcal{P}=(a_0,\ldots,|\mathcal{J}|)$. Note that, given that $\gamma_s^{(1)}$ is a normally distributed distributed random variable centered around the origin, we have $\mathbb{P}\left(\gamma_s^{(1)}\geq 0\right)=1/2$. Also,

$$\mathbb{P}\left(\gamma_\ell\geq 0\right) = \mathbb{P}\left(\theta_a\gamma_a+\theta_s\gamma_s^{\ell+}\geq 0\right) = \mathbb{P}\left(\left\{\sum_{P\in\mathcal{P}} f_P\cdot|P|\right\}^{1/2}\cdot\gamma_a+\mathbb{P}(\gamma_s^{(1)}\geq 0)^{-1/2}\cdot\gamma_s^+\geq 0\right),$$

which can be efficiently computed numerically.

(b) *Heavy-Tailed Queues:* Since the steady state does not exist for heavy-tailed queues, we propose to extend the formulas for θ_a and θ_s and obtain

$$\theta_a \approx \left(\frac{\alpha \sum\limits_{P \in \mathcal{P}} f_P \cdot |P|}{\sum\limits_{\ell \in \mathcal{J}} \mathbb{P}\left(\gamma_\ell \geq 0\right)} \right)^{(\alpha-1)/\alpha} \quad \text{and} \quad \theta_s \approx \left(\frac{\alpha}{\sum\limits_{\ell \in \mathcal{J}} \mathbb{P}\left(\gamma \geq 0\right) \cdot \mathbb{P}\left(\gamma_s^{(1)} \geq 0\right)} \right)^{(\alpha-1)/\alpha}, \tag{81}$$

where $\gamma = \theta_a \gamma_a + \theta_m \gamma_s^+ / m$, γ_s^+ is defined in Eq. (77) . Note that the probability $\mathbb{P}\left(\gamma_s^{(1)} \geq 0\right)$ and

$$\mathbb{P}\left(\gamma \geq 0\right) = \mathbb{P}\left(\{ \sum_{P \in \mathcal{P}} f_P \cdot |P| \, \}^{(\alpha-1)/\alpha} \cdot \gamma_a + \mathbb{P}(\gamma_s^{(1)} \geq 0)^{-(\alpha-1)/\alpha} \cdot \gamma_s^{\ell+} \geq 0 \right)$$

can be efficiently computed numerically given the distributions of γ_a and $\gamma_s^{\ell+}$.

Insights and Computational Tractability

The insights we draw from our analysis of light-tailed and heavy-tailed feed-forward queueing networks queues are similar to the ones obtained for single and tandem queues. Furthermore, simulating the expected overall system time of the n^{th} job in a feed-forward network requires simulating all queues in every path $P \in \mathcal{P}$ in the system for all n jobs. Our approach, on the other hand, involves (a) running a simulation to fit the distribution of $\gamma_s^{\ell+}$ as defined in Eq. (77), and (b) computing double integrals with respect to γ_a and $\gamma_s^{\ell+}$, for all nodes $\ell \in \mathcal{J}$. Note that extending the results to multi-server feed-forward networks does not affect the efficiency of our approach.

6 Concluding Remarks: Limitations and Future Directions

In this paper, we studied the problem of analyzing the transient system time in multi-server queueing systems and feed-forward networks. For such queueing systems, we presented an analytically tractable approach to analyzing the transient behavior with general, possibly heavy-tailed, arrival and service processes. This is achieved by modeling the system's inter-arrival and service times via polyhedral sets which are characterized by parameters that control the degree of conservatism. We obtain closed-form expressions for the worst case system time revealing qualitative insights on the dependence of the system time on the traffic intensity and the tail behavior of the inter-arrival and service times. We propose a novel algorithm to approximate the expected system time by averaging the worst case system times by treating the parameters characterizing the uncertainty sets as random variables. This proposed methodology provides a novel framework to study stochastic systems that combines the computational tractability of optimization and the notion of dimensional reduction of uncertainty.
As observed from the numerical results, this methodology yields accurate predictions with low errors relative to simulation, especially for queueing systems with general light-tailed primitives. However, the approximation errors are higher for the following systems which also leads to suggest future directions:

- *Exploring alternate approximations for multi-server queueing systems with heavy tailed services or arrivals*: Our approach currently suggests that the expected waiting time for multi server queues with heavy tailed primitives is infinite, which is not true for all multi server systems.
- *Exploring alternate ways to analyze early transient behavior*: Our approach, which is based on limit laws, leads to relatively higher errors when analyzing early transient regime of multi-server queueing systems.
- *Obtain performance bounds on the tail probability of the performance measure of interest*: Our approach allows us to analyze quantiles and expected values of the waiting times, but does not provide a direct way to calculate tail probabilities. Our approach could potentially be used for this purpose by constructing constraints implied by bounds on the tail probabilities of the underlying stochastic processes.
- *Analyze queueing systems with feedback*: In this paper, we analyzed queueing networks with feed-forward structure. A natural extension would be to also consider queueing networks where some of the customers are fed back into the system.

Overall, we believe that we are just beginning to understand the application of robust optimization based approaches to analyze the expected behavior of stochastic systems and we certainly expect that our approach can be strengthened and extended in various directions as discussed above.

Acknowledgements

We would like to thank the associate editor and the reviewers of our paper for their comments that improved the paper.

