Submitted to *Operations Research*
manuscript (Please, provide the mansucript number!)

Robust Transient Multi-Server Queues and Feedforward Networks

Chaithanya Bandi[*] Dimitris Bertsimas[†] Nataly Youssef[‡]

We propose an analytically tractable approach for studying the transient behavior of multi-server queueing systems and feedforward networks with deterministic routing. We model the queueing primitives via polyhedral uncertainty sets inspired by the limit laws of probability. These uncertainty sets are characterized by 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 feedforward networks with deterministic routing. These analytic formulas offer rich qualitative insights on the dependence of the system times as a function of fundamental quantities in the queueing system. Moreover, we break new grounds and present an algorithm which appropriately averages worst case values obtained at different degrees of conservatism. This methodology achieves significant computational tractability and provides good approximations for the expected system time relative to simulation.

Key words: Transient Queueing Theory, Robust Optimization, Heavy Tails, Relaxation Time, Steady State

1. Introduction

The origin of queueing theory dates back to the beginning of the 20th century, when Erlang (1909) 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.

[*] Assistant Professor of Economics and Decision Sciences, Kellogg School of Management, Northwestern University, Evanston, IL 60208, c-bandi@kellogg.northwestern.edu

[†] Boeing Professor of Operations Research, Co-director, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, dbertsim@mit.edu

[‡] Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, youssefn@mit.edu

Most of the applications mentioned above are characterized by time varying arrival and service patterns, and even if they have non time-varying such patterns, they experience substantially long transient regimes, especially under heavy-traffic conditions, and may not reach steady state within their operation time window. In addition, the arrival and service processes exhibit heavy-tailed behavior, as reported by Leland et al. (1995) and Crovella (1997) for the Internet; by Barabasi (2005) for call centers; and by Loboz (2012) and Benson et al. (2010) for data centers. In these situations, steady state is never reached. As a result, in many significant applications, steady state analysis is simply not relevant. Consequently, central research questions in this context are mainly concerned with **(a)** the evolution of waiting times over time, and **(b)** the time it takes a queueing system to reach steady state.

Despite the need for 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, Karlin and McGregor (1958) obtained the transition probabilities of the Markov chain describing the queue length as functions of Poisson-Charlier polynomials. Bailey (1954a,b) 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 Abate and Whitt (1987a,b), Choudhury et al. (1994), Choudhury and Whitt (1995), Abate and Whitt (1998)) to obtain additional insights on the queue length process. These analyses also provide insights on the usefulness of reflected Brownian motion approximations for queues. Bertsimas et al. (1991) 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, Bertsimas and Nakazato (1992) 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. Massey (2002), Hampshire et al. (2006) study the transient analysis problem for process sharing markovian queues with time-varying rates using a technique known as "uniform acceleration".

As discussed in Odoni and Roth (1983), there are multiple approximations available but a tractable theory of transient analysis of G/G/m queues is lacking (see also Gross and Harris (1974), Heyman and Sobel (1982), and Keilson (1979)). Further complicating the transient analysis is the effect of initial conditions, which gives rise to a significantly different behaviors as empirically investigated in Kelton and Law (1985) and Odoni and Roth (1983). 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., Koopman (1972), Neuts (2004), Moore (1975), Rider (1976), Grassmann (1977), Chang (1977), Kotiah (1978), Grassmann (1980), and Rothkopf and Oren (1979)). Newell (1971), 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. Mori (1976), 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 Bertsimas and Natarajan (2007), Osogami and Raymond (2013) 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. Xie et al. (2011) 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.

Contributions

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 Bandi et al. (2012) 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 then carry out an average case analysis and break new ground by treating the parameters characterizing the uncertainty sets as random variables and approximate the expected system time via averaging the worst case values. This averaging approach achieves significant tractability by reducing the problem of transient analysis to a two dimensional integral. Our framework combines qualitative insights via closed form expressions and produces accurate predictions of transient system times relative to simulation for heavy traffic queues with various interarrival and service time distributions, heavy tail coefficients and number of servers. Furthermore, our approach is generalizable to networks of queues in series (tandem queues) and feedforward networks with deterministic routing.

The motivation behind our idea stems from the rich development of optimization as a scientific field during the second part of the 20th century. From its early years (Dantzig (1949)), 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 Ben-Tal et al. (2009), and Bertsimas et al. (2011a). 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 Bandi and Bertsimas (2013)).

The structure of the paper is as follows. In Section 2, we present our uncertainty set modeling assumptions and motivate their construction via the probabilistic limit laws. In Section 3, present our worst case as well as average case analysis for single multi-server queues. In Section 4, we extend our approach to a tandem system of queues. In Section 5, we discuss the advantages of our approach in obtaining insights and being computationally tractable. In Section 6, we extend the approach to analyze feed-forward networks with deterministic routing. Section 7 concludes the paper.

2. Proposed Framework

In the traditional probabilistic study of queues, the interarrival times $\mathbf{T} = \{T_1, \ldots, T_n\}$ and service times $\mathbf{X} = \{X_1, \ldots, X_n\}$ are modeled as renewal processes. In a first-come first-serve (FCFS) single-server queue, the system time is defined by Lindley (1952) as

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

In the event where the queue starts its operation with $n_0 \ge 0$ initial jobs (for which $T_1, \ldots, T_{n_0} = 0$), the system time recursion becomes

$$\begin{aligned} S_n &= \max\left\{ \max_{1 \le k \le n_0} \sum_{i=k}^{n} X_i - \sum_{i=n_0+1}^{n} T_i, \max_{n_0+1 \le k \le n} \left(\sum_{i=k}^{n} X_i - \sum_{i=k+1}^{n} T_i \right) \right\} \\ &= \max\left\{ \sum_{i=1}^{n} X_i - \sum_{i=n_0+1}^{n} T_i, \max_{n_0+1 \le k \le n} \left(\sum_{i=k}^{n} X_i - \sum_{i=k+1}^{n} T_i \right) \right\}. \end{aligned} \tag{2}$$

Analyzing the system time entails the understanding of the complex relationships between the random variables associated with the interarrival 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.

In this section, we propose to extend the framework introduced in Bandi et al. (2012) by modeling the uncertainty in the arrival and service processes via parameterized polyhedral sets, rather than assuming probability distributions. This framework substantially reduces the dimensionality of the uncertainty, which results in closed-form expressions of the worst-case behavior of queueing systems. Furthermore, we break new grounds by taking advantage of the uncertainty dimensionality reduction and obtain analytical expressions describing the average-case system behavior.

2.1. Uncertainty Modeling

In this paper, we consider the framework proposed by Bandi et al. (2012). In particular, we construct polyhedral uncertainty sets inspired by the generalized Central Limit Theorem (CLT) reproduced below in Theorem 1.

Theorem 1. **Generalized CLT (Samorodnitsky and Taqqu (1994))**

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

$$\frac{\sum_{i=1}^{n} Y_i - n\mu}{C_\alpha n^{1/\alpha}} \sim Y, \tag{3}$$

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

As in Bandi et al. (2012), we constrain the quantities T_i and X_i to take values while satisfying

$$\frac{\sum\limits_{i=k+1}^{n} T_i - \dfrac{n-k}{\lambda}}{(n-k)^{1/\alpha_a}} \geq -\Gamma_a, \quad \text{and} \quad \frac{\sum\limits_{i=k}^{n} X_i - \dfrac{n-k+1}{\mu}}{(n-k+1)^{1/\alpha_s}} \leq \Gamma_s, \quad \forall k = 1, \ldots, n,$$

for some parameters Γ_a and Γ_s that we use to control the degree of conservatism. Note that Γ_a and Γ_s are used to constrain the normalized partial sums for all values that the index k can take on. Motivated by the expression of the system time in initially nonempty queues, we propose to constrain the total sums in Eq. (2) by

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

for some parameters γ_a and γ_s. Note that, for $\gamma_a < \Gamma_a$ and $\gamma_s < \Gamma_s$, the above inequalities allow a tighter bound on performance measures in the case of initially nonempty queues. Consequently, we make the following modeling assumptions.

Assumption 1. *We make the following assumptions on the interarrival and service times.*

(a) *The interarrival times* $(T_{n_0+1},\ldots,T_n)$ *belong to the parametrized uncertainty set*

$$\mathcal{U}^a = \mathcal{U}^a\left(\gamma_a,\Gamma_a\right) = \left\{ (T_{n_0+1},\ldots,T_n) \middle| \left\{ \begin{array}{l} \displaystyle\sum_{i=n_0+1}^{n} T_i - \frac{n-n_0}{\lambda} \geq -\gamma_a (n-n_0)^{1/\alpha_a} \\ \displaystyle\sum_{i=k+1}^{n} T_i - \frac{n-k}{\lambda} \geq -\Gamma_a (n-k)^{1/\alpha_a}, \quad \forall\ n_0 \leq k \leq n \end{array} \right\} \right\},$$

where $1/\lambda$ *is the expected interarrival time,* n_0 *is the initial buffer in the queue,* $\gamma_a,\Gamma_a \in \mathbb{R}$ *are parameters that control the degree of conservatism, and* $1 < \alpha_a \leq 2$ *models possibly heavy-tailed probability distributions.*

(b) *For a single-server queue, the service times* $(X_1,\ldots,X_n)$ *belong to the uncertainty set*

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

where $1/\mu$ *is the expected service time,* $\gamma_s,\Gamma_s \in \mathbb{R}$ *are parameters that control the degree of conservatism, and* $1 < \alpha_s \leq 2$ *models possibly heavy-tailed probability distributions.*

(c) *For an m-server queue,* $m \geq 2$*, we let* ν *be a non-negative integer such that* $\nu = \lfloor n/m \rfloor$*, where* n *is the index corresponding to the* n^{th} *arriving job. We partition the job indices into sets* $K_i = \{k \leq n : \lfloor k/m \rfloor = i\}$*, for* $i = 0,1,\ldots,\nu$*, i.e.,*

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

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

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

Remark: Note that the uncertainty sets that we consider in this paper are subsets of the ones introduced in Bandi et al. (2012). Furthermore, we allow (γ_a,Γ_a), (γ_a,Γ_s) and (γ_m,Γ_m) to take both positive and negative values. When these parameters are positive, our uncertainty set allows the sums of inter arrival times to take values below the mean and the sums of service times to take values exceeding the mean, which yields positive waiting times. On the other hand, when these parameters are negative, our uncertainty set constrains the sums of the inter arrival times to take values exceeding the mean and the sum of the service times to take values below the mean, in which case, the system yields zero waiting time.

2.2. Performance Analysis Metrics

To understand the performance of queueing systems, we seek in particular an analytical characterization of the expected system time, given by

$$\overline{S}_n = \mathbb{E}_{\mathbf{T},\mathbf{X}}\left[S_n\left(\mathbf{T},\mathbf{X}\right)\right]. \tag{4}$$

The above expression is challenging to compute by modeling the primitives in a probabilistic queue via stochastic processes, due to the high dimensionality of the uncertainty and the complex relationships between the random variables associated with the interarrival and service times. Using our uncertainty modeling framework, we obtain an approximation of the expected system time by **(a)** computing the *worst case value* assuming the primitives satisfy Assumption 1, then **(b)** *averaging* the results with respect to the parameters (γ_a,Γ_a), (γ_s,Γ_s), and (γ_m,Γ_m). We present next the details of our approach.

Worst Case Behavior

To characterize the worst case behavior, we formulate the related performance analysis question as a robust optimization problem. In particular, we seek the worst case system time $\widehat{S}_n$ in queues satisfying Assumption 1. The worst case analysis can be cast as optimization problems of the form

$$\widehat{S}_n\left(\mathbf{T}\right) = \max_{\mathbf{X}\in\mathcal{U}^m} S_n \quad \text{and} \quad \widehat{S}_n = \max_{\mathbf{T}\in\mathcal{U}^a} \widehat{S}_n\left(\mathbf{T}\right), \tag{5}$$

which give rise to a closed form characterization of the worst case system time. To illustrate, Theorem 2 presents the result for an initially empty single-server queue with light tailed primitives.

Theorem 2. **(Worst Case System Time in an Initially Empty Single-Server Queue)** *In a single-server FCFS queue with $n_0 = 0$, $\mathbf{T} \in \mathcal{U}^a$, $\mathbf{X} \in \mathcal{U}^s$, $\alpha_a = \alpha_s = 2$ and $\rho < 1$, the worst case system time for the n^{th} job is given by*

$$\widehat{S}_n \le \max \begin{cases} (\Gamma_a + \Gamma_s)\sqrt{n} - \dfrac{1-\rho}{\lambda} n + \left(\dfrac{1}{\mu} + \Gamma_s\right), & \textit{if} \;\; n < \dfrac{\lambda^2 \left[(\Gamma_a + \Gamma_s)^+\right]^2}{4(1-\rho)^2} \\ \dfrac{\lambda}{4} \cdot \dfrac{\left[(\Gamma_a + \Gamma_s)^+\right]^2}{1-\rho} + \left(\dfrac{1}{\mu} + \Gamma_s\right), & \textit{otherwise,} \end{cases} \tag{6}$$

where the notation $(a)^+ = \max(0, a)$.

The evolution of the worst case system time is characterized by two distinct states: **(a)** a *transient state* where the system time is dependent on n with the system time in an initially empty queue increasing at an order of $n^{1/\alpha}$ when $\Gamma_a + \Gamma_s > 0$; and **(b)** a *steady state* where the system time is independent of n. When $\Gamma_a + \Gamma_s < 0$, jobs do not experience any waiting time, and therefore the system time is equal to the service time.

The characterization of the worst case waiting time bears qualitative similarity to the bounds established by Osogami and Raymond (2013) and Kingman (1970) for the transient and steady state system time in a $GI/GI/1$ queue, respectively,

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

where $e = \exp(1) = 2.718$.

Sections 3 and 4 present extensions of Theorem 2 to the cases of initially nonempty multi-server single and tandem queues with heavy-tailed arrivals and services. We next discuss how we leverage the worst case expressions that we obtain to predict the average system behavior.

Average Case Behavior

Instead of taking the expectation of the system time over the random variables $\mathbf{T}$ and $\mathbf{X}$ to analyze the average case behavior, we propose to treat the parameters (γ_a, Γ_a), (γ_s, Γ_s) and (γ_m, Γ_m) as random variables and compute

$$\widetilde{S}_n = \mathbb{E}\left[\widehat{S}_n\right]. \tag{7}$$

Philosophically, this approach distills all the probabilistic information contained in the random variables X_i's and T_i's into the the parameters $\{(\gamma_a,\Gamma_a),(\gamma_s,\Gamma_s)\}$ by allowing them to behave as random variables. Capturing the "randomness" of the arrival and service stochastic processes via only few random variables allows a significant dimensionality reduction of the uncertainty.

This framework therefore yields a tractable analysis of the expected transient waiting time by reducing the problem to solving a low-dimensional integral. To illustrate the averaging idea, consider again the case of an initially empty single-server queue with light tailed primitives. We express the bound on the worst case system time $\widehat{S}_n$ in Eq. (6) as

$$\widehat{S}^t_n(\Gamma_a,\Gamma_s)\cdot\mathbb{1}^t_n(\Gamma_a,\Gamma_s)+\widehat{S}^s(\Gamma_a,\Gamma_s)\cdot\mathbb{1}^s_n(\Gamma_a,\Gamma_s), \tag{8}$$

where the indicator functions $\mathbb{1}^t_n$ and $\mathbb{1}^s_n$ reflect the condition for the system to be in the transient state and the steady state, respectively. By treating Γ_a and Γ_s as random variables, we compute the expected value of the expression in Eq. (8) to obtain $\widetilde{S}_n$. This computation involves a double integration, which can be solved efficiently using numerical integration techniques.

With a careful choice of the distributions of the parameters $\{(\gamma_a,\Gamma_a),(\gamma_s,\Gamma_s)\}$, we show in Sections 3.2 and 4.2 that this averaging approach provides numerical outputs which match the simulated values accurately with most errors below 10%.

3. Analysis of a Single Queue

In this section, we study the worst case and average behavior of a single queue with a FCFS scheduling policy and a traffic intensity $\rho=\lambda/(m\mu)<1$, where m denotes the number of servers. We also assume that the queue begins operation with a job buffer $n_0\geq 0$.

