Performance Analysis of Queueing Networks via Robust Optimization

Dimitris Bertsimas * David Gamarnik † Alexander Anatoliy Rikun ‡

April 28, 2010

Abstract

Performance analysis of queueing networks is one of the most challenging areas of queueing theory. Barring very specialized models such as product-form type queueing networks, there exist very few results which provide provable non-asymptotic upper and lower bounds on key performance measures.

In this paper we propose a new performance analysis method, which is based on the robust optimization. The basic premise of our approach is as follows: rather than assuming that the stochastic primitives of a queueing model satisfy certain probability laws, such as, for example, i.i.d. interarrival and service times distributions, we assume that the underlying primitives are deterministic and satisfy the *implications* of such probability laws. These implications take the form of simple linear constraints, namely, those motivated by the Law of the Iterated Logarithm (LIL). Using this approach we are able to obtain performance bounds on some key performance measures. Furthermore, these performance bounds imply similar bounds in the underlying stochastic queueing models.

*Operations Research Center and Sloan School of Management, MIT, Cambridge, MA, 02139, e-mail: dbertsim@mit.edu

†Operations Research Center and Sloan School of Management, MIT, Cambridge, MA, 02139, e-mail: gamarnik@mit.edu

‡Operations Research Center, MIT, Cambridge, MA, 02139, e-mail: arikun@mit.edu

We demonstrate our approach on two types of queueing networks: a) Tandem Single Class queueing network and b) Multiclass Single Server queueing network. In both cases, using the proposed robust optimization approach, we are able to obtain *explicit* upper bounds on some steady-state performance measures. For example, for the case of TSC system we obtain a bound of the form $C(1-\rho)^{-1}\ln\ln((1-\rho)^{-1})$ on the expected steady-state sojourn time, where C is an explicit constant and ρ is the bottleneck traffic intensity. This qualitatively agrees with the correct heavy traffic scaling of this performance measure up to the $\ln\ln((1-\rho)^{-1})$ correction factor.

1 Introduction

Performance analysis of queueing networks is one of the most challenging areas of queueing theory. The difficulty stems from the presence of network feedback, which introduces a complicated multidimensional structure into the stochastic processes underlying the key performance measures. Short of specialized cases, such as product form networks, which typically rely on Poisson arrival/exponential service time distributional assumptions, the problem is largely unresolved. Specifically, given the topological description of a queueing network and given the description of the underlying stochastic primitives such as interarrival and service times distributions, we do not have good tools for computing exactly or obtaining upper and lower bounds on key performance measures, such as, for example average queue lengths and waiting times. Some of results which provide non-asymptotic bounds on performance measures can be found in [BPT94],[KK94],[KM04], [JOK97],[BGT96],[BNM99], all of which require Markovian (Poisson arrival/exponential service time) distributional assumptions. Moreover, some of these bounds become quite weak as traffic intensity (of some of the network components) approach unity. For example, a bound of the form $O((1-\rho^*)^{-2})$ is obtained in [BGT01], where ρ^* is the bottleneck (real or virtual, see the reference) traffic intensity. The other references can lead to infinite upper bounds even in the cases where stationary distribution exists. The approaches in these papers also do not extend to the case of

non-Markovian systems. As a consequence, most of the known performance analysis results are of an asymptotic nature, which apply to queueing networks in various limiting regimes, such as the heavy traffic regime [Har90],[Whi02],[CY01], large deviations methods [GOW04],[SW95], approximations by phase-type distributions [Kle75],[LR87].

In this paper, we partially fill this gap by developing a new performance analysis approach based on robust optimization methods. The theory of robust optimizaiton emerged recently as a very successful and constructive approach for the analysis of certain stochastic modeling problems [Soy73],[BTN98], [BTN99], [BS03], [BS04]. The main premise of our approach in the queueing context is that, rather than assuming probabilistic laws for the underlying stochastic primitives, such as, for example, i.i.d. interarrival and service times, we consider a deterministic queueing model and we will assume only the *implications* of these laws. Specifically we consider implications of the Law of the Iterated Logarithm (LIL). The objective is to find laws which on the one hand hold in the underlying stochastic queueing model and, on the other hand, lead to linear constraints in the formulation of the robust optimization problem, and LIL accomplishes this. We illustrate our approach using two queueing models, namely the Tandem Single Class (TSC) queueing system operating under the First-In-First-Out (FIFO) scheduling policy, and the Multiclass Single Server (MCSS) queueing system operating under an arbitrary work-conserving policy. Motivated by the LIL, we consider constraints of the form $\sum_{1\leq i\leq k} U_i \leq \lambda^{-1}k + \Gamma\sqrt{k\ln\ln k}$, for all $k \geq 1$. Here $(U_k, k \geq 1)$ is any of the stochastic primitives of the underlying queueing system, such as, for example, the sequence of interarrival times and λ stands for the rate of this stochastic primitive. Using these bounds, we derive *explicit* bounds on some performance measures such as sojourn time in the TSC system, namely, the time it takes for a job to be processed by all the servers, and the virtual workload (virtual waiting time) in the MCSS system, namely, the time required to clear the current backlog in the absence of future arrivals. In both models we derive upper bounds on the aforementioned performance measures for the corresponding deterministic counterpart models