References

1. Joseph Abate and Ward Whitt. Transient behavior of regulated brownian motion, i: Starting at the origin. *Advances in Applied Probability*, 19(3):560–598, 1987.
2. Joseph Abate and Ward Whitt. Transient behavior of the *M/M/1* queue: Starting at the origin. *Queueing Systems*, 2(1):41–65, 1987.
3. Joseph Abate and Ward Whitt. Transient behavior of the *M/M/1* queue via laplace transforms. *Advances in Applied Probability*, 20(1):145–178, 1987.
4. Joseph Abate and Ward Whitt. Calculating transient characteristics of the erlang loss model by numerical transform inversion. *Stochastic Models*, 14(3):663–680, 1998.
5. Sren Asmussen, Klemens Binswanger, and Bjarne Hjgaard. Rare events simulation for heavy-tailed distributions. *Bernoulli*, 6(2):pp. 303–322, 2000.
6. N. T. J. Bailey. A continuos time treatment of a simple queue using generating functions. *Journal of Royal Statistical Society*, B16:288–291, 1954.
7. N. T. J. Bailey. some further results in the non-equilibrium theory of a simple queue. *Journal of Royal Statistical Society*, B19:326–333, 1954.
8. Chaithanya Bandi and Dimitris Bertsimas. Tractable stochastic analysis via robust optimization. *Mathematical Programming*, 134:23–70, 2013.
9. Chaithanya Bandi, Dimitris Bertsimas, and Nataly Youssef. Robust queueing theory. *Operations Research*, 63(3):676–700, 2015.
10. Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *NATURE*, 435:207, 2005.
11. Aharon Ben-Tal, Laurent El-Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009.
12. Theophilus Benson, Aditya Akella, and David A. Maltz. Network traffic characteristics of data centers in the wild. In *Proceedings of the 10th annual conference on Internet measurement*, IMC '10, pages 267–280, New York, NY, USA, 2010. ACM.
13. D. Bertsimas, D. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53:464–501, 2011.
14. D. Bertsimas, D. Gamarnik, and A. Rikun. Performance analysis of queueing networks via robust optimization. *Operations Research*, 3:68–93, 2011.
15. D. Bertsimas, J. Keilson, D. Nakazato, and H. Zhang. Transient and busy period analysis of the *GI/G/1* queue as a hilbert factorization problem. *Journal of Applied Probability*, 28:873–885, 1991.
16. D. Bertsimas and D. Nakazato. Transient and busy period analysis for the *GI/G/1* queue; the method of stages. *Queuing Systems and Applications*, 10:153–184, 1992.
17. Dimitris Bertsimas, David Gamarnik, and Alexander Anatoliy Rikun. Performance analysis of queueing networks via robust optimization. *Operations research*, 59(2):455–466, 2011.
18. Dimitris Bertsimas and Karthik Natarajan. A semidefinite optimization approach to the steady-state analysis of queueing systems. *Queueing Systems*, 56(1):27–39, 2007.
19. Jose Blanchet and Peter Glynn. Efficient rare-event simulation for the maximum of heavy-tailed random walks. *The Annals of Applied Probability*, 18(4):1351–1378, 08 2008.
20. O.J. Boxma and J.W. Cohen. The m/g/1 queue with heavy-tailed service time distribution. *IEEE Journal on Selected Areas in Communications*, 16(5):749–763, 1998.
21. S. S. L. Chang. Simulation of transient and time varying conditions in queueing networks. *Proceedings of the Seventh Annual Pittsburgh Conference on Modeling and Simulation*, pages 1075–1078, 1977.
22. Gagan L. Choudhury, David M. Lucantoni, and Ward Whitt. Multi-dimensional transform inversion with applications to the transient *M/G/1* queue. *Annals of Applied Probability*, 4:719–740, 1994.
23. Gagan L. Choudhury and Ward Whitt. Computing transient and steady-state distributions in polling models by numerical transform inversion. *IEEE International Conference on Communications*, pages 803–809, 1995.
24. M. Crovella. The relationship between heavy-tailed file sizes and self-similar network traffic. *INFORMS Applied Probability Conference*, 1997.
25. G. B. Dantzig. Programming of interdependent activities: II mathematical model. *Econometrica*, 17:200–211, 1949.
26. A. K. Erlang. The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik, B*, 20, 1909.
27. George S. Fishman and Ivo J. B. F. Adan. How heavy-tailed distributions affect simulation-generated time averages. *ACM Trans. Model. Comput. Simul.*, 16(2):152–173, April 2006.
28. Sergey Foss and Dmitry Korshunov. On large delays in multi-server queues with heavy tails. *Mathematics of Operations Research*, 37(2):201–218, 2012.
29. Serguei Foss and Dmitry Korshunov. Heavy tails in multi-server queue. *Queueing Systems*, 52(1):31–48, 2006.
30. W. K. Grassmann. Transient solutions in markovian queueing systems. *Comput. Opns. Res.*, 4:47–53, 1977.
31. W. K. Grassmann. Transient and steady state results for two parallel queues. *Omega*, 8:105–112, 1980.
32. D. Gross and C. M. Harris. Fundamentals of queueing theory. *John Wiley & Sons, New York.*, 1974.
33. RobertC. Hampshire, Mor Harchol-Balter, and WilliamA. Massey. Fluid and diffusion limits for transient sojourn times of processor sharing queues with time varying rates. *Queueing Systems*, 53(1-2):19–30, 2006.
34. J.M. Harrison and R.J. Williams. Brownian models of feedforward queueing networks: Quasireversibility and product form solutions. *The Annals of Applied Probability*, 2(2):263–293, 1992.
35. D. P. Heyman and M. J. Sobel. *Stochastic Models in Operations Research: Vol. 1.* McGraw-Hill, New York, 1982.
36. Samuel Karlin and James McGregor. Many server queueing processes with poisson input and exponential service times. *Pacific Journal of Mathematics*, 8(1):87–118, 1958.
37. J. Keilson. Markov chain models-rarity and exponentiality. *Springer-Verlag*, 1979.
38. W. David Kelton and Averill M. Law. The transient behavior of the *M/M/s* queue, with implications for steady-state simulation. *Operations Research*, 33(2):378–396, 1985.