3.1. Worst Case Behavior

We study the system time using the worst case approach as proposed in Bandi et al. (2012). We consider an m-server queueing system with n_0 initial jobs. 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 (Krivulin (1994))

$$C_n = \max\left(A_n, C_{(n-m)}\right) + X_n \quad \text{and} \quad S_n = C_n - A_n = \max\left(C_{(n-m)} - A_n, 0\right) + X_n, \tag{9}$$

where $A_n = \sum_{i=1}^{n} T_i$ denotes the 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 Bandi et al. (2012), 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 waiting time for the n^{th} job, for any $\mathbf{T}$. In particular, Bandi et al. (2012) consider the case of uncertainty sets where $\gamma_s \geq \Gamma_s \geq 0$ and show that the worst case system time is equal to the one achieved in a queue where no overtaking is allowed, i.e., where jobs leave the queue in the same order of their arrival, yielding

$$\widehat{S}_n(\mathbf{T}) = \max_{0 \leq k \leq \nu} \left(\max_{\mathcal{U}^m} \sum_{i=k}^{\nu} X_{r(i)} - \sum_{i=r(k)+1}^{n} T_i \right), \tag{10}$$

where $r(i) = n - (\nu - i)m \in K_i$, for all $\mathbf{T}$. We extend this result to the case where $\gamma_m \leq \Gamma_m$ and $\Gamma_m \geq 0$ and obtain Proposition 1, and the proof is presented in Appendix A3.

Proposition 1. *Given a sequence of inter-arrival times $\mathbf{T} = \{T_1, \ldots, T_n\}$, the worst case system time $\widehat{S}_n(\mathbf{T})$ in an FCFS queue modeled by $\mathcal{U}^a(\gamma_a, \Gamma_a)$, $\mathcal{U}^m(\gamma_m, \Gamma_m)$, where $\Gamma_m \geq 0$, is such that*

$$\widehat{S}_n(\mathbf{T}) = \max_{0 \leq k \leq \nu} \left(\max_{\mathcal{U}^s_m} \sum_{i=k}^{\nu} X_{r(i)} - \sum_{i=r(k)+1}^{n} T_i \right), \tag{11}$$

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

To handle the case of $\Gamma_m < 0$, we propose an upper bound on the worst case system time. We let $\Gamma_m^+ = \max(0, \Gamma_m)$, and since $\mathcal{U}^m(\Gamma_m) \subseteq \mathcal{U}^m(\Gamma_m^+)$,

$$\widehat{S}_n(\mathbf{T}) = \max_{\mathcal{U}^m(\Gamma_m)} S_n(\mathbf{T}, \mathbf{X}) \leq \max_{\mathcal{U}^m(\Gamma_m^+)} S_n(\mathbf{T}, \mathbf{X}) = \max_{0 \leq k \leq \nu} \left(\max_{\mathcal{U}^m(\Gamma_m^+)} \sum_{i=k}^{\nu} X_{r(i)} - \sum_{i=r(k)+1}^{n} T_i \right). \tag{12}$$

Initially Empty Queues

We apply Assumption 1 and the fact that $\gamma_m \le \Gamma_m$, to translate Eqs. (11) and (12) into solving the following one-dimensional nonlinear optimization problem

$$\widehat{S}_n \le \max_{0\le k\le\nu} \left\{ \frac{\nu-k+1}{\mu} + \Gamma_m^+ (\nu-k+1)^{1/\alpha_s} - \frac{m(\nu-k)}{\lambda} + \Gamma_a \left[m(\nu-k)\right]^{1/\alpha_a} \right\}. \tag{13}$$

This bound can be solved efficiently for the general case where $\alpha_s \neq \alpha_a$. Theorem 3 provides a closed form expression for the upper bound on the worst case system time in an initially empty queue for the special case where $\alpha_a = \alpha_s = \alpha$.

Theorem 3. **(Highest System Time in an Initially Empty Multi-Server Queue)** *In an initially empty m-server FCFS queue where* $\mathbf{T} \in \mathcal{U}^a$, $\mathbf{X} \in \mathcal{U}^m$ *where* $\Gamma = m^{1/\alpha}\Gamma_a + \Gamma_m^+ > 0$, $\alpha_a = \alpha_s = \alpha$ *and* $\rho < 1$, *the worst-case system time for all* $n \in K$ *given by*

$$\widehat{S}_n \le \begin{cases} \Gamma \cdot \nu^{1/\alpha} - \dfrac{m(1-\rho)}{\lambda} \cdot \nu + \left(\dfrac{1}{\mu} + \Gamma_m^+\right), & \text{if } \nu < \left(\dfrac{\lambda\Gamma/m}{\alpha(1-\rho)}\right)^{\alpha/(\alpha-1)} \\ \dfrac{\alpha-1}{\alpha^{\alpha/(\alpha-1)}} \cdot \dfrac{\lambda^{1/(\alpha-1)} \cdot \Gamma^{\alpha/(\alpha-1)}}{\left[m(1-\rho)\right]^{1/(\alpha-1)}}, & \text{otherwise.} \end{cases} \tag{14}$$

Proof of Theorem 3. Since $(\nu-k+1)^{1/\alpha_s} \le (\nu-k)^{1/\alpha_s} + 1$, and given $\Gamma_m^+ \ge 0$, we bound Eq. (13) by

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

which follows from the fact that $(\nu-k+1)^{1/\alpha_s} \le (\nu-k)^{1/\alpha_s} + 1$. By making the transformation $x = \nu - k$, where $x \in \mathbb{N}$, we can represent this maximization problem as

$$\max_{0\le x\le\nu, x\in\mathbb{N}} \left(\beta \cdot x^{1/\alpha} - \delta \cdot x\right) \le \max_{0\le x\le\nu, x\in\mathbb{R}} \left(\beta \cdot x^{1/\alpha} - \delta \cdot x\right), \tag{15}$$

where $\beta = m^{1/\alpha}\Gamma_a + \Gamma_m^+$ 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 $\widehat{S}_n = 1/\mu + \Gamma_m^+$. For $\beta > 0$, the function h is concave in x with an unconstrained maximizer

$$x^* = \left(\frac{\beta}{\alpha\delta}\right)^{\alpha/(\alpha-1)} = \left(\frac{\lambda(\Gamma_m + m^{1/\alpha}\Gamma_a)}{\alpha m(1-\rho)}\right)^{\alpha/(\alpha-1)}. \tag{16}$$

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,

$$\widehat{S}_n \leq \beta\left(\frac{\beta}{\alpha\delta}\right)^{1/(\alpha-1)} - \delta\left(\frac{\beta}{\alpha\delta}\right)^{\alpha/(\alpha-1)} + \left(\frac{1}{\mu} + \Gamma_m^+\right) = \frac{\alpha-1}{\alpha^{\alpha/(\alpha-1)}} \cdot \frac{\beta^{\alpha/(\alpha-1)}}{\delta^{1/(\alpha-1)}} + \left(\frac{1}{\mu} + \Gamma_m^+\right). \tag{17}$$

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

$$\widehat{S}_n = \beta(\nu)^{1/\alpha} - \delta(\nu) + \left(\frac{1}{\mu} + \Gamma_m^+\right). \tag{18}$$

We obtain Eq. (14) by substituting β and δ by their expressions in parts (a) and (b). □

Note that, for the case where $\Gamma = m^{1/\alpha}\Gamma_a + \Gamma_m^+ \leq 0$, the function in Eq. (14) is increasing in k over the interval $k \in [0,\nu]$, for $\rho = \lambda/\mu < 1$. It is therefore maximized at $k = \nu$, which yields

$$\widehat{S}_n = \widehat{X}_n \leq \frac{1}{\mu} + \Gamma_m^+.$$

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 $\Gamma \leq 0$ involves typically long inter arrival times and short service times.

Initially Nonempty Queues

We next analyze the case where $n_0 > 0$ with $T_i = 0$ for all $i = 1,\ldots,n_0$. The first m jobs in the queue are routed immediately to the servers without any delays. We are interested in the behavior for $n > m$. For $\Gamma_m \geq 0$ and assuming $n_0 \in K_\phi$, i.e., $\phi = \lfloor n_0/m \rfloor$, we can rewrite Eq. (11) as

$$\textbf{(a)}\ \text{for } n \leq n_0: \quad \widehat{S}_n = \max_{\mathbf{X}\in\mathcal{U}^m}\left(\max_{0\leq k\leq\nu\leq\phi} \sum_{i=k}^{\nu} X_{r(i)}\right) = \max_{\mathbf{X}\in\mathcal{U}^m}\left(\sum_{i=0}^{\nu} X_{r(i)}\right) \tag{19}$$

$$\textbf{(b)}\ \text{for } n > n_0: \quad \widehat{S}_n = \max\left\{\begin{array}{l} \max\limits_{0\leq k\leq\phi}\left(\max\limits_{\mathbf{X}\in\mathcal{U}^m} \sum\limits_{i=k}^{\nu} X_{r(i)}\right) - \max\limits_{\mathbf{T}\in\mathcal{U}^a} \sum\limits_{i=n_0+1}^{n} T_i, \\ \max\limits_{\phi<k\leq\nu}\left(\max\limits_{\mathbf{X}\in\mathcal{U}^m} \sum\limits_{i=k}^{\nu} X_{r(i)} - \max\limits_{\mathbf{T}\in\mathcal{U}^a} \sum\limits_{i=r(k)+1}^{n} T_i\right) \end{array}\right\}. \tag{20}$$

Similarly to the empty case, the worst case system time involves one-dimensional nonlinear optimization problems which can be solved efficiently. In particular, for $n \leq n_0$, we have

$$\widehat{S}_n\left(\Gamma_m\right) = \max_{\mathbf{X}\in\mathcal{U}^m}\left(\sum_{i=0}^{\nu} X_{r(i)}\right) = \left(\frac{\nu+1}{\mu} + \gamma_m(\nu+1)^{1/\alpha_s}\right)^+, \tag{21}$$

where $r = r(0) = n - \nu m$ and $\nu = \lfloor n/m \rfloor$. For $n > n_0$, we have

$$\widehat{S}_n\left(\Gamma_a,\Gamma_m\right) \leq \max\left\{\begin{array}{l} \left(\dfrac{\nu-k+1}{\mu} + \gamma_m\left(\nu-k+1\right)^{1/\alpha_s}\right)^+ - \dfrac{n-n_0}{\lambda} + \gamma_a\left(n-n_0\right)^{1/\alpha_a}, \\ \max\limits_{\phi<k\leq\nu}\left(\dfrac{\nu-k+1}{\mu} + \Gamma_m^+\left(\nu-k+1\right)^{1/\alpha_s} - \dfrac{m(\nu-k)}{\lambda} + \Gamma_a\left[m(\nu-k)\right]^{1/\alpha_a}\right) \end{array}\right\}, \tag{22}$$

As for initially empty queues, the optimization problem in Eq. (22) can be solved efficiently for the general case where $\alpha_a \neq \alpha_s$. Theorem 4 provides a closed form expression for the upper bound on the worst case system time for the special case where $\alpha_a = \alpha_s = \alpha$.

Theorem 4. **(Highest System Time in an Initially Nonempty Multi-Server Queue)** *In an m-server FCFS queue with $n_0 \in K_\phi$ and $\mathbf{T}\in\mathcal{U}^a$, $\mathbf{X}\in\mathcal{U}^m$ such that $\Gamma = m^{1/\alpha} + \Gamma_m^+ > 0$, $\alpha_a = \alpha_s = \alpha$ and $\rho < 1$, the worst case system time for $n_0 < n \in K_\nu$ is given by*

$$\widehat{S}_n \leq \max\left\{\begin{array}{l} \left(\dfrac{\nu+1}{\mu} + \gamma_m\left(\nu+1\right)^{1/\alpha_s}\right)^+ - \dfrac{n-n_0}{\lambda} + \gamma_a\left(n-n_0\right)^{1/\alpha_a}, \\ \left\{\begin{array}{ll} \Gamma\left(\nu-\phi\right)^{1/\alpha} - \dfrac{m(1-\rho)}{\lambda}\left(\nu-\phi\right) + \left(\dfrac{1}{\mu} + \Gamma_m^+\right), & \textit{if } \nu-\phi < \left(\dfrac{\lambda\Gamma/m}{\alpha(1-\rho)}\right)^{\alpha/(\alpha-1)} \\ \dfrac{\alpha-1}{\alpha^{\alpha/(\alpha-1)}}\dfrac{\lambda^{1/(\alpha-1)}\cdot\Gamma^{\alpha/(\alpha-1)}}{\left[m(1-\rho)\right]^{1/(\alpha-1)}} + \left(\dfrac{1}{\mu} + \Gamma_m^+\right), & \textit{otherwise.} \end{array}\right. \end{array}\right\}. \tag{23}$$

Proof of Theorem 4. To solve bound the maximization problem in Eq. (22), we take a similar approach to that presented in the proof of Theorem 3 and cast the problem in the form

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

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. □

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

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

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 $1/\mu + \Gamma_m^+$.

3.2. Average Case Behavior

To analyze the average behavior of a queueing system, we treat the parameters (γ_a, Γ_a) and (γ_m, Γ_m) as random variables and compute the expected value of the worst case system time $\widehat{S}_n$. For ease of notation, we express the worst case system time in Eq. (23) as

$$\widehat{S}_n \leq \max\left\{\widehat{S}_n^b(\gamma_a,\gamma_s),\ \begin{cases} \widehat{S}_n^t(\Gamma_a,\Gamma_m), & \text{if } \left\lfloor \frac{n}{m} \right\rfloor - \left\lfloor \frac{n_0}{m} \right\rfloor < \left[\frac{\lambda\left(m^{1/\alpha}\Gamma_a + \Gamma_s^+\right)^+}{\alpha m(1-\rho)}\right]^{\alpha/(\alpha-1)} \\ \widehat{S}^s(\Gamma_a,\Gamma_m), & \text{otherwise} \end{cases}\right\}, \tag{24}$$

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. We would like to rewrite the above upper bound on the worst case system time as

$$\max\left\{\widehat{S}_n^b(\gamma_a,\gamma_s),\ \widehat{S}_n^t(\Gamma_a,\Gamma_m)\cdot \mathbb{1}_n^t(\Gamma_a,\Gamma_m) + \widehat{S}^s(\Gamma_a,\Gamma_m)\cdot \mathbb{1}_n^s(\Gamma_a,\Gamma_m)\right\},$$

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. Specifically, by writing the condition in Eq. (24) as an interval over Γ_a and Γ_m, we obtain

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

By positing some assumptions on the distributions of the parameters (γ_a, Γ_a) and (γ_m, Γ_m), we compute $\widehat{S}_n$ via numerical integration. We next discuss our choice of the distributions of the parameters $\{(\gamma_a, \Gamma_a), (\gamma_s, \Gamma_s)\}$ inspired by the limit laws of probability.

Choice of Variability Distribution

From Assumption 1, 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,

$$\gamma_a = -\left(\frac{\sum_{i=n_0+1}^{n} T_i - \dfrac{n-n_0}{\lambda}}{(n-n_0)^{1/\alpha}} \right) \propto -Z_a \quad \text{and} \quad \gamma_s = \left(\frac{\sum_{i=1}^{n} X_i - \dfrac{n}{\mu}}{n^{1/\alpha}} \right) \propto Z_s, \tag{25}$$

which approximately behave as a random variable following a limiting distribution. In a multi-server queue, and assuming without loss of generality that $n = \nu m$, we obtain

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

where the last inequality is due to Assumption 1(c). We can therefore express γ_m as

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

(a) Light-Tailed Primitives: For light tails, γ_a and γ_s obey the normal distribution, i.e.,

$$\gamma_a \sim \mathcal{N}(0, \sigma_a) \quad \text{and} \quad \gamma_s \sim \mathcal{N}(0, \sigma_s),$$

where σ_a and σ_s denote the standard deviations associated with the inter-arrival and service processes, respectively.

(b) Heavy-Tailed Primitives: 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. (25) are such that

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

where C_α is a normalizing constant as introduced in Eq. (3). As a concrete example, for Pareto distributed inter arrivals and service times,

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

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 $Z_a = S_\alpha(1, C_\alpha, 0)$.

