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

Robust Queueing Theory

Chaithanya Bandi*, Dimitris Bertsimas†, Nataly Youssef‡

We propose an alternative approach for studying queues based on robust optimization. We model the uncertainty in the arrivals and services via polyhedral uncertainty sets which are inspired from the limit laws of probability. Using the generalized central limit theorem, this framework allows to model heavy-tailed behavior characterized by bursts of rapidly occurring arrivals and long service times. We take a worst-case approach and obtain closed form upper bounds on the system time in a multi-server queue. These expressions provide qualitative insights which mirror the conclusions obtained in the probabilistic setting for light-tailed arrivals and services and generalize them to the case of heavy-tailed behavior. We also develop a calculus for analyzing a network of queues based on the following key principle: **(a)** the departure from a queue, **(b)** the superposition, and **(c)** the thinning of arrival processes have the same uncertainty set representation as the original arrival processes. The proposed approach **(a)** yields results with error percentages in single digits relative to simulation, and **(b)** is to a large extent insensitive to the number of servers per queue, network size, degree of feedback, traffic intensity, and somewhat sensitive to the degree of diversity of external arrival distributions in the network.

Key words: Queueing Theory, Robust Optimization, Heavy Tails, Stochastic Networks

1. Introduction

The origin of queueing theory dates back to the beginning of the 20^{th} century, when Erlang (1909) published his fundamental paper on congestion in telephone traffic. In addition to formulating and solving several practical problems arising in telephony, Erlang laid the foundations for queueing theory in terms of the nature of assumptions and techniques of analysis that are being used to this day. Given the modeling power of probability theory, a substantial literature of queueing theory was developed which views queueing primitives as renewal processes. In particular, the Poisson process has played a privileged role in modeling the arrival process of a queue. When combined with

* Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139, cbandi@mit.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

exponentially distributed service times, the resulting $M/M/m$ queue with m servers is tractable to analyze in steady-sate.

While exponentiality leads to a tractable theory, assuming general distributions, on the other hand, yields considerable difficulty with respect to performing a near-exact analysis of the system. In fact, the analysis of the $GI/GI/m$ queue with independent and generally distributed arrivals and services is, by and large, intractable. The most general method, due to Pollaczek (1957), analyzes the performance of the $GI/GI/m$ queue by formulating a multi-dimensional problem in the complex plane. Gall (1998) portrays the difficulty of explicitly characterizing the equations for the $GI/GI/m$ queue given that their "partial solution can only be derived after long and complex calculations involving multiple contour integrals in a multi-dimensional complex plane". When arrival and service distributions have rational Laplace transforms of order p (for example Coxian distributions with p phases), the $GI/GI/m$ problem becomes intractable for higher order p values. Bertsimas (1990) reports numerical results for queues with up to 100 servers and $p=2$ by finding all $h=\binom{m+p-1}{m}$ complex roots to distinct polynomial equations and solving a linear system of dimension h. The system's dimension, however, increases to 4.5 million when $p=5$, hence illustrating the complexity of the problem under these assumptions.

The situation becomes even more challenging if one considers analyzing the performance of queueing networks. A key result that allows generalizations to networks of queues is Burke's theorem (Burke (1956)) which states that the departure process from an $M/M/m$ queue in steady-state is Poisson. This property allows one to analyze queueing networks and leads to product form solutions as in Jackson (1957). However, when the queueing system is not $M/M/m$, the departure process is no longer a renewal process, i.e., the inter-departure times are dependent. With the departure process lacking the renewal property, it is difficult to determine performance measures exactly, even for a simple network with queues in tandem. The two avenues in such cases are *simulation* and *approximation*. Simulation provides an accurate depiction of the system's performance, but can take a considerable amount of time in order for the results to be statistically significant,

especially for heavy-tailed systems in heavy traffic. In addition, simulation models are often complex, which makes it difficult to isolate and understand key qualitative insights. On the other hand, approximation methods, such as QNA developed by Whitt (1983) and QNET developed by J. G. Dai and J. M. Harrison (1992), provide a fair estimation of performance, but suffer from a lack of generalizability to model heavy-tailed behavior.

Given these challenges, the key problem of performance analysis of queueing networks has remained open under the probabilistic framework. In his opening lecture at the conference entitled "100 Years of Queueing–The Erlang Centennial", Kingman (2009), one of the pioneers of queueing theory in the 20^{th} century, writes, "*If a queue has an arrival process which cannot be well modeled by a Poisson process or one of its near relatives, it is likely to be difficult to fit any simple model, still less to analyze it effectively. So why do we insist on regarding the arrival times as random variables, quantities about which we can make sensible probabilistic statements? Would it not be better to accept that the arrivals form an irregular sequence, and carry out our calculations without positing a joint probability distribution over which that sequence can be averaged?*". In practice, probability distributions are not inherent to the queueing system; they represent a modeling choice of the modeler that attempts to approximate the actual underlying behavior of the arrival and service processes.

We propose an alternative framework to model queueing systems based on optimization theory. The motivation behind our idea stems from the rich development of optimization as a scientific field during the second part of the 20^{th} 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 key idea of our approach is to make the limit laws of probability theory the primitive assumptions and formulate

the problems arising in queueing systems as robust optimization problems. An initial effort along these lines includes the work by Bertsimas et al. (2011b) where probabilistic guarantees on the length of a busy period and the waiting time are provided through robust optimization. Herein, we build upon this work and present a new approach for modeling the primitives of queueing systems by uncertainty sets. This framework allows us to derive exact performance analysis of the underlying stochastic system. 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 (2012a, 2013, 2012b, 2014)).

Our robust optimization approach to queueing theory bears philosophical similarity with the *deterministic network calculus* approach which was pioneered by Cruz (1991a,b) (see also Gallager and Parekh (1994), El-Taha and Stidham (1999), C.S.Chang (2001), Boudec and Thiran (2001)). Both methods (a) take a non-probabilistic approach by placing deterministic constraints on the traffic flow and (b) derive bounds on key queueing performance measures via a worst case paradigm. There has also been a significant literature on what is called *stochastic network calculus* (see Jiang and Liu (2008), Jiang (2012), Ciucu et al. (2005), Burchard et al. (2011) for an overview). We note, however, that the primitives of stochastic network calculus are in fact probabilistic, so the similarity, even at the philosophical level, is significantly smaller. To a lesser degree, there is also philosophical similarity (in that it is a deterministic and worst case approach) with *adversarial queueing theory* (Borodin et al. (2001), Gamarnik (2003, 2000), Goel (1999)) which was developed for stability analysis in multi-class queueing networks. In contrast, our aspiration in this work is to develop a theory of performance analysis, and thus there is no overlap between adversarial and robust queueing theory beyond the philosophical level. Beyond their deterministic and worst case paradigms, significant differences can be noted when comparing our framework to the network calculus approach.

(a) Different Underlying Assumptions: While both methods postulate deterministic constraints over the arrival process, the assumptions are different in nature. The deterministic network calculus bounds the number of external arrivals n_t up to time t by $n_t \leq \lambda \cdot t + B$,

where λ denotes the traffic rate and B is a constant accounting for burstiness. In contrast, our assumption on the arrival process yields different bounds on the number of arrivals n_t. In fact, denoting the arrival time of the n_t^{th} job by t, i.e., $\sum_{i=1}^{n_t} T_i = t$, and applying Assumption 1(a) with tail coefficient $\alpha_a = 2$, we obtain $n_t - \lambda\Gamma_a\sqrt{n_t} \leq \lambda t \leq n_t + \lambda\Gamma_a\sqrt{n_t}$, where Γ_a represents the effect of variability. Writing $\delta^2 = n_t$ yields $\delta^2 - \lambda\Gamma_a\delta \leq \lambda t \leq \delta^2 + \lambda\Gamma_a\delta$. This implies that $\delta \geq \left(-\lambda\Gamma_a + \sqrt{\lambda^2\Gamma_a^2 + 4\lambda t}\right)/2$, leading to $n_t \geq \lambda t - t^{\frac{1}{2}}\lambda^{\frac{3}{2}}\Gamma_a$. Similarly, we obtain $n_t \leq \lambda t + t^{\frac{1}{2}}\lambda^{\frac{3}{2}}\Gamma_a$, which results in the following bounds on the number of arrivals by time t

$$|n_t - \lambda \cdot t| \leq \Gamma_a \lambda^{3/2} t^{1/2}. \tag{1}$$

Note that the way we handle variability is different from the deterministic network calculus, and is motivated and indeed consistent with the limit laws of probability (see Section 2.2).

(b) Tighter Bounds for single server queues: It is widely believed that the network calculus approach can provide overly conservative bounds for single-server queues. In the words of Ciucu and Hohlfeld (2010) "The deterministic network calculus can lead to conservative bounds because many of the statistical properties of the arrivals are not accounted for," and for the stochastic network calculus "in M/M/1 and M/D/1 queuing scenarios where exact results are available, the stochastic network calculus bounds are reasonably accurate," (see also Ciucu (2007)). Our approach, however, provides a bound on the system times for single-server queues that is qualitatively similar to its probabilistic counterpart (see Section 3.3). Our computations further show that, by constraining nature via bounding the variability allowed in our uncertainty sets, we obtain results within often 4-6%, and at most 8% in stochastic queueing networks (see Section 5).

(c) Generalizability: Our approach extends to more complex queueing systems such as multi-server queues (see Section 3.2) and queueing networks with feedback (see Section 4). However, "*for $GI/GI/m, (m > 1)$, stochastic network calculus based analysis remains plain blank*" and "*feedback analysis is perhaps the most critical open challenge for stochastic network calculus*", as remarked by Jiang (2012). Furthermore, while the stochastic network calculus has recently

addressed heavy tails in a single-server setting (see Burchard et al. (2012)), our framework is capable of providing closed-form upper bounds on the system time, while maintaining deterministic assumptions. In probabilistic queues, Kelly et al. (1998) considers this problem for markovian processes, and in network calculus setting, Xie and Jiang (2009), Jiang and Liu (2008) have obtained some preliminary results in queues under priority disciplines. We plan to investigate such disciplines under our framework in future work.

Specifically, our contributions and structure of the paper are as follows:

(a) In Section 2, we introduce the uncertainty model and propose to replace the renewal process primitives with uncertainty sets that the arrival and service processes satisfy.

(b) In Section 3, we study single and multi-server queues operating under a first-come first-serve (FCFS) scheduling policy. Taking a worst case approach, we obtain closed form upper bounds on the system time, which not only carry the same qualitative insights found via traditional queueing theory, but also extend the analysis to include heavy-tailed arrivals and services.

(c) In Section 4, we analyze the departure process under the assumption that servers act adversarially so as to maximize the system time in the queue. We show that the departure times belong to the arrival uncertainty set. This result is asymptotically akin to Burke's theorem and therefore forms the cornerstone of the proposed steady-state network analysis.

(d) In Section 5, we develop a calculus describing the three operations which affect the arrival process in queueing networks: passing through a queue, superposition and thinning. This allows an analytic characterization of the steady-state performance of queueing networks under the assumption of adversarial servers.

(e) In Section 6, we present extensions of the results in Sections 3-5 to accommodate the case where arrival and service times possess different tail behaviors.

(f) In Section 7, we show that the proposed network analysis provides a good approximation for the analysis of a stochastic queueing network. The computations suggest that the robust approach can be adapted to be within 4-6% from simulation. We also investigate the sensitivity of the

results in terms of the number of servers per queue, network size, degree of feedback, traffic intensity, and the degree of diversity of external arrival distributions in the network.

2. Proposed Framework

In the traditional probabilistic study of queues, the inter-arrival times $\mathbf{T} = \{T_1, T_2, \ldots, T_n\}$ and service times $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ are modeled as renewal processes. Understanding the behavior of time spent by the n^{th} job in a queueing system entails the understanding of the complex relationships between the random variables associated with the inter-arrival and service times. For instance, in a single-server first-come first-serve (FCFS) queue, the system time S_n is given by (Lindley (1952)) as

$$S_n = W_n + X_n = \max\left(W_{n-1} + X_{n-1} - T_n, 0\right) + X_n = \max_{1\le k\le n}\left(\sum_{\ell=k}^{n} X_\ell - \sum_{\ell=k+1}^{n} T_\ell\right), \tag{2}$$

where W_n denotes the waiting time, i.e., the time spent waiting to enter service. The high dimensional nature of the performance analysis problem makes the probabilistic approach by and large intractable. The study of multi-server queues is even more challenging.

Instead, we assume inter-arrival and service times belong to uncertainty sets. We take a robust optimization approach and seek the worst case system time experienced by the n^{th} job under the uncertainty set assumptions. In this section, we present our model of uncertainty, motivated by the probabilistic limit laws.

2.1. Motivation via the Limit Laws

Motivated by the expression in Eq. (2), we propose to bound the partial sums over the inter-arrival and service times. We guide our bounding procedure by the conclusions of probability theory, namely the probabilistic weak convergence theorems. These theorems express the distribution of the sum of many independent and identically distributed random variables as converging to one of a small set of stable distributions.

Light Tailed Distributions: Suppose that the inter-arrival and service times are independent and identically distributed (i.i.d.) with means $1/\lambda$ and $1/\mu$, and finite standard deviations σ_a and σ_s, respectively. By the central limit theorem, as $n \to \infty$, the random variables

$$\frac{\sum_{i=k+1}^{n} T_i - \frac{n-k}{\lambda}}{\sigma_a (n-k)^{1/2}} \quad \text{and} \quad \frac{\sum_{i=k+1}^{n} X_i - \frac{n-k}{\mu}}{\sigma_s (n-k)^{1/2}}$$

are asymptotically standard normal. We know that a standard normal Z satisfies $\mathbb{P}(Z \leq 2) \approx 0.975$, $\mathbb{P}(Z \leq 3) \approx 0.995$. We therefore assume that the quantities T_i and X_i take values such that

$$\sum_{i=k+1}^{n} T_i - \frac{n-k}{\lambda} \geq -\Gamma_a (n-k)^{1/2} \quad \text{and} \quad \sum_{i=k+1}^{n} X_i - \frac{n-k}{\mu} \leq \Gamma_s (n-k)^{1/2}, \tag{3}$$

where the variability parameters Γ_a and Γ_s can be chosen to ensure that the inter-arrival times and the service times satisfy the corresponding inequalities with high enough probability.

Heavy Tailed Distributions: Under a probabilistic framework, a sequence of random variables $\{Y_i\}_{i \geq 1}$ whose variance is undefined, are associated with heavy-tailed distributions. Such random variables satisfy the generalized central limit theorem (Samorodnitsky and Taqqu (1994)).

THEOREM 1. **Generalized Central Limit Theorem**

Let $Y_1, Y_2, \ldots$ be a sequence of i.i.d. random variables, with mean μ and undefined variance. Then

$$\frac{\sum_{i=1}^{n} Y_i - n\mu}{C_\alpha n^{1/\alpha}} \sim Y, \tag{4}$$

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

To illustrate, the normalized sum of a large number of positive Pareto random variables with common distribution may be approximated by a random variable Y following a standard stable distribution with a tail coefficient α and $C_\alpha = [\Gamma(1-\alpha) cos(\pi\alpha/2)]^{1/\alpha}$, where $\Gamma(\cdot)$ denotes the gamma function. For a tail coefficient of $\alpha = 1.5$, we obtain $\mathbb{P}(Y \leq 6.5) \approx 0.975$ and $\mathbb{P}(Y \leq 19) \approx 0.995$ via the tail probability approximations given by Nolan (1997). We therefore assume that the quantities T_i and X_i take values such that the partial sums

$$\sum_{i=k+1}^{n} T_i - \frac{n-k}{\lambda} \geq -\Gamma_a (n-k)^{1/\alpha} \quad \text{and} \quad \sum_{i=k+1}^{n} X_i - \frac{n-k}{\mu} \leq \Gamma_s (n-k)^{1/\alpha}, \tag{5}$$

where the variability parameters Γ_a and Γ_s are chosen to ensure that the inter-arrival times and the service times satisfy the corresponding inequality with high enough probability. Since $\mathcal{O}\left(n^{1/\alpha}\right) > \mathcal{O}\left(n^{1/2}\right)$ for $1 < \alpha < 2$, the scaling by $(n-k)^{1/\alpha}$ in Eq. (5) allows the selection of smaller inter-arrival times and larger service times compared to Eq. (3) with the scaling by $(n-k)^{1/2}$.

2.2. Our Model of Uncertainty

Our model of uncertainty is primarily driven by our desire to analyze the worst case system time. Guided by the system time's expression in Eq. (2) for a single-server queue, we lower bound the partial sums over the inter-arrival times and upper bound the partial sums over the service times.

ASSUMPTION 1. **(Queueing Primitives Assumptions)**

(a) *The inter-arrival times* $\{T_1, T_2, \ldots, T_n\}$ *belong to the parametrized uncertainty set*

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

where $1/\lambda$ *is the expected inter-arrival time,* Γ_a *is a parameter that captures variability information, and* $1 < \alpha_a \leq 2$ *models possibly heavy-tailed probability distributions.*

(b) *The service times* $\{X_1, X_2, \ldots, X_n\}$ *for a single-server queue belong to the parametrized uncertainty set*

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

where $1/\mu$ *is the expected service time,* Γ_s *is a parameter that captures variability information, and* $1 < \alpha_s \leq 2$ *models possibly heavy-tailed probability distributions.*

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

$$J_0 = \{1, \ldots, m\}, J_1 = \{m+1, \ldots, 2m\}, \ldots, J_\nu = \{\nu m + 1, \ldots, n\}.$$

Let $j_i \in J_i$ 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

$$\mathcal{U}_m^s = \left\{ (X_1, \ldots, X_n) \left| \frac{\sum_{i \in \mathcal{I}} X_{j_i} - \dfrac{|\mathcal{I}|}{\mu}}{|\mathcal{I}|^{1/\alpha_s}} \leq \Gamma_s, \ \forall \ j_i \in J_i, \ \textit{and} \ \ i \in \mathcal{I} \subseteq \{0, \ldots, \nu\} \right. \right\}.$$

Note that $\mathcal{U}_1^s \subset \mathcal{U}^s$.

We present the following remarks regarding the proposed uncertainty set assumptions.

(a) Modeling Dependence: While the uncertainty sets are motivated by i.i.d. assumptions on the underlying random variables, $(T_1, T_2, \ldots, T_n) \in \mathcal{U}^a$ does not necessarily imply that $(T_1, T_2, \ldots, T_n)$ are independent.

(b) Modeling Heavy-Tailed Behavior: Assumption 1 presents another modeling approach for heavy-tailed behavior, inspired by Theorem 1. Unlike the probabilistic setting where heavy-tailed distributions imply unboundedness and infinite variance, our assumption implies that the service times are bounded. Assumption 1 allows, however, the service times to be substantially large by appropriately selecting the parameter Γ_s. For instance, for a Pareto distribution with $\alpha_s = 1.5$, $1/\mu = 2.85$, and $\Gamma_s = 19 C_\alpha = 35.055$, we have $\mathbb{P}(X_n \leq 1/\mu + \Gamma_s) \approx 0.996$, that is, with high probability the service times are large but bounded.