39. J.F.C. Kingman. Inequalities in the theory of queues. *Journal of the Royal Statistical Society*, 32:102–110, 1970.
40. B. O. Koopman. Revenue maximization when bidders have budgets. *Operations Research*, pages 1089–1114, 1972.
41. T. C. T. Kotiah. Approximate transient analysis of some queuing systems. *Operations Research*, 26(2):333–346, 1978.
42. N.K. Krivulin. A recursive equations based representation of the $G/G/m$ queue. *Applied Math Letters*, 7(3):73–77, 1994.
43. W.E. Leland, M.S. Taqqu, and D.V. Wilson. On the self-similar nature of Ethernet traffic. *ACM SIGCOMM Computer Communication Review*, 25(1):202–213, 1995.
44. D. V. Lindley. The theory of queues with a single server. *Mathematical Proceedings of the Cambridge Philosophical Society*, 1952.
45. Charles Loboz. Cloud resource usage - heavy tailed distributions invalidating traditional capacity planning models. *J. Grid Comput.*, 10(1):85–108, 2012.
46. WilliamA. Massey. The analysis of queues with time-varying rates for telecommunication models. *Telecommunication Systems*, 21(2-4):173–204, 2002.
47. S. C. Moore. Approximating the behavior of non-stationary single server queues. *Operations Research*, 23:1011–1032, 1975.
48. M. Mori. Transient behavior of the mean waiting time and its exact forms in *M/M/1* and *M/D/1*. *Journal of the Operations Research Society of Japan*, 19:14–31, 1976.
49. M. Neuts. The single server queue in discrete time: Numerical analysis I. *Naval Research Logistics*, 20:297–304, 2004.
50. G.F. Newell. *Applications of Queueing Theory.* Chapman & Hall, 1971.
51. A.R. Odoni and E. Roth. An empirical investigation of the transient behavior of stationary queueing systems. *Operations Research*, 31(3):432–455, 1983.
52. Takayuki Osogami and Rudy Raymond. Analysis of transient queues with semidefinite optimization. *Queueing Systems*, pages 195–234, 2013.
53. K. L. Rider. A simple approximation to the average queue size in the time-dependent *M/M/1* queue. *Journal of the ACM*, 23(2):361–367, 1976.
54. M. H. Rothkopf and S. S. Oren. A closure approximation for the nonstationary *M/M/s* queue. *Management Science*, 25:522–534, 1979.
55. G. Samorodnitsky and M.S. Taqqu. *Stable non-Gaussian Random Processes: Stochastic Models with Infinite Variance.* Chapman & Hall, 1994.
56. Martin Schlather. Limit distributions of norms of vectors of positive iid random variables. *Annals of probability*, pages 862–881, 2001.
57. H. Vasquez-Leal, R. Castaneda-Sheissa, U. Filobello-Nino, A. Sarmiento-Reyes, and J. Sanchez Orea. High accurate simple approximation of normal distribution related integrals. *Mathematical Problems in Engineering*, 2012.
58. Ward Whitt. The impact of a heavy-tailed service-time distribution upon the m/gi/s waiting-time distribution. *Queueing Systems*, 36(1-3):71–87, 2000.
59. Jing Xie, Yuming Jiang, and Min Xie. A temporal approach to stochastic network calculus. *CoRR*, abs/1112.2822, 2011.

Appendix: All Proofs

Proof of Theorem 2. Since $(\nu - k + 1)^{1/\alpha} \le (\nu - k)^{1/\alpha} + 1$, and given $\mathit{\Gamma}_m^+ \ge 0$, we bound Eq. (23) by

$$\widehat{S}_n \le \max_{0\le k\le \nu} \left\{ \frac{\nu - k}{\mu} + \mathit{\Gamma}_m^+ (\nu - k)^{1/\alpha} - \frac{m(\nu - k)}{\lambda} + \mathit{\Gamma}_a \left[m(\nu - k) \right]^{1/\alpha} \right\} + \left(\frac{1}{\mu} + \mathit{\Gamma}_m^+ \right).$$

By making the transformation $x = \nu - k$, where $x \in \mathbb{N}$, we can represent this problem as

$$\max_{0\le x\le \nu, x\in\mathbb{N}} \left(\beta \cdot x^{1/\alpha} - \delta \cdot x \right) \quad \le \quad \max_{0\le x\le \nu, x\in\mathbb{R}} \left(\beta \cdot x^{1/\alpha} - \delta \cdot x \right), \tag{82}$$

where $\beta = m^{1/\alpha}\mathit{\Gamma}_a + \mathit{\Gamma}_m^+$ and $\delta = m(1-\rho)/\lambda > 0$, given $\rho < 1$. If $\beta \le 0$, the function $h(x) = \beta \cdot x^{1/\alpha} - \delta \cdot x \le 0$ for all values of x, implying $\widehat{S}_n = 1/\mu + \mathit{\Gamma}_m^+$. For $\beta > 0$, the function h is concave in x with an unconstrained maximizer

$$x^* = \left(\frac{\beta}{\alpha\delta} \right)^{\alpha/(\alpha-1)} = \left(\frac{\lambda(\mathit{\Gamma}_m + m^{1/\alpha}\mathit{\Gamma}_a)}{\alpha m(1-\rho)} \right)^{\alpha/(\alpha-1)}. \tag{83}$$

Maximizing the function $h(\cdot)$ over the interval $[0, \nu]$ involves a constrained one-dimensional concave maximization problem whose solution gives rise to closed-form solutions.