The characterization of the exact distribution of the parameters $(\Gamma_a, \Gamma_s, \Gamma_m)$ is however challenging. Instead, we propose a simplification where we express Γ_a, Γ_s and Γ_m as linear functions of γ_a, γ_s and γ_m, respectively. Specifically, we let

$$\Gamma_a = \theta_a \gamma_a, \;\; \Gamma_s = \theta_s \gamma_s, \;\; \text{and } \Gamma_m = \theta_m \gamma_m,$$

where $(\theta_a, \theta_s, \theta_m)$ are some scalars which we select as follows.

(a) Light-Tailed Primitives: We choose $(\theta_a, \theta_s, \theta_m)$ are scalars chosen so that the average worst case steady-state system time matches the bound provided by Kingman (1970), which is particularly tight in heavy traffic. In other words, we ensure that $\widetilde{S}_n = \overline{S}_n$ for a large enough value of n. To illustrate, for a multi-server queue, we ensure that

$$\frac{\lambda}{4(1-\rho)} \cdot \mathbb{E}\left[\left[\left(\theta_a \gamma_a + \theta_m \gamma_m^+ / m^{1/2}\right)^+\right]^2\right] = \frac{\lambda}{2(1-\rho)} \cdot \left(\sigma_a^2 + \sigma_s^2/m^2\right). \tag{26}$$

Let $\gamma = \theta_a \gamma_a + \theta_m \gamma_m^+ / m^{1/2} = \theta_a \gamma_a + \theta_m \gamma_s^+ / m$. We approximate the expectation in Eq. (26) by

$$\mathbb{E}\left[(\gamma^+)^2\right] \approx \mathbb{P}(\gamma \geq 0) \cdot \mathbb{E}\left[\gamma^2\right] = \mathbb{P}(\gamma \geq 0) \cdot \left(\theta_a^2 \sigma_a^2 + \mathbb{P}(\gamma_s \geq 0) \cdot \theta_m^2 \sigma_s^2 / m^2\right)$$

By pattern matching the two expressions in Eq. (26), the parameters θ_a and θ_m are such that

$$\theta_a \approx \left(\frac{2}{\mathbb{P}(\gamma \geq 0)}\right)^{1/2} \quad \text{and} \quad \theta_m \approx \left(\frac{2}{\mathbb{P}(\gamma \geq 0) \cdot \mathbb{P}(\gamma_s \geq 0)}\right)^{1/2}. \tag{27}$$

Note that, given that γ_s is a normally distributed distributed random variable centered around the origin, we have $\mathbb{P}(\gamma_s \geq 0) = 1/2$. Also, $\mathbb{P}(\gamma \geq 0)$ can be efficiently computed numerically, with

$$\mathbb{P}(\gamma \geq 0) = \mathbb{P}(\theta_a \gamma_a + \theta_m \gamma_s^+ / m \geq 0) = \mathbb{P}\left(\gamma_a + 2^{1/2} \cdot \gamma_s^+ / m \geq 0\right).$$

(b) The steady state in heavy-tailed queues does not exist, which does not allow us to use a similar matching procedure to the one introduced for light-tailed queues. Instead, we propose to extend the formulas in Eq. (27) to obtain

$$\theta_a \approx \left(\frac{\alpha}{\mathbb{P}(\gamma \geq 0)}\right)^{(\alpha-1)/\alpha} \quad \text{and} \quad \theta_m \approx \left(\frac{\alpha}{\mathbb{P}(\gamma \geq 0) \cdot \mathbb{P}(\gamma_s \geq 0)}\right)^{(\alpha-1)/\alpha}. \tag{28}$$

where the probabilities can be efficiently computed numerically. Specifically,

$$\mathbb{P}(\gamma \geq 0) = \mathbb{P}\left(\theta_a \gamma_a + \theta_m \gamma_s^+ / m \geq 0\right) = \mathbb{P}\left(\gamma_a + \mathbb{P}(\gamma_s \geq 0)^{-(\alpha-1)/\alpha} \cdot \gamma_s^+ / m \geq 0\right).$$

We note that, by expressing Γ_a and Γ_m in terms of γ_a and γ_s, the average system time $\widetilde{S}_n$ can be computed by taking expectations with respect to γ_a and γ_s

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

The above double integral can be efficiently computed using numerical techniques. We next compare the performance of the proposed averaging technique with simulated values.

Computational Results

We investigate the performance of our approach relative to simulation and examine the effect of the system's parameters (traffic intensity, tail coefficient, initial buffer and number of servers) on its accuracy. Specifically, we run simulations for single and multi-server queues with normally and Pareto distributed inter arrival and service times.

We present the average percent error between simulated expected values $\overline{S}_n$ and the predictions $\widetilde{S}_n$. In particular, we report the following quantity

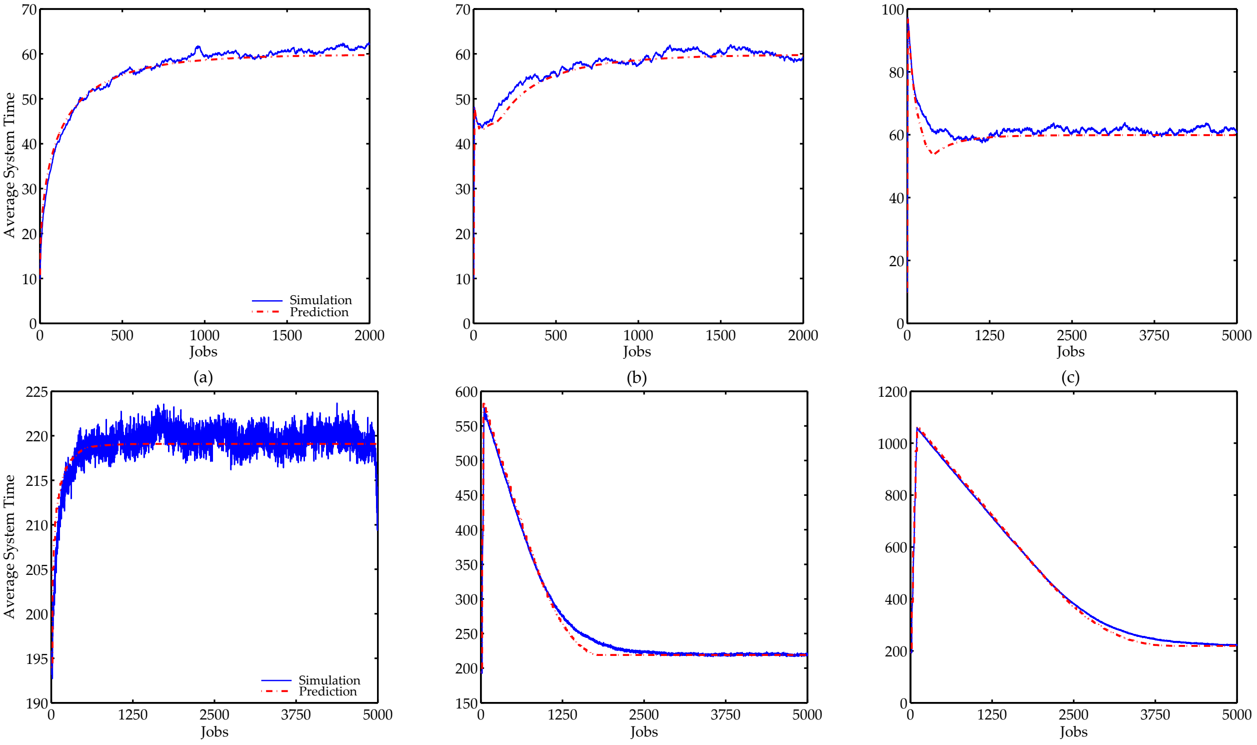
$$\text{Average Percent Error} = \frac{1}{N-1} \cdot \sum_{n=2}^{N} \left| \frac{\overline{S}_n - \widetilde{S}_n}{\overline{S}_n} \right| \times 100\%.$$

The term N corresponds to either the number of jobs observed until relaxation is reached (observed from simulations) or the maximum number of jobs for which the simulations was run ($N = 5,000$ for both light-tailed and heavy-tailed distributions). Tables 1 and 2 report the errors for multi-server queues with normally and Pareto distributed inter arrival and service times, respectively.

Figure 1 compares our approximation (dotted line) with simulation (solid line) for queues with normally distributed inter arrival and service times with $m = 1$ (top panels) and $m = 20$ (bottom panels). Figure 2 presents a graphical snapshot of our approximation (dotted line) in comparison to simulation (solid line) for queues with Pareto distributed primitives with $m = 1$ (top panels) and $m = 20$ (bottom panels).

Table 1 Errors relative to simulations for multi-server queues with normally distributed primitives.

		1 Server			10 Servers			20 Servers		
	ρ	$n_0 = 0$	$n_0 = 5$	$n_0 = 20$	$n_0 = 0$	$n_0 = 20$	$n_0 = 50$	$n_0 = 0$	$n_0 = 50$	$n_0 = 100$
$\sigma_a = \sigma_s = 2.5$	0.95	5.25	3.51	5.46	0.90	3.09	2.21	0.92	1.51	1.07
	0.96	5.22	3.28	6.85	0.60	2.45	2.11	0.75	1.53	1.03
	0.97	4.02	2.26	5.29	0.43	2.83	2.28	0.60	1.92	1.10
	0.98	3.51	1.95	7.09	0.76	3.38	3.73	0.34	2.57	1.65
	0.99	3.54	1.54	8.77	1.90	4.90	2.73	0.98	2.89	0.62
$\sigma_a = \sigma_s = 4.0$	0.95	1.81	3.52	8.91	0.75	3.78	3.52	1.20	2.21	1.97
	0.96	2.32	4.40	9.41	1.07	3.92	4.32	0.95	2.44	2.50
	0.97	1.75	2.16	9.74	1.14	4.36	5.34	0.62	2.73	3.42
	0.98	3.69	5.13	7.25	3.31	6.90	8.94	0.89	4.00	2.52
	0.99	5.05	4.09	8.51	4.47	7.74	5.09	2.83	5.08	1.50



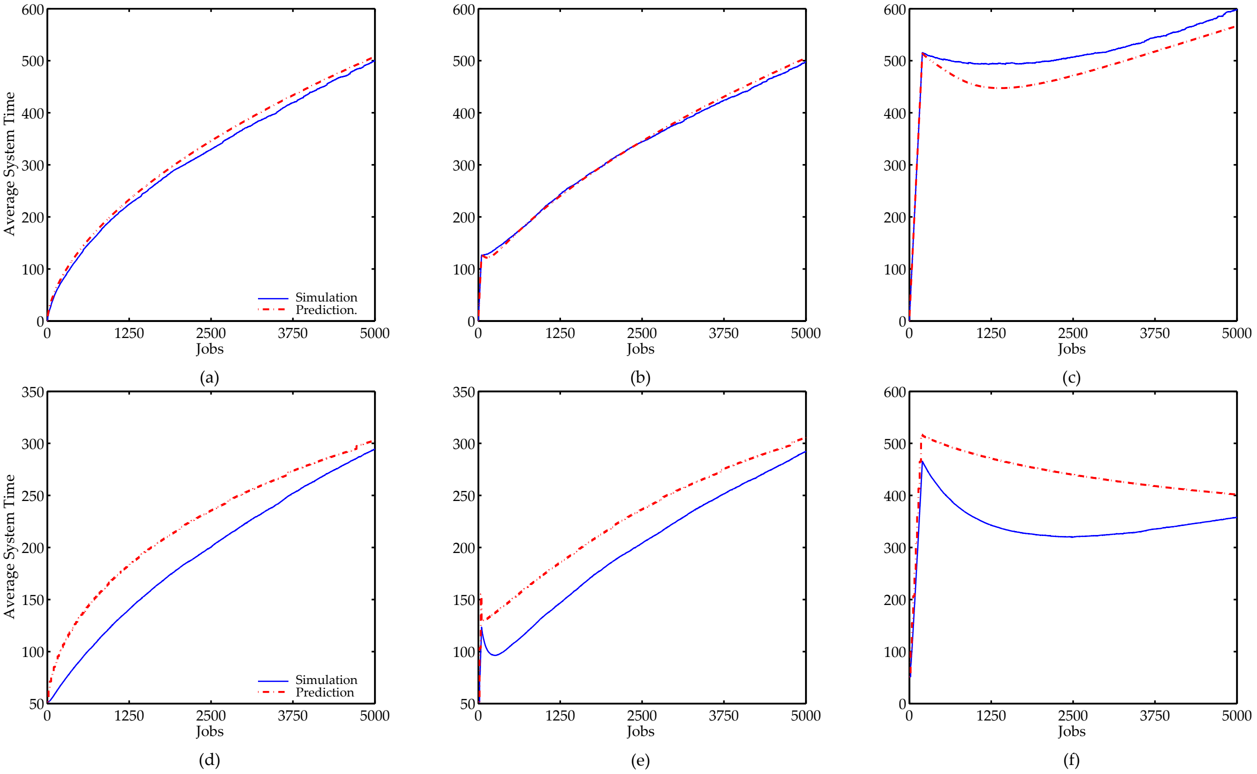
(a) (b) (c)

(d) (e) (f)

Figure 1 Simulated (solid line) versus predicted values (dotted line) for a single queue with normally distributed primitives ($\sigma_a = \sigma_s = 4.0$) and $\rho = 0.97$. Panels (a)–(c) correspond to an instance with $m = 1$ and $n_0 = 0, 5, 10$. Panels (d)–(f) correspond to an instance with $\rho = 0.99$ and $n_0 = 0, 50, 100$.

Table 2 Errors relative to simulations for multi-server queues with Pareto distributed primitives.

		1 Server			10 Servers			20 Servers		
	ρ	$n_0 = 0$	$n_0 = 50$	$n_0 = 200$	$n_0 = 0$	$n_0 = 50$	$n_0 = 200$	$n_0 = 0$	$n_0 = 50$	$n_0 = 200$
$\alpha_a = \alpha_s = 1.6$	0.95	9.59	7.18	1.78	12.5	9.49	13.9	17.9	15.9	25.5
	0.96	6.39	2.62	5.36	12.4	9.52	13.1	18.5	16.5	27.3
	0.97	4.86	1.49	5.98	12.1	9.56	13.7	19.6	17.8	28.6
	0.98	3.39	1.13	6.40	11.7	10.0	14.9	21.8	19.9	29.2
	0.99	2.59	2.08	6.63	11.9	11.9	15.6	24.5	22.6	29.3
$\alpha_a = \alpha_s = 1.7$	0.95	9.59	7.18	1.78	9.22	7.85	5.44	21.6	18.5	17.4
	0.96	10.4	4.69	2.24	12.4	9.04	9.68	21.3	17.6	18.5
	0.97	8.75	3.14	2.92	12.7	9.63	9.76	21.7	17.7	19.8
	0.98	7.18	1.94	3.39	13.0	11.2	10.8	22.8	18.7	20.4
	0.99	5.72	1.17	3.66	13.9	13.5	11.4	24.4	20.3	20.4



(a) (b) (c)

(d) (e) (f)

Figure 2 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 $n_0 = 0, 50, 200$.

Our approach generally yields errors within 10% for multi-server queues with normally distributed inter arrival and service times. Errors for heavy-tailed multi-server queues seem to increase with the number of servers, with magnitudes within 15% for 10-server queues and 30% for 20-server queues. However, we still capture the general behavior of the system time as shown in Figure 2.

Furthermore, as shown by simulations and empirical studies performed by Odoni and Roth (1983) 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 1(a),(d).

(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 1(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 1(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 1(e),(f).

4. Analysis of Queues in Series

In this section, we extend our analysis of single queues to the analysis of tandem queues. Consider a network of J queues in series and let $\mathbf{X}^{(j)} = \left\{X_1^{(j)}, \ldots, X_n^{(j)}\right\}$ denote the service times at the j^{th} queue. We make the following assumptions.