(c) Richness of the Service Uncertainty Set: In order to illustrate the set $\mathcal{U}_m^s$, we consider the example for $n = 5$ and $m = 2$:

$$(|\mathcal{I}| = 3) \quad \left\{ \begin{array}{ll} X_1 + X_3 + X_5 \leq 3/\mu + \Gamma_s \cdot 3^{1/\alpha_s} & X_2 + X_3 + X_5 \leq 3/\mu + \Gamma_s \cdot 3^{1/\alpha_s} \\ X_1 + X_4 + X_5 \leq 3/\mu + \Gamma_s \cdot 3^{1/\alpha_s} & X_2 + X_4 + X_5 \leq 3/\mu + \Gamma_s \cdot 3^{1/\alpha_s} \end{array} \right\},$$

$$(|\mathcal{I}| = 2) \quad \left\{ \begin{array}{ll} X_1 + X_3 \leq 2/\mu + \Gamma_s \cdot 2^{1/\alpha_s} & X_2 + X_3 \leq 2/\mu + \Gamma_s \cdot 2^{1/\alpha_s} \\ X_1 + X_4 \leq 2/\mu + \Gamma_s \cdot 2^{1/\alpha_s} & X_2 + X_4 \leq 2/\mu + \Gamma_s \cdot 2^{1/\alpha_s} \\ X_1 + X_5 \leq 2/\mu + \Gamma_s \cdot 2^{1/\alpha_s} & X_2 + X_5 \leq 2/\mu + \Gamma_s \cdot 2^{1/\alpha_s} \\ X_3 + X_5 \leq 2/\mu + \Gamma_s \cdot 2^{1/\alpha_s} & X_4 + X_5 \leq 2/\mu + \Gamma_s \cdot 2^{1/\alpha_s} \end{array} \right\},$$

$$(|\mathcal{I}| = 1) \quad \left\{ X_1, \ X_2, \ X_3, \ X_4, \ X_5 \leq \frac{1}{\mu} + \Gamma_s \right\}.$$

In general, the inequalities associated with the set $\mathcal{I}$ involve the sum of $|\mathcal{I}|$ service times, where each service time is selected out of a set J_i, for $i \in \mathcal{I}$, yielding $\mathcal{O}\left(m^{|\mathcal{I}|}\right)$ such inequalities.

Though the number of constraints in the set is exponential, we will show later that the problem of finding the worst case system time given $\mathbf{T} \in \mathcal{U}^a$ and $\mathbf{X} \in \mathcal{U}_m^s$ is efficiently solvable and yields analytic bounds (refer to Section 3.2). Currently, the uncertainty set includes constraints involving jobs from different sets in the partition $J_0, J_1, \ldots, J_\nu$. While we could have also added constraints with jobs selected from the same set J_i, the set $\mathcal{U}_m^s$ represents a minimal set of inequalities for our bounds on the worst case system time to be valid.

(d) Limiting the Adversary: Despite taking a worst-case approach, one can obtain accurate results that compare with simulations of average behavior by bounding the power of the adversary. Parameterizing the uncertainty sets $\mathcal{U}^a$ and $\mathcal{U}_m^s$ by the variability parameters Γ_a and Γ_s allows us to control the degree of robustness of the approach.

In summary, the key data primitives characterizing **(a)** the arrival process in the queue are $(\lambda, \Gamma_a, \alpha_a)$, and **(b)** the service process in the queue are $(\mu, \Gamma_s, \alpha_s)$. In Sections 3-5, we assume that arrival and service processes have symmetric tail behavior, i.e., $\alpha_a = \alpha_s = \alpha$. We provide bounds for the case of asymmetric tail coefficients in Section 6.

3. Optimization-Based Performance Analysis

In this section, we study the worst case behavior of a single queue with an FCFS scheduling policy and a traffic intensity $\rho = \lambda/(m\mu) < 1$, where m denotes the number of servers. For a given sequence of inter-arrival times $\mathbf{T} = (T_1, \ldots, T_n)$, we let

$$\widehat{S}_n(\mathbf{T}) = \max_{\mathbf{X} \in \mathcal{U}_m^s} S_n. \tag{6}$$

We seek the highest system time that the n^{th} job can experience in the queue, assuming the arrivals are governed by Assumption 1(a), by solving the following optimization problem

$$\widehat{S}_n = \max_{\mathbf{T} \in \mathcal{U}^a} \widehat{S}_n(\mathbf{T}). \tag{7}$$

The above optimization problem is tractable given the choice of polyhedral uncertainty sets. In fact, we show in this section that this problem effectively reduces to one-dimensional nonlinear optimization problem that can be solved efficiently. We further provide a closed-form upper bound on the worst case system time, which is particularly tight for large values of n.

3.1. Worst-Case Performance in a Single-Server Queue

Given a realization $\mathbf{T}$, and using Eq. (2), the worst case system time of the n^{th} job in a single-server queue is given by

$$\widehat{S}_n(\mathbf{T}) = \max_{\mathbf{X}\in\mathcal{U}^s} \max_{1\le k\le n} \left(\sum_{i=k}^{n} X_i - \sum_{i=k+1}^{n} T_i\right) \le \max_{1\le k\le n} \left(\max_{\mathbf{X}\in\mathcal{U}^s} \sum_{i=k}^{n} X_i - \sum_{i=k+1}^{n} T_i\right) \tag{8}$$

where the second inequality is due to exchanging the order of the maximization. Proposition 1 shows that the bound in Eq. (8) is tight, and that there there exists a sample path which achieves the worst case value with nondecreasing service times.

PROPOSITION 1. *In a single-server FCFS queue, there exists a sample path* $\widehat{\mathbf{X}} \in \mathcal{U}^s$ *with non-decreasing service times achieving*

$$\widehat{S}_n(\mathbf{T}) = \max_{1\le k\le n} \left(\max_{\mathbf{X}\in\mathcal{U}^s} \sum_{i=k}^{n} X_i - \sum_{i=k+1}^{n} T_i\right). \tag{9}$$

Proof of Proposition 1. We show that there exists a sequence of service times $\widehat{\mathbf{X}} \in \mathcal{U}^s$ which achieves the bound in Eq. (8), such that

$$\sum_{i=k}^{n} \widehat{X}_i = \max_{\mathbf{X}\in\mathcal{U}^s} \sum_{i=k}^{n} X_i = \frac{n-k+1}{\mu} + \Gamma_s (n-k+1)^{1/\alpha_s}, \ \forall\ k = 1,\ldots,n.$$

Given the triangular structure of the above system of equalities, this solution is unique and can be computed via backward substitution. Specifically,

$$\widehat{X}_i = \frac{1}{\mu} + \Gamma_s \left[(n-i+1)^{1/\alpha_s} - (n-i)^{1/\alpha_s}\right], \text{ for all } i = 1,\ldots,n. \tag{10}$$

Since the function $f(i) = (n-i+1)^{1/\alpha_s} - (n-i)^{1/\alpha_s}$ is increasing in i, we conclude that the obtained service times are nondecreasing, i.e., $\widehat{X}_1 \le \ldots \le \widehat{X}_n$. □

Assuming $\mathbf{T} \in \mathcal{U}^a$, and given Eqs. (7) and (9), the worst case system time can be written as

$$\widehat{S}_n = \max_{\mathbf{T}\in\mathcal{U}^a} \max_{1\le k\le n} \left(\max_{\mathbf{X}\in\mathcal{U}^s} \sum_{i=k}^{n} X_i - \sum_{i=k+1}^{n} T_i\right) \le \max_{1\le k\le n} \left(\max_{\mathbf{X}\in\mathcal{U}^s} \sum_{i=k}^{n} X_i - \min_{\mathbf{T}\in\mathcal{U}^a} \sum_{i=k+1}^{n} T_i\right).$$

By a similar argument to the one in the proof of Proposition 1, we can show that the above bound is tight, and that there exists a sequence of interarrival times $\widehat{\mathbf{T}} \in \mathcal{U}^a$ such that

$$\sum_{i=k+1}^{n} \widehat{T}_i = \min_{\mathbf{T}\in\mathcal{U}^a} \sum_{i=k+1}^{n} T_i = \frac{n-k}{\lambda} - \Gamma_a (n-k)^{1/\alpha_a}, \text{ for all } k = 1,\ldots,n-1, \tag{11}$$

which achieves the worst case value. This yields the following *exact characterization* of the worst case system time as

$$\widehat{S}_n = \max_{1\le k\le n}\left\{\frac{n-k+1}{\mu} + \Gamma_s\left(n-k+1\right)^{1/\alpha_s} - \frac{n-k}{\lambda} + \Gamma_a\left(n-k\right)^{1/\alpha_a}\right\}. \tag{12}$$

The worst case performance analysis hence reduces to a one-dimensional nonlinear optimization problem, which can be solved efficiently. Theorem 2 provides a closed form upper bound on the worst case system time for the case where $\alpha_a = \alpha_s = \alpha$.

THEOREM 2 **(Worst Case System Time in a Single-Server FCFS Queue)**.

In a single-server FCFS queue with $\mathbf{T}\in\mathcal{U}^a$, $\mathbf{X}\in\mathcal{U}^s$, $\alpha_a=\alpha_s=\alpha$ *and* $\rho<1$,

$$\widehat{S}_n \le \frac{\alpha-1}{\alpha^{\alpha/(\alpha-1)}}\cdot\frac{\lambda^{1/(\alpha-1)}\left(\Gamma_a+\Gamma_s\right)^{\alpha/(\alpha-1)}}{(1-\rho)^{1/(\alpha-1)}} + \frac{1}{\lambda}, \tag{13}$$

Proof of Theorem 2. Since $\Gamma_a\left(n-k\right)^{1/\alpha} \le \Gamma_a\left(n-k+1\right)^{1/\alpha}$, we can bound Eq. (12) by

$$\begin{aligned}\widehat{S}_n &\le \max_{1\le k\le n}\left\{\left(\Gamma_a+\Gamma_s\right)\left(n-k+1\right)^{1/\alpha} - \frac{n-k}{\lambda} + \frac{n-k+1}{\mu}\right\}\\ &= \max_{1\le k\le n}\left\{\left(\Gamma_a+\Gamma_s\right)\left(n-k+1\right)^{1/\alpha} - \frac{1-\rho}{\lambda}\left(n-k+1\right)\right\} + \frac{1}{\lambda}.\end{aligned} \tag{14}$$

By making the transformation $x = n-k+1$, where $x\in\mathbb{N}$, Eq. (14) becomes of the form

$$\max_{1\le x\le n}\beta\cdot x^{1/\alpha} - \delta\cdot x \;\le\; \max_{x\in\mathbb{R}^+}\beta\cdot x^{1/\alpha} - \delta\cdot x = \frac{\alpha-1}{\alpha^{\alpha/(\alpha-1)}}\cdot\frac{\beta^{\alpha/(\alpha-1)}}{\delta^{1/(\alpha-1)}}, \tag{15}$$

where $\beta = \Gamma_a+\Gamma_s > 0$ and $\delta = (1-\rho)/\lambda > 0$, given $\rho<1$. Note that the bound in Eq. (15) is independent of n, and is therefore true for all values of n. The continuous maximizer of the unconstrained maximization problem in Eq. (15) is given by

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

We obtain Eq. (13) by substituting β and δ by their respective expressions in the optimal objective function value given in Eq. (15). □

Tightness of the Bound: We note that the bound in Eq. (13) is nearly tight for heavy-traffic systems operating under steady state. In the process of obtaining the closed form expressions, the bounding procedure in the proof of Theorem 2 involved three steps: **(1)** bounding the term $\Gamma_a(n-j)^{1/\alpha}$ by $\Gamma_a(n-j+1)^{1/\alpha}$, **(2)** relaxing the integer requirement for the index j and treating it as a continuous variable, and **(3)** bounding the constrained maximization by an unconstrained maximization in Eq. (15). We note that under heavy-traffic assumptions (i.e., ρ is close to unity), steps **(1)** and **(2)** produce nearly tight bounds, both in terms of achievability within the uncertainty sets and numerical accuracy. Specifically, there exist sequences of inter-arrival and service times that lead to a system time within an error

$$\Delta = \mathcal{O}\left((1-\rho)^{\alpha/(\alpha-1)} + \left(\left(1+(1-\rho)^{\alpha/(\alpha-1)}\right)^{1/\alpha} - 1\right)\right),$$

from the bound in Eq. (13), where $\Delta \to 0$ as $\rho \to 1$ (please see the appendix for details). Moreover, step **(3)** is tight for values of n exceeding the maximizer value in Eq. (16), i.e.,

$$n > \left(\frac{\lambda\left(\Gamma_a+\Gamma_s\right)}{\alpha(1-\rho)}\right)^{\alpha/(\alpha-1)}.$$

3.2. Worst-Case Performance in a Multi-Server Queue

We now analyze the case of an FCFS queue with m parallel servers and consider job $\ell \le n$, where $\ell \in J_\gamma$. The central difficulty in analyzing multi-server queues lies in the fact that overtaking may occur, i.e., the ℓ^{th} departure may not correspond to the ℓ^{th} job arriving to the queue. Let C_ℓ denote the completion time of the ℓ^{th} job, i.e., the time the ℓ^{th} job leaves the system (including service), and $C_{(\ell)}$ denote the time of the ℓ^{th} departure from the system. In general, the following recursions describe the dynamics in a multi-server queue (Krivulin (1994))

$$C_\ell = \max\left(A_\ell, C_{(\ell-m)}\right) + X_\ell \quad \text{and} \quad S_\ell = C_\ell - A_\ell = \max\left(C_{(\ell-m)} - A_\ell, 0\right) + X_\ell, \tag{17}$$

where $A_\ell = \sum_{i=1}^{\ell} T_i$ denotes the the time of arrival of the ℓ^{th} job.

Taking a worst case approach allows us to overcome the challenges of multi-server queue dynamics and obtain an exact characterization of the worst case system time for the n^{th} job, for any $\mathbf{T}$, as

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

where $r(i) = n-(\nu-i)m$. To prove this result, we use the following procedure:

(1) We introduce a set of policies $\mathcal{P}$ that do not allow overtaking until some $\ell \le n$, and obtain an analytic expression of the system time under such policies (see Proposition 2),

(2) Then, for any $\mathbf{T}$, we obtain an exact characterization of the the worst case system time under $\mathcal{P}$, which can be achieved via a sequence of nondecreasing service times (see Proposition 3),

(3) Last, we show that, for any $\mathbf{T}$, the worst case system time for an FCFS queue is equal to the worst case system time for a multi-server queue under $\mathcal{P}$ (see Proposition 4).

We next present the proofs of Propositions 2-4.

No-Overtaking Behavior: For all policies in $\mathcal{P}$, no overtaking occurs until ℓ. Hence, until ℓ, the jobs depart in the same order they arrive, i.e., $C^{\mathcal{P}}_{(k)} = C^{\mathcal{P}}_k$, for all $1\le k\le \ell$. Under $\mathcal{P}$, the recursion in Eq. (17) therefore simplifies to

$$C^{\mathcal{P}}_\ell = \max\left(C^{\mathcal{P}}_{\ell-m}, A_\ell\right) + X_\ell, \quad \text{and} \quad S^{\mathcal{P}}_\ell = C^{\mathcal{P}}_\ell - A_\ell = \max\left(C^{\mathcal{P}}_{\ell-m} - A_\ell, 0\right) + X^{\mathcal{P}}_\ell. \tag{18}$$

Using this recursive formula, Proposition 2 gives an explicit expression of the system time $S^{\mathcal{P}}_\ell$ in a multi-server queue operating under $\mathcal{P}$.

PROPOSITION 2. *Under a set of polices $\mathcal{P}$ that do not allow overtaking until job $\ell \le n$, where $\ell \in J_\gamma$, the system time of the ℓ^{th} job in an m-server queue is given by*

$$S^{\mathcal{P}}_\ell = \max_{0\le k\le \gamma}\left(\sum_{i=k}^{\gamma} X_{s(i)} - \sum_{i=s(k)+1}^{\ell} T_i\right), \tag{19}$$

where $s(i) = \ell-(\gamma-i)m$.

Proof of Proposition 2. Utilizing Eq. (18), and since $C^{\mathcal{P}}_{\ell-m} = S^{\mathcal{P}}_{\ell-m} + A_{\ell-m}$, we obtain

$$S^{\mathcal{P}}_\ell = \max\left(S^{\mathcal{P}}_{\ell-m} + A_{\ell-m} - A_\ell, 0\right) + X^{\mathcal{P}}_\ell = \max\left(S^{\mathcal{P}}_{\ell-m} + X^{\mathcal{P}}_\ell - \left(A_\ell - A_{\ell-m}\right), X^{\mathcal{P}}_\ell\right).$$

Applying the recursion expression to the term $S^{\mathcal{P}}_{\ell-m}$ above yields

$$\begin{aligned}
S^{\mathcal{P}}_{\ell} &= \max\left(\max\left(S^{\mathcal{P}}_{\ell-2m}+X_{\ell-m}-(A_{\ell-m}-A_{\ell-2m}),X_{\ell-m}\right)+X_{\ell}-(A_{\ell}-A_{\ell-m}),X_{\ell}\right)\\
&= \max\left(S^{\mathcal{P}}_{\ell-2m}+(X_{\ell-m}+X_{\ell})-(A_{\ell-m}-A_{\ell-2m})-(A_{\ell}-A_{\ell-m}),X_{\ell-m}+X_{\ell}-(A_{\ell}-A_{\ell-m}),X_{\ell}\right)\\
&= \max\left(S^{\mathcal{P}}_{\ell-2m}+(X_{\ell-m}+X_{\ell})-(A_{\ell}-A_{\ell-2m}),(X_{\ell-m}+X_{\ell})-(A_{\ell}-A_{\ell-m}),X_{\ell}\right)
\end{aligned}$$

Since $\ell\in J_{\gamma}=\{\gamma m+1,\ldots,(\gamma+1)m\}$, we have $\ell\le(\gamma+1)m$, implying $1\le\ell-\gamma m\le m$. Hence, we can apply the recursion until $S^{\mathcal{P}}_{\ell-\gamma m}$ and obtain

$$S^{\mathcal{P}}_{\ell}=\max\left(S^{\mathcal{P}}_{\ell-\gamma m}+\sum_{i=0}^{\gamma-1}X_{\ell-im}-(A_{\ell}-A_{\ell-\gamma m}),\sum_{i=0}^{\gamma-1}X_{\ell-im}-\left(A_{\ell}-A_{\ell-(\gamma-1)m}\right),\ldots,X_{\ell}\right).$$

Note that the first m jobs enter service without waiting, implying that their system time is equal to their service time. Since $\ell-\gamma m\le m$, we have $S^{\mathcal{P}}_{\ell-\gamma m}=X_{\ell-\gamma m}$. And expressing the arrival times A_j as the sum of the interarrival times $T_1,\ldots,T_j$, the system time can then be written as

$$\begin{aligned}
S^{\mathcal{P}}_{\ell} &= \max\left(X_{\ell-\gamma m}+\sum_{i=0}^{\gamma-1}X_{\ell-im}-\sum_{i=\ell-\gamma m+1}^{\ell}T_i,\ \sum_{i=0}^{\gamma-1}X_{\ell-im}-\sum_{i=\ell-(\gamma-1)m+1}^{\ell}T_i,\ \ldots\ ,\ X_{\ell}\right)\\
&= \max\left(\sum_{i=0}^{\gamma}X_{\ell-im}-\sum_{i=\ell-\gamma m+1}^{\ell}T_i,\ \sum_{i=0}^{\gamma-1}X_{\ell-im}-\sum_{i=\ell-(\gamma-1)m+1}^{\ell}T_i,\ \ldots\ ,\ X_{\ell}\right)\\
&= \max\left(\sum_{i=0}^{\gamma}X_{\ell-(\gamma-i)m}-\sum_{i=\ell-\gamma m+1}^{\ell}T_i,\ \sum_{i=1}^{\gamma}X_{\ell-(\gamma-i)m}-\sum_{i=\ell-(\gamma-1)m+1}^{\ell}T_i,\ \ldots\ ,\ X_{\ell}\right).
\end{aligned}$$