and prove that similar bounds also hold for the same performance measures in the underlying stochastic models. In both cases the bounds are of the order $O(\frac{1}{1-\rho}\ln\ln\frac{1}{1-\rho})$, where ρ is the (bottleneck for the case of TSC model) traffic intensity. This matches the correct $O(\frac{1}{1-\rho})$ order short of $\ln\ln((1-\rho)^{-1})$ error. While the technical derivation of these bounds is involved, the conceptual approach is very simple. An interesting distinction of our approach from other robust optimization type results is that our results are explicit, as opposed to numeric results one typically obtains from the formulating and solving a robust optimization model. These explicit bounds however, come at a price of not caring much for the constants corresponding to the leading coefficient. In order to keep things simple we sometimes use very crude estimates for such constants.

Our approach bears similarity with some earlier works in the queueing literature. Specifically, the pioneering work of Cruz [Cru91a],[Cru91b] used a similar non-probabilistic approach to performance analysis by deriving bounds based on placing deterministic constraints on the flow of traffic called "burstiness constraints". The method could be applied to fairly general network topologies and led to more research in the area. In [GP93],[GP94], tighter performance bounds were obtained assuming a "Leaky Bucket" rate admission control from [Tur86] and particular service disciplines. In addition, there is some similarity between the philosophy of our approach and the *adversarial queueing network approach* [BKR$^+$01],[AAF$^+$96],[Gam03],[Gam00],[Goe99], which emerged in the last decade in the computer science literature and also replaces the stochastic assumptions with adversarial deterministic ones. The deterministic constraints used in the aforementioned works are of the form of $A(t) \leq \lambda t + B$ where $A(t)$ is the number of external arrivals into the queueing system up to time t and λ represents the arrival rate. As it turns out, these types of assumptions are too restrictive from the probabilistic point of view and do not lead to bounds on the underlying stochastic network: observe that every renewal process $A(t)$ arising from an i.i.d. sequence with positive variance violates this assumption almost surely for every B for large enough t. As we demonstrate in this paper, the constraints motivated by the LIL,

namely $A(t) \leq \lambda t + B\sqrt{t \ln \ln t}$, can indeed be served to obtain performance bounds, which can be translated into the underlying stochastic network. In fact, the key contribution of our approach is that the deterministic constraints we place on the service and arrival processes are rich enough to lead to stochastic results. The results based on "Leaky Buckets", bounded burstiness and adversarial queueing theory address very general queueing networks. It would be an interesting research project to extend our results based on robust optimization to these general network structures.

The rest of the paper is structured as follows. In the following section we describe two queueing models under the consideration, namely the tandem single class queueing network and the single server multiclass queueing network, as well as their robust optimization counterpart models. Our main results, namely the performance bounds in robust optimization type queueing systems and their implications for stochastic queueing systems are stated in Section 3. The proofs of our main results are in Sections 4 and 5. Some concluding thoughts and directions for further research are outlined in Section 6. Several technical results necessary for proofs of main theorems are delayed untill the Appendix section.

We close this section with some notational conventions. ln stands for the logarithm with natural base. The notation $(x)^{\frac{1}{2}}$ for a non-negative vector $x \in \mathbb{R}^d$ means applying the square root operator coordinate-wise: $(x)^{\frac{1}{2}} = (x_i^{\frac{1}{2}}, 1 \leq i \leq d)$. A^T denotes a transposition operator applied to the matrix A.

2 Model description

We now describe the two queueing models analyzed in this paper, both very well studied models in the literature. We begin by describing these models in the stochastic setting, and then we describe their deterministic robust optimization counterparts.

2.1 A tandem single class (TSC) queueing network. Stochastic model

The model is a tandem of single servers $S_1, \ldots, S_J$ processing a single stream of jobs arriving from outside and requiring services at $S_1, \ldots, S_J$ in this order. The jobs arrive from outside according to an i.i.d. renewal process. Let $U_1, U_2, U_3, \ldots$ denote i.i.d. interarrival times with a common distribution function $F_a(t) = \mathbb{P}(U_1 \leq t)$, where U_1 is the time at which the first job arrives. The external arrival rate is defined to be $\lambda \triangleq 1/\mathbb{E}[U_1]$ and the variance of U_1 is denoted by σ_a^2.

The jobs arriving externally join the buffer corresponding to server S_1 where they are served using First-In-First-Out (FIFO) scheduling policy. We assume that all buffers are of infinite capacity. After service completion, jobs are routed to the buffer of server S_2, where they are also served using FIFO scheduling policy, then they are routed to servers S_3, S_4, etc. After service completion in server S_J the jobs depart from the network. Let V_k^j denote the service time requirement for job k in server j. We assume that the sequence $(V_k^j, k \geq 1)$ is i.i.d. for each j, and is independent from all other random variables in the network. The distribution of the service time in server j is $F_{s,j}(t) = \mathbb{P}(V_1^j \leq t), t \geq 0$. The service rate in server S_j is defined to be $\mu_j \triangleq 1/\mathbb{E}[V_1^j]$, and we denote by $\mu_{\min} = \min_{1 \leq j \leq J} \mu_j$ the rate of the slowest server. $\sigma_{s,j}^2$ denotes the variance of V_1^j for each $j = 1, \ldots, J$. The traffic intensity in server S_j is defined to be $\rho_j = \lambda/\mu_j$, and the bottleneck traffic intensity is defined to be $\rho^* = \max_j \rho_j = \lambda/\mu_{\min}$.