Assumption 2. *We make the following assumptions for the service times at the j^{th} queue in a tandem network. Let $1/\mu_j$ be the expected service time, $\Gamma_s^{(j)} \in \mathbb{R}$ a parameter that controls the degree* *of conservatism, and* $1 < \alpha_s^{(j)} \leq 2$ *modeling possibly heavy-tailed probability distributions.*

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

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

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

$$\mathcal{U}_j^m = \left\{ \left(X_1^{(j)}, \ldots, X_n^{(j)} \right) \middle| \sum_{i \in \mathcal{I}} X_{k_i}^{(j)} - \frac{|\mathcal{I}|}{\mu_j} \leq \Gamma_m^{(j)} |\mathcal{I}|^{1/\alpha_s^{(j)}}, \ \forall \ k_i \in K_i, \ \textit{and} \ \ i \in \mathcal{I} \subseteq \{0, \ldots, \nu\} \right\}.$$

We further assume that the inter arrival times $\mathbf{T} = (T_1, \ldots, T_n)$ to the tandem network satisfy the uncertainty set $\mathcal{U}^a$, as described in Assumption 1.

We consider, for the purpose of the discussion, a tandem network with J single-server queues. The system time of the n^{th} job at the j^{th} queue is such that

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

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 can be cast as

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

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 $\mathbf{X}^{(j-1)}$. For an isolated queue, Bandi et al. (2012) show that the interdeparture times belong to the inter-arrival uncertainty set $\mathcal{U}^a$. However, this characterization is only tight under steady-state conditions. Obtaining an exact transient characterization of the inter-departure process is challenging. Instead of going this route, we propose to use the recursive formulas that define the dynamics in a network of queues in series to study the overall system time

$$S_n = S_n^{(1)} + \ldots S_n^{(J)}.$$

Bertsimas et al. (2011b) obtain an exact characterization of the system time in a tandem network of single-server queues where

$$S_n = \max_{1 \leq k_1 \leq \ldots \leq k_J \leq n} \left(\sum_{i=k_1}^{k_2} X_i^{(1)} + \sum_{i=k_2}^{k_3} X_i^{(2)} + \ldots + \sum_{i=k_J}^{n} X_i^{(J)} - \sum_{i=k_1+1}^{n} T_i \right). \tag{29}$$

Similarly to the analysis of a single queue, we propose an analysis of the worst case overall system time under Assumption 2 which provides closed form bounds. We then leverage the analytic expressions of the worst case system time to understand the average behavior of tandem queueing networks with multiple servers.

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

$$\widehat{S}_n(\mathbf{T}) = \max_{1\le k_1\le\ldots\le k_J\le n}\left(\max_{\mathcal{U}_1^s}\sum_{i=k_1}^{k_2} X_i^{(1)} + \max_{\mathcal{U}_2^s}\sum_{i=k_2}^{k_3} X_i^{(2)} + \ldots + \max_{\mathcal{U}_J^s}\sum_{i=k_J}^{n} X_i^{(J)} - \sum_{i=k_1+1}^{n} T_i\right). \tag{30}$$

Theorem 5 provides a similar result for multi-server queues in series, under the assumption that each queue acts adversarially in view of maximizing its system time, for all $\mathbf{T}$.

Theorem 5. **(Worst Case System Time in a Tandem Queue with Multiple Servers)** *In a network of J multi-server queues in series with inter arrival times $\mathbf{T} = \{T_1,\ldots,T_n\}$, the overall system time of the n^{th} job is given by*

$$\widehat{S}_n(\mathbf{T}) = \max_{0\le k_1\le\ldots\le k_J\le \nu}\left(\max_{\mathcal{U}_1^s}\sum_{i=k_1}^{k_2} X_{r(i)}^{(1)} + \max_{\mathcal{U}_2^s}\sum_{i=k_2}^{k_3} X_{r(i)}^{(2)} + \ldots + \max_{\mathcal{U}_J^s}\sum_{i=k_J}^{n} X_{r(i)}^{(J)} - \sum_{i=r(k_1)+1}^{n} T_i\right), \tag{31}$$

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

The proof is similar to the proof presented in Bandi et al. (2012), and presented in Appendix A1. By minimizing the partial sum of the interarrival times, we obtain an exact characterization of the worst case system time in a tandem queue as

$$\widehat{S}_n = \max_{0\le k_1\le\ldots\le k_J\le \nu}\left(\max_{\mathcal{U}_1^s}\sum_{i=k_1}^{k_2} X_{r(i)}^{(1)} + \ldots + \max_{\mathcal{U}_J^s}\sum_{i=k_J}^{n} X_{r(i)}^{(J)} - \max_{\mathbf{T}\in\mathcal{U}^a}\sum_{i=r(k_1)+1}^{n} T_i\right). \tag{32}$$

Initially Empty Queues in Tandem

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

$$\widehat{S}_n\left(\Gamma_a,\Gamma_s^{(1)},\ldots,\Gamma_s^{(J)}\right)\le \max_{0\le k_1\le\ldots\le k_J\le\nu}\left\{\begin{array}{c}\displaystyle\sum_{j=1}^{J}\frac{k_{j+1}-k_j+1}{\mu_j}+\Gamma_m^{(j)}\left(k_{j+1}-k_j+1\right)^{1/\alpha_s^{(j)}}\\[2ex] \displaystyle-\frac{m(\nu-k_1)}{\lambda}+\Gamma_a\left[m\left(\nu-k_1\right)\right]^{1/\alpha_a}\end{array}\right\}, \tag{33}$$

which involves a J-dimensional nonlinear optimization problem. Theorem 6 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 6. **(Highest System Time in an Initially Empty Tandem Queue)** *In an initially empty network of J multi-server queues in series with $\mathbf{T}\in\mathcal{U}^a$, $\mathbf{X}^{(j)}\in\mathcal{U}_j^s$, for all $j=1,\ldots,J$, $\alpha_a=\alpha_s^{(1)}=\ldots=\alpha_s^{(J)}=\alpha$, and $\Gamma=m^{1/\alpha}\Gamma_a+\Gamma_m>0$, where*

$$\Gamma_m=\left(\sum_{j=1}^{J}(\Gamma_m^{(j)+})^{\alpha/\alpha-1}\right)^{\alpha-1/\alpha}, \tag{34}$$

the worst-case system time for $\mu_1=\ldots=\mu_J$ and $\rho<1$ is given by

$$\widehat{S}_n\le\begin{cases}\Gamma\cdot\nu^{1/\alpha}-\dfrac{m(1-\rho)}{\lambda}\nu+\dfrac{J}{\mu}+\displaystyle\sum_{i=1}^{J}\Gamma_m^{(i)+}, & \textit{if}\;\; \nu+J\le\left[\dfrac{\lambda\Gamma}{\alpha m(1-\rho)}\right]^{\alpha/(\alpha-1)}\\[3ex] \dfrac{\alpha-1}{\alpha^{\alpha/(\alpha-1)}}\cdot\dfrac{\lambda^{1/(\alpha-1)}\cdot\Gamma^{\alpha/(\alpha-1)}}{\left[m(1-\rho)\right]^{1/(\alpha-1)}}+\dfrac{J}{\mu}+\displaystyle\sum_{i=1}^{J}\Gamma_m^{(i)+}, & \textit{otherwise.}\end{cases} \tag{35}$$

Proof of Theorem 6. From Eq. (33), we have that the worst case system time is given by

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

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

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

We will isolate the problem of maximizing $\left[\Gamma_m^{(1)+}\left(k_2-k_1\right)^{1/\alpha}+\ldots+\Gamma_m^{(J)+}\left(\nu-k_J\right)^{1/\alpha}\right]$ 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

$$\max_{0 \le k_1 \le \nu, k_1 \in \mathbb{N}} \left(m^{1/\alpha} \Gamma_a \left(\nu - k_1\right)^{1/\alpha} - \frac{m(1-\rho)}{\lambda} \left(\nu - k_1\right) + \left\{ \begin{array}{ll} \max & \left[\Gamma_m^{(1)+} x_1^{1/\alpha} + \ldots + \Gamma_m^{(J)+} x_J^{1/\alpha} \right] \\ \text{s.t.} & x_1 + \ldots + x_J = \nu - k_1 \\ & x_j \in \mathbb{N}, \forall j = 2, \ldots, J \end{array} \right\} \right). \tag{36}$$

It is easy to see, based on first order optimality conditions (see Appendix B), that the optimal solution to the inner optimization problem satisfies

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

Using the additional condition that $\sum_{j=1}^{J} x_j^* = 1$, we obtain

$$x_k^* = \frac{(\nu - k_1)(\Gamma_m^{(k)+})^{\alpha/(\alpha-1)}}{\sum\limits_{j=1}^{J} (\Gamma_m^{(j)+})^{\alpha/(\alpha-1)}} \quad \forall k = 1, 2, \ldots, J,$$

leading to an optimal value of

$$\Gamma_m^{(1)+} (x_1^*)^{1/\alpha} + \ldots + \Gamma_m^{(J)+} (x_1^*)^{1/\alpha} = \left(\sum_{j=1}^{J} (\Gamma_m^{(j)+})^{\alpha/(\alpha-1)} \right)^{(\alpha-1)/\alpha}.$$

Substituting the optimal solution of the inner problem in Eq. (36), the performance analysis reduces to solving the following one-dimensional optimization problem

$$\max_{0 \le k_1 \le \nu} \left\{ \left(m^{1/\alpha} \Gamma_a + \left[\sum_{j=1}^{J} (\Gamma_m^{(j)+})^{\alpha/(\alpha-1)} \right]^{(\alpha-1)/\alpha} \right) \cdot \left(\nu - k_1\right)^{1/\alpha} - \frac{m(1-\rho)}{\lambda} \left(\nu - k_1\right) \right\}, \tag{37}$$

which can be cast in the form of the optimization problem in Eq. (15), 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 3, the solution to Eq. (37) is

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

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

The case where $\Gamma = m^{1/\alpha}\Gamma_a + \Gamma_m \leq 0$ arises when $\Gamma_a < 0$, since $\Gamma_m > 0$ as defined in Eq. (34). This scenario is characterized by long inter-arrival times yielding zero waiting times. The worst case system time therefore reduces to

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

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 $n_0 > 0$. 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. (32) as

$$\textbf{(a)} \text{ for } n \leq n_0 : \widehat{S}_n = \max_{0 \leq k_1 \leq \ldots \leq k_J \leq \nu \leq \gamma} \left(\max_{\mathcal{U}_1^s} \sum_{i=k_1}^{k_2} X_{r(i)}^{(1)} + \ldots + \max_{\mathcal{U}_J^s} \sum_{i=k_J}^{n} X_{r(i)}^{(J)} \right) \tag{38}$$

$$\textbf{(b)} \text{ for } n > n_0 : \widehat{S}_n = \max \left\{ \begin{array}{c} \max\limits_{\substack{0 \leq k_1 \leq \ldots \leq k_J \leq \nu \\ k_1 \leq \gamma}} \left(\max\limits_{\mathcal{U}_1^s} \sum\limits_{i=k_1}^{k_2} X_{r(i)}^{(1)} + \ldots + \max\limits_{\mathcal{U}_J^s} \sum\limits_{i=k_J}^{n} X_{r(i)}^{(J)} \right) - \max\limits_{\mathbf{T} \in \mathcal{U}^a} \sum\limits_{i=n_0+1}^{n} T_i, \\ \\ \max\limits_{\gamma < k_1 \leq \ldots \leq k_J \leq \nu} \left(\max\limits_{\mathcal{U}_1^s} \sum\limits_{i=k_1}^{k_2} X_{r(i)}^{(1)} + \ldots + \max\limits_{\mathcal{U}_J^s} \sum\limits_{i=k_J}^{n} X_{r(i)}^{(J)} - \max\limits_{\mathbf{T} \in \mathcal{U}^a} \sum\limits_{i=r(k_1)+1}^{n} T_i \right) \end{array} \right\}. \tag{39}$$

By Assumption 1, the worst case system time involves solving J-dimensional nonlinear optimization problems. Theorem 7 provides a closed form bound on the worst case system time in an initially nonempty 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 7. **(Highest Waiting Time in an Initially Nonempty Tandem Queue)** *In an initially nonempty network of J multi-server queues in series with $n_0 > m$, $\mathbf{T} \in \mathcal{U}^a$, $\mathbf{X}^{(j)} \in \mathcal{U}_j^s$, for all $j = 1, \ldots, J$, $\alpha_a = \alpha_s^{(1)} = \ldots = \alpha_s^{(J)} = \alpha$, and $\Gamma = m^{1/\alpha}\Gamma_a + \Gamma_m > 0$, where Γ_m is defined in Eq. (34), the worst-case system time for $\mu_1 = \ldots = \mu_J$ and $\rho < 1$ is given by*

$$\widehat{S}_n \leq \max \left\{ \begin{array}{l} \dfrac{\nu+J}{\mu} + \Gamma_m \cdot (\nu+J)^{1/\alpha} - \dfrac{(n-n_0)^+}{\lambda} + \gamma_a \left[(n-n_0)^+ \right]^{1/\alpha}, \\ \left\{ \begin{array}{ll} \Gamma \left[(\nu-\phi)^+ \right]^{1/\alpha} - \dfrac{m(1-\rho)}{\lambda} (\nu-\phi)^+, & \textit{if} \;\; (\nu-\phi)^+ < \left[\dfrac{\lambda \Gamma/m}{\alpha(1-\rho)} \right]^{\alpha/(\alpha-1)} \\ \dfrac{\alpha-1}{\alpha^{\alpha/(\alpha-1)}} \dfrac{\lambda^{1/(\alpha-1)} \cdot \Gamma^{\alpha/(\alpha-1)}}{\left[m(1-\rho) \right]^{1/(\alpha-1)}}, & \textit{otherwise} \end{array} \right\} \end{array} \right\}. \tag{40}$$

The proof is similar to the proof of Theorem 6, and presented in Appendix A2. Note that, for the case where $\Gamma = m^{1/\alpha}\Gamma_a + \Gamma_m \leq 0$, the worst case system time

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

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 sum of the service times.

4.2. Average Case Behavior

To analyze the average behavior of a queueing system, we treat the parameters (γ_a, Γ_a) and $(\gamma_m^{(j)}, \Gamma_m^{(j)})$ as random variables and compute $\widetilde{S}_n$, the expected value of the worst case system time. Similarly to the case of a single queue, we express $\widetilde{S}_n$ as

$$\widetilde{S}_n = \mathbb{E}\left[\max \left\{ \widehat{S}_n^b(\gamma_a, \Gamma_m), \; \widehat{S}_n^t(\Gamma_a, \Gamma_m) \cdot \mathbb{1}_n^t(\Gamma_a, \Gamma_m) + \widehat{S}^s(\Gamma_a, \Gamma_m) \cdot \mathbb{1}_n^s(\Gamma_a, \Gamma_m) \right\} \right],$$

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 Γ_m is a function of $\Gamma_m^{(j)}$, for $j = 1, \ldots, J$, as depicted in Eq. (34). Also 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

$$\left\{ \begin{array}{ll} \mathbb{1}_n^t(\Gamma_a, \Gamma_m) = 1 & \text{if} \;\; m^{1/\alpha}\Gamma_a + \Gamma_m > \dfrac{\alpha m(1-\rho)}{\lambda} \cdot \left(\left\lfloor \dfrac{n}{m} \right\rfloor - \left\lfloor \dfrac{n_0}{m} \right\rfloor \right)^{(\alpha-1)/\alpha} \\ \mathbb{1}_n^s(\Gamma_a, \Gamma_m) = 1 & \text{otherwise.} \end{array} \right.$$

By positing some assumptions on the distributions of the parameters $\{(\gamma_a, \Gamma_a), (\gamma_m, \Gamma_m)\}$, we compute $\widetilde{S}_n$ via numerical integration.

Choice of Variability Distributions

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

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

where γ_a and $\gamma_s^{(j)}$ follow limiting distributions as defined in the case of a single queue, for $j = 1, \ldots, J$. More specifically, $\gamma_a \sim \mathcal{N}(0, \sigma_a)$ and $\gamma_s^{(j)} \sim \mathcal{N}(0, \sigma_s)$ 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 captured as a function of $\Gamma_m^{(j)}$, for $j = 1, \ldots, J$. Specifically, by Eq. (34),

$$\Gamma_m = \left(\sum_{j=1}^{J} (\Gamma_m^{(j)+})^{\alpha/\alpha-1} \right)^{\alpha-1/\alpha} = \frac{\theta_m}{m^{(\alpha-1)/\alpha}} \cdot \gamma_s^+ \quad \text{where} \quad \gamma_s^+ = \left(\sum_{j=1}^{J} (\gamma_s^{(j)+})^{\alpha/\alpha-1} \right)^{\alpha-1/\alpha}. \tag{41}$$

We propose an approximation of the distribution of γ_s^+ by fitting generalized extreme value distribution to the sampled distribution with a shape parameter ψ_s, scale parameter ξ_s and a location parameter ϕ_s. Table 3 summarizes the parameters defining the generalized extreme value distribution for light-tailed service times with $\sigma_s = 1$ and heavy-tailed queues for $J = 10, 25$ and 50. Figure 3 shows that this fit provides a good approximation of the sampled distribution for the example of $J = 25$ queues in series.