(a) If $x^* \in [0, \nu]$, then x^* is the maximizer of the function h over the interval $[0, \nu]$, leading to an expression that is independent of ν,

$$\widehat{S}_n \le \beta \left(\frac{\beta}{\alpha\delta} \right)^{1/(\alpha-1)} - \delta \left(\frac{\beta}{\alpha\delta} \right)^{\alpha/(\alpha-1)} + \left(\frac{1}{\mu} + \mathit{\Gamma}_m^+ \right) = \frac{(\alpha-1)}{\alpha^{\alpha/(\alpha-1)}} \cdot \frac{\beta^{\alpha/(\alpha-1)}}{\delta^{1/(\alpha-1)}} + \left(\frac{1}{\mu} + \mathit{\Gamma}_m^+ \right). \tag{84}$$

(b) If $x^* > \nu$, the function h is non-decreasing over the interval $[0, \nu]$, with $h(\nu) \ge h(x)$ for all $x \in [0, \nu]$, leading to an expression that is dependent on ν,

$$\widehat{S}_n = \beta(\nu)^{1/\alpha} - \delta(\nu) + \left(\frac{1}{\mu} + \mathit{\Gamma}_m^+ \right). \tag{85}$$

We obtain Eq. (24) by substituting β and δ by their expressions in parts (a) and (b). □

Proof of Theorem 3. To bound the maximization problem in Eq. (32), we take a similar approach to that presented in the proof of Theorem 2 and cast the problem in the form

$$\max_{0\leq x\leq \nu-\phi, x\in\mathbb{R}} \left(\beta\cdot x^{1/\alpha}-\delta\cdot x\right)=\begin{cases}\beta\cdot(\nu-\phi)^{1/\alpha}-\delta\cdot(\nu-\phi), & \text{if } \nu-\phi\leq\left(\frac{\beta}{\alpha\delta}\right)^{\alpha/(\alpha-1)},\\ \dfrac{(\alpha-1)}{\alpha^{\alpha/(\alpha-1)}}\cdot\dfrac{\beta^{\alpha/(\alpha-1)}}{\delta^{1/(\alpha-1)}}, & \text{otherwise,}\end{cases}$$

where $\beta=m^{1/\alpha}\Gamma_a+\Gamma_m^+$ and $\delta=m(1-\rho)/\lambda$. Substituting the terms β and ϕ by their respective values in the above expression yields the desired result. □

*Proof of Theorem 5.*From Eq. (50), we have that the worst case system time is given by

$$\widehat{S}_n=\frac{J}{\mu}+\max_{0\leq k_1\leq\ldots\leq k_J\leq\nu}\left\{\begin{array}{l}\left[\Gamma_m^{(1)+}(k_2-k_1+1)^{1/\alpha}+\ldots+\Gamma_m^{(J)+}(\nu-k_J+1)^{1/\alpha}\right]+\\ \Gamma_a\left[m(\nu-k_1)\right]^{1/\alpha}-\dfrac{m(1-\rho)}{\lambda}(\nu-k_1)\end{array}\right\}.$$

Furthermore, since $(k_{j+1}-k_j+1)^{1/\alpha}\leq(k_{j+1}-k_j)^{1/\alpha}+1$, for all j=1,..., J, we obtain

$$\widehat{S}_n\leq\frac{J}{\mu}+\sum_{j=1}^{J}\Gamma_m^{(j)+}+\max_{0\leq k_1\leq\ldots\leq k_J\leq\nu}\left\{\begin{array}{l}\left[\Gamma_m^{(1)+}(k_2-k_1)^{1/\alpha}+\ldots+\Gamma_m^{(J)+}(\nu-k_J)^{1/\alpha}\right]+\\ \Gamma_a\left[m(\nu-k_1)\right]^{1/\alpha}-\dfrac{m(1-\rho)}{\lambda}(\nu-k_1)\end{array}\right\}.$$

We will isolate the problem of maximizing $\left[\Gamma_m^{(1)+}(k_2-k_1)^{1/\alpha}+\ldots+\Gamma_m^{(J)+}(\nu-k_J)^{1/\alpha}\right]$ for fixed values of k_1,ν, and make the transformations $x_1=k_2-k_1,\ldots,x_J=\nu-k_J$, where $x_j\in\mathbb{N}$, for all $j=1,\ldots,J$. With these transformations, the optimization problem simplifies to

$$\max_{0\leq k_1\leq\nu, k_1\in\mathbb{N}}\left(m^{1/\alpha}\Gamma_a(\nu-k_1)^{1/\alpha}-\frac{m(1-\rho)}{\lambda}(\nu-k_1)+\left\{\begin{array}{ll}\max & \left[\Gamma_m^{(1)+}x_1^{1/\alpha}+\ldots+\Gamma_m^{(J)+}x_J^{1/\alpha}\right]\\ \text{s.t.} & x_1+\ldots+x_J=\nu-k_1\\ & x_j\in\mathbb{N},\forall j=2,\ldots,J\end{array}\right\}\right) \tag{86}$$