The compact representation of the above expression becomes

$$S^{\mathcal{P}}_{\ell}=\max_{0\le k\le\gamma}\left(\sum_{i=k}^{\gamma}X_{\ell-(\gamma-i)m}-\sum_{i=\ell-(\gamma-i)m+1}^{\ell}T_i\right).$$

Substituting $s(i)=\ell-(\gamma-i)m$ yields Eq. (19). □

We next introduce some notation that will be used in the remaining part of this section. Let us fix the vector of service times $\mathbf{X}^{\ell+}=(X_{\ell+1},\ldots,X_n)$. Let $\mathbf{T}^{\ell}=(T_1,\ldots,T_{\ell})$ and $\mathbf{X}^{\ell}=(X_1,\ldots,X_{\ell})$. By Assumption 1(c), the vector $(\mathbf{X}^{\ell},\mathbf{X}^{\ell+})\in\mathcal{U}^s_m$. For some realization of inter-arrival times $\mathbf{T}^{\ell}$ and service times $\mathbf{X}^{\ell+}$, we define the worst case system time under $\mathcal{P}$ as

$$\begin{aligned}
\widehat{S}^{\mathcal{P}}_{\ell}(\mathbf{T}^{\ell},\mathbf{X}^{\ell+}) = \max_{\mathbf{X}^{\ell}}\quad & S^{\mathcal{P}}_{\ell}(\mathbf{T}^{\ell},\mathbf{X}^{\ell})\\
\text{s.t.}\quad & (\mathbf{X}^{\ell},\mathbf{X}^{\ell+})\in\mathcal{U}^s_m.
\end{aligned}\tag{20}$$

By Proposition 2, the worst case system time under $\mathcal{P}$ for a given sequence $(\mathbf{T}^\ell, \mathbf{X}^{\ell+})$ is given by

$$\begin{aligned}\widehat{S}_\ell^{\mathcal{P}}\left(\mathbf{T}^\ell, \mathbf{X}^{\ell+}\right) &= \max_{\left(\mathbf{x}^\ell, \mathbf{X}^{\ell+}\right)\in\mathcal{U}_m^s} \max_{0\le k\le\gamma} \left(\sum_{i=k}^{\gamma} X_{s(i)} - \sum_{i=s(k)+1}^{\ell} T_i\right) \\ &\le \max_{0\le k\le\gamma} \left(\max_{\left(\mathbf{x}^\ell, \mathbf{X}^{\ell+}\right)\in\mathcal{U}_m^s} \sum_{i=k}^{\gamma} X_{s(i)} - \sum_{i=s(k)+1}^{\ell} T_i\right), \end{aligned} \tag{21}$$

where $s(i) = \ell - (\gamma - i)m$. Proposition 3 shows that the bound in Eq. (21) is tight and that there exists a sample path which achieves the worst case value with nondecreasing service times.

PROPOSITION 3. *In an m-server queue, under a set of policies $\mathcal{P}$ that do not allow overtaking until job $\ell \le n$, where $\ell \in J_\gamma$, and given a realization $X^{\ell+} \in \mathcal{U}_m^s$, there exists a sample path $\left(\widehat{X}_1^{\mathcal{P}}, \ldots, \widehat{X}_\ell^{\mathcal{P}}\right)$ with non-decreasing service times achieving*

$$\widehat{S}_\ell^{\mathcal{P}}\left(\mathbf{T}^\ell, \mathbf{X}^{\ell+}\right) = \max_{0\le k\le\gamma} \left(\max_{\mathcal{U}_m^s} \sum_{i=k}^{\gamma} X_{s(i)} - \sum_{i=s(k)+1}^{\ell} T_i\right), \tag{22}$$

where $s(i) = \ell - (\gamma - i)m$.

Proof of Proposition 3. The index $s(i) = \ell - (\gamma - i)m = (\ell - \gamma m) + im$. And, since $\ell \in J_\gamma = \{\gamma m + 1, \ldots, (\gamma+1)m\}$, we have $\gamma m + 1 \le \ell \le (\gamma+1)m$, implying $1 \le \ell - \gamma m \le m$. Therefore,

$$im + 1 \le s(i) = (\ell - \gamma m) + im \le (i+1)m,$$

yielding $s(i) \in J_i$. Since, for $i \ne j$, the indices $s(i)$ and $s(j)$ belong to different sets in the partition $J_0, \ldots, J_\gamma$. Hence, we can use Assumption 1(c) for $\mathcal{I} = \{k, \ldots, \gamma\} \cup \mathcal{I}'$, where $\mathcal{I}' \subseteq \{\gamma+1, \ldots, \nu\}$ and $|\mathcal{I}| = \gamma - k + |\mathcal{I}'| + 1$, to obtain

$$\sum_{i=k}^{\gamma} X_{s(i)} + \sum_{i\in\mathcal{I}'} X_{j_i} \le \frac{\gamma - k + |\mathcal{I}'| + 1}{\mu} + \Gamma_s \left[\gamma - k + |\mathcal{I}'| + 1\right]^{1/\alpha_s}.$$

This implies the following bound the partial sums of the service times in Eq. (21)

$$\sum_{i=k}^{\gamma} X_{s(i)} \le \frac{\gamma - k + |\mathcal{I}'| + 1}{\mu} + \Gamma_s \left(\gamma - k + |\mathcal{I}'| + 1\right)^{1/\alpha_s} - \sum_{i\in\mathcal{I}'} X_{j_i}, \tag{23}$$

for all $k = 0, \ldots, \gamma$. Since Eq. (23) is true for all $\mathcal{I}' \subset \{\gamma+1, \ldots, \nu\}$, then

$$\sum_{i=k}^{\gamma} X_{s(i)} \le \min_{\mathcal{I}'\subseteq\{\gamma+1,\ldots,\nu\}} \left\{\frac{\gamma - k + |\mathcal{I}'| + 1}{\mu} + \Gamma_s \left(\gamma - k + |\mathcal{I}'| + 1\right)^{1/\alpha_s} - \sum_{i\in\mathcal{I}'} X_{j_i}\right\}, \tag{24}$$

$$= \frac{\gamma - k + |\mathcal{I}_k^*| + 1}{\mu} + \Gamma_s \left(\gamma - k + |\mathcal{I}_k^*| + 1\right)^{1/\alpha_s} - \sum_{i \in \mathcal{I}_k^*} X_{j_i}, \tag{25}$$

where $\mathcal{I}_k^*$ is the minimizer in Eq. (24). Eq. (25) implies, for all $k = 0, \ldots, \gamma$, that

$$\max_{(\mathbf{X}^\ell, \mathbf{X}^{\ell+}) \in \mathcal{U}_m^s} \sum_{i=k}^{\gamma} X_{s(i)} = \frac{\gamma - k + |\mathcal{I}_k^*| + 1}{\mu} + \Gamma_s \left(\gamma - k + |\mathcal{I}_k^*| + 1\right)^{1/\alpha_s} - \sum_{i \in \mathcal{I}_k^*} X_{j_i}.$$

We next show that there exists a sequence $\left(\widehat{X}_1^{\mathcal{P}}, \ldots, \widehat{X}_\ell^{\mathcal{P}}\right)$ that achieves

$$\sum_{i=k}^{\gamma} \widehat{X}_{s(i)}^{\mathcal{P}} = \max_{\mathcal{U}_m^s} \sum_{i=k}^{\gamma} X_{s(i)} = \frac{\gamma - k + |\mathcal{I}_k^*| + 1}{\mu} + \Gamma_s \left(\gamma - k + |\mathcal{I}_k^*| + 1\right)^{1/\alpha_s} - \sum_{i \in \mathcal{I}_k^*} X_{j_i}, \tag{26}$$

for all $k = 0, \ldots, \gamma$. Due to its triangular structure, the above system of equalities yields a unique solution $\left(\widehat{X}_{s(0)}^{\mathcal{P}}, \ldots, \widehat{X}_{s(\gamma-1)}^{\mathcal{P}}, \widehat{X}_{s(\gamma)}^{\mathcal{P}}\right)$, which can be computed via backward substitution. Specifically,

$$\begin{cases} \widehat{X}_{s(\gamma)}^{\mathcal{P}} = \widehat{X}_\ell^{\mathcal{P}} = \dfrac{|\mathcal{I}_\gamma^*| + 1}{\mu} + \Gamma_s \left(|\mathcal{I}_\gamma^*| + 1\right)^{1/\alpha_s} - \displaystyle\sum_{i \in \mathcal{I}_\gamma^*} X_{j_i}, \\ \widehat{X}_{s(k)}^{\mathcal{P}} = \dfrac{|\mathcal{I}_k^*| - |\mathcal{I}_{k+1}^*| + 1}{\mu} + \Gamma_s \left[\left(\gamma - k + |\mathcal{I}_k^*| + 1\right)^{1/\alpha_s} - \left(\gamma - k + |\mathcal{I}_{k+1}^*|\right)^{1/\alpha_s} \right] - \displaystyle\sum_{i \in \mathcal{I}_k^*} X_{j_i} + \sum_{i \in \mathcal{I}_{k+1}^*} X_{j_i}, \end{cases}$$

for all $k = 0, \ldots, \gamma - 1$. To complete the sequence, we propose to set the service times of all jobs belonging to a partition J_i to have the same value as job $s(i) \in J_i$, for all $i = 0, \ldots, \gamma$, i.e.,

$$\widehat{X}_{j_i}^{\mathcal{P}} = \widehat{X}_{s(i)}^{\mathcal{P}}, \text{ for all } j_i \in J_i, \text{ where } i = 0, \ldots, \gamma. \tag{27}$$

(a) We next show that, given $\mathbf{X}^{\ell+}$, the chosen sequence of service times satisfies the inequalities of set $\mathcal{U}_m^s$. Since the service times are nondecreasing, the sum of service times selected from a set $\mathcal{I}'' \subseteq \{0, \ldots, \gamma\}$, such that $|\mathcal{I}''| = \gamma - k + 1$, can be upper-bounded by

$$\sum_{i \in \mathcal{I}''} \widehat{X}_{j_i}^{\mathcal{P}} \leq \sum_{i=k}^{\gamma} \widehat{X}_{s(i)}^{\mathcal{P}}.$$

And given Eqs. (23)-(26), we obtain

$$\sum_{i \in \mathcal{I}} \widehat{X}_{j_i}^{\mathcal{P}} = \sum_{i \in \mathcal{I}'} \widehat{X}_{j_i}^{\mathcal{P}} + \sum_{i \in \mathcal{I}''} \widehat{X}_{j_i}^{\mathcal{P}} \leq \frac{|\mathcal{I}'| + |\mathcal{I}''|}{\mu} + \Gamma_s \left(|\mathcal{I}'| + |\mathcal{I}''| \right)^{1/\alpha_s},$$

for all $\mathcal{I} = \mathcal{I}' \cup \mathcal{I}'' \subseteq \{0, \ldots, \nu\}$. The sequence of service times $\left(\widehat{X}_1^{\mathcal{P}}, \ldots, \widehat{X}_\ell^{\mathcal{P}}\right)$ therefore satisfies the inequalities of the uncertainty set $\mathcal{U}_m^s$, for any realization $\mathbf{X}^{\ell+}$, and is hence feasible. As a result, the bound in Eq. (21) can be achieved with equality.

(b) The chosen sequence of service times is also nondecreasing.

(1) Given the optimality of set $\mathcal{I}_k^*$ from Eq. (25), we have

$$\frac{|\mathcal{I}_k^*|}{\mu}+\Gamma_s\left[\gamma-k+|\mathcal{I}_k^*|+1\right]^{1/\alpha_s}-\sum_{i\in\mathcal{I}_k^*}X_{j_i}\le\frac{|\mathcal{I}_{k+1}^*|}{\mu}+\Gamma_s\left[\gamma-k+|\mathcal{I}_{k+1}^*|+1\right]^{1/\alpha_s}-\sum_{i\in\mathcal{I}_{k+1}^*}X_{j_i}.$$

Rearranging the terms in the above inequality yields

$$\frac{|\mathcal{I}_k^*|-|\mathcal{I}_{k+1}^*|}{\mu}+\Gamma_s\left[\gamma-k+|\mathcal{I}_k^*|+1\right]^{1/\alpha_s}-\sum_{i\in\mathcal{I}_k^*}X_{j_i}+\sum_{i\in\mathcal{I}_{k+1}^*}X_{j_i}\le\Gamma_s\left[\gamma-k+|\mathcal{I}_{k+1}^*|+1\right]^{1/\alpha_s}. \quad (28)$$

By Eq. (27) and using the characterization of $\widehat{X}^{\mathcal{P}}_{s(k)}$, Eq. (28) leads to the following upper bound on the service times

$$\widehat{X}^{\mathcal{P}}_{j_k}\le\frac{1}{\mu}+\Gamma_s\left[\left(\gamma-k+|\mathcal{I}_{k+1}^*|+1\right)^{1/\alpha_s}-\left(\gamma-k+|\mathcal{I}_{k+1}^*|\right)^{1/\alpha_s}\right],\ \forall\ j_k\in J_k. \quad (29)$$

(2) Moreover, as in Eq. (26), we have

$$\sum_{i=k+1}^{\gamma}\widehat{X}^{\mathcal{P}}_{s(i)}=\frac{\gamma-(k+1)+|\mathcal{I}_{k+1}^*|+1}{\mu}+\Gamma_s\left(\gamma-(k+1)+|\mathcal{I}_{k+1}^*|+1\right)^{1/\alpha_s}-\sum_{i\in\mathcal{I}_{k+1}^*}X_{j_i},$$

which simplifies to

$$\widehat{X}^{\mathcal{P}}_{s(k+1)}=\frac{\gamma-k+|\mathcal{I}_{k+1}^*|}{\mu}+\Gamma_s\left(\gamma-k+|\mathcal{I}_{k+1}^*|\right)^{1/\alpha_s}-\left(\sum_{i=k+2}^{\gamma}\widehat{X}^{\mathcal{P}}_{s(i)}+\sum_{i\in\mathcal{I}_{k+1}^*}X_{j_i}\right). \quad (30)$$

By Assumption 1(c), for $\{k+2,\ldots,\gamma\}\cup\mathcal{I}_{k+1}^*$, we obtain

$$\sum_{i=k+2}^{\gamma}\widehat{X}^{\mathcal{P}}_{s(i)}+\sum_{i\in\mathcal{I}_{k+1}^*}X_{j_i}\le\frac{\gamma-(k+1)+|\mathcal{I}_{k+1}^*|}{\mu}+\Gamma_s\left(\gamma-(k+1)+|\mathcal{I}_{k+1}^*|\right)^{1/\alpha_s}.$$

Applying the above bound to Eq. (30), we obtain

$$\widehat{X}^{\mathcal{P}}_{j_{k+1}}=\widehat{X}^{\mathcal{P}}_{s(k+1)}\ge\frac{1}{\mu}+\Gamma_s\left[\left(\gamma-(k+1)+|\mathcal{I}_{k+1}^*|+1\right)^{1/\alpha_s}-\left(\gamma-(k+1)+|\mathcal{I}_{k+1}^*|\right)^{1/\alpha_s}\right]. \quad (31)$$

Combining the bounds obtained in Eqs. (29) and (31), we obtain for all $k=0,\ldots,\gamma-1$

$$\begin{aligned}\widehat{X}_{j_k}&\le\frac{1}{\mu}+\Gamma_s\left[\left(\gamma-k+|\mathcal{I}_{k+1}^*|+1\right)^{1/\alpha_s}-\left(\gamma-k+|\mathcal{I}_{k+1}^*|\right)^{1/\alpha_s}\right]\\&\le\frac{1}{\mu}+\Gamma_s\left[\left(\gamma-(k+1)+|\mathcal{I}_{k+1}^*|+1\right)^{1/\alpha_s}-\left(\gamma-(k+1)+|\mathcal{I}_{k+1}^*|\right)^{1/\alpha_s}\right]\le\widehat{X}_{j_{k+1}},\end{aligned}$$

where the first and last inequalities are due to Eqs. (29) and (31), respectively, and the second inequality holds since the function $f(i) = (\nu - i + 1)^{1/\alpha_s} - (\nu - i)^{1/\alpha_s}$ is increasing in i. Hence,

$$\widehat{X}^{\mathcal{P}}_{j_0} \leq \widehat{X}^{\mathcal{P}}_{j_1} \leq \ldots \leq \widehat{X}^{\mathcal{P}}_{j_\gamma}.$$

By the construction in Eq. (27), we conclude that the sequence of service times is nondecreasing. This completes the proof. □

In the special case where $\ell = n$, Eq. (22) implies that the worst case system time for the n^{th} job under $\mathcal{P}$ can be written as

$$\widehat{S}^{\mathcal{P}}_n(\mathbf{T}) = \max_{0 \leq k \leq \nu} \left(\max_{\mathbf{X} \in \mathcal{U}^s_m} \sum_{i=k}^{\nu} X_{r(i)} - \sum_{i=r(k)+1}^{n} T_i \right), \tag{32}$$

where $r(i) = n - (\nu - i)m$. Additionally, there exists a nondecreasing sequence of service times that achieves the worst case value, such that

$$\widehat{X}^{\mathcal{P}}_{j_k} = \frac{1}{\mu} + \Gamma_s \left[(\nu - k + 1)^{1/\alpha_s} - (\nu - k)^{1/\alpha_s} \right], \quad \forall j_k \in J_k \quad \text{and} \quad k = 0, \ldots, \nu. \tag{33}$$

FCFS Behavior: We next relate the worst case behavior under $\mathcal{P}$ to the worst case behavior in a multi-server FCFS queue.

PROPOSITION 4. *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 is such that*

$$\widehat{S}_n(\mathbf{T}) = \widehat{S}^{\mathcal{P}}_n(\mathbf{T}) = \max_{0 \leq k \leq \nu} \left(\max_{\mathcal{U}^s_m} \sum_{i=k}^{\nu} X_{r(i)} - \sum_{i=r(k)+1}^{n} T_i \right), \tag{34}$$

where $r(i) = n - (\nu - i)m$ *and* $\nu = \lfloor (n-1)/m \rfloor$.

Proof of Proposition 4. 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

$$\begin{aligned} \widehat{S}_i(\mathbf{T}^i, \mathbf{X}^{i+}) = \max_{\mathbf{X}^i} \quad & S_i(\mathbf{T}^i, \mathbf{X}^i) \\ \text{s.t.} \quad & (\mathbf{X}^i, \mathbf{X}^{i+}) \in \mathcal{U}^s_m. \end{aligned} \tag{35}$$

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 $\widehat{X}_1 \leq \ldots \leq \widehat{X}_i$, 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, 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, yielding $\widehat{S}_i\left(\mathbf{T}^i, \mathbf{X}^{i+}\right) = \widehat{S}_i^{\mathcal{P}}\left(\mathbf{T}^i, \mathbf{X}^{i+}\right)$.