Table 3 Generalized extreme value distributions for γ_s^+ for light ($\sigma_s = 1$) and heavy-tailed services.

	10 Queues			25 Queues			50 Queues		
Parameters	$\alpha = 2$	$\alpha = 1.6$	$\alpha = 1.7$	$\alpha = 2$	$\alpha = 1.6$	$\alpha = 1.7$	$\alpha = 2$	$\alpha = 1.7$	$\alpha = 1.6$
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

This step allows us to reduce the computational effort to obtain $\widehat{S}_n$ 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^+. We next take a similar approach to choose $(\theta_a, \theta_s, \theta_m)$ as in the case of a single queue.

(a) *Light-Tailed Queues:* The overall expected system time in a tandem network of identical queues

$$\overline{S}_n = \overline{S}_n^{(1)} + \ldots + \overline{S}_n^{(J)} = J \cdot \overline{S}_n^{(j)} \leq J \cdot \frac{\lambda}{2(1-\rho)} \cdot \left(\sigma_a^2 + \sigma_s^2 / m^2 \right),$$

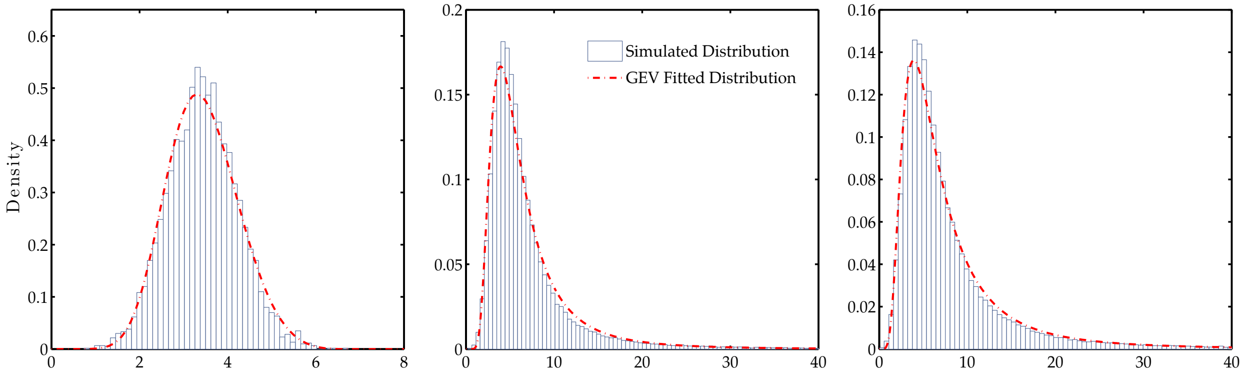


Figure 3 Sampled distribution and fitted generalized extreme value distribution for the effective service parameter γ_s^+ for the case of $J = 25$ queues in series with (a) $\alpha = 2$, (b) $\alpha = 1.7$, and (c) $\alpha = 1.6$.

where the last equality stems from the bound by Kingman (1970). We then ensure that $\widetilde{S}_n$ matches the above expression, i.e.,

$$\frac{\lambda}{4(1-\rho)} \cdot \mathbb{E}\left[\left(\gamma^+\right)^2\right] = \frac{\lambda}{2(1-\rho)} \cdot \left(J\sigma_a^2 + J\sigma_s^2/m^2\right), \tag{42}$$

where $\gamma = \theta_a\gamma_a + \theta_m\gamma_s^+/m$ and γ_s^+ is defined in Eq. (41). The expected value in Eq. (42)

$$\mathbb{E}\left[\left(\gamma^+\right)^2\right] \approx \mathbb{P}\left(\gamma \geq 0\right) \cdot \mathbb{E}\left[\gamma^2\right] = \mathbb{P}\left(\gamma \geq 0\right) \cdot \left(\theta_a^2\sigma_a^2 + \theta_m^2 \mathbb{E}\left[\left(\gamma_s^+\right)^2\right]/m^2\right).$$

Given the symmetry of the normal distribution, we express the second moment of γ_s^+ as

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

By pattern matching the two expressions in Eq. (42), the parameters θ_a and θ_m are such that

$$\theta_a \approx \left(\frac{2J}{\mathbb{P}\left(\gamma \geq 0\right)}\right)^{1/2} \quad \text{and} \quad \theta_m \approx \left(\frac{2}{\mathbb{P}\left(\gamma \geq 0\right) \cdot \mathbb{P}(\gamma_s^{(j)} \geq 0)}\right)^{1/2}. \tag{43}$$

Note that, given that γ_s is a normally distributed distributed random variable centered around the origin, we have $\mathbb{P}\left(\gamma_s^{(j)} \geq 0\right) = 1/2$. Also,

$$\mathbb{P}\left(\gamma \geq 0\right) = \mathbb{P}\left(\theta_a\gamma_a + \theta_m\gamma_s^+/m \geq 0\right) = \mathbb{P}\left(J^{1/2} \cdot \gamma_a + 2^{1/2} \cdot \gamma_s^+/m \geq 0\right),$$

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_m and obtain

$$\theta_a \approx \left(\frac{\alpha J}{\mathbb{P}\left(\gamma \geq 0\right)}\right)^{(\alpha-1)/\alpha} \quad \text{and} \quad \theta_m \approx \left(\frac{\alpha}{\mathbb{P}\left(\gamma \geq 0\right) \cdot \mathbb{P}(\gamma_s^{(j)} \geq 0)}\right)^{(\alpha-1)/\alpha}, \tag{44}$$

where $\gamma = \theta_a \gamma_a + \theta_m \gamma_s^+ / m$, γ_s^+ is defined in Eq. (41) . Note that the probabilities

$$\mathbb{P}\left(\gamma_s^{(j)} \geq 0\right) \quad \text{and} \quad \mathbb{P}\left(\gamma \geq 0\right) = \mathbb{P}\left(J^{(\alpha-1)/\alpha} \cdot \gamma_a + \mathbb{P}(\gamma_s^{(j)} \geq 0)^{-(\alpha-1)/\alpha} \cdot \gamma_s^+ / m \geq 0\right)$$

can be efficiently computed numerically given the distributions of γ_a, $\gamma_s^{(j)}$ and $\gamma_s^{(+)}$.

Computational Results

We investigate the performance of our approach relative to simulation and examine the effect of the system's parameters (traffic intensity, tail coefficient, initial buffer and number of servers) on its accuracy. Specifically, we run simulations for single and multi-server queues with normally and Pareto distributed inter arrival and service times. Tables 4 and 5 report the errors relative to simulation until either relaxation is reached or the maximum number of jobs the simulation was run for ($N = 20,000$ for both light-tailed and heavy-tailed queues). Figure 4 compares our approximation (dotted line) with simulation (solid line) for normally and Pareto distributed queues in series with a traffic intensity $\rho = 0.90$.

Table 4 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	5.83	3.59	6.36	0.74	3.38	0.85	2.39
	0.92	4.85	2.82	5.58	0.81	3.41	0.82	2.41
	0.94	4.28	2.28	5.68	0.92	3.26	0.81	2.33
	0.96	4.68	2.22	4.54	1.10	3.57	0.77	2.26
$\sigma_a = \sigma_s = 4.0$	0.90	1.67	2.66	2.75	1.68	3.18	1.77	2.62
	0.92	1.98	1.52	5.91	1.93	3.46	1.73	2.32
	0.94	2.32	2.26	4.07	2.45	4.37	1.80	2.23
	0.96	2.12	3.46	2.79	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.

Table 5 Errors for single-server tandem queues with Pareto distributed primitives.

		10 Queues		25 Queues		50 Queues	
	ρ	$n_0 = 0$	$n_0 = 2000$	$n_0 = 0$	$n_0 = 3500$	$n_0 = 0$	$n_0 = 5000$
$\alpha_a = \alpha_a = 1.6$	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

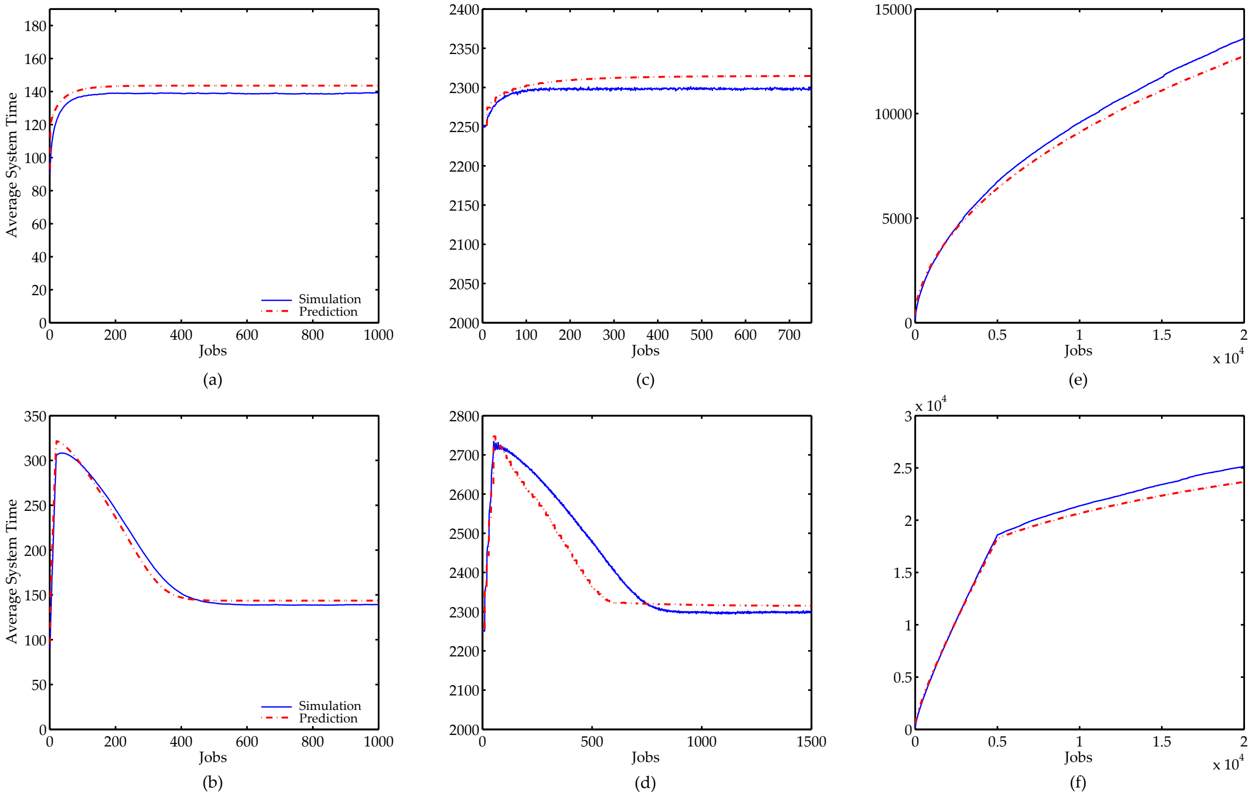


Figure 4 Simulated (solid line) versus predicted values (dotted line). Panels (a)-(d) correspond to normally distributed queues in series with $\sigma_a = 2.5$ and $\rho = 0.90$ with $J = 10$, $m = 1$, and $n_0 = 0, 20$ (panels (a) and (b), respectively) and $J = 25$, $m = 10$, and $n_0 = 0, 50$ (panels (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.

Our approach generally yields errors within 10% for multi-server queues with normally distributed inter arrival and service times. Errors for heavy-tailed multi-server queues are mostly within 10% with occasional outliers. We notice a difference in the behavior of tandem queues with $n_0 > 0$, namely the system exhibits slower recovery for the initial perturbation, as shown in Figure 4(b),(c) for light-tailed queues and Figure 4(f) for heavy-tailed queues. We note that we omit the results for tandem networks of multi-server Pareto queues in this draft due to the high computational cost of simulations.

5. Insights and Computational Tractability

In this section, we discuss the implications of our framework. In particular, our approach (a) provides *analytical tractability* leading to insights that meet the conclusions of probabilistic queueing theory and (b) ensures *computational tractability* leading to reasonably accurate predictions.

5.1. Insights

We draw the following insights from our analysis for both light-tailed and heavy-tailed systems.

Light-Tailed Multi-server Queueing Systems

In a multi-server queue characterized by the set of parameters $\{(\gamma_a, \Gamma_a), (\gamma_s, \Gamma_s)\}$, 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 5 shows a graphical representation of the evolution of the worst case system time under our modeling assumptions.

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* under the worst case setting as the number of jobs observed by the queue before reaching steady state, for a given set of parameters $\{(\gamma_a, \Gamma_a), (\gamma_s, \Gamma_s)\}$. For an initially empty queue, the robust relaxation time is defined as

$$\widehat{n}_{\mathrm{r}} = m \cdot \left[\frac{\lambda \Gamma^{+}}{2m(1-\rho)} \right]^{2},$$

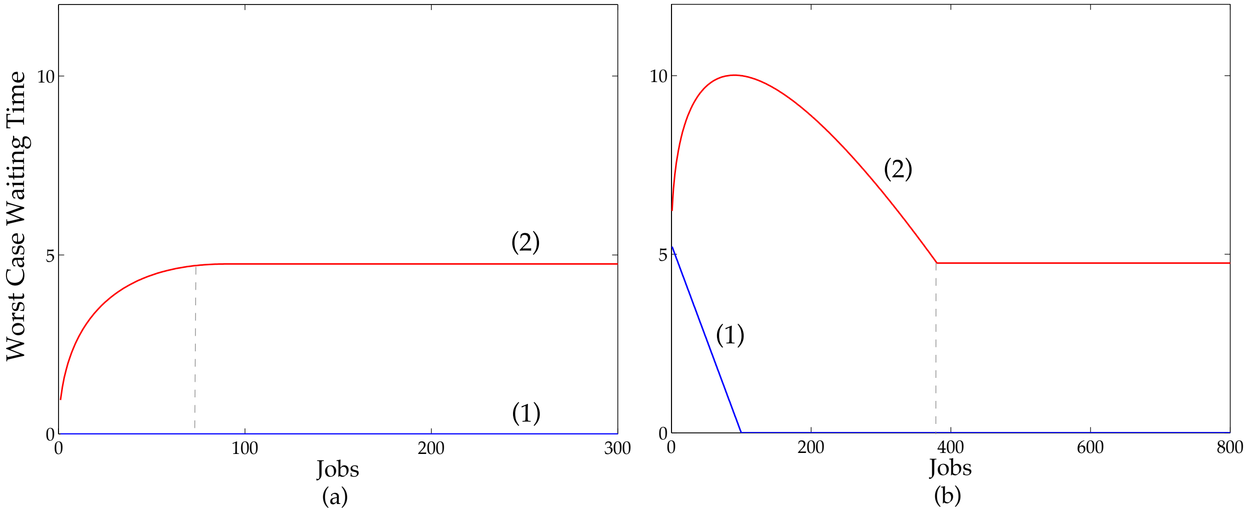


Figure 5 Worst case system time for a single-server queue with $\rho = 0.95$, $\Gamma_a = 0$ and $\Gamma_s = 0, 1$ (respectively curves (1) and (2)), for (a) zero initial jobs, i.e., $n_0 = 0$, and (b) 5 initial jobs, i.e., $n_0 = 5$. The dotted lines indicate the phase change from transient to steady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

where $\Gamma = m^{1/\alpha}\Gamma_a + \Gamma_s$.

Table 6 summarizes the effect of the traffic intensity on the steady-state system time and the robust relaxation time.

Table 6 System time behavior in a light-tailed multi-server queue.