The optimal solution to the inner optimization problem satisfies

$$\Gamma_m^{(1)+}(x_1^*)^{1/(\alpha-1)}=\Gamma_m^{(2)+}(x_2^*)^{1/(\alpha-1)}=\ldots=\Gamma_m^{(J)+}(x_J^*)^{1/(\alpha-1)},$$

by the first order optimality conditions. Using the additional condition that $\sum_{j=1}^{J}x_j^*=\nu-k_1$, the optimal solution can be found analytically as

$$x_i^*=\frac{(\Gamma_m^{(i)+})^{\alpha/(\alpha-1)}}{\displaystyle\sum_{j=1}^{J}(\Gamma_m^{(j)+})^{\alpha/(\alpha-1)}}\cdot(\nu-k_1)\quad\forall i=1,2,\ldots,J,$$

leading to an optimal value of

$$\Gamma_m^{(1)+}(x_1^*)^{1/\alpha}+\ldots+\Gamma_m^{(J)+}(x_1^*)^{1/\alpha}=(\nu-k_1)^{1/\alpha}\cdot\left(\sum_{j=1}^{J}(\Gamma_m^{(j)+})^{\alpha/(\alpha-1)}\right)^{(\alpha-1)/\alpha} \tag{87}$$

Substituting the optimal solution of the inner problem in Eq. (86), the performance analysis reduces to solving the following one-dimensional optimization problem

$$\max_{0\leq k_1\leq\nu}\left\{\left(m^{1/\alpha}\Gamma_a+\left[\sum_{j=1}^{J}(\Gamma_m^{(j)+})^{\alpha/(\alpha-1)}\right]^{(\alpha-1)/\alpha}\right)\cdot(\nu-k_1)^{1/\alpha}-\frac{m(1-\rho)}{\lambda}(\nu-k_1)\right\}, \tag{88}$$

which can be cast in the form of the optimization problem in Eq. (82), with

$$\beta=m^{1/\alpha}\Gamma_a+\left(\sum_{j=1}^{J}(\Gamma_m^{(j)+})^{\alpha/(\alpha-1)}\right)^{(\alpha-1)/\alpha}\quad\text{and}\quad\delta=\frac{m(1-\rho)}{\lambda}.$$

Referring to the proof of Theorem 2, the solution to Eq. (88) is

$$\max_{0 \le x \le \nu} \quad \beta \cdot x^{1/\alpha} - \delta \cdot x = \begin{cases} \beta \cdot \nu^{1/\alpha} - \delta \cdot \nu, & \text{if } \nu \le \left(\frac{\beta}{\alpha\delta}\right)^{\alpha/(\alpha-1)} \\ \dfrac{(\alpha-1)}{\alpha^{\alpha/(\alpha-1)}} \cdot \dfrac{\beta^{\alpha/(\alpha-1)}}{\delta^{1/(\alpha-1)}}, & \text{otherwise.} \end{cases}$$

We obtain the desired result by substituting β and δ by their respective values. □

Proof of Theorem 6. We maximize both terms in Eq. (54) separately as follows.

(a) By Assumption 1 and applying similar arguments to those presented in the proof of Theorem 5, the first term in Eq. (54) is bounded by

$$\max_{\substack{0 \le k_1 \le \phi, \\ k_1 \in \mathbb{N}}} \left(\frac{\nu - k_1}{\mu} + \left\{ \begin{array}{l} \max \left[\mathit{\Gamma}_m^{(1)+} x_1^{1/\alpha} + \ldots + \mathit{\Gamma}_m^{(J)+} x_J^{1/\alpha} \right] \\ \text{s.t. } \; x_1 + \ldots + x_J = \nu - k_1 \\ \qquad x_j \in \mathbb{N}, \forall j = 2, \ldots, J \end{array} \right\} \right) + \frac{J}{\mu} + \sum_{j=1}^{J} \mathit{\Gamma}_m^{(j)+} - \frac{n - n_0}{\lambda} + \gamma_a (n - n_0)^{1/\alpha}. \tag{89}$$

The optimal objective function of the inner optimization problem in Eq. (89) is given by Eq. (87). Hence, the bound on the first term in Eq. (54) becomes

$$\max_{0 \le k_1 \le \phi} \left(\frac{\nu - k_1}{\mu} + \mathit{\Gamma}_m \cdot (\nu - k_1)^{1/\alpha} \right) + \frac{J}{\mu} + \sum_{j=1}^{J} \mathit{\Gamma}_m^{(j)+} - \frac{n - n_0}{\lambda} + \gamma_a (n - n_0)^{1/\alpha},$$

where $\mathit{\Gamma}_m$ is defined in Eq. (51). Since $\mathit{\Gamma}_m \ge 0$, the term $x/\mu + \mathit{\Gamma}_m x^{1/\alpha}$ is increasing in x, yielding

$$\max_{0 \le k_1 \le \phi} \left(\frac{\nu - k_1}{\mu} + \mathit{\Gamma}_m \cdot (\nu - k_1)^{1/\alpha} \right) = \frac{\nu}{\mu} + \mathit{\Gamma}_m \cdot \nu^{1/\alpha}.$$

(b) To bound the second term in Eq. (54), we take a similar approach to that presented in the proof of Theorem 5 and cast the problem in the form

$$\max_{0 \le x \le \nu - \phi, x \in \mathbb{R}} \left(\beta \cdot x^{1/\alpha} - \delta \cdot x \right) = \left\{ \begin{array}{ll} \beta \cdot (\nu - \phi)^{1/\alpha} - \delta \cdot (\nu - \gamma) & \text{if } \nu - \phi \le \left(\frac{\beta}{\alpha\delta}\right)^{\alpha/(\alpha-1)} \\ \dfrac{(\alpha-1)}{\alpha^{\alpha/(\alpha-1)}} \cdot \dfrac{\beta^{\alpha/(\alpha-1)}}{\delta^{1/(\alpha-1)}} & \text{otherwise} \end{array} \right\}.$$

Substituting $\beta = m^{1/\alpha} \mathit{\Gamma}_a + \mathit{\Gamma}_m$ and $\delta = m(1 - \rho)/\lambda$ yields the desired result.