(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 $i \in J_0 = \{1, \ldots, m\}$. 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

$$\begin{aligned}\widehat{S}_n(\mathbf{T}) &= \max_{\mathbf{X}\in\mathcal{U}_m^s} \max\left(S_\ell + X_n - \sum_{j=\ell+1}^{n} T_j, X_n\right) \\ &\le \max\left(\max_{\mathbf{X}\in\mathcal{U}_m^s} S_\ell + X_n - \sum_{j=\ell+1}^{n} T_j, \max_{\mathbf{X}\in\mathcal{U}_m^s} X_n\right). \end{aligned} \tag{36}$$

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

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

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

$$S_\ell\left(\mathbf{T}^\ell, \widetilde{\mathbf{X}}^\ell\right) \le \widehat{S}_\ell\left(\mathbf{T}^\ell, \widetilde{\mathbf{X}}^{\ell+}\right) = \widehat{S}_\ell^{\mathcal{P}}\left(\mathbf{T}^\ell, \widetilde{\mathbf{X}}^{\ell+}\right).$$

Hence, we bound the expression in Eq. (36) by

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

where the second inequality expresses $\widehat{S}_\ell^{\mathcal{P}}\left(\mathbf{T}^\ell, \widetilde{\mathbf{X}}^{\ell+}\right)$ explicitly using Eq. (22). Rearranging the terms, and since $\left(\widehat{\mathbf{X}}^i, \widetilde{\mathbf{X}}^{i+}\right) \in \mathcal{U}_m^s$, we obtain

$$\begin{aligned}\widehat{S}_n(\mathbf{T}) &\le \max\left\{\max_{0\le k\le\gamma}\left(\sum_{i=k}^{\gamma} \widehat{X}_{s(i)} + \widetilde{X}_n - \sum_{i=s(k)+1}^{\ell} T_i - \sum_{i=\ell+1}^{n} T_i\right),\ \max_{\mathcal{U}_m^s} X_n\right\} \\ &\le \max\left\{\max_{0\le k\le\gamma}\left(\max_{\mathcal{U}_m^s}\left\{\sum_{i=k}^{\gamma} X_{s(i)} + X_n\right\} - \sum_{i=s(k)+1}^{n} T_i\right),\ \max_{\mathcal{U}_m^s} X_n\right\}.\end{aligned} \tag{37}$$

Recall that $s(k) = \ell - (\gamma - k)m \in J_k$. 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 $n-\ell \ge m$. Consider the partition $J_0, J_1, \ldots, J_\nu$ 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 $J_0, J_1, \ldots, J_\nu$. With $\ell \in J_\gamma$, and $n \in J_\nu$, this implies $\nu \ge \gamma+1$. We consider the following two cases.

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

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

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

$$\max_{\mathcal{U}_m^s} \left\{ \sum_{i=k}^{\gamma} X_{s(i)} + X_n \right\} = \max_{\mathcal{U}_m^s} \left\{ \sum_{i=k}^{\nu} X_{s(i)} \right\}. \tag{38}$$

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

$$\sum_{i=s(k)+1}^{n} T_i \ge \sum_{i=r(k)+1}^{n} T_i, \text{ for all } 0 \le k \le \gamma. \tag{39}$$

Combining Eqs. (38) and (39), Eq. (37) becomes

$$\widehat{S}_n(\mathbf{T}) \le \max \left\{ \max_{0 \le k \le \nu-1} \left(\max_{\mathcal{U}_m^s} \sum_{i=k}^{\nu} X_{r(i)} - \sum_{i=r(k)+1}^{n} T_i \right), \max_{\mathcal{U}_m^s} X_n \right\}. \tag{40}$$

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

$$\max_{\mathcal{U}_m^s} \left\{ \sum_{i=k}^{\gamma} X_{s(i)} + X_n \right\} = \max_{\mathcal{U}_m^s} \left\{ \sum_{i=k+1}^{\gamma+1} X_{r(i)} + X_n \right\} \le \max_{\mathcal{U}_m^s} \left\{ \sum_{i=k+1}^{\nu} X_{r(i)} \right\}. \tag{41}$$

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

$$\sum_{i=s(k)+1}^{n} T_i \ge \sum_{i=r(k+1)+1}^{n} T_i, \text{ for all } 0 \le k \le \gamma. \tag{42}$$

Applying the bounds in Eqs. (41) and (42), Eq. (37) becomes

$$\begin{aligned}\widehat{S}_n(\mathbf{T}) &\le \max\left\{\max_{0\le k\le\gamma}\left(\max_{\mathcal{U}_m^s}\sum_{i=k+1}^{\nu}X_{r(i)}-\sum_{i=r(k+1)+1}^{n}T_i\right),\ \max_{\mathcal{U}_m^s}X_n\right\}\\ &= \max\left\{\max_{1\le k\le\gamma+1}\left(\max_{\mathcal{U}_m^s}\sum_{i=k}^{\nu}X_{r(i)}-\sum_{i=r(k)+1}^{n}T_i\right),\ \max_{\mathcal{U}_m^s}X_n\right\}. \qquad (43)\end{aligned}$$

Since $\nu \ge \gamma+2$, we can further bound Eq. (43) to obtain

$$\widehat{S}_n(\mathbf{T}) \le \max\left\{\max_{0\le k\le\nu-1}\left(\max_{\mathcal{U}_m^s}\sum_{i=k}^{\nu}X_{r(i)}-\sum_{i=r(k)+1}^{n}T_i\right),\ \max_{\mathcal{U}_m^s}X_n\right\}. \qquad (44)$$

Combining the results in Eqs. (40) and (44) from cases (1) and (2), we conclude that the worst case system time under FCFS is bounded by the worst case system time under $\mathcal{P}$, i.e.,

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

This bound is in fact tight and can be achieved under a scenario where the service times are chosen such that $\left(\widehat{X}_1,\ldots,\widehat{X}_n\right)=\left(\widehat{X}_1^{\mathcal{P}},\ldots,\widehat{X}_n^{\mathcal{P}}\right)\in\mathcal{U}_m^s$ (see Eq. (33)). Note that this optimal solution consists of nondecreasing service times, hence proving the inductive hypothesis. □

Given Propositions 3 and 4, the worst case system time of the n^{th} job is given by

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

where $r(i)=n-(\nu-i)m$. The above bound is in fact tight, as it can be achieved for the sequence of interarrivals presented in Eq. (11). As a result, by applying Assumption 1, we obtain an exact characterization of the worst case system time as

$$\widehat{S}_n=\max_{0\le k\le\nu}\left\{\frac{\nu-k+1}{\mu}+\Gamma_s(\nu-k+1)^{1/\alpha_s}-\frac{m(\nu-k)}{\lambda}+\Gamma_a[m(\nu-k)]^{1/\alpha_a}\right\}. \qquad (45)$$

The worst case performance analysis problem reduces to a one-dimensional nonlinear optimization problem, which can be solved efficiently. Theorem 3 provides a closed form upper bound on the worst case system time for the case where $\alpha_a=\alpha_s=\alpha$.

THEOREM 3 **(Worst Case System Time in a Multi-Server FCFS Queue).**

In an m-server FCFS queue with $\mathbf{T} \in \mathcal{U}^a$, $\mathbf{X} \in \mathcal{U}^s_m$, $\alpha_a = \alpha_s = \alpha$ and $\rho < 1$,

$$\widehat{S}_n \leq \frac{\alpha - 1}{\alpha^{\alpha/(\alpha-1)}} \cdot \frac{\lambda^{1/(\alpha-1)} \left(\Gamma_a + \Gamma_s/m^{1/\alpha}\right)^{\alpha/(\alpha-1)}}{(1-\rho)^{1/(\alpha-1)}} + \frac{m}{\lambda}, \tag{46}$$

Proof of Theorem 3. The maximization problem in Eq. (45) can be written in the same form as in Eq. (15) by substituting $x = \nu - j + 1$, with $\beta = m^{1/\alpha}\Gamma_a + \Gamma_s > 0$ and $\delta = m(1-\rho)/\lambda > 0$, given $\rho < 1$. Substituting β and δ by their respective values in Eq. (15) yields the desired bound. □

Similarly to the single-server queue, the closed form bound on the system time is nearly tight for heavy-traffic systems operating in steady state (please see the appendix for details).

3.3. Implications and Insights

To summarize, we obtain closed form upper bounds on the system time in an FCFS queue, with

$$\widehat{S}_n \leq \begin{cases} \dfrac{\alpha - 1}{\alpha^{\alpha/(\alpha-1)}} \cdot \dfrac{\lambda^{1/(\alpha-1)} \left(\Gamma_a + \Gamma_s\right)^{\alpha/(\alpha-1)}}{(1-\rho)^{1/(\alpha-1)}} + \dfrac{1}{\lambda} & \text{(single-server queue)} \\ \dfrac{\alpha - 1}{\alpha^{\alpha/(\alpha-1)}} \cdot \dfrac{\lambda^{1/(\alpha-1)} \left(\Gamma_a + \Gamma_s/m^{1/\alpha}\right)^{\alpha/(\alpha-1)}}{(1-\rho)^{1/(\alpha-1)}} + \dfrac{m}{\lambda} & \text{(multi-server queue)} \end{cases}$$

These bounds are nearly tight for heavy-traffic systems operating under steady-state. We present next the implications and insights that follow from the analysis.

(a) Qualitative Insights: *Our approach leads to the same qualitative conclusions as stochastic queueing theory* with respect to the behavior of the system time in terms of the traffic intensity and uncertainty on the inter-arrival and service times. In fact, the classical i.i.d. arrival and service processes with finite variance can be modeled by setting $\alpha = 2$. Eq. (13) becomes

$$\widehat{S}_n \leq \frac{\lambda}{4} \cdot \frac{\left(\Gamma_a + \Gamma_s\right)^2}{1-\rho} + \frac{1}{\lambda} \quad \text{and} \quad \widehat{S}_n \leq \frac{\lambda}{4} \cdot \frac{\left(\Gamma_a + \Gamma_s/m^{1/2}\right)^2}{1-\rho} + \frac{m}{\lambda}, \tag{47}$$

for single server and multi-server queues, respectively. Kingman (1970) provides insightful bounds on the expected waiting time in steady state for the $GI/GI/1$ and $GI/GI/m$ queues. Given that $\mathbb{E}[S_n] = \mathbb{E}[W_n] + \mathbb{E}[X_n]$, where $\mathbb{E}[X_n] = 1/\mu$, the bounds on the expected system times translate to

$$\mathbb{E}[S_n] \leq \frac{\lambda}{2} \cdot \frac{\sigma_a^2 + \sigma_s^2}{1-\rho} + \frac{1}{\mu} \quad \text{and} \quad \mathbb{E}[S_n] \leq \frac{\lambda}{2} \cdot \frac{\sigma_a^2 + \sigma_s^2/m + (1/m - 1/m^2)/\mu^2}{1-\rho} + \frac{1}{\mu}. \tag{48}$$

The bounds in the proposed framework share the same functional dependence on $\lambda/(1-\rho)$ and on the variability parameters Γ_a^2, Γ_s^2/m, (correspondingly σ_a^2, σ_s^2/m) as probabilistic bounds. Note that the bounds in Eq. (47) depend on the magnitude of the variability parameters.

(b) Heavy Tails Behavior: Our approach allows a closed-form expression for the steady-state system time for all values of $\alpha \in (1,2)$, which include heavy tailed random variables. We observe that heavier the tail, i.e., the smaller the tail coefficient α, the higher the order of the waiting time and the system time, given its dependence on $1/(1-\rho)^{1/(\alpha-1)}$. To illustrate, a decrease in the tail coefficient from $\alpha = 2$ to $\alpha = 1.5$ increases the waiting time by one order of magnitude. This is in agreement with the stochastic queueing theory literature, where it is known that the waiting time exhibits a heavy-tailed distribution under heavy tailed services (see Whitt (2000), Crovella (2000)).

4. The Departure Process with Adversarial Servers

In this section, we study the output of a single queue under the assumption that servers act adversarially to maximize the time spent in the queue. Specifically, we show that, with adversarial servers, the inter-departure times $\mathbf{D} = \{D_1, D_2, \ldots, D_n\}$ belong to the arrival uncertainty set $\mathcal{U}^a$. The characterization of the departure uncertainty set $\mathcal{U}^d$ as a subset of the arrival uncertainty set $\mathcal{U}^a$ is increasingly tighter with larger values of n. This result is akin to the Burke theorem and forms the cornerstone of our network analysis.

4.1. Adversarial Servers

Fixing the value of n, we view the queueing system from an adversarial perspective, where the servers act so as to maximize the system time of the n^{th} job, for all possible sequences of inter-arrival times. This assumption is reminiscent of the service curves approach of the stochastic network calculus, see Jiang and Liu (2008). In other words, the servers choose their adversarial service times $\widehat{\mathbf{X}} = \left(\widehat{X}_1, \ldots, \widehat{X}_n\right)$ to achieve $\widehat{S}_n(\mathbf{T})$, for all $\mathbf{T}$. Given the results of Propositions 1, and 4, the servers choose their service times according to Eqs. (10) and (33), respectively, i.e.,

$$\widehat{X}_i = \frac{1}{\mu} + \Gamma_s \left[(n-i+1)^{1/\alpha_s} - (n-i)^{1/\alpha_s} \right], \quad \text{for all } i = 1, \ldots, n. \tag{49}$$

$$\widehat{X}_{j_i} = \frac{1}{\mu} + \Gamma_s \left[(\nu-i+1)^{1/\alpha_s} - (\nu-i)^{1/\alpha_s} \right], \quad \text{for all } j_i \in J_i \text{ and } i = 0, \ldots, \nu. \tag{50}$$

and achieve the worst case system time

$$\widehat{S}_n(\mathbf{T}) = \begin{cases} \max\limits_{1\le k\le n} \left(\max\limits_{\mathbf{X}\in\mathcal{U}^s} \sum\limits_{i=k}^{n} X_i - \sum\limits_{i=k+1}^{n} T_i \right) = \max\limits_{1\le k\le n} \left(\sum\limits_{i=k}^{n} \widehat{X}_i - \sum\limits_{i=k+1}^{n} T_i \right), \\ \max\limits_{0\le k\le \nu} \left(\max\limits_{\mathcal{U}_m^s} \sum\limits_{i=k}^{\nu} X_{r(i)} - \sum\limits_{i=r(k)+1}^{n} T_i \right) = \max\limits_{0\le k\le \nu} \left(\sum\limits_{i=k}^{\nu} \widehat{X}_{r(i)} - \sum\limits_{i=r(k)+1}^{n} T_i \right), \end{cases} \tag{51}$$

for all $\mathbf{T}$, for single-server and m-server queues, respectively. Note that the adversarial service times are nondecreasing, implying $\widehat{X}_1 \le \widehat{X}_2 \le \ldots \le \widehat{X}_n$. In a multi-server setting, the monotonicity of the adversarial service times ensures no overtaking can occur, and as a result, jobs leave in the same order of their arrival. We note that the adversarial service times depend on the value of n, i.e., $\widehat{\mathbf{X}} = \widehat{\mathbf{X}}^{(n)}$. We dropped the superscript n in our analysis, for ease of notation. We next study the departure process in a multi-server queue with adversarial servers.

4.2. Departure Times

For a multi-server queue, the time between the k^{th} and n^{th} departures is the difference between $C_{(n)}$ and $C_{(k)}$. Assuming servers act adversarially, no overtaking is allowed to occur. As a result, the k^{th} and n^{th} departures correspond to the k^{th} and n^{th} jobs, respectively. In this case,

$$\sum_{i=k+1}^{n} D_i = C_{(n)} - C_{(k)} = C_n - C_k = A_n + \widehat{S}_n(\mathbf{T}) - A_k - \widehat{S}_k(\mathbf{T}) = \sum_{i=k+1}^{n} T_i + \widehat{S}_n(\mathbf{T}) - \widehat{S}_k(\mathbf{T}). \tag{52}$$

Characterizing the exact departure uncertainty set in an queue with adversarial servers can be made via minimizing Eq. (52) with respect to $\mathbf{T} \in \mathcal{U}^a$, for all $1 \le k \le n-1$. Theorem 4 obtains a lower bound over these minimization problems

$$\sum_{i=k+1}^{n} D_i \ge \frac{n-k}{\lambda} - \Gamma_a (n-k)^{1/\alpha}, \quad \text{for all } 0 \le k \le n-1,$$

implying that, in an adversarial setting, the departure times belong to the arrival uncertainty set.

THEOREM 4. **(Passing through a Queue With Adversarial Servers)**

For a multi-server queue with inter-arrival times $\mathbf{T} \in \mathcal{U}^a$*, adversarial service times* $\widehat{\mathbf{X}}$*, and* $\rho < 1$*, the inter-departure times* $\mathbf{D} = \{D_1, D_2, \ldots, D_n\}$ *belongs to the set* $\mathcal{U}^d$ *satisfying*

$$\mathcal{U}^d \subseteq \mathcal{U}^a = \left\{ (D_1, D_2, \ldots, D_n) \left| \frac{\sum_{i=k+1}^{n} D_i - \frac{n-k}{\lambda}}{(n-k)^{1/\alpha_a}} \geq -\Gamma_a, \ \forall\ 0 \leq k \leq n-1 \right. \right\}. \tag{53}$$

Proof of Theorem 4. We note that, for $k = 0$, Eq. (52) results in $C_n \geq A_n$, yielding the desired bound. In the remainder of this proof we assume $k \geq 1$. We first consider the case of a single-server queue which illustrates the main intuition of the proof.