Steady System Time*	Relaxation Time*
$\mathcal{O}\left(\dfrac{(\Gamma^+)^2}{m(1-\rho)}\right)$	$\mathcal{O}\left(\dfrac{n_0}{1-\rho}\right) + \mathcal{O}\left(\dfrac{(\Gamma^+)^2}{m(1-\rho)^2}\right)$

[*] $\Gamma = m^{1/2}\Gamma_a + \Gamma_s$.

Increasing the traffic intensity from ρ_1 to ρ_2 and keeping all other parameters unchanged yields an increase of the order of

$$\left[\left(\frac{1-\rho_1}{1-\rho_2}\right)^2 - 1\right] \cdot 100\% \quad \text{and} \quad \left[\left(\frac{1-\rho_1}{1-\rho_2}\right) - 1\right] \cdot 100\%$$

in the robust relaxation time and the steady state worst case system time, respectively, for $\Gamma > 0$. For instance, increasing the traffic intensity from 0.97 to 0.98 and from 0.98 to 0.99 results in a 50% and 100% increase in the steady state worst case system time, while the robust relaxation time experiences an increase by 120% and 300%, respectively. This asserts the observation by Morse (1955) that, as saturation of the system is approached, the waiting time increases, but the length of the time for the system time to reach steady state increases even more markedly.

The insights from Table 6 extend to the average steady-state system time and relaxation time. While the bound on the expected steady-state system time provided by Kingman (1970) provides a good estimate of $\overline{S}_\infty$, our averaging technique helps determine the average time it takes until steady state is reached in closed form, without having to compute the system time for every job in the queue. In fact, we know that, for any given job n, the system is in steady state as long as

$$n \geq m \cdot \left[\frac{\lambda \Gamma^+}{2m(1-\rho)}\right]^2, \text{ or equivalently } \Gamma^+ \leq \frac{2m(1-\rho)}{\lambda} \left(n/m\right)^{1/2}.$$

Therefore, the system reaches steady state by the n^{th} job with probability

$$\mathbb{P}\left(\Gamma^+ \leq \frac{2(1-\rho)}{\lambda} \cdot (n-r)^{1/2}\right) = p.$$

Knowing the probability distribution of the effective parameter Γ, we can find the desired value for any percentile. We can then compute an approximation of the number of jobs on average that the system needs to service to be p-close to S_∞. Specifically,

$$\frac{2(1-\rho)}{\lambda} \cdot \left(\overline{n}_r^p/m\right)^{1/2} = F^{-1}(p) \quad \text{yielding} \quad \overline{n}_r^p = m \cdot \left(\frac{\lambda \cdot F^{-1}(p)}{2(1-\rho)}\right)^2, \tag{45}$$

where $F(\cdot)$ denotes the cumulative distribution of Γ. This simple calculation provides a powerful tool for practitioners who are concerned with estimating how long it takes the system to come reasonably close to its steady state.

Heavy-Tailed Multi-server Queueing Systems

Under probabilistic assumptions, heavy-tailed queues are characterized by an infinitely long transient state as they never reach steady state. However, in our robust framework, for a given set of parameters $\{(\gamma_a, \Gamma_a), (\gamma_s, \Gamma_s)\}$, 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 interarrival and service times implied by Assumption 1, which involves 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 7.

Table 7 Effect of traffic intensity and heavy tails on worst case behavior of multi-server queues.

Worst Case Steady System Time*	Robust Relaxation Time*
$\mathcal{O}\left(\dfrac{(\Gamma^{+})^{\alpha/\alpha-1}}{m(1-\rho)^{1/(\alpha-1)}}\right)$	$\mathcal{O}\left(\dfrac{n_0}{1-\rho}\right)+\mathcal{O}\left(m\cdot\left[\dfrac{\Gamma^{+}}{m(1-\rho)}\right]^{\alpha/(\alpha-1)}\right)$

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

To illustrate this, we consider an initially empty single server queue and incrementally decrease the tail coefficient from $\alpha = 2$ to $\alpha = 1.75$ and from $\alpha = 1.75$ to $\alpha = 1.5$. For a traffic intensity $\rho = 0.95$, and a choice of $\Gamma_a = 0$ and $\Gamma_s = 1$, the steady state worst case system time experiences an increase by 115% and 420%, respectively, and the relaxation time increases by 190% and 680% respectively.

Our averaging technique allows us to reconcile our conclusions with probabilistic queueing theory. From Table 7, the average system time is proportional to $\mathbb{E}\left[(\Gamma^{+})^{\alpha/(\alpha-1)}\right]$. When Γ_a and Γ_s are heavy-tailed random variables, then moments of Γ_a and Γ_s higher than or equal to the second moment are infinite, implying that $\mathbb{E}\left[(\Gamma^{+})^{\alpha/(\alpha-1)}\right]$ is infinite for $\alpha < 2$. This leads to the conclusion that the average steady-state system time $\widetilde{S}_\infty$ and the relaxation time are infinite for heavy-tailed queues, which is in agreement with conclusions of probabilistic analysis.

5.2. Computational Tractability

Another key feature of our approach is the computational tractability in computing the system time evolution with time for a variety of queueing systems. Our approach, for all the queueing systems we considered, consists of a double integral which we compute by discretization. As we have demonstrated in Sections 3 and 4, to achieve errors of less than 20%, we use a discretization of a thousand. In particular, our approach takes on average on the order of milli-seconds for computing the expected system time or waiting time. Moreover, our approach takes the same time irrespective of the system parameters: traffic ratio (ρ), number of servers (m), light or heavy tailed nature (α), and the number of queues in the tandem system (J). This is in contrast to simulation which is the traditional way of calculating the average system time in these queues. On the other hand, in our approach, we are required to simulate only two random variables Γ_a, Γ_s, and is therefore efficient. We elaborate on this further.

(a) Computational complexity of calculating $\mathbb{E}[S_n]$: 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) Presence 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 Fishman and Adan (2006), Asmussen et al. (2000), Blanchet and Glynn (2008)). 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 a multi-server system: For simulating a FCFS multi-server queueing system, one of the key steps involves sorting the workloads at each server to assign a job to a server. This sorting process is required for each sample path. On the other hand, we present a closed form solution to the case of multi-servers and this sorting step is not required.

(d) Simulation of a tandem queueing system: For simulating a tandem queueing system, we need to simulate each queue in the system for all the n jobs, which is avoided in our approach.

6. FeedForward Networks with Deterministic Routing

In this section, we extend our analysis approach to analyze new queueing systems. As a case study, we consider the the case of feed-forward networks. In feed-forward queueing networks each job is passed from one queue to the next without returning to an earlier passed one, making them acyclic.

We exploit this acyclic or a tree structure to view a feed-forward network as a combination of tandem queues. In particular, we will illustrate using our approach in the context of a feed-forward network of single servers with deterministic routing.

Consider a feed-forward network with node set $\mathcal{N}$, and routing adjacency matrix $\mathbf{A}$, with the interpretation that $A_{ij} = 1$ if and only if there is a link from i to j. Under a deterministic routing, such a network is characterized by the sets $\mathcal{L}_i$, $\forall i \in \mathcal{N}$ where $\mathcal{L}_i$ is the sequence of all jobs that pass through node i. Note that for a feed-forward network, the sets $\mathcal{L}_i$, $\forall i \in \mathcal{N}$ satisfy the following properties:

$$\textbf{(1)}\ \mathcal{L}_i = \bigcup_{j\in\mathcal{N}|Aij=1} \mathcal{L}_j, \qquad \textbf{(2)}\ \mathcal{L}_i \cap \mathcal{L}_j = \varnothing, \quad \forall i,j \text{ that do not share a parent node.}$$

Given such a network, Proposition 2 characterizes the worst case system of an n^{th} job in a feed-forward network composed of single-server queues with deterministic routing. Proposition 2 also generalizes to the case of feed-forward network of multi-server queues, based on the same intuition. Note that for a tandem queue of J nodes, this reduces to our previous results.

Proposition 2. **(System Time in a Feed-forward Network with Deterministic Routing)** *In a feed-forward network of single server queues characterized by $\{\mathcal{N}, \mathbf{A}\}$ with inter arrival times $\mathbf{T} = \{T_1, \dots, T_n\}$, and service times $\left\{\mathbf{X}^{(i)} \in \mathcal{U}_i^s\right\}_{i\in\mathcal{N}}$. Suppose the set of all nodes that job n passes through M nodes given by $\{a_1, a_2, \dots, a_M\}$, then the overall system time of the n^{th} job is given by*

$$\widehat{S}_n(\mathbf{T}) = \max_{k_1,\dots,k_M}\left(\max_{\mathcal{U}^s_{a_1}} \sum_{i=k_1, i\in\mathcal{L}_{a_1}}^{k_2} X_i^{(a_1)} + \max_{\mathcal{U}^s_{a_2}} \sum_{\substack{i=k_2\\ i\in\mathcal{L}_{a_2}}}^{k_3} X_i^{(a_2)} + \dots + \max_{\mathcal{U}^s_{a_M}} \sum_{\substack{i=k_M\\ i\in\mathcal{L}_{a_M}}}^{n} X_i^{(a_M)} - \sum_{i=k_1+1}^{n} T_i\right),$$

where $\{k_i\}_{i=1}^M$ satisfy $k_1 \le k_2 \le \dots \le k_M$ and $k_i \in \mathcal{L}_{a_i}$ for all $i = 1, \dots, M$.

As an illustration, we consider the sample feedforward network depicted in Figure 8. This network is characterized by $\mathcal{N} = \{1,2,3,4,5,6,7\}$, and $\mathbf{A}$ given by

$$A_{12} = A_{13} = A_{24} = A_{25} = A_{36} = A_{37} = 1, \ A_{ij} = 0 \text{ otherwise.}$$

Furthermore, the sets $\mathcal{L}_i$ are given by

$$\mathcal{L}_1 = \{1,2,3,4,\dots\}, \ \mathcal{L}_2 = \{2,4,6,\dots\}, \ \mathcal{L}_3 = \{1,3,5,\dots\}, \ \mathcal{L}_4 = \{4,8,12,\dots\}, \text{ and so on.}$$

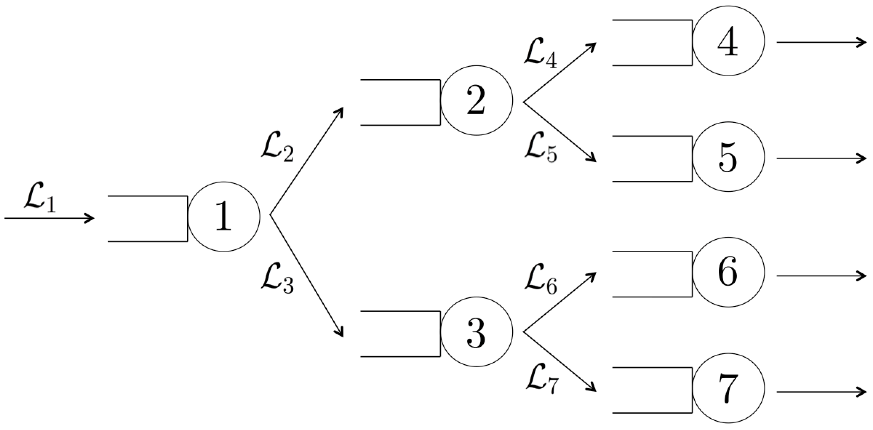


Figure 6 Simple Feed-forward network with deterministic routing.

Consider job n and suppose it goes through nodes $1, 2$, and 4 which happens when n is a multiple of four. Applying Proposition 2, we obtain

$$\widehat{S}_n(\mathbf{T}) = \max_{k_1,k_2,k_3}\left(\max_{\mathcal{U}_1^s}\sum_{i=k_1, i\in\mathcal{L}_1}^{k_2} X_i^{(1)} + \max_{\mathcal{U}_2^s}\sum_{i=k_2, i\in\mathcal{L}_2}^{k_3} X_i^{(2)} + \max_{\mathcal{U}_4^s}\sum_{i=k_3, i\in\mathcal{L}_4}^{n} X_i^{(4)} - \sum_{i=k_1+1}^{n} T_i\right).$$

By applying the bounds from the uncertainty assumptions over the service times,

$$\widehat{S}_n(\mathbf{T}) \leq \max_{k_1,k_2,k_3}\left\{\begin{array}{l}\dfrac{k_2-k_1+1}{\mu_1} + \Gamma_s^{1+}\cdot(k_2-k_1+1)^{1/\alpha} + \dfrac{(k_3-k_2)/2+1}{\mu_2} + \Gamma_s^{2+}\cdot\left(\dfrac{k_3-k_2}{2}+1\right)^{1/\alpha} \\ + \dfrac{(n-k_3)/4+1}{\mu_4} + \Gamma_s^{4+}\cdot\left(\dfrac{n-k_3}{4}+1\right)^{1/\alpha} - \displaystyle\sum_{i=k_1+1}^{n} T_i\end{array}\right\},$$

which follows from the fact that there are $(k_3-k_2)/2+1$ jobs between k_2^{th} and k_3^{th} job in node 2, and $(n-k_3)/4+1$ jobs between k_3^{th} and n^{th} job in node 4. Following the same approach, Proposition 3 presents the expression of the worst case system time in a feed-forward network composed of single-server queues with deterministic routing.

Proposition 3. **(Highest System Time in a Feed-forward Network with Single Servers)** *In a feed-forward network of single server queues characterized by $\{\mathcal{N}, \mathbf{A}\}$ with inter arrival times $\mathbf{T} = \{T_1, \ldots, T_n\}$, and service times $\left\{\mathbf{X}^{(i)} \in \mathcal{U}_i^s\right\}_{i\in\mathcal{N}}$. Suppose the set of all nodes that job n passes through M nodes given by $\{a_1, a_2, \ldots, a_M\}$, then the overall system time of the n^{th} job is given by*

$$\widehat{S}_n \le \max_{k_i} \left\{ \begin{array}{l} \dfrac{k_2-k_1}{\mu_1} + \Gamma_s^{1+} \cdot (k_2-k_1)^{1/\alpha} + \dfrac{p_2(k_3-k_2)}{\mu_2} + \Gamma_s^{2+} \cdot [p_2(k_3-k_2)+1]^{1/\alpha} + \ldots + \\ \dfrac{p_M(n-k_M)+1}{\mu_M} + \Gamma_s^{M+} \cdot [p_M(n-k_M)+1]^{1/\alpha} + \displaystyle\sum_{i=1}^{M} \left(\frac{1}{\mu_i} + \Gamma_s^{i+} \right) - \frac{n-k_1}{\lambda} + \Gamma_a (n-k_1)^{1/\alpha} \end{array} \right\},$$

where $p_i = |\mathcal{L}_{a_i}| / |\mathcal{L}_{a_1}|$ *and* $\{k_i\}_{i=1}^{M}$ *satisfy* $k_1 \le k_2 \le \ldots \le k_M, \quad k_i \in \mathcal{L}_{a_i}, \quad \forall i = 1, \ldots, M.$

We perform a change of variable by let $x_1 = k_2 - k_1,\ x_2 = k_3 - k_2, \ldots,\ x_M = n - k_M$, which yields

$$\widehat{S}_n \le h(x_1, x_2, \ldots, x_M) = \max_{x_i} \left\{ \begin{array}{l} \dfrac{x_1}{\mu_1} + \Gamma_s^{1+} \cdot (x_1)^{1/\alpha} + \dfrac{p_2 x_2}{\mu_2} + \Gamma_s^{2+} \cdot (p_2 x_2)^{1/\alpha} + \ldots + \\ \dfrac{P_M x_M}{\mu_M} + \Gamma_s^{M+} \cdot (P_M x_M)^{1/\alpha} + \displaystyle\sum_{i=1}^{M} \left(\frac{1}{\mu_i} + \Gamma_s^{i+} \right) - \frac{n-k_1}{\lambda} + \Gamma_a (n-k_1)^{1/\alpha} \end{array} \right\},$$

where $p_i = |\mathcal{K}_{a_i}| / |\mathcal{K}_{a_1}|$ for all $i = 1, \ldots, M$. By letting $\tilde{\mu}_i = \mu_i / p_i$ and $\tilde{\Gamma}_s^i = \Gamma_s^i \cdot p_i^{1/\alpha}$, for all $i = 1, \ldots, M$, the worst case system time becomes

$$\widehat{S}_n \le h(x_1, x_2, \ldots, x_M) = \max_{x_1, \ldots, x_M} \left(\sum_{i=1}^{M} \frac{x_i}{\tilde{\mu}_i} + \sum_{i=1}^{M} \tilde{\Gamma}_s^{M+} \cdot (x_M)^{1/\alpha} + \sum_{i=1}^{M} \left(\frac{1}{\mu_i} + \Gamma_s^{i+} \right) - \frac{\sum_{i=1}^{M} x_i}{\lambda} + \Gamma_a (n-k_1)^{1/\alpha} \right).$$

The function $h(\cdot)$ is concave and therefore allows efficient solution. We next provide an upper bound. Note that $\sum_{i=1}^{M} x_i = n - k_1$, and by using the approach taken in the proof of Theorem 6 in Eq. (36), we obtain

$$\begin{aligned} \widehat{S}_n &\le \max_{x_1, \ldots, x_M} \left\{ \sum_{i=1}^{M} \frac{x_i}{\tilde{\mu}_i} + (n-k_1)^{1/\alpha} \left(\sum_{i=1}^{M} (\tilde{\Gamma}_s^{(i)+})^{\alpha/\alpha-1} \right)^{\alpha-1/\alpha} + \sum_{i=1}^{M} \left(\frac{1}{\mu_i} + \Gamma_s^{i+} \right) - \frac{n-k_1}{\lambda} + \Gamma_a (n-k_1)^{1/\alpha} \right\} \\ &\le \max_{k_1} \left\{ \frac{n-k_1}{\min\{\tilde{\mu}_i\}} + (n-k_1)^{1/\alpha} \left(\sum_{i=1}^{M} (\tilde{\Gamma}_s^{(i)+})^{\alpha/\alpha-1} \right)^{\alpha-1/\alpha} + \sum_{i=1}^{M} \left(\frac{1}{\mu_i} + \Gamma_s^{i+} \right) - \frac{n-k_1}{\lambda} + \Gamma_a (n-k_1)^{1/\alpha} \right\} \end{aligned}$$

After this point, the structure of the worst case system time is akin to what we obtained in our previous sections. To compute the average system time $\widehat{S}_n$, we use an averaging algorithm similar to the averaging algorithm we used in Sections 3 and 4 as follows.