□

Proof of Theorem 7. The desired result is obtained by maximizing the system time for each path $P \in \mathcal{P}_\ell$. In order to apply the bounds on the system times from Assumption 1 to the quantity in Eq. (67), we need to account for the number of jobs that pass through node a_j between the arrivals of job k_{a_j} which belongs to $\mathcal{E}_{a_{j-1}a_j} \subseteq \mathcal{L}_{a_j}$ and job $k_{a_{j+1}}$ which belongs to $\mathcal{E}_{a_j a_{j+1}} \subseteq \mathcal{L}_{a_j}$. Mathematically, we let $\mathit{\Delta}_{a_j}$ denote this number, i.e.,

$$\mathit{\Delta}_{a_j} = \left| \left\{ k : k_{a_j} \le k \le k_{a_{j+1}}, k \in \mathcal{L}_{a_j} \right\} \right|. \tag{90}$$

By Eq. (61), the fraction of jobs passing through queue a_j is ϕ_{a_j}, yielding

$$\mathit{\Delta}_{a_j} = \phi_{a_j} \cdot \left(k_{a_{j+1}} - k_{a_j} + 1 \right).$$

By Assumption 1, and given that $\tilde{\mathit{\Gamma}}_s^{(j)} \le \tilde{\mathit{\Gamma}}_s^{(j)+}$, for all $j \in \mathcal{J}$, we bound the service times by

$$\max_{\mathcal{U}_{a_j}^s} \sum_{i=k_{a_j}}^{k_{a_{j+1}}} X_i^{(a_j)} = \frac{\mathit{\Delta}_{a_j}}{\mu_{a_j}} + \mathit{\Gamma}_s^{(a_j)+} \cdot \mathit{\Delta}_{a_j}^{1/\alpha} = \frac{\phi_{a_j} \cdot \left(k_{a_{j+1}} - k_{a_j} + 1 \right)}{\mu_{a_j}} + \mathit{\Gamma}_s^{(a_j)+} \cdot \left[\phi_{a_j} \cdot \left(k_{a_{j+1}} - k_{a_j} + 1 \right) \right]^{1/\alpha}.$$

By applying Assumptions 1, Eq. (67) becomes

$$\max_{P\in\mathcal{P}_\ell}\left[\sum_{j\in P}\left(\frac{1}{\tilde{\mu}_j}+\tilde{\Gamma}_s^{(j)+}\right)+\max_{1\le k_{a_1}\le\ldots\le k_\ell\le n}\left\{\begin{array}{l}\dfrac{k_{a_2}-k_{a_1}}{\tilde{\mu}_{a_1}}+\tilde{\Gamma}_s^{(a_1)+}\cdot(k_{a_2}-k_{a_1})^{1/\alpha}+\ldots+\dfrac{n-k_\ell}{\tilde{\mu}_\ell}+\\ \tilde{\Gamma}_s^{(\ell)+}\cdot(n-k_\ell)^{1/\alpha}-\dfrac{n-k_{a_1}}{\lambda}+\Gamma_a(n-k_{a_1})^{1/\alpha}\end{array}\right\}\right] \tag{91}$$

where $\tilde{\mu}_j=\mu_j/\phi_j$ and $\tilde{\Gamma}_s^{(j)}=\Gamma_s^{(j)}\cdot\phi_j^{1/\alpha}$, for all $j\in\mathcal{J}$. We let $\tilde{\mu}_P=\min\left\{\tilde{\mu}_{a_j},a_j\in P\right\}$, $\rho_P=\lambda/\tilde{\mu}_P$. By making the change of variable $x_{a_j}=k_{a_{j+1}}-k_{a_j}$, for all $a_j\in P$, we bound the maximization problem in Eq. (91) by

$$\max_{1\le k_{a_1}\le n}\left(\Gamma_a\left(n-k_{a_1}\right)^{1/\alpha}-\frac{1-\rho_P}{\lambda}\left(n-k_{a_1}\right)+\left\{\begin{array}{ll}\max & \left[\tilde{\Gamma}_s^{(a_1)+}\cdot x_{a_1}^{1/\alpha}+\ldots+\tilde{\Gamma}_s^{(\ell)+}\cdot x_{a_J}^{1/\alpha}\right]\\ \text{s.t.} & x_{a_1}+\ldots+x_\ell=n-k_{a_1}\end{array}\right\}\right). \tag{92}$$

The optimal objective function for the inner optimization problem is given in Eq. (87). The performance analysis reduces to solving the following one-dimensional optimization problem

$$\max_{1\le k_{a_1}\le n}\left\{\left(\Gamma_a+\left[\sum_{j\in P}\left(\tilde{\Gamma}_s^{(j)+}\right)^{\alpha/(\alpha-1)}\right]^{(\alpha-1)/\alpha}\right)\cdot\left(n-k_{a_1}\right)^{1/\alpha}-\frac{1-\rho_P}{\lambda}\left(n-k_{a_1}\right)\right\}, \tag{93}$$

which can be cast in the form of the optimization problem in Eq. (82), with

$$\beta=\Gamma_a+\left(\sum_{j\in P}\left(\tilde{\Gamma}_s^{(j)+}\right)^{\alpha/(\alpha-1)}\right)^{(\alpha-1)/\alpha}\quad\text{and}\quad\delta=\frac{1-\rho_P}{\lambda}.$$