Single-Server Queue. In a single-server queue with adversarial servers, we can express the system time of the k^{th} job as

$$\begin{aligned}
\widehat{S}_k(\mathbf{T}) &= \max_{1 \leq j \leq k} \left(\sum_{i=j}^{k} \widehat{X}_i - \sum_{i=j+1}^{k} T_i \right) = \max_{1 \leq j \leq k} \left(\sum_{i=j}^{n} \widehat{X}_i - \sum_{i=k+1}^{n} \widehat{X}_i - \sum_{i=j+1}^{n} T_i + \sum_{i=k+1}^{n} T_i \right) \\
&= \sum_{i=k+1}^{n} T_i - \sum_{i=k+1}^{n} \widehat{X}_i + \max_{1 \leq j \leq k} \left(\sum_{i=j}^{n} \widehat{X}_i - \sum_{i=j+1}^{n} T_i \right),
\end{aligned}$$

where we obtain the last equality by extracting the partial sums that are independent of the index j out of the maximum term. Eq. (52) therefore becomes

$$\sum_{i=k+1}^{n} D_i = \sum_{i=k+1}^{n} \widehat{X}_i + \widehat{S}_n(\mathbf{T}) - \max_{1 \leq j \leq k} \left(\sum_{i=j}^{n} \widehat{X}_i - \sum_{i=j+1}^{n} T_i \right). \tag{54}$$

We next consider the following two cases and analyze them separately:

$$\textbf{Case 1.} \sum_{i=k+1}^{n} \widehat{X}_i \geq \frac{n-k}{\lambda} - \Gamma_a (n-k)^{1/\alpha_a}.$$

$$\textbf{Case 2.} \sum_{i=k+1}^{n} \widehat{X}_i < \frac{n-k}{\lambda} - \Gamma_a (n-k)^{1/\alpha_a}.$$

Case 1. In a single-server queue, we note that for $k \leq n$, we have

$$\max_{1 \leq j \leq k} \left(\sum_{i=j}^{n} \widehat{X}_i - \sum_{i=j+1}^{n} T_i \right) \leq \max_{1 \leq j \leq n} \left(\sum_{i=j}^{n} \widehat{X}_i - \sum_{i=j+1}^{n} T_i \right) = \widehat{S}_n(\mathbf{T}).$$

This results in the partial sum of inter-departure times to be lower bounded by the partial sum of service times, and given the assumption in Case (a), we obtain

$$\sum_{i=k+1}^{n} D_i \geq \sum_{i=k+1}^{n} \widehat{X}_i + \widehat{S}_n(\mathbf{T}) - \widehat{S}_n(\mathbf{T}) = \sum_{i=k+1}^{n} \widehat{X}_i \geq \frac{n-k}{\lambda} - \Gamma_a (n-k)^{1/\alpha}.$$

Case 2. For a single-server queue, we can bound the maximum term in Eq. (59) by

$$\max_{1\leq j\leq k} \left(\sum_{i=j}^{n} \widehat{X}_i - \sum_{i=j+1}^{n} T_i \right) \leq \widehat{X}_k + \max_{1\leq j\leq k} \left(\sum_{i=j+1}^{n} \widehat{X}_i - \sum_{i=j+1}^{n} T_i \right),$$

where the inequality is due to $\widehat{X}_j \leq \widehat{X}_k$ for $j \leq k$, since the adversarial service times are nondecreasing. Given that $\widehat{S}_n(\mathbf{T}) \geq \widehat{X}_n \geq \widehat{X}_k$, the partial sum of inter-departure times in Eq. (59) is then lower-bounded by

$$\sum_{i=k+1}^{n} D_i \geq \sum_{i=k+1}^{n} \widehat{X}_i - \max_{1\leq j\leq k} \left(\sum_{i=j+1}^{n} \widehat{X}_i - \sum_{i=j+1}^{n} T_i \right). \tag{55}$$

Substituting the value of the adversarial service times and upper bounding the partial sum of inter-arrival times according to Assumption 1(a),

$$\max_{1\leq j\leq k} \left(\sum_{i=j+1}^{n} \widehat{X}_i - \sum_{i=j+1}^{n} T_i \right) \leq \max_{1\leq j\leq k} g(n-j),$$

where the function $g(\cdot)$ is such that

$$g(x) = \frac{x}{\mu} + \Gamma_s \cdot x^{1/\alpha_s} - \frac{x}{\lambda} + \Gamma_a \cdot x^{1/\alpha_a}. \tag{56}$$

The function $g(\cdot)$ is concave, monotonically increasing from zero to a positive maximum value after which it becomes monotonically decreasing. Negative function values belong to the phase where the function is decreasing. The assumption of Case (b) translates to

$$\sum_{i=k+1}^{n} \widehat{X}_i = \frac{n-k}{\mu} + \Gamma_s (n-k)^{1/\alpha} < \frac{n-k}{\lambda} - \Gamma_a (n-k)^{1/\alpha_a}, \text{ implying that } \; g(n-k) < 0.$$

Since $g(n-k) < 0$, the function $g(\cdot)$ is decreasing. Therefore, for $j \leq k$, i.e., $n-j \geq n-k$, we have $g(n-j) \leq g(n-k)$, yielding

$$\max_{1\leq j\leq k} \left(\sum_{i=j+1}^{n} \widehat{X}_i - \sum_{i=j+1}^{n} T_i \right) \leq \max_{1\leq j\leq k} g(n-j) = g(n-k). \tag{57}$$

Applying the bound obtained in Eq. (57) to Eq. (55), we obtain

$$\begin{aligned}\sum_{i=k+1}^{n} D_i &\geq \sum_{i=k+1}^{n} \widehat{X}_i - \frac{n-k}{\mu} - \Gamma_s (n-k)^{1/\alpha_s} + \frac{n-k}{\lambda} - \Gamma_a (n-k)^{1/\alpha_a} \\ &= \frac{n-k}{\lambda} - \Gamma_a \left(n-k\right)^{1/\alpha_a}.\end{aligned}$$

We now extend the proof to the more complex case of a multi-server queue.

Multi-Server Queue. Suppose $k \in J_\gamma$. With adversarial service times and by Eq. (51),

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

where $s(i) = k - (\gamma - i)m$. We analyze the cases where $\gamma \leq \nu - 1$ and $\gamma = \nu$ separately.

(a) Suppose that $\gamma \leq \nu - 1$. Rewriting the partial sums in terms of $\nu - 1$ and n, we obtain

$$\begin{aligned}\widehat{S}_k\left(\mathbf{T}\right) &= \max_{0 \leq j \leq \gamma} \left(\sum_{i=j}^{\nu-1} \widehat{X}_{s(i)} - \sum_{i=\gamma+1}^{\nu-1} \widehat{X}_{s(i)} - \sum_{i=s(j)+1}^{n} T_i + \sum_{i=k+1}^{n} T_i \right) \\ &= \sum_{i=k+1}^{n} T_i - \sum_{i=\gamma+1}^{\nu-1} \widehat{X}_{s(i)} + \max_{0 \leq j \leq \gamma} \left(\sum_{i=j}^{\nu-1} \widehat{X}_{s(i)} - \sum_{i=s(j)+1}^{n} T_i \right). \end{aligned} \tag{58}$$

By replacing the system time $\widehat{S}_k\left(\mathbf{T}\right)$ in Eq. (52) by its value from Eq. (58), the bound on the sum of inter-departure times becomes

$$\sum_{i=k+1}^{n} D_i \geq \sum_{i=\gamma+1}^{\nu-1} \widehat{X}_{s(i)} + \widehat{S}_n\left(\mathbf{T}\right) - \max_{0 \leq j \leq \gamma} \left(\sum_{i=j}^{\nu-1} \widehat{X}_{s(i)} - \sum_{i=s(j)+1}^{n} T_i \right). \tag{59}$$

We consider the following two cases

$$\textbf{Case 1.} \quad \sum_{i=\gamma+1}^{\nu-1} \widehat{X}_{s(i)} \geq \frac{n-k}{\lambda} - \Gamma_a (n-k)^{1/\alpha}.$$

$$\textbf{Case 2.} \quad \sum_{i=\gamma+1}^{\nu-1} \widehat{X}_{s(i)} < \frac{n-k}{\lambda} - \Gamma_a (n-k)^{1/\alpha}.$$

Case 1. Since $s(i) \in J_i$ and $r(i+1) \in J_{i+1}$, we have $s(i) < r(i+1)$ for all $i = 0, \ldots, \nu - 1$. By the monotonicity of the adversarial service times, we have $\widehat{X}_{s(i)} \leq \widehat{X}_{r(i+1)}$, and

$$\sum_{i=s(j)+1}^{n} T_i \geq \sum_{i=r(j+1)+1}^{n} T_i,$$

for all $0 \le i, j \le \gamma \le \nu - 1$. Hence, we can bound the maximum term in Eq. (59) by

$$\max_{0\le j\le\gamma}\left(\sum_{i=j}^{\nu-1}\widehat{X}_{s(i)} - \sum_{i=s(j)+1}^{n} T_i\right) \le \max_{0\le j\le\gamma}\left(\sum_{i=j}^{\nu-1}\widehat{X}_{r(i+1)} - \sum_{i=r(j+1)+1}^{n} T_i\right)$$
$$= \max_{1\le j\le\gamma+1}\left(\sum_{i=j}^{\nu}\widehat{X}_{r(i)} - \sum_{i=r(j)+1}^{n} T_i\right). \tag{60}$$

Since $\gamma \le \nu - 1$, then $\gamma + 1 \le \nu$, and we can further bound Eq. (60) to obtain

$$\max_{0\le j\le\gamma}\left(\sum_{i=j}^{\nu-1}\widehat{X}_{s(i)} - \sum_{i=s(j)+1}^{n} T_i\right) \le \max_{0\le j\le\nu}\left(\sum_{i=j}^{\nu}\widehat{X}_{r(i)} - \sum_{i=r(j)+1}^{n} T_i\right) = \widehat{S}_n(\mathbf{T}), \tag{61}$$

where the last equality is due to Eq. (51). Applying the bound in Eq. (61) to Eq. (59), and given the assumption in Case 1.,

$$\sum_{i=k+1}^{n} D_i \ge \sum_{i=\gamma+1}^{\nu-1}\widehat{X}_{r(i)} + \widehat{S}_n(\mathbf{T}) - \widehat{S}_n(\mathbf{T}) = \sum_{i=\gamma+1}^{\nu-1}\widehat{X}_{r(i)} \ge \frac{n-k}{\lambda} - \Gamma_a(n-k)^{1/\alpha}.$$

Case 2. Since $\widehat{S}_n(\mathbf{T}) \ge 0$, Eq. (59) becomes

$$\sum_{i=k+1}^{n} D_i \ge \sum_{i=\gamma+1}^{\nu-1}\widehat{X}_{s(i)} - \max_{0\le j\le\gamma}\left(\sum_{i=j}^{\nu-1}\widehat{X}_{s(i)} - \sum_{i=s(j)+1}^{n} T_i\right).$$

By substituting the values of the adversarial service times and bounding the sum of inter-arrival times by Assumption 1(a), the maximum term in the above equation can be upper bounded by

$$\max_{0\le j\le\gamma}\left(\sum_{i=j}^{\nu-1}\widehat{X}_{s(i)} - \sum_{i=s(j)+1}^{n} T_i\right) \le \max_{0\le j\le\gamma} h(\nu - j), \tag{62}$$

where the function $h(\cdot)$ is such that

$$h(x) = \frac{x}{\mu} + \Gamma_s \cdot x^{1/\alpha_s} - \frac{m\cdot x + c}{\lambda} + \Gamma_a \cdot (m\cdot x + c)^{1/\alpha_a}, \tag{63}$$

and c is a constant with $c = (n - \nu m) - (k - \gamma m)$. The function $h(\cdot)$ is concave, monotonically increasing to some positive maximum value, after which it becomes monotonically decreasing. Negative function values belong to the phase where $h(\cdot)$ is decreasing. Note that, since $n = r(n) = n - (\nu - \nu)m$ and $k = s(\gamma) = k - (\gamma - \gamma)m$, we can write

$$n - k = r(\nu) - s(\gamma) = [n - (\nu - \nu)m] - [k - (\gamma - \gamma)m] = m\cdot(\nu - \gamma) + c.$$

As a result, the assumption of Case 2. translates to

$$\sum_{i=\gamma+1}^{\nu-1} \widehat{X}_{s(i)} = \frac{\nu-\gamma}{\mu} + \Gamma_s (\nu-\gamma)^{1/\alpha_s} < \frac{n-k}{\lambda} - \Gamma_a (n-k)^{1/\alpha_a}$$
$$= \frac{m\cdot(\nu-\gamma)+c}{\lambda} - \Gamma_a (m\cdot(\nu-\gamma)+c)^{1/\alpha_a},$$

implying $h(\nu-\gamma)<0$, and the function $h(\cdot)$ is decreasing beyond $\nu-\gamma$. For $j\le\gamma$, we have $\nu-j\ge\nu-\gamma$, and since $h(\cdot)$ is decreasing beyond $\nu-\gamma$, we obtain $h(\nu-j)\le h(\nu-\gamma)$. Therefore the bound in Eq. (62) becomes

$$\max_{0\le j\le\gamma}\left(\sum_{i=j}^{\nu-1}\widehat{X}_{s(i)} - \sum_{i=s(j)+1}^{n} T_i\right) \le \max_{0\le j\le\gamma} h(\nu-j) = h(\nu-\gamma).$$

Given the values of the adversarial service times and the fact that $n-k=m\cdot(\nu-\gamma)+c$,

$$h(\nu-\gamma) = \frac{\nu-\gamma}{\mu} + \Gamma_s(\nu-\gamma)^{1/\alpha_s} - \frac{m\cdot(\nu-\gamma)+c}{\lambda} + \Gamma_a(m\cdot(\nu-\gamma)+c)^{1/\alpha_a} \quad (64)$$
$$= \sum_{i=\gamma+1}^{\nu-1}\widehat{X}_{s(i)} - \frac{n-k}{\lambda} + \Gamma_a(n-k)^{1/\alpha_a}. \quad (65)$$

As a result, the bound in Eq. (59) becomes

$$\sum_{i=k+1}^{n} D_i \ge \sum_{i=\gamma+1}^{\nu}\widehat{X}_{r(i)} - h(\nu-\gamma) = \frac{n-k}{\lambda} - \Gamma_a(n-k)^{1/\alpha}.$$

(b) Suppose that $\gamma=\nu$, i.e. $k,n\in J_\nu$. Rewriting the partial sums in terms of ν and n, we obtain

$$\widehat{S}_k(\mathbf{T}) = \max_{0\le j\le\nu}\left(\sum_{i=j}^{\nu}\widehat{X}_{s(i)} - \sum_{i=s(j)+1}^{n} T_i + \sum_{i=k+1}^{n} T_i\right)$$
$$= \sum_{i=k+1}^{n} T_i + \max_{0\le j\le\nu}\left(\sum_{i=j}^{\nu}\widehat{X}_{s(i)} - \sum_{i=s(j)+1}^{n} T_i\right). \quad (66)$$

By replacing the system time $\widehat{S}_k(\mathbf{T})$ in Eq. (52) by its value from Eq. (66), the bound on the sum of inter-departure times becomes

$$\sum_{i=k+1}^{n} D_i \ge \widehat{S}_n(\mathbf{T}) - \max_{0\le j\le\nu}\left(\sum_{i=j}^{\nu}\widehat{X}_{s(i)} - \sum_{i=s(j)+1}^{n} T_i\right). \quad (67)$$

We consider the following two cases

Case 1. $0 \ge \dfrac{n-k}{\lambda} - \Gamma_a(n-k)^{1/\alpha}.$

Case 2. $0 < \dfrac{n-k}{\lambda} - \Gamma_a(n-k)^{1/\alpha}.$

Case 1. Under the assumption of Case 1., and since the inter-departure times are non-negative,

$$\sum_{i=k+1}^{n} D_i \geq 0 \geq \frac{n-k}{\lambda} - \Gamma_a (n-k)^{1/\alpha}.$$

Case 2. Given that $k = s(\nu)$, the maximum term in Eq. (67) can be rewritten as

$$\begin{aligned}\max_{0\leq j\leq \nu}\left(\sum_{i=j}^{\nu}\widehat{X}_{s(i)} - \sum_{i=s(j)+1}^{n} T_i\right) &= \max_{0\leq j\leq \nu}\left(\widehat{X}_{s(\nu)} + \sum_{i=j}^{\nu-1}\widehat{X}_{s(i)} - \sum_{i=s(j)+1}^{n} T_i\right)\\ &= \widehat{X}_k + \max_{0\leq j\leq \nu}\left(\sum_{i=j}^{\nu-1}\widehat{X}_{s(i)} - \sum_{i=s(j)+1}^{n} T_i\right). \end{aligned} \tag{68}$$

Using Eq. (68), and since $\widehat{S}_n(\mathbf{T}) \geq \widehat{X}_n \geq \widehat{X}_k$, by the monotonicity of the adversarial service times, Eq. (67) becomes

$$\begin{aligned}\sum_{i=k+1}^{n} D_i &\geq \widehat{S}_n(\mathbf{T}) - \widehat{X}_k - \max_{0\leq j\leq \nu}\left(\sum_{i=j}^{\nu-1}\widehat{X}_{s(i)} - \sum_{i=s(j)+1}^{n} T_i\right)\\ &\geq -\max_{0\leq j\leq \nu}\left(\sum_{i=j}^{\nu-1}\widehat{X}_{s(i)} - \sum_{i=s(j)+1}^{n} T_i\right) = -\max_{0\leq j\leq \nu} h(\nu - j), \end{aligned} \tag{69}$$

where the function $h(\cdot)$ is defined in Eq. (63). Note that, since $\gamma = \nu$, we obtain $n - k = c$. As a result, the assumption of Case 2. translates to

$$0 < \frac{n-k}{\mu} - \Gamma_a (n-k)^{1/\alpha_a} = \frac{c}{\lambda} - \Gamma_a \cdot c^{1/\alpha_a} = -h(0),$$

implying $h(0) < 0$, and the function is decreasing beyond 0. For $j \leq \nu$, we have $\nu - j \geq 0$, and since $h(\cdot)$ is decreasing beyond 0, we obtain $h(\nu - j) \leq h(0)$. Therefore the bound in Eq. (69) becomes

$$\sum_{i=k+1}^{n} D_i \geq -\max_{0\leq j\leq \nu} h(\nu - j) = -h(0) = \frac{n-k}{\lambda} - \Gamma_a (n-k)^{1/\alpha_a}.$$

This completes the proof. □

4.3. Implications and Insights

We present next the implications and insights that follow from the analysis of the departure times for queues with adversarial servers.

(a) Tightness of the Departure Characterization: The characterization $\mathcal{U}^d \subseteq \mathcal{U}^a$ is true for all values of n, though its tightness improves for increasing values of n. In other words, in a queue with adversarial servers, the inequality

$$\min_{\mathbf{T}\in\mathcal{U}^a} \sum_{i=k+1}^{n} D_i \geq \frac{n-k}{\lambda} - \Gamma_a (n-k)^{1/\alpha}$$

becomes tighter as n increases. To illustrate this point, Figure 1 shows the percent error between the left hand side and the right hand ride of the above inequality for various values of k and n. We note that, the higher the value of n, the lower the error is for all k values.

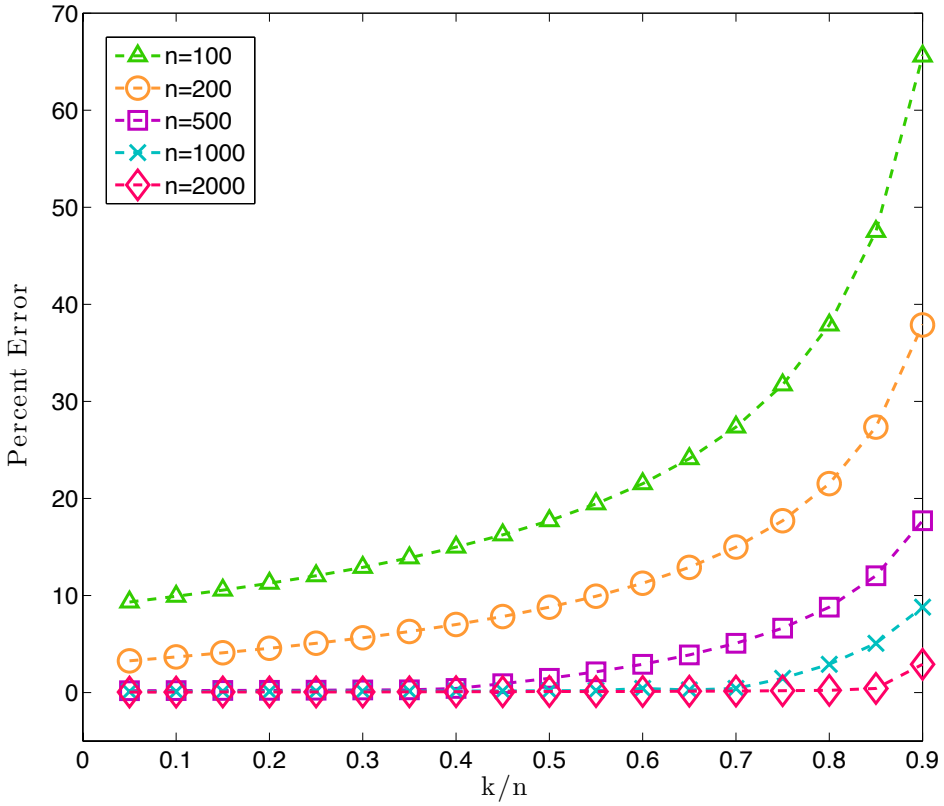

Figure 1 Percent error values generated by comparing the minimum value of the sum $\sum_{i=k+1}^{n} D_i$ (computed numerically by an optimization solver) and the expression $\frac{n-k}{\lambda} - \Gamma_a (n-k)^{1/\alpha}$ for various values of k and n. The instance shown corresponds to a single-server queue with adversarial servers, traffic intensity $\rho = 0.9$, service rate $\mu = 1$, variability parameters $\Gamma_a = \Gamma_s = 1$, and tail coefficient $\alpha = 2$.