Step 1. *Estimate parameters* θ_a, θ_m: To do this, we compare the steady state system time with the one obtained in simulation. To compute the steady state system time, we use the closed form bounds we obtained in Bandi et al. (2012) for general queueing networks.

Step 2. *Estimate averaging probability distributions*: To do this, we follow the methodology of previous sections and fit a distribution to the the parameters Γ_a and

$$\Gamma_{ffwd} = \left(\sum_{i=1}^{M} (\tilde{\Gamma}_s^{(i)+})^{\alpha/\alpha-1} \right)^{\alpha-1/\alpha}.$$

Step 3. *Compute double integrations*: Finally, we compute the average system time of the n^{th} job by solving a double integral with respect to Γ_a and Γ_{ffwd}.

This algorithm reduces to the one we presented in Section 4 for a tandem system of queues.

7. Concluding Remarks

We study the problem of analyzing the transient system time in multi-server queueing systems and feedforward networks with deterministic routing. We model the system's interarrival 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 as a function of the traffic intensity and the tail behavior of interarrival and service times in multi-server queues and feedforward networks with deterministic routing.

We propose a novel algorithm to approximate the expected system time by averaging the worst case system times obtained by choosing different levels of conservatism. This is done by treating the parameters characterizing the uncertainty sets as random variables. This methodology achieves tractability given that we collapse the dimension of uncertainty to two parameters. Furthermore, our approach yields accurate predictions with low errors relative to simulation. The 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.

Acknowledgements

We would like to thank the area editor, the associate editor and the reviewers of the paper for their constructive feedback which improved the paper significantly.

References

J. Abate and W. Whitt. Transient behavior of the *M/M/1* queue: Starting at the origin. *Queueing Systems*, 2(1):41–65, 1987a.

J. Abate and W. Whitt. Transient behavior of the *M/M/1* queue via laplace transforms. *Advances in Applied Probability*, 20(1):145–178, 1987b.

J. Abate and W. Whitt. Calculating transient characteristics of the erlang loss model by numerical transform inversion. *Stochastic Models*, 14(3):663–680, 1998.

S. Asmussen, K. Binswanger, and B. Hjgaard. Rare events simulation for heavy-tailed distributions. *Bernoulli*, 6(2):pp. 303–322, 2000. ISSN 13507265. URL `http://www.jstor.org/stable/3318578`.

N. T. J. Bailey. A continuos time treatment of a simple queue using generating functions. *Journal of Royal Statistical Society*, B16:288–291, 1954a.

N. T. J. Bailey. some further results in the non-equilibrium theory of a simple queue. *Journal of Royal Statistical Society*, B19:326–333, 1954b.

C. Bandi and D. Bertsimas. Tractable stochastic analysis via robust optimization. *Mathematical Programming*, 134:23–70, 2013.

C. Bandi, D. Bertsimas, and N. Youssef. Robust queueing theory. Submitted for publication, 2012.

A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *NATURE*, 435:207, 2005.

A. Ben-Tal, L. El-Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

T. Benson, A. Akella, and D. 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.

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.

D. Bertsimas and K. Natarajan. A semidefinite optimization approach to the steady-state analysis of queueing systems. *Queueing Systems*, 56(1):27–39, 2007.

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.

D. Bertsimas, D. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53:464–501, 2011a.

D. Bertsimas, D. Gamarnik, and A. Rikun. Performance analysis of queueing networks via robust optimization. *Operations Research*, 3:68–93, 2011b.

J. Blanchet and P. Glynn. Efficient rare-event simulation for the maximum of heavy-tailed random walks. *The Annals of Applied Probability*, 18(4):1351–1378, 08 2008. doi: 10.1214/07-AAP485. URL `http://dx.doi.org/10.1214/07-AAP485`.

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.

G. L. Choudhury and W. Whitt. Computing transient and steady-state distributions in polling models by numerical transform inversion. *IEEE International Conference on Communications*, pages 803–809, 1995.

G. L. Choudhury, D. M. Lucantoni, and W. Whitt. Multi-dimensional transform inversion with applications to the transient *M/G/1* queue. *Annals of Applied Probability*, 4:719–740, 1994.

M. Crovella. The relationship between heavy-tailed file sizes and self-similar network traffic. *INFORMS Applied Probability Conference*, 1997.

G. B. Dantzig. Programming of interdependent activities: II mathematical model. *Econometrica*, 17:200–211, 1949.

A. K. Erlang. The theory of probabilities and telephone conversations. *Nyt Tidsskrift for Matematik, B*, 20, 1909.

G. S. Fishman and I. J. B. F. Adan. How heavy-tailed distributions affect simulation-generated time averages. *ACM Trans. Model. Comput. Simul.*, 16(2):152–173, Apr. 2006. ISSN 1049-3301. doi: 10.1145/1138464.1138467. URL `http://doi.acm.org/10.1145/1138464.1138467`.

W. K. Grassmann. Transient solutions in markovian queueing systems. *Comput. Opns. Res.*, 4:47–53, 1977.

W. K. Grassmann. Transient and steady state results for two parallel queues. *Omega*, 8:105–112, 1980.

D. Gross and C. M. Harris. Fundamentals of queueing theory. *John Wiley & Sons, New York.*, 1974.

R. Hampshire, M. Harchol-Balter, and W. 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. ISSN 0257-0130. URL `http://dx.doi.org/10.1007/s11134-006-7584-x`.

D. P. Heyman and M. J. Sobel. *Stochastic Models in Operations Research: Vol. 1*. McGraw-Hill, New York, 1982.

S. Karlin and J. McGregor. Many server queueing processes with poisson input and exponential service times. *Pacific Journal of Mathematics*, 8(1):87–118, 1958. URL `http://projecteuclid.org/euclid.pjm/1103040247`.

J. Keilson. Markov chain models-rarity and exponentiality. *Springer-Verlag*, 1979.

W. D. Kelton and A. M. Law. The transient behavior of the *M/M/s* queue, with implications for steady-state simulation. *Operations Research*, 33(2):378–396, 1985.

J.F.C.. Kingman. Inequalities in the theory of queues. *Journal of the Royal Statistical Society*, 32:102–110, 1970.

B. O. Koopman. Revenue maximization when bidders have budgets. *Operations Research*, pages 1089–1114, 1972.

T. C. T. Kotiah. Approximate transient analysis of some queuing systems. *Operations Research*, 26(2): 333–346, 1978.

N. Krivulin. A recursive equations based representation of the *G/G/m* queue. *Applied Math Letters*, 7(3): 73–77, 1994.

W. Leland, M. Taqqu, and D. Wilson. On the self-similar nature of Ethernet traffic. *ACM SIGCOMM Computer Communication Review*, 25(1):202–213, 1995.

D. V. Lindley. The theory of queues with a single server. *Mathematical Proceedings of the Cambridge Philosophical Society*, 1952.

C. Loboz. Cloud resource usage - heavy tailed distributions invalidating traditional capacity planning models. *J. Grid Comput.*, 10(1):85–108, 2012.

W. Massey. The analysis of queues with time-varying rates for telecommunication models. *Telecommunication Systems*, 21(2-4):173–204, 2002. ISSN 1018-4864. doi: 10.1023/A:1020990313587. URL `http://dx.doi.org/10.1023/A%3A1020990313587`.

S. C. Moore. Approximating the behavior of non-stationary single server queues. *Operations Research*, 23: 1011–1032, 1975.

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.

P. M. Morse. Stochastic processes of waiting lines. *Operations Research*, 3:255–261, 1955.

M. Neuts. The single server queue in discrete time: Numerical analysis I. *Naval Research Logistics*, 20: 297–304, 2004.

G. Newell. *Applications of Queueing Theory*. Chapman & Hall, 1971.

A. Odoni and E. Roth. An empirical investigation of the transient behavior of stationary queueing systems. *Operations Research*, 31(3):432–455, 1983.

T. Osogami and R. Raymond. Analysis of transient queues with semidefinite optimization. *Queueing Systems*, pages 195–234, 2013.

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.

M. H. Rothkopf and S. S. Oren. A closure approximation for the nonstationary *M/M/s* queue. *Management Science*, 25:522–534, 1979.

G. Samorodnitsky and M. Taqqu. *Stable non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. Chapman & Hall, 1994.

J. Xie, Y. Jiang, and M. Xie. A temporal approach to stochastic network calculus. *CoRR*, abs/1112.2822, 2011.

Appendix A

A1. Proof of Theorem 5

Proof of Theorem 5. We prove the result using the technique of mathematical induction.

(a) Initial Step: As presented in Bandi et al. (2012), we can express the system time in an m-server queue as

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

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

$$\widehat{S}_n^{(2)}(\mathbf{T}) + \ldots + \widehat{S}_n^{(J)}(\mathbf{T}) = \max_{0\le k_2\le\ldots\le k_J\le\nu}\left(\max_{\mathcal{U}_2^s}\sum_{i=k_2}^{k_3} X_{r(i)}^{(2)} + \ldots + \max_{\mathcal{U}_J^s}\sum_{i=k_J}^{n} X_{r(i)}^{(J)} - \sum_{i=r(k_2)+1}^{n} T_i^{(2)}\right), \quad (46)$$

where $\mathbf{T}^{(2)} = \left\{T_1^{(2)},\ldots,T_n^{(2)}\right\}$ denotes the sequence of interarrival 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 interderpature times from the first queue by $\mathbf{D}^{(1)} = \left\{D_1^{(1)},\ldots,D_n^{(1)}\right\}$, we have

$$\sum_{i=r(k_2)+1} T_i^{(2)} = \sum_{i=r(k_2)+1} D_i^{(1)} = \sum_{i=(k_2)+1}^{n} T_i + \widehat{S}_n^{(1)}(\mathbf{T}) - \widehat{S}_{r(k_2)}^{(1)}(\mathbf{T}), \quad (47)$$

where the last equality is due to the fact that no overtaking occurs at the first queue in the worst case approach. Combining Eqs. (46)-(47), we obtain

$$\begin{aligned}\widehat{S}_n(\mathbf{T}) &= \widehat{S}_n^{(1)}(\mathbf{T}) + \widehat{S}_n^{(2)}(\mathbf{T}) + \ldots + \widehat{S}_n^{(J)}(\mathbf{T})\\ &= \max_{0\le k_2\le\ldots\le k_J\le\nu}\left(\max_{\mathcal{U}_2^s}\sum_{i=k_2}^{k_3} X_{r(i)}^{(2)} + \ldots + \max_{\mathcal{U}_J^s}\sum_{i=k_J}^{n} X_{r(i)}^{(J)} - \sum_{i=r(k_2)+1}^{n} T_i^{(2)} + \widehat{S}_{r(k_2)}^{(1)}(\mathbf{T})\right). \quad (48)\end{aligned}$$

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. (48), 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 yields the desired result. □

A2. Proof of Theorem 7

Proof of Theorem 7. We show that, after bounding the terms in Eqs. (38) and (39), we obtain simple one-dimensional optimization problems for which an analytic solution exists.

(a) We first consider the case where $n\le n_0$. By following the same procedure for bounding the partial sums of the service times applied in the proof of Theorem 6, we obtain the following bound on the worst case system time

$$\widehat{S}_n\left(\Gamma_m\right)=\max_{0\le k_1\le\ldots\le k_J\le\nu\le\phi}\left(\frac{\nu-k_1+J}{\mu}+\Gamma_m^+\left[\left(k_2-k_1+1\right)^{1/\alpha}+\ldots+\left(\nu-k_J+1\right)^{1/\alpha}\right]\right),$$

where $\Gamma_m=\max\left(\Gamma_s^{(1)},\ldots,\Gamma_s^{(J)}\right)$. By making the transformation $x_1=k_2-k_1+1,\ldots,x_J=\nu-k_J+1$, where $x_j\in\mathbb{N}$, for all $j=1,\ldots,J$, we can represent the maximization problem in the above expression as which simplifies to

$$\max_{0\le k_1\le\nu\le\phi,k_1\in\mathbb{N}}\left(\frac{\nu-k_1+J}{\mu}+\left\{\begin{array}{ll}\max & \Gamma_m^+\left[x_1^{1/\alpha}+\ldots x_J^{1/\alpha}\right]\\ \text{s.t.} & x_1+\ldots+x_J=\nu-k_1+J\\ & k_j\in\mathbb{N},\forall j=2,\ldots,J\end{array}\right\}\right). \tag{49}$$

We relax the above optimization problem by assuming all variables are continuous. Note that, since $\Gamma_m^+\ge 0$, the inner optimization problem is maximized for

$$x_1^*=x_2^*=\ldots=x_J^*=\frac{\nu-k_1+J}{J}.$$

Substituting the optimal solution of the inner problem in Eq. (49), the performance analysis reduces to solving the following one-dimensional optimization problem

$$\max_{0\le k_1\le\nu\le\phi}\left(\frac{\nu-k_1+J}{\mu}+J^{1-1/\alpha}\Gamma_m^+\cdot\left(\nu-k_1+J\right)^{1/\alpha}\right)=\frac{\nu+J}{\mu}+J^{1-1/\alpha}\Gamma_m^+\cdot\left(\nu+J\right)^{1/\alpha}. \tag{50}$$

(b) We next consider the case where $n > n_0$. We proceed by maximizing both terms in Eq. (39).