Referring to the proof of Theorem 2, the solution to Eq. (93) is

$$\max_{0\le x\le n}\ \beta\cdot n^{1/\alpha}-\delta\cdot n=\begin{cases}\beta\cdot n^{1/\alpha}-\delta\cdot n, & \text{if } n\le\left(\frac{\beta}{\alpha\delta}\right)^{\alpha/(\alpha-1)}\\ \dfrac{(\alpha-1)}{\alpha^{\alpha/(\alpha-1)}}\cdot\dfrac{\beta^{\alpha/(\alpha-1)}}{\delta^{1/(\alpha-1)}}, & \text{otherwise.}\end{cases}$$

We obtain the desired result by substituting β and δ by their respective values. □

Proof of Proposition 2. We prove the result using the technique of mathematical induction.
(a) Initial Step: As presented in [9], the system time in an m-server queue

$$\widehat{S}_n\left(\mathbf{T}\right)=\widehat{S}_n^{(1)}\left(\mathbf{T}\right)=\max_{0\le k_1\le\nu}\left(\max_{\mathbf{X}^{(1)}\in\mathcal{U}_m^s}\sum_{i=k_1}^{\nu}X_{r(i)}^{(1)}-\sum_{i=r(k_1)+1}^{n}T_i\right),$$

and therefore the result holds for $J=1$.
(b) Inductive Step: We now suppose that the result holds for $J-1$ queues in series, which expresses the system time across queues 2 through J as

$$\widehat{S}_n^{(2)}\left(\mathbf{T}\right)+\ldots+\widehat{S}_n^{(J)}\left(\mathbf{T}\right)=\max_{0\le k_2\le\ldots\le k_J\le\nu}\left(\max_{\mathcal{U}_2^s}\sum_{i=k_2}^{k_3}X_{r(i)}^{(2)}+\ldots+\max_{\mathcal{U}_J^s}\sum_{i=k_J}^{n}X_{r(i)}^{(J)}-\sum_{i=r(k_2)+1}^{n}T_i^{(2)}\right), \tag{94}$$

where $\mathbf{T}^{(2)}=\left\{T_1^{(2)},\ldots,T_n^{(2)}\right\}$ denotes the sequence of inter-arrival times to the second queue. Note that the arrival to the second queue is simply the departure from the first queue, and therefore, denoting the inter-departure times from the first queue by $\mathbf{D}^{(1)}=\left\{D_1^{(1)},\ldots,D_n^{(1)}\right\}$, we have

$$\sum_{i=r(k_2)+1}T_i^{(2)}=\sum_{i=r(k_2)+1}D_i^{(1)}=\sum_{i=(k_2)+1}^{n}T_i+\widehat{S}_n^{(1)}\left(\mathbf{T}\right)-\widehat{S}_{r(k_2)}^{(1)}\left(\mathbf{T}\right), \tag{95}$$

where the last equality is due to the fact that no overtaking occurs at the first queue in the worst case approach. Combining Eqs. (94)-(95), we obtain

$$\begin{aligned}\widehat{S}_n(\mathbf{T}) &= \widehat{S}_n^{(1)}(\mathbf{T}) + \widehat{S}_n^{(2)}(\mathbf{T}) + \ldots + \widehat{S}_n^{(J)}(\mathbf{T}) \\ &= \max_{0 \le k_2 \le \ldots \le k_J \le \nu} \left(\max_{\mathcal{U}_2^s} \sum_{i=k_2}^{k_3} X_{r(i)}^{(2)} + \ldots + \max_{\mathcal{U}_J^s} \sum_{i=k_J}^{n} X_{r(i)}^{(J)} - \sum_{i=r(k_2)+1}^{n} T_i^{(2)} + \widehat{S}_{r(k_2)}^{(1)}(\mathbf{T}) \right). \quad (96)\end{aligned}$$

Since no overtaking occurs in the first queue, and given that $\lfloor r(k_2)/m \rfloor = k_2$, the system time of the $r(k_2)^{\text{th}}$ job can be expressed as

$$S_{r(k_2)}^{(1)}(\mathbf{T}) = \max_{0 \le k_1 \le k_2} \left(\max_{\mathcal{U}_1^s} \sum_{i=k_1}^{k_2} X_{r(i)}^{(1)} - \sum_{i=r(k_1)+1}^{r(k_2)} T_i \right).$$

Substituting the above expression in Eq. (96), the overall system time becomes

$$S_n = \max_{0 \le k_2 \le \ldots \le k_J \le \nu} \left(\max_{\mathcal{U}_2^s} \sum_{i=k_2}^{k_3} X_{r(i)}^{(2)} + \ldots + \max_{\mathcal{U}_J^s} \sum_{i=k_J}^{n} X_{r(i)}^{(J)} - \sum_{i=r(k_2)+1}^{n} T_i + \max_{0 \le k_1 \le k_2} \left(\max_{\mathcal{U}_1^s} \sum_{i=k_1}^{k_2} X_{r(i)}^{(1)} - \sum_{i=r(k_1)+1}^{r(k_2)} T_i \right) \right).$$

Rearranging the terms in the above expression proves the inductive result. This concludes the inductive step, and by mathematical induction, we have the desired result. □