(b) Robust Burke Theorem: Asymptotically, the characterization of the departure process in Theorem 4 is tight, which implies that the departure uncertainty set is therefore approximately

equal to the arrival uncertainty set for large values of n. This is akin to the Burke Theorem from the stochastic queueing literature, which states that, asymptotically, the departure process in an $M/M/m$ queue is a Poisson process with a rate equal to that of the arrival process. By looking at asymptotics, Theorem 4 can be thought of as a generalization of the Burke's theorem to more general setting such as heavy-tailed behavior. This result allows us to decompose a network of queues with adversarial servers and provides the cornerstone of our network analysis, as we shall cover next.

5. The Robust Queueing Network Analyzer

Consider a network of J queues serving a single class of jobs. Each job enters the network through some queue j, and either leaves the network or departs towards another queue right after completion of his service. The primitive data in the queueing network are:

(a) External arrival processes with parameters $(\lambda_j, \Gamma_{a,j}, \alpha_{a,j})$ that arrive to each node $j = 1, \ldots, J$.

(b) Service processes with parameters $(\mu_j, \Gamma_{s,j}, \alpha_{s,j})$, and the number of servers $m_j,\ j = 1, \ldots, J$.

(c) Routing matrix $\mathbf{F} = [f_{ij}]$, $i, j = 1, \ldots, J$, where f_{ij} denotes the fraction of jobs passing through queue i and are routed to queue j. The fraction of jobs leaving the network from queue i is $1 - \sum_j f_{ij}$.

In this section, we assume the arrival and service processes have symmetric tail behavior, i.e., $\alpha_{a,j} = \alpha_{s,j} = \alpha$, for all $j = 1, \ldots, J$. In order to analyze the system time in a particular queue j in the network, we need to characterize the overall arrival process to queue j and then apply Theorem 2 for single-server and Theorem 3 for multi-server queues. The arrival process in queue j is the superposition of different processes, each of which is either an external arrival process, or a departure process from another queue, or a thinning of a departure process from another queue, or a thinning of an external arrival process. Correspondingly, in order to analyze the network, we need to characterize the effect that the following operations have on the arrival process:

(a) Passing through a queue: Under this operation, the jobs exit the queue with inter-departure times $\mathbf{D} = \{D_1, \ldots, D_n\}$. For queues with adversarial servers, Theorem 4 shows that the inter-

departure times satisfy the arrival uncertainty set. This characterization is tighter in steady-state and is akin to the Burke's theorem.

(b) Superposition of arrival processes: Under this operation, p arrival processes $\mathbf{T}^j \in \mathcal{U}_j^a$, $j = 1, \ldots, p$ combine to form a single arrival process. Theorem 5 characterizes the uncertainty set of the combined arrival process.

(c) Thinning of an arrival process: Under this operation, a fraction f of arrivals from a given arrival process is classified as type I while the remaining arrivals are classified as type II. In Theorem 6, we characterize the uncertainty set of the resulting thinned type I process.

We note that the analysis of the departure times entails a queueing behavioral assumption, namely that servers act adversarially so as to maximize the system time. However, the results for the superposition and thinning operations do not make assumptions regarding the behavior of servers. Taken together, our network analysis provides an exact characterization of the steady-state performance of queueing networks under the assumption of adversarial servers. This analysis provides a good approximation of the performance in stochastic queueing networks as shown in Section 7.

5.1. The Superposition Process

Let us consider a queue j that is fed by q arrival processes. Let $\mathcal{U}_j^a$ denote the uncertainty set representing the inter-arrival times $\mathbf{T}^j = \{T_1^j, \ldots, T_n^j\}$ from arrival process $j = 1, \ldots, p$. We denote the uncertainty set of the combined arrival process by $\mathcal{U}_{sup}^a$. Given the primitives $(\lambda_j, \Gamma_{a,j}, \alpha)$, $j = 1, \ldots, p$, we define the *superposition operator* $(\lambda_{sup}, \Gamma_{a,sup}, \alpha_{sup}) = Combine\Big\{ (\lambda_j, \Gamma_{a,j}, \alpha), j = 1, \ldots, p \Big\}$, where $(\lambda_{sup}, \Gamma_{a,sup}, \alpha_{sup})$ characterize the merged arrival process $\mathbf{T}^{sup} = \{T_1^{sup}, \ldots, T_n^{sup}\}$.

THEOREM 5 **(Superposition Operator)**. *The superposition of arrival processes characterized by the uncertainty sets*

$$\mathcal{U}_j^a = \left\{ (T_1^j, \ldots, T_n^j) \;\middle|\; \frac{\sum_{i=k+1}^{n} T_i - \frac{n-k}{\lambda_j}}{(n-k)^{1/\alpha}} \geq -\Gamma_{a,j}\,, \forall k \leq n-1 \right\},\ j = 1, \ldots, p, \tag{70}$$

results in a merged arrival process characterized by the uncertainty set

$$\mathcal{U}^a_{sup} \subseteq \left\{ (T_1^{sup}, \ldots, T_n^{sup}) \;\middle|\; \frac{\sum_{i=k+1}^{n} T_i - \frac{n-k}{\lambda_{sup}}}{(n-k)^{1/\alpha}} \geq -\Gamma_{a,sup}\,, \forall\ 0 \leq k \leq n-1 \right\},$$

where the effective arrival rate, tail coefficient and variability parameter are such that

$$\lambda_{sup} = \sum_{j=1}^{p} \lambda_j,\ \ \alpha_{sup} = \alpha,\ \ \Gamma_{a,sup} = \frac{1}{\sum_{j=1}^{p} \lambda_j} \cdot \left(\sum_{j=1}^{p} \left(\lambda_j \Gamma_{a,j}\right)^{\alpha/(\alpha-1)} \right)^{(\alpha-1)/\alpha}. \tag{71}$$

Proof of Theorem 5. We first consider $p = 2$ and then generalize the result by induction.

(a) By Eq. (70), the inter-arrival times $\mathbf{T}^1 = \{T_i^1, \ldots, T_n^1\}$ and $\mathbf{T}^2 = \{T_i^2, \ldots, T_n^2\}$ are such that

$$\lambda_j \sum_{i=k_j+1}^{n_j} T_i^j \geq (n_j - k_j) - \lambda_j \Gamma_{a,j} \left(n_j - k_j\right)^{1/\alpha},\ j = 1, 2.$$

Summing over index $j = 1, 2$, we obtain

$$\lambda_1 \sum_{i=k_1+1}^{n_1} T_i^1 + \lambda_2 \sum_{i=k_2+1}^{n_2} T_i^2 \geq (n_1 - k_1 + n_2 - k_2) - \lambda_1 \Gamma_{a,1} \left(n_1 - k_1\right)^{1/\alpha} - \lambda_2 \Gamma_{a,2} \left(n_2 - k_2\right)^{1/\alpha}. \tag{72}$$

We consider the time window $\mathcal{T}$ between the arrival of the k_1^{th} and the n_1^{th} jobs from the first arrival process. We assume that, within period $\mathcal{T}$, the queue sees arrivals of jobs $(k_2 + 1)$ up to n_2 from the second arrival process. Therefore, period $\mathcal{T}$ can be written in terms of the combined inter-arrival times $\mathbf{T}^{sup} = \{T_1^{sup}, \ldots, T_n^{sup}\}$ as

$$\mathcal{T} = \sum_{i=k_1+1}^{n_1} T_i^1 = \sum_{i=k+1}^{n} T_i^{sup},\ \text{where } k = k_1 + k_2,\ \text{and } n = n_1 + n_2. \tag{73}$$

Without loss of generality, we assume that $\sum_{i=k_1+1}^{n_1} T_i^1 \geq \sum_{i=k_2+1}^{n_2} T_i^2$ and by Eqs. (73),

$$(\lambda_1 + \lambda_2) \sum_{i=k+1}^{n} T_i^{sup} \geq \lambda_1 \sum_{i=k_1+1}^{n_1} T_i^1 + \lambda_2 \sum_{i=k_2+1}^{n_2} T_i^2 \geq (n-k) - \lambda_1 \Gamma_{a,1} \left(n_1 - k_1\right)^{1/\alpha} - \lambda_2 \Gamma_{a,2} \left(n_2 - k_2\right)^{1/\alpha},$$

where the last inequality is obtained by applying the bound in Eq. (72) and substituting $n_1 + n_2 = n$ and $k_1 + k_2 = k$. By rearranging and dividing both sides by $(\lambda_1 + \lambda_2)$ and $(n-k)^{1/\alpha}$,

$$\frac{\sum_{i=k+1}^{n} T_i^{sup} - \frac{n-k}{\lambda_{sup}}}{(n-k)^{1/\alpha}} \geq -\gamma_{a,sup},\ \text{where } \lambda_{sup} = \lambda_1 + \lambda_2, \alpha_{sup} = \alpha,\ \text{and}$$

$$\gamma_{a,sup} = \frac{\lambda_1}{\lambda_1+\lambda_2}\Gamma_{a,1}\left(\frac{n_1-k_1}{n_1-k_1+n_2-k_2}\right)^{1/\alpha} + \frac{\lambda_2}{\lambda_1+\lambda_2}\Gamma_{a,2}\left(\frac{n_2-k_2}{n_1-k_1+n_2-k_2}\right)^{1/\alpha}.$$

We let the fraction of arrivals from the first process be denoted by

$$x = \frac{n_1-k_1}{n_1-k_1+n_2-k_2}, \quad \text{with} \quad x \in [0,1]. \tag{74}$$

The maximum value that $\gamma_{a,sup}$ achieves over $x \in [0,1]$ can be determined by optimizing the following one-dimensional concave maximization problem

$$\max_{x\in(0,1)} \left\{\beta x^{1/\alpha} + \delta\left(1-x\right)^{1/\alpha}\right\} = \left(\beta^{\alpha/(\alpha-1)} + \delta^{\alpha/(\alpha-1)}\right)^{(\alpha-1)/\alpha}, \tag{75}$$

where $\beta = \frac{\lambda_1}{\lambda_1+\lambda_2}\Gamma_{a,1}$, and $\delta = \frac{\lambda_2}{\lambda_1+\lambda_2}\Gamma_{a,2}$. Substituting β and δ by their respective values in Eq. (75) completes the proof for $p=2$. We refer to this procedure of combining two arrival processes by the operator $(\lambda_{sup}, \Gamma_{a,sup}, \alpha_{sup}) = Combine\left\{(\lambda_1, \Gamma_{a,1}, \alpha), (\lambda_2, \Gamma_{a,2}, \alpha)\right\}$.

(b) Suppose that the arrivals to a queue come from arrival processes 1 through $(p-1)$. We assume that the combined arrival process belongs to the proposed uncertainty set, with

$$\bar{\lambda} = \sum_{j=1}^{p-1}\lambda_j \text{ and } \overline{\Gamma}_a = \frac{1}{\bar{\lambda}}\cdot\left(\sum_{j=1}^{p-1}\left(\lambda_j\Gamma_{a,j}\right)^{\alpha/(\alpha-1)}\right)^{(\alpha-1)/\alpha}$$

Extending the proof to p sources can be easily done by repeating the procedure shown in part **(a)** through the operator $(\lambda_{sup}, \Gamma_{a,sup}, \alpha_{sup}) = Combine\left\{\left(\bar{\lambda}, \overline{\Gamma}_a, \alpha\right), (\lambda_p, \Gamma_{a,p}, \alpha)\right\}$. □

5.2. The Thinning Process

We consider an arrival process in which a fraction f of arrivals is classified as type I and the remaining arrivals are classified as type II, where $f = p/q$ is assumed rational and $p \geq 0$ and $q > 0$ are integers, with $p \leq q$. We note that the assumption on the rationality of the fraction f is not very restrictive, since any irrational number can be arbitrarily closely approximated by a rational number. We consider the following routing scheme: (a) we first thin the original arrival process $\mathbf{T} = \{T_1, \ldots, T_n\}$ into q split processes such that jobs j, $j+q$, $j+2q$, etc. are selected to form the split process j, where $1 \leq j \leq q$, (b) we then superpose p of these split processes to form the desired thinned process. Our computational results suggest that this routing policy provides a

good approximation of the probabilistic routing policy. Given the primitives (λ, Γ_a) of the original process and the fraction f, we define the *thinning operator* $(\lambda_{split}, \Gamma_{a,split}, \alpha) = Split\Big\{ (\lambda, \Gamma_a, \alpha), f \Big\}$, where $(\lambda_{split}, \Gamma_{a,split}, \alpha)$ characterizes the thinned arrival process $\mathbf{T}^{split} = \{T_1^{split}, \ldots, T_n^{split}\}$.

THEOREM 6 **(Thinning Operator)**. *The thinned arrival process of a rational fraction f of arrivals belonging to $\mathcal{U}^a$ is described by the uncertainty set*

$$\mathcal{U}^a_{split} \subseteq \left\{ (T_1^{split}, \ldots, T_n^{split}) \;\middle|\; \frac{\sum_{i=k+1}^{n} T_i^{split} - \frac{n-k}{\lambda_{split}}}{(n-k)^{1/\alpha}} \geq -\Gamma_{a,split},\ \forall\ 0 \leq k \leq n-1 \right\}, \tag{76}$$

where $\lambda_{split} = \lambda \cdot f$ *and* $\Gamma_{a,split} = \Gamma_a \cdot \left(\frac{1}{f}\right)^{1/\alpha}$.

Proof of Theorem 6. We denote the rational fraction $f = p/q$, where $p \geq$ and $q > 0$ are integers, with $p \leq q$. By our routing mechanism, we first split the original arrival process into q split processes $\mathbf{T}^j = \left\{T_i^j\right\}_{i \geq 1}$, each associated with a thinning fraction $f_j = 1/q$, where $j = 1, \ldots, q$. We then combine p split processes and employ the results from Theorem 5 to obtain the desired characterization for the thinned process $\mathbf{T}^{split} = \left\{T_i^{split}\right\}_{i \geq 1}$.

(a) The split process $\left\{T_i^j\right\}_{i \geq 1}$ is formed by selecting jobs j, $j+q$, $j+2q$, etc. In other words, the $(k_j+1)^{\text{th}}$ job in the split process corresponds to the $(j+k_jq)^{\text{th}}$ job in the original process. Consider the time window T between the $(k_j+1)^{\text{th}}$ and the $(n_j+1)^{\text{th}}$ arrivals in the split process $\left\{T_i^j\right\}_{i \geq 1}$. This time window corresponds to the time elapsed between the $(j+k_jq)^{\text{th}}$ and the $(j+n_jq)^{\text{th}}$ arrivals in the original process, yielding

$$T = \sum_{i=(k_j+1)+1}^{n_j+1} T_i^j = \sum_{i=j+k_jq+1}^{j+n_jq} T_i \geq \frac{n_jq - k_jq}{\lambda} - \Gamma_a \left(n_jq - k_jq\right)^{1/\alpha} = \frac{n_j - k_j}{\lambda_j} - \Gamma_{a,j} \left(n_j - k_j\right)^{1/\alpha},$$

where $\lambda_j = \lambda \cdot 1/q = \lambda \cdot f_j$ and $\Gamma_{a,j} = \Gamma_a \cdot q^{1/\alpha} = \Gamma_a \cdot (1/f_j)^{1/\alpha}$, and this characterization is identical to all q split processes. Eq. (76) holds for fractions of the type $f_j = 1/q$, where $q \in \mathbb{N}^+$.

(b) We next show that the above result can be extended for any rational fraction $f = p/q$. The corresponding split process $\left\{T_i^{split}\right\}_{i \geq 1}$ can be seen as a superposition of p out of the q split processes characterized by an uncertainty set of the form described in Assumption 1 with

parameters λ_j and $\Gamma_{a,j}$, as obtained in part (a). Without loss of generality, suppose we combine split processes 1 through p. Utilizing the findings of Theorem 5, we obtain Eq. (76) with

$$\lambda_{split} = \sum_{j=1}^{p} \lambda_j = p\lambda/q = \lambda \cdot f, \quad \text{and} \quad \Gamma_{a,split} = \frac{1}{\lambda_{split}} \cdot \left(\sum_{j=1}^{p} (\lambda_j \Gamma_{a,j})^{\alpha/(\alpha-1)} \right)^{(\alpha-1)/\alpha}.$$

Substituting the values of λ_j and $\Gamma_{a,j}$ obtained in part (a) in the above expression yields $\Gamma_{a,split} = \Gamma_a \cdot (1/f)^{1/\alpha}$, hence concluding the proof. □

Remark: The superposition and thinning operators are consistent. In fact, it is easy to check that, for splitting fractions f_j such that $\sum_{j=1}^{m} f_j = 1$,

$$Combine\Big\{ Split\Big\{ (\lambda, \Gamma_a, \alpha), f_j \Big\}, j = 1, \ldots, m \Big\} = (\lambda, \Gamma_a, \alpha).$$

5.3. The Overall Network Characterization

We perceive the queueing network as a collection of independent queues that could be analyzed separately. The servers in each queue behave in an adversarial manner to maximize the time jobs spend in the queue. We employ the *Combine* and *Split* operators in view of characterizing the effective arrival process to each queue in the network. Knowledge of the effective arrival process allows to study the system time spent at the queue through Theorems 2 and 3 for a single-server and multi-sever queue, respectively. The output of the queue belongs to the effective arrival uncertainty set as shown in Theorem 4 . Theorem 7 characterizes the effective arrival process perceived at each queue in the network.