(1) The first term involves a similar optimization problem to the one solved in part (a), hence

$$\max_{\substack{0\le k_1\le\ldots\le k_J\le\nu\\ k_1\le\phi}} \left(\max_{\mathcal{U}_1^s} \sum_{i=k_1}^{k_2} X_{r(i)}^{(1)} + \ldots + \max_{\mathcal{U}_J^s} \sum_{i=k_J}^{n} X_{r(i)}^{(J)} \right) \le \frac{\nu+J}{\mu} + J^{1-1/\alpha}\Gamma_m^+ \cdot (\nu+J)^{1/\alpha}.$$

Also, note that, by Assumption 1(a), we have

$$\max_{\mathbf{T}\in\mathcal{U}^a} \sum_{i=n_0+1}^{n} T_i = \frac{n-n_0}{\lambda} - \gamma_a (n-n_0)^{1/\alpha}.$$

(2) Following the same bounding procedure as in the proof of Theorem 6, the second term in Eq. (39) can be bounded by the following one-dimensional optimization problem

$$\max_{\phi\le k_1\le\nu} \left(\left(m^{1/\alpha}\Gamma_a + J^{1-1/\alpha}\Gamma_m^+\right)\cdot(\nu-k_1)^{1/\alpha} - \frac{m(1-\rho)}{\lambda}(\nu-k_1) \right) = \max_{0\le x\le\nu-\phi, x\in\mathbb{R}} \left(\beta\cdot x^{1/\alpha} - \delta\cdot x\right).$$

Referring to the proof of Theorem 3, the optimal solution to the above expression is

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

Substituting $\beta = m^{1/\alpha}\Gamma_a + J^{1-1/\alpha}\Gamma_m^+$ and $\delta = m(1-\rho)/\lambda$ yields the desired result. □

A3. Proof of Proposition 1

Proof of Proposition 1. Consider job i. In an FCFS queue, jobs enter service in the order of their arrival. Hence, job i enters service prior to all future incoming jobs. As a result, the system time of job i depends on $\mathbf{T}^i = (T_1, \ldots, T_i)$ and $\mathbf{X}^i = (X_1, \ldots, X_i)$. For some realization of inter-arrival times $\mathbf{T}^i$ and service times $\mathbf{X}^{i+}$, we define the worst case system time in an FCFS queue as

$$\begin{aligned} \widehat{S}_i\left(\mathbf{T}^i, \mathbf{X}^{i+}\right) = \max_{\mathbf{X}^i} \quad & S_i\left(\mathbf{T}^i, \mathbf{X}^i\right) \\ \text{s.t.} \quad & \left(\mathbf{X}^i, \mathbf{X}^{i+}\right) \in \mathcal{U}_m^s. \end{aligned} \tag{51}$$

We next prove our result using the technique of mathematical induction. We postulate and verify the following inductive hypothesis: Under an FCFS policy, there exists a sequence of service times $\widehat{\mathbf{X}}^i$ that achieves the worst case system time $\widehat{S}_i\left(\mathbf{T}^i, \mathbf{X}^{i+}\right)$, with $\widehat{X}_1 \le \ldots \le \widehat{X}_i$, for any given $\mathbf{T}$ and $\mathbf{X}^{i+}$, such that $\left(\widehat{\mathbf{X}}^i, \mathbf{X}^{i+}\right) \in \mathcal{U}_m^s$.

Note that for $i \geq j > k$, job k enters service before job j under an FCFS policy. Given the nondecreasing service times because of the inductive hypothesis, we have $\widehat{X}_j \geq \widehat{X}_k$, implying that job j cannot depart the queue before job k. As a result, under our inductive hypothesis, in an FCFS queue with $\widehat{X}_1 \leq \ldots \leq \widehat{X}_i$, no overtaking occurs until job i. By using Proposition 2 in Bandi et al. (2012), we obtain

$$\widehat{S}_i\left(\mathbf{T}^i, \mathbf{X}^{i+}\right) = \max_{0 \leq k \leq \delta}\left(\sum_{j=k}^{\delta} X_{s(j)} - \sum_{j=s(k)+1}^{i} T_j\right),$$

where δ is such that $i \in K_\delta$.

(a) Initial Step: We first show that the inductive hypothesis holds for $i = 1, \ldots, m$. Since we address the steady-state, we assume, without loss of generality, that the queue is initially empty. Hence, the first m jobs enter service immediately with $S_i = X_i$, for $i \in K_0 = \{1, \ldots, m\}$. Applying Assumption 1(c) for $\mathcal{I} = \{0\} \cup \mathcal{I}'$, for all sets $\mathcal{I}' \subseteq \{1, \ldots, \nu\}$, we obtain

$$X_i + \sum_{k \in \mathcal{I}'} X_{j_k} \leq \frac{|\mathcal{I}'| + 1}{\mu} + \Gamma_s\left(|\mathcal{I}'| + 1\right)^{1/\alpha_s}.$$

This implies that

$$\begin{aligned} X_i &\leq \frac{|\mathcal{I}'| + 1}{\mu} + \Gamma_s\left(|\mathcal{I}'| + 1\right)^{1/\alpha_s} - \sum_{k \in \mathcal{I}'} X_{j_k}, \ \forall\ \mathcal{I}' \subseteq \{1, \ldots, \nu\} \\ &\leq \min_{\mathcal{I}' \subseteq \{1, \ldots, \nu\}} \ \frac{|\mathcal{I}'| + 1}{\mu} + \Gamma_s\left(|\mathcal{I}'| + 1\right)^{1/\alpha_s} - \sum_{k \in \mathcal{I}'} X_{j_k}. \end{aligned}$$

Let $\mathcal{I}^*$ be the minimizer. Thus, to maximize their system time for given $(\mathbf{T}, X_{m+1}, \ldots, X_n)$, it suffices to set their service time to their highest value, i.e.,

$$\widehat{X}_i = \frac{|\mathcal{I}^*| + 1}{\mu} + \Gamma_s\left(|\mathcal{I}^*| + 1\right)^{1/\alpha_s} - \sum_{k \in \mathcal{I}^*} X_{j_k}, \text{ for all } i = 1, \ldots, m.$$

This results in $\widehat{X}_1 = \ldots = \widehat{X}_m$, which satisfies the inductive hypothesis for $i = 1, \ldots, m$.

(b) Inductive Step: We suppose that the inductive hypothesis is true until $i = n - 1$ and prove it for $i = n$. Let $\ell < n$ be the last job that was served by the server which is currently serving job n. Then, the system time S_n is given by

$$S_n = \max\left(C_\ell - A_n, 0\right) + X_n = \max\left(S_\ell + A_\ell - A_n, 0\right) + X_n$$

$$= \max\left(S_\ell - \sum_{j=\ell+1}^{n} T_j, 0\right) + X_n = \max\left(S_\ell + X_n - \sum_{j=\ell+1}^{n} T_j, X_n\right).$$

For any given realization $\mathbf{T}$, the worst case system time is bounded by

$$\begin{aligned} \widehat{S}_n(\mathbf{T}) &= \max_{\mathbf{X}\in\mathcal{U}_m^s} \max\left(S_\ell + X_n - \sum_{j=\ell+1}^{n} T_j, X_n\right) \\ &\le \max\left(\max_{\mathbf{X}\in\mathcal{U}_m^s} S_\ell + X_n - \sum_{j=\ell+1}^{n} T_j, \max_{\mathbf{X}\in\mathcal{U}_m^s} X_n\right). \end{aligned} \tag{52}$$

Let $\left(\widetilde{X}_1, \ldots, \widetilde{X}_n\right)$ be some sequence of service times that maximizes $S_\ell + X_n$, i.e.,

$$\max_{\mathbf{X}\in\mathcal{U}_m^s} S_\ell + X_n = S_\ell\left(\mathbf{T}^\ell, \mathbf{X}^\ell\right) + \widetilde{X}_n.$$

From the induction hypothesis, given a realization $\mathbf{T}$ and $\mathbf{X}^{\ell+}$, there a sequence of non - decreasing service times $\mathbf{X}^\ell$ that achieves the worst case system time, implying

$$S_\ell\left(\mathbf{T}^\ell, \mathbf{X}^\ell\right) \le \widehat{S}_\ell\left(\mathbf{T}^\ell, \mathbf{X}^{\ell+}\right) = \max_{0\le k\le\gamma}\left(\sum_{i=k}^{\gamma} \widehat{X}_{s(i)} - \sum_{i=s(k)+1}^{\ell} T_i\right),$$

where $\ell \in K_\delta$. Hence, we bound the expression in Eq. (52) by

$$\begin{aligned} \widehat{S}_n(\mathbf{T}) &\le \max\left\{\widehat{S}_\ell\left(\mathbf{T}^\ell, \mathbf{X}^{\ell+}\right) + \widetilde{X}_n - \sum_{i=\ell+1}^{n} T_i,\ \max_{\mathcal{U}_m^s} X_n\right\} \\ &\le \max\left\{\max_{0\le k\le\gamma}\left(\sum_{i=k}^{\gamma} \widehat{X}_{s(i)} - \sum_{i=s(k)+1}^{\ell} T_i\right) + \widetilde{X}_n - \sum_{i=\ell+1}^{n} T_i,\ \max_{\mathcal{U}_m^s} X_n\right\}. \end{aligned}$$

Rearranging the terms, and since $\left(\widehat{\mathbf{X}}^i, \widetilde{\mathbf{X}}^{i+}\right) \in \mathcal{U}_m^s$, we obtain

$$\begin{aligned} \widehat{S}_n(\mathbf{T}) &\le \max\left\{\max_{0\le k\le\gamma}\left(\sum_{i=k}^{\gamma} \widehat{X}_{s(i)} + \widetilde{X}_n - \sum_{i=s(k)+1}^{\ell} T_i - \sum_{i=\ell+1}^{n} T_i\right),\ \max_{\mathcal{U}_m^s} X_n\right\} \\ &\le \max\left\{\max_{0\le k\le\gamma}\left(\max_{\mathcal{U}_m^s}\left\{\sum_{i=k}^{\gamma} X_{s(i)} + X_n\right\} - \sum_{i=s(k)+1}^{n} T_i\right),\ \max_{\mathcal{U}_m^s} X_n\right\}. \end{aligned} \tag{53}$$

Recall that $s(k) = \ell - (\gamma - k)m \in K_k$. Given that no overtaking occurrs until ℓ, at the time job n enters service, the jobs served by the remaining $(m-1)$ servers should have arrived after job ℓ and before job n, i.e., they belong to the set $\mathcal{I} = \{\ell+1, \ldots, n-1\}$. Since there are $(m-1)$ such jobs, we have

$$m - 1 \le |\mathcal{I}| = n - 1 - (\ell + 1) + 1 = n - \ell - 1,$$

yielding $n-\ell\geq m$. Consider the partition $K_0,K_1,\ldots,K_\nu$ that we considered in Assumption 1(c). Since two jobs j and k in the same set satisfy $|j-k|<m$, jobs n and ℓ belong to two distinct sets in the partition $K_0,K_1,\ldots,K_\nu$. With $\ell\in K_\gamma$, and $n\in K_\nu$, this implies $\nu\geq\gamma+1$. We consider the following two cases.

(1) If $\nu=\gamma+1$, then by Assumption 1(c),

$$\max_{\mathcal{U}_m^s}\left\{\sum_{i=0}^{\gamma}X_{s(i)}+X_n\right\}=\frac{\nu+1}{\mu}+\gamma_m\left(\nu+!\right)^{1/\alpha_s}, \tag{54}$$

$$\max_{\mathcal{U}_m^s}\left\{\sum_{i=0}^{\nu}X_{s(i)}\right\}=\frac{\nu+1}{\mu}+\gamma_m\left(\nu+1\right)^{1/\alpha_s}, \tag{55}$$

$$\max_{\mathcal{U}_m^s}\left\{\sum_{i=k}^{\gamma}X_{s(i)}+X_n\right\}=\frac{\nu-k+1}{\mu}+\Gamma_s\left(\nu-k+1\right)^{1/\alpha_s},\quad\forall k\geq 1, \tag{56}$$

$$\max_{\mathcal{U}_m^s}\left\{\sum_{i=k}^{\nu}X_{s(i)}\right\}=\frac{\nu-k+1}{\mu}+\Gamma_s\left(\nu-k+1\right)^{1/\alpha_s},\quad\forall k\geq 1, \tag{57}$$

where $r(i)=n-(\nu-i)m$. Therefore, we have

$$\max_{\mathcal{U}_m^s}\left\{\sum_{i=k}^{\gamma}X_{s(i)}+X_n\right\}=\max_{\mathcal{U}_m^s}\left\{\sum_{i=k}^{\nu}X_{s(i)}\right\}. \tag{58}$$

Also, the index $r(k)=n-(\nu-k)m=n-(\gamma+1-k)m$. Given that $n\geq\ell+m$, we have $r(k)\geq\ell-(\gamma-k)m=s(k)$, which results in

$$\sum_{i=s(k)+1}^{n}T_i\geq\sum_{i=r(k)+1}^{n}T_i,\ \text{for all } 0\leq k\leq\gamma. \tag{59}$$

Combining Eqs. (58) and (59), Eq. (53) becomes

$$\widehat{S}_n\left(\mathbf{T}\right)\leq\max\left\{\max_{0\leq k\leq\nu-1}\left(\max_{\mathcal{U}_m^s}\sum_{i=k}^{\nu}X_{r(i)}-\sum_{i=r(k)+1}^{n}T_i\right),\ \max_{\mathcal{U}_m^s}X_n\right\}. \tag{60}$$

(2) If $\nu\geq\gamma+2$, then by Assumption 1(c),

$$\max_{\mathcal{U}_m^s}\left\{\sum_{i=k}^{\gamma}X_{s(i)}+X_n\right\}=\max_{\mathcal{U}_m^s}\left\{\sum_{i=k+1}^{\gamma+1}X_{r(i)}+X_n\right\}\leq\max_{\mathcal{U}_m^s}\left\{\sum_{i=k+1}^{\nu}X_{r(i)}\right\}. \tag{61}$$

Also, since $s(k)\in J_k$ and $r(k+1)\in J_{k+1}$, we have $s(k)\leq r(k+1)$, which implies

$$\sum_{i=s(k)+1}^{n}T_i\geq\sum_{i=r(k+1)+1}^{n}T_i,\ \text{for all } 0\leq k\leq\gamma. \tag{62}$$

Applying the bounds in Eqs. (61) and (62), Eq. (53) becomes

$$\begin{aligned}\widehat{S}_n\left(\mathbf{T}\right) &\leq \max\left\{\max_{0\leq k\leq\gamma}\left(\max_{\mathcal{U}_m^s}\sum_{i=k+1}^{\nu}X_{r(i)}-\sum_{i=r(k+1)+1}^{n}T_i\right),\ \max_{\mathcal{U}_m^s}X_n\right\}\\ &= \max\left\{\max_{1\leq k\leq\gamma+1}\left(\max_{\mathcal{U}_m^s}\sum_{i=k}^{\nu}X_{r(i)}-\sum_{i=r(k)+1}^{n}T_i\right),\ \max_{\mathcal{U}_m^s}X_n\right\}. \qquad (63)\end{aligned}$$

Since $\nu\geq\gamma+2$, we can further bound Eq. (63) to obtain

$$\widehat{S}_n\left(\mathbf{T}\right) \leq \max\left\{\max_{0\leq k\leq\nu-1}\left(\max_{\mathcal{U}_m^s}\sum_{i=k}^{\nu}X_{r(i)}-\sum_{i=r(k)+1}^{n}T_i\right),\ \max_{\mathcal{U}_m^s}X_n\right\}. \qquad (64)$$

Combining the results in Eqs. (60) and (64) from cases (1) and (2), we conclude that the worst case system time under FCFS is bounded by

$$\widehat{S}_n\left(\mathbf{T}\right) \leq \max_{0\leq k\leq\nu}\left(\max_{\mathcal{U}_m^s}\sum_{i=k}^{\nu}X_{r(i)}-\sum_{i=r(k)+1}^{n}T_i\right).$$

This bound is tight and is achieved under a scenario where the service times are chosen such that Eqs. (54)–(54) are tight. This leads to

$$\begin{aligned}\widehat{X}_{j_k} &= \frac{1}{\mu}+\Gamma_m\left[\left(\nu-k+1\right)^{1/\alpha_s}-\left(\nu-k\right)^{1/\alpha_s}\right],\quad \forall j_k\in K_k \quad\text{and}\quad k=1,\ldots,\nu \qquad (65)\\ \widehat{X}_{j_0} &= \frac{1}{\mu}+\gamma_m\left(\nu+1\right)^{1/\alpha_s}-\Gamma_m\left(\nu\right)^{1/\alpha_s},\quad \forall j_0\in K_0 \quad\text{and}\\ &\leq \frac{1}{\mu}+\Gamma_m\left[\left(\nu-k+1\right)^{1/\alpha_s}-\left(\nu-k\right)^{1/\alpha_s}\right],\quad \forall j_0\in K_0. \qquad (66)\end{aligned}$$

Note that this optimal solution consists of nondecreasing service times, hence proving the inductive hypothesis. □