Proof of Proposition 3. We use the principle of mathematical induction to prove this result. Specifically, we assume that the result is true for any job $j \le n-1$ passing by some node q from the feed-forward network (disregarding where the j^{th} job goes next in the network after q), i.e.,

$$S_j(\mathcal{P}_q) = \max_{P \in \mathcal{P}_q} \left\{ \max_{\substack{1 \le k_{a_1} \le k_{a_2} \le \ldots \le k_q \le j \\ k_{i+1} \in \mathcal{E}_{a_i a_{i+1}}}} \left(\sum_{\substack{i=k_{b_1} \\ i \in \mathcal{L}_{a_1}}}^{k_{a_2}} X_i^{(a_1)} + \sum_{\substack{i=k_{a_2} \\ i \in \mathcal{L}_{a_2}}}^{k_{a_3}} X_i^{(a_2)} + \ldots + \sum_{\substack{i=k_q \\ i \in \mathcal{L}_q}}^{j} X_i^{(q)} - \sum_{i=k_{a_1}+1}^{j} T_i \right) \right\}, \quad (97)$$

where $\mathcal{P}_q$ denotes the set of all paths $P = (a_0, a_1, \ldots, q)$ that pass by q (disregarding the network after q). We next proceed to show that the result holds for job n exiting the network at queue ℓ.

The system time of the n^{th} job at queue ℓ can be expressed as

$$S_n^{(\ell)} = \max_{\substack{1 \le k_\ell \le n \\ k_\ell \in \mathcal{L}_\ell}} \left(\sum_{\substack{i=k_\ell \\ i \in \mathcal{L}_\ell}}^{n} X_i^{(\ell)} - \sum_{\substack{i=k_\ell+1 \\ i \in \mathcal{L}_\ell}}^{n} T_i^{(\ell)} \right) \quad (98)$$

Suppose $k_\ell \in \mathcal{E}_{q\ell}$, i.e., job k_ℓ enters queue ℓ from queue q, and without loss of generality, suppose that job n enters queue ℓ from queue r, i.e., $n \in \mathcal{E}_{r\ell}$. Then,

$$\sum_{\substack{i=k_\ell+1 \\ i \in \mathcal{L}_\ell}}^{n} T_i^{(\ell)} = \left(\sum_{i=1}^{n} T_i + S_n(\mathcal{P}_r) \right) - \left(\sum_{i=1}^{k_\ell} T_i + S_{k_\ell}(\mathcal{P}_q) \right). \quad (99)$$

Combining Eqs. (98) and (99), we obtain

$$S_n^{(\ell)} + S_n(\mathcal{P}_r) = \max_{\substack{1 \le k_\ell \le n \\ k_\ell \in \mathcal{L}_\ell}} \left(\sum_{\substack{i=k_\ell \\ i \in \mathcal{L}_\ell}}^{n} X_i^{(\ell)} + S_{k_\ell}(\mathcal{P}_q) - \sum_{i=1}^{n} T_i + \sum_{i=1}^{k_\ell} T_i \right) = \max_{\substack{1 \le k_\ell \le n \\ k_\ell \in \mathcal{L}_\ell}} \left(\sum_{\substack{i=k_\ell \\ i \in \mathcal{L}_\ell}}^{n} X_i^{(\ell)} + S_{k_\ell}(\mathcal{P}_q) - \sum_{i=k_\ell+1}^{n} T_i \right)$$

By the induction hypothesis, we substitute the value of $S_{k_\ell}(\mathcal{P}_q)$ in the above equation and obtain

$$S_n^{(\ell)} + S_n(\mathcal{P}_r) = S_n(\mathcal{P}_{r\ell}) = \max_{P \in \mathcal{P}_{q\ell}} \left\{ \max_{\substack{1 \le k_{a_1} \le \ldots \le k_q \le k_\ell \le n \\ k_{i+1} \in \mathcal{E}_{a_i a_{i+1}}}} \left(\sum_{\substack{i=k_{a_1} \\ i \in \mathcal{L}_{a_1}}}^{k_{a_2}} X_i^{(a_1)} + \ldots + \sum_{\substack{i=k_q \\ i \in \mathcal{L}_q}}^{k_\ell} X_i^{(q)} + \sum_{\substack{i=k_\ell \\ i \in \mathcal{L}_\ell}}^{n} X_i^{(\ell)} - \sum_{i=k_{b_1}+1}^{n} T_i \right) \right\},$$

where $\mathcal{P}_{r\ell}$ and $\mathcal{P}_{q\ell}$ are the sets of paths that end at node r and q, respectively, and then feed in to node ℓ (disregarding what comes next in the network). Given that q and r were chosen arbitrarily, the result holds for any nodes q and r that feed into queue ℓ, i.e. for all $q, r \in \mathcal{P}_\ell$. Hence,

$$S_n\left(\mathcal{P}_\ell\right) = \max_{P \in \mathcal{P}_\ell} \left\{ \max_{\substack{1 \le k_{a_1} \le \ldots \le k_q \le k_\ell \le n \\ k_{i+1} \in \mathcal{E}_{a_i a_{i+1}}}} \left(\sum_{\substack{i=k_{a_1} \\ i \in \mathcal{L}_{a_1}}}^{k_{a_2}} X_i^{(a_1)} + \ldots + \sum_{\substack{i=k_q \\ i \in \mathcal{L}_q}}^{k_\ell} X_i^{(q)} + \sum_{\substack{i=k_\ell \\ i \in \mathcal{L}_\ell}}^{n} X_i^{(\ell)} - \sum_{i=k_{a_1}+1}^{n} T_i \right) \right\}. \quad (100)$$

This concludes the inductive step and proves the result for job n. Next considering the base case of $n = 1$, it is trivial to check the validity of inductive hypothesis. Therefore, the result follows from induction. This concludes the proof. □