THEOREM 7 **(Queueing Network Characterization)**. *The behavior of a single class queueing network is equivalent to that of a collection of independent queues with adversarial servers, where the arrival process to node j characterized by the uncertainty set*

$$\mathcal{U}_j^a \subseteq \left\{ (T_1^j, \ldots, T_n^j) \;\middle|\; \frac{\sum_{i=k+1}^{n} T_i^j - \frac{n-k}{\bar{\lambda}_j}}{(n-k)^{1/\alpha}} \geq -\overline{\Gamma}_{a,j}, \; \forall\, 0 \leq k \leq n-1 \right\}, \; j = 1 \ldots, J,$$

where $\{\bar{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_J\}$ and $\{\overline{\Gamma}_{a,1}, \overline{\Gamma}_{a,2}, \ldots, \overline{\Gamma}_{a,j}\}$ *satisfy the set of equations for all* $j = 1, \ldots, J$

$$\bar{\lambda}_j = \lambda_j + \sum_{i=1}^{J} (\bar{\lambda}_i f_{ij}), \tag{77}$$

$$\overline{\Gamma}_{a,j} = \frac{1}{\bar{\lambda}_j} \cdot \left[\left(\lambda_j \cdot \Gamma_{a,j}\right)^{\alpha/(\alpha-1)} + \sum_{i=1}^{J} \left(\bar{\lambda}_i \cdot \overline{\Gamma}_{a,i}\right)^{\alpha/(\alpha-1)} \cdot f_{ij} \right]^{(\alpha-1)/\alpha}. \tag{78}$$

Proof of Theorem 7. Let us consider a queue j receiving jobs from **(a)** external arrivals described by parameters $(\lambda_j, \Gamma_{a,j}, \alpha)$, and **(b)** internal arrivals routed from queues i, where $i = 1, \ldots, J$, resulting from splitting the effective departure process from queue i by f_{ij}. By Theorem 4, the effective departure process from queue i belongs to the uncertainty set satisfied by the effective arrival process to queue i and described by the parameters $\left(\bar{\lambda}_i, \overline{\Gamma}_{a,i}, \alpha\right)$. The effective arrival process to queue j can therefore be represented as

$$\left(\bar{\lambda}_j, \overline{\Gamma}_{a,j}, \alpha\right) = Combine\left\{ \left(\lambda_j, \Gamma_{a,j}, \alpha\right), \left(Split\left\{ \left(\bar{\lambda}_i, \overline{\Gamma}_{a,i}, \alpha\right), f_{ij} \right\}\right), i = 1, \ldots, J \right\} \tag{79}$$

By Theorem 6, we substitute the split processes by their resulting parameters and obtain the superposition of $J+1$ arrival processes

$$\left(\bar{\lambda}_j, \overline{\Gamma}_{a,j}, \alpha\right) = Combine\left\{ \left(\lambda_j, \Gamma_{a,j}, \alpha\right), \left(f_{ij}\bar{\lambda}_i, \overline{\Gamma}_{a,i} \left(\frac{1}{f_{ij}}\right)^{1/\alpha}, \alpha \right), i = 1, \ldots, J \right\} \tag{80}$$

Applying now Theorem 5 yields Eqs. (77) and (78). □

Note that in our analysis, we have assumed that each queue in the network perceives one stream of external arrivals. However, Theorem 7 can be extended in the case where external arrivals are thinned among different queues in the network. This can be done by adding a node in the network for each thinned external arrival process and appending its thinning probabilities to the transition matrix **F**. We next provide the main insights and implications that arise from Theorem 7.

(a) Network Performance Analysis: Theorem 7 allows us to compute performance measures in a queueing network by considering the queues separately. For instance, the system time $\widehat{S}_n^j$ at queue j can be determined through Theorems 2 and 3 with an effective arrival parameters $(\bar{\lambda}_j, \overline{\Gamma}_{a,j}, \alpha)$ and service parameters (μ, Γ_s, α).

(b) Tractable System Solution: Determining the overall network parameters $(\bar{\lambda}, \overline{\Gamma})$ amounts to solving a set of linear equations. To see this, substitute $x_j = \left(\bar{\lambda}_j \overline{\Gamma}_{a,j}\right)^{\alpha/(\alpha-1)}$, for all $j = 1, \ldots, J$, in Eqs. (77) and (78) to obtain the following linear system of equations

$$\begin{cases} \bar{\lambda}_j = \lambda_j + \sum_{i=1}^{J} \bar{\lambda}_i f_{ij} & j = 1, \ldots, J, \\ x_j = \left(\lambda_j \Gamma_{a,j}\right)^{\alpha/(\alpha-1)} + \sum_{i=1}^{J} f_{ij} x_i & j = 1, \ldots, J. \end{cases}$$

Given that the routing matrix $\mathbf{F} = \{f_{ij}\}$ is sub-stochastic, the linear system of equations solves for $(\bar{\lambda}_j, x_j)$, hence allowing to determine $\overline{\Gamma}_{a,j}$, for all $j = 1, \ldots, J$.

6. Queues with Asymmetric Heavy-tailed Arrival and Service Processes

In this section, we extend our results for systems characterized by an asymmetry between the arrival and service tail coefficients. Note that we assume that, within a given queue j, all servers have the same tail coefficient $\alpha_{s,j}$. We omit the proofs as they are straightforward generalizations of the proofs in Theorems 2-7.

System Time: Similarly to Eq. (45), the worst-case system time for asymmetric heavy-tailed arrival and service processes is given by

$$\widehat{S}_n = \max_{0 \le j \le \nu} \left\{ m^{1/\alpha_a} \Gamma_a \left(\nu - j\right)^{1/\alpha_a} + \Gamma_s \left(\nu - j + 1\right)^{1/\alpha_s} - \frac{m(\nu - j)}{\lambda} + \frac{\nu - j + 1}{\mu} \right\}.$$

While the above one-dimensional nonlinear integer optimization problem can be solved efficiently to obtain exact values, we present Theorem 8 which presents a closed-form upper bound on the worst-case system time. The bound is not tight, but provides useful insights as to the effect of the heavy-tailed processes on the system time.

THEOREM 8. **System Time with Asymmetric Tails**

In an m-server FCFS queue with $\mathbf{T} \in \mathcal{U}^a$, $\mathbf{X} \in \mathcal{U}_m^s$, $\alpha_a \neq \alpha_s$ *such that* $\rho < 1$,

$$\widehat{S}_n \le \frac{\bar{\alpha} - 1}{\bar{\alpha}^{\bar{\alpha}/(\bar{\alpha}-1)}} \cdot \frac{\lambda^{1/(\bar{\alpha}-1)} \left(\Gamma_a + \Gamma_s / m^{1/\bar{\alpha}}\right)^{\bar{\alpha}/(\bar{\alpha}-1)}}{(1-\rho)^{1/(\bar{\alpha}-1)}} + \frac{m}{\lambda}, \text{ for } n = r + \nu m, \text{ and } r = 1, \ldots, m,$$

where $\bar{\alpha} = \min\left(\alpha_a, \alpha_s\right)$.

Departure Process: The arguments in the proof of Theorem 4 generalize to the asymmetric case, and the characterization is asymptotically tight.

Superposition Process: We next consider the case when the arrival streams are characterized by different tail coefficient $\alpha_{a,j}$. For notational convenience, we let $\mathbf{1}_{\{\alpha_{a,j}=\alpha\}}$ denote the indicator variable defined by $\mathbf{1}_{\{\alpha_{a,j}=\alpha\}}=1$ if $\alpha_{a,j}=\alpha$ and zero otherwise.

THEOREM 9. **Superposition Operator for Asymmetric Tails**

The superposition of arrival processes characterized by the uncertainty sets

$$\mathcal{U}_j^a = \left\{ (T_1^j,\ldots,T_n^j) \;\middle|\; \frac{\sum_{i=k+1}^{n} T_i - \frac{n-k}{\lambda_j}}{(n-k)^{1/\alpha_{a,j}}} \geq -\Gamma_{a,j}\,,\ \forall\ 0\leq k\leq n-1 \right\},\ j=1,\ldots,N, \tag{81}$$

results in a merged arrival process characterized by the uncertainty set

$$\mathcal{U}_{sup}^a \subseteq \left\{ (T_1^{sup},\ldots,T_n^{sup}) \;\middle|\; \frac{\sum_{i=k+1}^{n} T_i - \frac{n-k}{\lambda_{sup}}}{(n-k)^{1/\bar{\alpha}}} \geq -\Gamma_{a,sup}\,,\ \forall\ 0\leq k\leq n-1 \right\},$$

where the effective tail coefficient $\bar{\alpha}=\min_j \alpha_{a,j}$, *and*

$$\lambda_{sup}=\sum_{j=1}^{N}\lambda_j \ \textit{and}\ \ \Gamma_{a,sup}=\frac{1}{\sum_{j=1}^{N}\lambda_j}\cdot\left(\sum_{j=1}^{N}\mathbf{1}_{\{\alpha_{a,j}\ =\ \bar{\alpha}\}}\cdot\left(\lambda_j\Gamma_{a,j}\right)^{\alpha_{a,j}/(\alpha_{a,j}-1)}\right)^{(\bar{\alpha}-1)/\bar{\alpha}}$$

Theorem 9 implies that the effective tail behavior is dominated by the arrival process with the heaviest tail. While this lower bound may not be tight, the related inaccuracies do not seem to induce large discrepancies within the network according to our computations. In fact, Tables 9 and 10 in Section 7 suggest that having external arrival processes with various tail behavior does not worsen the performance of our algorithm and errors are within 8.7%.

Thinning Process: We note that the split arrival process inherits the tail coefficient α_a corresponding to the thinned arrival process. Hence, Theorem 6 still holds in this case.

The Generalized Queueing Network: We characterize the parameters of the effective arrival processes to each queueing node in the network under the assumption of asymmetric tail behavior. We observe that the parameter $\bar{\alpha}_{a,j}$ describing the tail behavior of the effective arrival process depends on the tail behavior of all the queueing nodes that communicate with node j.

THEOREM 10. **Queueing Network Characterization with Asymmetric Tails**

Consider a queueing network with J queues and external arrival processes characterized by

$(\lambda_j, \Gamma_{a,j}, \alpha_{a,j})$. *The behavior of this network is equivalent to that of a collection of independent queues, with the arrival process to node j characterized by the uncertainty set*

$$\overline{\mathcal{U}}_j^a \subseteq \left\{ (T_1,\ldots,T_n) \;\middle|\; \frac{\sum_{i=k+1}^{n} T_i - \frac{n-k}{\bar{\lambda}_j}}{(n-k)^{1/\bar{\alpha}_{a,j}}} \geq -\overline{\Gamma}_{a,j}\,,\ \forall\ 0\leq k\leq n-1 \right\},\ j=1\ldots,J,$$

where $\left\{\bar{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_J\right\}$ *and* $\left\{\overline{\Gamma}_{a,1}, \overline{\Gamma}_{a,2}, \ldots, \overline{\Gamma}_{a,J}\right\}$ *satisfy the set of equations for all* $j=1,\ldots,J$

$$\bar{\lambda}_j = \lambda_j + \sum_{i=1}^{J} \left(\bar{\lambda}_i f_{ij}\right),$$

$$\overline{\Gamma}_{a,j} = \frac{1}{\bar{\lambda}_j} \cdot \left[\mathbf{1}_{\{\alpha_{a,j}=\bar{\alpha}_{a,j}\}} \cdot \left(\lambda_j \Gamma_{a,j}\right)^{\alpha_{a,j}/(\alpha_{a,j}-1)} + \sum_{i=1}^{J} \mathbf{1}_{\{\bar{\alpha}_{a,i}=\bar{\alpha}_{a,j}\}} \cdot \left(\bar{\lambda}_i \overline{\Gamma}_{a,i}\right)^{\bar{\alpha}_{a,i}/(\bar{\alpha}_{a,i}-1)} \cdot f_{ij} \right]^{(\bar{\alpha}_{a,j}-1)/\bar{\alpha}_{a,j}},$$

with $\bar{\alpha}_{a,j} = \min_{i:\{i\to j\}} \alpha_{a,i}$, *where* $\{i\to j\}$ *means that node i communicates with node j in the network with routing matrix* $\mathbf{F}$.

We next provide the main insights and implications that arise from Theorem 10.

(a) Effect of Heavier Tails: Theorem 10 implies that the tail behavior of the effective arrival process at a given queue is determined by the "heaviest" tail among all departure processes arriving to this queue including the external arrival process to the queue. If all nodes communicate with each other, the tail behavior of the queueing network is then determined by the heaviest tail among the external arrival processes.

(b) Tractable System Solution: Note that the set of equations that characterize the effective arrival process are similar to Eqs. (77) and (78). The only difference in the system for the asymmetric case is the presence of indicator variables $\mathbf{1}_{\{\alpha_{a,j}=\bar{\alpha}_{a,j}\}}$ which isolate the heaviest tail among the merged arrival processes at any given node. Given that these indicator variables are known from data, one could think of this system as a linear system of equations (as for Eqs. (77) and (78)) with $\tilde{f}_{ij} = \mathbf{1}_{\{\bar{\alpha}_{a,i}=\bar{\alpha}_{a,j}\}} \cdot f_{ij}$. The modified routing matrix with entries $\tilde{f}_{ij}$ remains sub-stochastic allowing a unique solution to this system of linear equations.

7. Computational Results

We propose a Robust Queueing Network Analyzer (RQNA) algorithm to approximate the performance of *steady-state* stochastic queueing networks with the following primitive data:

(a) The distributions of the external arrival processes with parameters $(\lambda_j, \sigma_{a,j}, \alpha_{a,j})$ with coefficients of variation $c_{a,j}^2 = \lambda_j^2 \sigma_{a,j}^2,\ j = 1, \ldots, J$.

(b) The distributions of the service processes with parameters $(\mu_j, \sigma_{s,j}, \alpha_{s,j})$ with coefficients of variation $c_{s,j}^2 = \mu_j^2 \sigma_{s,j}^2$ and the number of servers $m_j,\ j = 1, \ldots, J$.

(c) Routing matrix $\mathbf{F} = [f_{ij}]$, $i, j = 1, \ldots, J$, where f_{ij} denotes the fraction of jobs passing through queue i routed queue j. The fraction of jobs leaving the network from queue i is $1 - \sum_j f_{ij}$.

We note that, for heavy-tailed arrivals and services with infinite variance, we truncate the probability distribution and compute the corresponding first and second moments. While the truncation achieves finite moments, it is still a fair depiction of heavy-tailed behavior. In fact, as observed in simulations, these systems do not reach steady-state until seeing a very large number of arrivals. We run simulations for a large number of arrivals ($n = 30,000$) to ensure steady-state is reached.

Our computations aim at providing a numerical validation that our network analysis framework, namely the key elements of network decomposition presented in Sections 4 and 5, provides a good approximation for the performance of stochastic queueing networks. We **(a)** compare our results with simulation and the Queueing Network Analyzer (QNA) proposed by Whitt (1983), and **(b)** investigate the relative performance of RQNA with respect to system's network size, degree of feedback, maximum traffic intensity, and diversity of external arrival distributions.

7.1. Derived Variability Parameters

To apply RQNA on stochastic queueing networks, we first need to translate the stochastic primitive data into uncertainty sets with appropriate variability parameters $(\Gamma_{a,j}, \Gamma_{s,j})$ for each $j = 1, \ldots, J$. Along the lines of QNA, we construct appropriate functions to describe the variability parameters Γ_a and Γ_s in terms of the distributions' first and second-order data, namely the arrival and service rates and their corresponding variances. We then simulate multiple isolated instances of a single queue with various arrival and service distributions and use regression to compute the variability parameters associated with the primitives' distributions. This allows us to build a dictionary or a look-up table of variability parameters values for given arrival and service distributions. We note

that this step is done prior to observing a network instance, and is therefore independent of the network analysis.

We consider a single queue with m servers characterized by $(\rho, \sigma_a, \sigma_s, \alpha)$ and model its variability parameters as $\Gamma_a = \sigma_a$ and $\Gamma_s = f(\rho, \sigma_a, \sigma_s, \alpha)$, where the functional form for $f(\cdot)$ is motivated by the Kingman's bound (Kingman (1970))

$$f(\rho, \sigma_s, \sigma_a, \alpha) = \left(\theta_0 + \theta_1 \cdot \sigma_s^2/m + \theta_2 \cdot \sigma_a^2 \rho^2 m\right)^{(\alpha-1)/\alpha} - \sigma_a m^{(\alpha-1)/\alpha}.$$

We simulate multiple instances of the queue for various parameters of $(\rho, \sigma_a, \sigma_s, \alpha_a, \alpha_s)$ and different arrival and service distributions. We employ linear regression to generate appropriate values for θ_0, θ_1 and θ_2 to adapt the value $\widehat{S}_n$ obtained inTheorem 3 to the expected value of the simulated system time. We propose two different adaptation regimes:

(a) *Service Distribution Dependent Adaptation* where we allow the set of values $(\theta_0, \theta_1, \theta_2)$ to depend on the service distribution, and

(b) *Service Distribution Independent Adaptation* where we obtain a single set of values $(\theta_0, \theta_1, \theta_2)$.

The motivation for considering the service independent adaptation regime is that often we might not know the service time distributions. We also note that we do not perform an adaptation of the values of $(\theta_0, \theta_1, \theta_2)$ for each arrival distribution, since in the network, we have no prior knowledge of the arrival distribution at a given queue. The only known distribution at each queue is in fact the service distribution, hence the proposed adaptation methods. Table 1 provides the resulting $(\theta_0, \theta_1, \theta_2)$ for each of the adaptation regimes.

Table 1 Service adaptation regimes.

$(\theta_0, \theta_1, \theta_2)$	Service Dependent		Service Independent
	Pareto	Normal	
θ_0	-0.05	-0.02	-0.06
θ_1	1.09	1.03	1.07
θ_2	1.11	1.04	1.07

When presented with an instance of a queue, we readily plug the values of $(\theta_0, \theta_1, \theta_2)$ into the proposed functional form to derive the variability parameters and apply Theorem 3 to compute

the steady-state system time. Tables 2 and 3 compare the system time we obtain by the above procedure relative to the simulated values for single queues. We observe that errors are within 9.5% of simulation. As expected, knowledge of the service distribution leads to more accurate answers.

In summary, the adaptation of the variability parameters allows a mapping of the expected system time obtained by simulations to the worst case system time under our approach. In other words, the dictionary we populated in this pre-algorithm step chooses variability parameters Γ_a and Γ_s that allow us to make the following approximation $\mathbb{E}[S(\mathbf{T},\mathbf{X})] \approx \widehat{S}(\Gamma_a,\Gamma_s)$.

Table 2 Multi-Server Single Queue: System time percent errors (service independent adaptation).

Case	1 server		3 servers		6 servers		10 servers	
(c_a^2, c_s^2)	Normal	Pareto**	Normal	Pareto	Normal	Pareto	Normal	Pareto
$(0.25, 0)$	6.55	6.81	8.82	-9.36	8.82	-9.60	8.45	-9.269
$(0.25, 1)$	6.48	-7.31	7.24	-7.96	8.67	-9.16	9.264	-8.839
$(0.25, 4)$	-7.27	-6.58	-8.20	9.85	-8.88	8.26	-8.84	8.89
$(1, 1)$	-6.58	6.42	-7.98	8.93	-9.04	7.96	-8.58	9.36
$(1, 4)$	6.07	6.12	8.55	8.26	8.63	8.66	7.98	9.05
$(4, 0)$	6.05	-5.70	-7.30	7.86	-8.61	8.48	-9.21	9.514
$(4, 1)$	-7.43	-7.84	-7.59	-7.91	-9.31	-9.01	-8.74	-9.553
$(4, 4)$	7.40	-5.89	-9.37	-8.78	-9.03	-9.36	-8.96	-9.34

Table 3 Multi-Server Single Queue: System time percent errors (service dependent adaptation).

Case	1 server		3 servers		6 servers		10 servers	
$(c_a^2, c_s^2)^*$	Normal	Pareto**	Normal	Pareto	Normal	Pareto	Normal	Pareto
$(0.25, 0)$	1.18	2.91	6.03	-6.27	6.27	-7.40	6.45	-6.88
$(0.25, 1)$	3.25	-3.08	5.44	-4.90	6.34	-6.58	7.38	-5.69
$(0.25, 4)$	-3.17	-2.71	-5.46	6.31	-5.97	5.18	-5.89	6.50
$(1, 1)$	-2.11	1.36	-5.75	7.25	-6.16	6.04	-6.56	7.22
$(1, 4)$	-1.53	2.48	6.66	6.02	6.91	6.04	6.87	7.14
$(4, 0)$	1.38	-1.73	-5.71	5.68	-6.50	5.34	-7.48	7.70
$(4, 1)$	-3.08	-1.82	-6.03	-4.07	-6.11	-7.76	-6.52	-7.12
$(4, 4)$	1.33	-2.23	-7.40	-7.82	-6.22	-5.81	-6.10	-7.18

* $c_a = \lambda\sigma_a$ and $c_s = \mu\sigma_s$ represent the coefficients of variation for the inter-arrival and service times.

** Truncated Pareto distribution $f_\alpha = \frac{\alpha x^{-\alpha-1}}{1-(1/H)^\alpha}$, choosing H to achieve the desired coefficient of variation, with mean $1/\mu = \frac{1}{1-(1/H)^\alpha} \cdot \frac{\alpha}{\alpha-1} \cdot \left(1 - 1/H^{\alpha-1}\right)$ and variance $\frac{1}{1-(1/H)^\alpha} \cdot \frac{\alpha}{\alpha-2} \cdot \left(1 - 1/H^{\alpha-2}\right) - 1/\mu^2$.

Note: The model parameters Γ_a and Γ_s may be derived from other available information about

the inter-arrival and service times. For instance, Bertsimas et al. (2014) propose a schema for constructing uncertainty sets and obtaining model parameters using available data and statistical hypothesis tests, and Bandi and Bertsimas (2012a) present approaches that use distributional and correlation information to inform the model parameters.

7.2. The RQNA Algorithm

Having derived the required primitive data for our robust approach, we next describe the RQNA algorithm we employ to compute performance measures of a given network of queues. To do this, we keep track of all possible paths that a job may follow throughout the network. A path p consists of a list of queues visited by some job from entering until leaving the network. We denote the set of all possible paths by $\mathcal{P}$. Let f_p be the fraction of jobs routed through each path $p \in \mathcal{P}$ across the network. The expected overall system time in a network can then be written as

$$\mathbb{E}[S_{tot}] = \sum_{p \in \mathcal{P}} f_p \mathbb{E}[S_p],$$

where S_p is the system time across each path $p \in \mathcal{P}$. Note that $\mathbb{E}[S_p]$ can be obtained by summing the individual expected system times at all nodes associated with this path. Using our adaptation technique presented in Section 7.1, we estimate the the expected system time at each node in path p by the worst case system expression using the generated variability parameters. Using this process, we estimate the expected system time of the network by computing a weighted sum of the worst case system times at each node. This is made explicit in the algorithm presented below.

ALGORITHM **(Robust Queueing Network Analyzer)**

Input: External arrival parameters $(\lambda_j, \sigma_{a,j}, \alpha_{a,j})$, service parameters $(\mu_j, \sigma_{s,j}, \alpha_{s,j})$, and routing matrix $\mathbf{F} = [f_{ij}]$, for $i, j = 1, \ldots, J$. Input also the service times distributions for the case of service dependent adaptation regime.

Output: System times at each node j, $j = 1, \ldots, J$.

1. For each external arrival process i in the network, set $\Gamma_{a,i} = \sigma_{a,i}$.

2. For each queue j in the network with parameters $(\mu_j, \sigma_{s,j}, \alpha_{s,j})$, compute

- **(a)** the effective parameters $\left(\overline{\lambda}_j, \overline{\Gamma}_{a,j}, \bar{\alpha}_{a,j}\right)$ according to Theorem 10 and set $\rho_j = \overline{\lambda}_j/\mu_j$,
- **(b)** the variability parameter $\Gamma_{s,j} = f\left(\rho_j, \overline{\Gamma}_{a,j}, \sigma_{s,j}, \bar{\alpha}_{a,j}, \alpha_{s,j}\right)$, and
- **(c)** the system time $\widehat{S}$ at node j using Theorem 3.

3. Compute the total system time of the network by computing

- **(a)** the set of all possible paths $\mathcal{P}$ in the network,
- **(b)** the fraction f_p of jobs routed through each path $p \in \mathcal{P}$,
- **(c)** the corresponding total system time $\widehat{S}^p$ across each path $p \in \mathcal{P}$ by summing the individual system times at all nodes associated with this path,
- **(d)** the total system time in the network $\widehat{S} = \sum_{p\in\mathcal{P}} f_p \widehat{S}^p$.

Note that, in Step 2(b), we treat each queue j in the network separately as a single isolated queue with an effective arrival process described by the variability parameter $\overline{\Gamma}_{a,j}$. We note that use $\overline{\Gamma}_{a,j}$ as an input to $f(.)$ in place of the standard deviation. Deriving the variability parameter $\Gamma_{s,j}$ can be done using wither the service independent or the service dependent adaptation regime based on whether we know the specific service time distribution at each queue.

7.3. Performance of RQNA in Comparison to QNA and Simulation

We consider the network shown in Figure 2 and perform computations assuming queues have either single or multiple servers, with normal or Pareto distributed service times.

Table 4 reports the percentage errors between the expected system times calculated by simulation and those obtained by each of QNA and RQNA, assuming all nine queues in the network have a single server. RQNA produces results that are often significantly closer to simulated values compared to QNA. Improvements generally range one order of magnitude better in favor of RQNA. Tables 5 and 6 summarize the percentage errors for RQNA relative to simulation for queues with 3, 6, and 10 servers using the service independent and service dependent adaptation regimes, respectively. We make the following observations.

(a) RQNA is fairly insensitive to the heavy-tailed nature of the service distributions. In fact, the percentage errors for the Pareto and normally distributed services are within the same order.

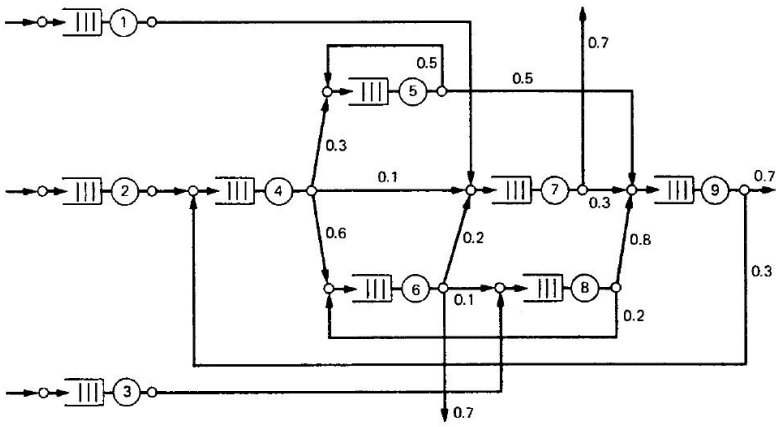

Figure 2 The Kuehn's Network (see Kuehn (1979)).

(b) Adaptation of $(\theta_0, \theta_1, \theta_2)$ to service distribution yields smaller errors up to 6%.

(c) RQNA's performance is generally stable with respect to the number of servers at each queue, yielding errors within the same range for instances with 3 to 10 servers per queue.

Table 4 Single-Server Network: System time percent errors relative to simulation.

Case	Pareto Distribution			Normal Distribution		
$(c_{a,j}^2, c_{s,j}^2)$	QNA	RQNA*	RQNA**	QNA	RQNA*	RQNA**
$(0.25, 0)$	22.78	8.28	3.29	15.28	7.79	1.39
$(0.25, 1)$	18.48	-8.82	-3.48	12.08	8.33	3.87
$(0.25, 4)$	20.13	-7.12	-3.05	11.57	-7.92	-3.88
$(1, 1)$	14.06	6.83	1.80	5.84	-7.12	-2.56
$(1, 4)$	10.15	6.88	2.89	-10.45	7.91	-0.68
$(4, 0)$	21.82	-7.24	-1.93	10.95	6.74	1.29
$(4, 1)$	23.71	-8.73	-2.14	14.18	-9.28	-3.51
$(4, 4)$	17.51	-7.17	-2.97	11.55	9.25	1.67

* Service Independent ** Service Dependent

7.4. Performance of RQNA as a Function of Network Parameters

We investigate the performance of RQNA (for the service dependent adaptation regime) as a function of the system's parameters (network size, degree of feedback, maximum traffic intensity among all queues and number of distinct distributions for the external arrival processes) in families of randomly generated queueing networks. We note that we randomly assign 3, 6 or 10 servers to each of the multi-server queues in the network independently of each other. Tables 7 and 8 report

Table 5 Multi-Server Network: System time percent errors (service independent adaptation).

Case	3 servers		6 servers		10 servers	
$(c_{a,j}^2, c_{s,j}^2)$	Normal	Pareto	Normal	Pareto	Normal	Pareto
$(0.25, 0)$	6.88	-7.21	7.33	-9.42	7.84	-8.00
$(0.25, 1)$	6.17	-5.82	7.72	-7.18	8.48	-6.87
$(0.25, 4)$	-6.33	7.39	-7.53	6.28	-7.14	7.82
$(1, 1)$	-7.26	8.02	-6.98	7.16	-7.98	9.22
$(1, 4)$	7.38	6.93	7.37	6.94	8.21	10.91
$(4, 0)$	-6.86	7.42	-7.26	6.76	-8.22	9.71
$(4, 1)$	-6.60	-5.01	-7.57	-8.99	-8.46	-9.34
$(4, 4)$	-8.14	-8.40	-7.53	-6.84	-7.49	-8.87

Table 6 Multi-Server Network: System time percent errors (service dependent adaptation).

Case	3 servers		6 servers		10 servers	
$(c_{a,j}^2, c_{s,j}^2)$	Normal	Pareto	Normal	Pareto	Normal	Pareto
$(0.25, 0)$	2.10	-2.73	2.63	-3.48	2.84	-3.66
$(0.25, 1)$	3.26	-0.80	4.03	-1.06	4.42	-0.99
$(0.25, 4)$	-2.07	1.42	-2.56	1.79	-2.76	1.91
$(1, 1)$	-3.18	1.72	-4.13	2.23	-3.98	2.23
$(1, 4)$	3.86	1.53	4.98	1.94	5.12	2.00
$(4, 0)$	-3.85	4.63	-5.82	5.36	-5.43	5.31
$(4, 1)$	-3.27	-4.28	-4.37	-4.83	-4.23	-5.67
$(4, 4)$	-3.28	-4.12	-5.82	-5.82	-5.83	-6.13

the system time percentage errors of RQNA relative to simulation as a function of the size of the network and the degree of feedback for queues with single and multiple servers, respectively.

(a) Errors are slightly higher for multi-server networks compared to single-server networks.

(b) RQNA's performance is generally stable for higher degrees of feedback with errors below 6.2%.

(c) RQNA is fairly insensitive to network size with a slight increase in percent errors between 10-node and 30-node networks.

Tables 9 and 10 present the system time percentage errors for RQNA relative to simulation as a function of the maximum traffic intensity among all queues in the network and the number of distinct distributions for the external arrival processes. Specifically, we design four sets of experiments in which we use one type (normal), two types (Pareto and normal), three types (Pareto, normal and Erlang) and four types (Pareto, normal, Erlang and exponential) of arrival distributions.

Table 7 Single-Server Networks: RQNA percent error as a function of network size and degree of feedback.

% Feedback loops / No of nodes	10	15	20	25	30
Feed-forward networks 0%	2.86	2.94	3.03	2.92	3.21
20%	3.12	3.25	3.29	3.71	3.64
35%	3.74	3.81	4.02	4.07	4.14
50%	4.42	4.63	4.84	5.23	5.65
70%	4.85	5.16	5.34	5.68	5.86

Table 8 Multi-Server Networks: RQNA percent error as a function of network size and degree of feedback.

% Feedback loops / No of nodes	10	15	20	25	30
Feed-forward networks 0%	3.59	3.55	3.76	3.43	3.85
20%	3.70	4.01	4.02	4.39	4.45
35%	4.32	4.78	4.95	5.03	4.88
50%	4.95	4.81	5.36	5.67	6.19
70%	5.02	5.56	5.93	5.96	6.03

(a) RQNA presents slightly improved results for lower traffic intensity levels. It is nevertheless fairly stable with respect to higher traffic intensity levels.

(b) The percentage errors generally increase with diversity of external arrival distributions, but still are below 8.5% relative to simulation.

Table 9 Single-Server Networks: RQNA percent error as a function of traffic intensity and variety of external arrival distributions.

No of different distributions	$\rho = 0.95$	$\rho = 0.9$	$\rho = 0.8$	$\rho = 0.65$	$\rho = 0.5$
1	3.34	3.26	3.17	3.02	2.72
2	6.38	5.85	5.47	4.87	3.24
3	7.43	7.09	6.04	5.88	4.53
4	7.56	6.98	6.81	6.29	5.18

8. Concluding Remarks

This paper revisited the problem of analyzing the performance measures of a single-class queue with multiple servers by proposing a new approach for modeling uncertainty. We proposed a robust optimization approach yielding closed-form solutions expressions for the system time in multi-server queues with possibly heavy-tailed arrival and service processes that are not available under

Table 10 Multi-Server Networks: RQNA percent error as a function of traffic intensity and variety of external arrival distributions.

No of different distributions	$\rho = 0.95$	$\rho = 0.9$	$\rho = 0.8$	$\rho = 0.65$	$\rho = 0.5$
1	4.05	4.09	3.62	3.68	3.23
2	5.08	7.10	6.42	6.11	3.71
3	5.92	6.32	6.90	7.34	5.68
4	7.67	8.64	7.28	6.85	5.37

traditional queueing theory. We extended our analysis to the study of arbitrary networks of queues with adversarial servers via the following key principle: **(a)** the departure from a queue, **(b)** the superposition, and **(c)** the thinning of arrival processes have the same uncertainty set representation as the original arrival processes. Our computations validated our model with error percentages in single digits (for all experiments we performed) relative to simulation and performs significantly better than QNA. Moreover, our approach is to a large extent insensitive to the number of servers per queue, network size, degree of feedback and traffic intensity, and somewhat sensitive to the degree of diversity of external arrival distributions in the network.

We are currently investigating extensions of the present framework in two directions: **(a)** a tractable transient analysis of multi-server FCFS queues (please see Bandi et al. (2014)), and **(b)** a tractable steady-state analysis of priority disciplines in multi-server queues.

Acknowledgements

We would like to thank the area editor, the associate editor and the reviewers of the paper for their very constructive and substantive comments that improved the paper significantly.

References

C. Bandi and D. Bertsimas. Tractable stochastic analysis via robust optimization. *Mathematical Programming, Series B*, 134:2370, 2012a.

C. Bandi and D. Bertsimas. Network information theory via robust optimization. Working Paper, 2012b.

C. Bandi and D. Bertsimas. Robust option pricing. European Journal of Operational Research, 2013.

C. Bandi and D. Bertsimas. Optimal design for multi-item auctions: A robust optimization approach. *Mathematics of Operations Research*, 39(4):1012–1038, 2014.

C. Bandi, D. Bertsimas, and N. Youssef. Robust transient multi-server queues and feed-forward networks. Submitted to Operations Research, 2014.

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

D. Bertsimas. An analytic approach to a general class of *G/G/s* queueing systems. *Operations Research*, 38: 139–155, 1990.

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.

D. Bertsimas, V. Gupta, and N. Kallus. Data-driven robust optimization. Submitted for Publication, 2014.

A. Borodin, J. Kleinberg, P. Raghavan, M. Sudan, and D. Williamson. Adversarial queueing theory. *Journal of ACM*, 2001.

J. L. Boudec and P. Thiran. *Network Calculus: A Theory of Deterministic Queuing Systems for the Internet*. LNCS, Springer, 2001.

A. Burchard, F. Ciucu, and J. Liebeherr. On superlinear scaling of network delays. *IEEE/ACM Transactions on Networking*, 19(4):1043–1056, 2011.

A. Burchard, F. Ciucu, and J. Liebeherr. Delay bounds in communication networks with heavy-tailed and self-similar traffic. *IEEE Transactions on Information Theory*, 58(2):1010–1024, 2012.

P. Burke. The output of a queueing system. *Operations Research*, 4(6):699–704, 1956.

F. Ciucu. Network calculus delay bounds in queueing networks with exact solutions. In L. Mason, T. Drwiega, and J. Yan, editors, *Managing Traffic Performance in Converged Networks*, volume 4516 of *Lecture Notes in Computer Science*, pages 495–506. Springer Berlin / Heidelberg, 2007.

F. Ciucu and O. Hohlfeld. On computing bounds on average backlogs and delays with network calculus. In *2010 IEEE International Conference on Communications*, pages 1–5, 2010.

F. Ciucu, A. Burchard, and J. Liebeherr. A network service curve approach for the stochastic analysis of networks. *ACM Sigmetrics*, 2005.

M. Crovella. Performance evaluation with heavy tailed distributions. *Computer Performance Evaluation, LNCS, Springer*, 1786:1–9, 2000.

R. L. Cruz. A calculus for network delay, Part I: Network elements in isolation. *IEEE Transactions on Information Theory*, 37:114–131, 1991a.

R. L. Cruz. A calculus for network delay, Part II: Network analysis. *IEEE Transactions on Information Theory*, 37:132–141, 1991b.

C.S.Chang. *Performance Guarantees in Communication Networks*. Springer, 2001.

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

M. El-Taha and S. Stidham. *Sample-Path Analysis of Queueing Systems*. Springer, 1999.

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

P. L. Gall. The stationary *G/G/s* queue. *Journal of Applied Mathematics and Stochastic Analysis*, 11:59–71, 1998.

G. Gallager and A. Parekh. A generalized processor sharing approach to flow control in integrated services networks: the multiple node case. *IEEE/ACM Transactions on Networking*, 2:137–150, 1994.

D. Gamarnik. Using fluid models to prove stability of adversarial queueing networks. *IEEE Transactions on Automatic Control*, 4:741–747, 2000.

D. Gamarnik. Stability of adaptive and non-adaptive packet routing policies in adversarial queueing networks. *SIAM Journal on Computing*, pages 371–385, 2003.

A. Goel. Stability of networks and protocols in the adversarial queueing model for packet routing. *Proc. 10th ACM-SIAM Symposium on Discrete Algorithms*, 1999.

J. Jackson. Networks of waiting lines. *Operations Research*, 5:518–521, 1957.

Y. Jiang. Stochastic network calculus for performance analysis of internet networks an overview and outlook. In *2012 International Conference on Computing, Networking and Commincations*, pages 638–644, 2012.

Y. Jiang and Y. Liu. *Stochastic Network Calculus*. Springer Publishing Company, Incorporated, 1 edition, 2008. ISBN 1848001266, 9781848001268.

F. P. Kelly, A. K. Maulloo, and D. K. Tan. Rate control for communication networks: shadow prices, proportional fairness and stability. *Journal of the Operational Research society*, pages 237–252, 1998.

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

J.F.C. Kingman. 100 years of queueing. *Proceedings of Conference on The Erlang Centennial*, pages 3–13, 2009.

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

P. Kuehn. Approximate analysis of general queueing networks by decomposition. *IEEE Trans. Comm.*, 1979.

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

J. Nolan. Numerical calculation of stable densities and distribution functions. *Stochastic Models*, pages 759–774, 1997.

F. Pollaczek. *Problèmes stochastiques posés par le phénomène de formation d'une queue d'attente à un guichet et par des phénomènes apparentés.* Mémorial des Sciences Mathématiques, Paris, 1957.

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

J. G. Dai and J. M. Harrison. Reflected brownian motion in an orthant: Numerical methods for steady-state analysis. *The Annals of Applied Probability*, 2:66–86, 1992.

W. Whitt. The queueing network analyzer. *Bell System Technical Journal*, pages 2779–2813, 1983.

W. Whitt. The impact of a heavy-tailed service-time distribution upon the $M/GI/s$ waiting-time distribution. *Queueing Systems*, 36:71–87, 2000.

J. Xie and Y. Jiang. Stochastic network calculus models under max-plus algebra. In *Global Telecommunications Conference, 2009. GLOBECOM 2009. IEEE*, pages 1–6, Nov 2009.