Denote by W_k^j the waiting time experienced by job k in server j not including the service time V_k^j. Let $W_k = \sum_j (W_k^j + V_k^j)$ be the sojourn time of the job k. Namely, this is time between the arrival of job k into buffer 1 and service completion of the same job in buffer J. Denote by $Q_j(t)$ the queue length in server j (the number of jobs in buffer j) at time t. We assume that initially all queues are empty: $Q_j(0) = 0, 1 \leq j \leq J$, although most of our results can either be easily adopted to the case of non-zero queues at time zero, or apply to the steady-state measures where the initializations of the

queues is irrelevant. Let I_k^j denote the idle time of server j in between servicing jobs $k-1$ and k for $k = 2, ..., N$. We define $I_1^j = 0 \ \ \forall j = 1, \ldots, J$.

The model just described will be denoted by TSC(St) (Tandem Single Class Stochastic) for short. It is known [Sig90],[Dai95],[DM95],[CY01] that as long as $\rho^* < 1$, and some additional mild conditions hold, such as finiteness of moments, TSC(St) is stable and the stochastic processes underlying the performance measures such as queue lengths, workloads, sojourn times are mixing. Namely, these processes are positive Harris recurrent [Dai95],[MT93], and the transient performance measures converge to the (unique) steady-state performance measures both in distributions and in moments. Computing these performance measures is a different matter, however. We denote by W_∞^j, W_∞ the steady state versions of the random variables W_k^j, W_k. Thus provided that $\rho^* < 1$ and some additional technical assumptions hold, we have

$$\lim_{n\to\infty} \mathbb{E}[W_n] = \mathbb{E}[W_\infty]. \tag{1}$$

We will assume that $\rho^* < 1$ holds without explicitly stating it. Rather than describing the assumptions required to make (1) true, we will simply assume when stating our results that (1) holds as well.

2.2 A multiclass single server (MCSS) queueing system. Stochastic model

We now describe our second queueing model. Consider a single server queueing system which processes J classes of jobs. The jobs of class $j = 1, 2, \ldots, J$ arrive from outside according to a renewal process with i.i.d. interarrival times $U_k^j, k \geq 1$ and distribution function $F_{a,j}(t) = \mathbb{P}(U_1^j \leq t)$. The arrival rate for class j jobs is $\lambda_j \triangleq 1/\mathbb{E}[U_1^j]$. It is possible that some classes j do not have an external arrival process, in which case $U_k^j = \infty$ almost surely and $\lambda_j = 0$. Let $\sigma_{a,j}^2$ be the variance of U_1^j. The sequences $(U_k^j, k \geq 1)$ are also assumed to be independent for different j. Let $\lambda = (\lambda_j)$ denote the J-vector of arrival rates. We let $\lambda_{\max} = \max_{1\leq j\leq J} \lambda_j$ and $\lambda_{\min} = \min_{1\leq j\leq J} \lambda_j$. We let $A(t) = (A_j(t))$ denote the vector of cumulative number of external arrivals up to time t where $A_j(t) = \max\{k : \sum_{1\leq i\leq k} U_i^j \leq t\}$.

The jobs corresponding to class j are stored in buffer B_j until served. As in the single class case, we assume all buffers are of infinite capacity. The service time for the k-th job arriving to buffer B_j is denoted by V_k^j and the sequence $(V_k^j, k \geq 1)$ is assumed to be i.i.d. with a common distribution function $F_{s,j}(t) = \mathbb{P}(V_1^j \leq t)$. Additionally, these sequences are assumed to be independent for all j and independent from the interarrival times sequences $(U_k^j, k \geq 1)$. The average service time for class j is $m_j \triangleq E[V_1^j]$ and the service rate is $\mu_j \triangleq 1/\mathbb{E}[V_1^j]$. $\sigma_{s,j}^2$ denotes the variance of V_1^j. Let $\bar{m} = (m_j)$ denote the J-vector of average service times and let $\mu = (\mu_j)$ be the J-vector of service rates. Let M denote the diagonal matrix with j-th entry equal to μ_j and let $\mu_{\max} = \max_{1 \leq j \leq J} \mu_j$.

We assume that the jobs in buffer B_j are served using FIFO rule, but prioritizing jobs between different buffers B_j is done using some scheduling policy θ. The only assumption we make about θ is that it is a *work-conserving* policy. Namely, the server is working full time as long as there is at least one job in one of the buffers $B_j,\ 1 \leq j \leq J$. The only performance measure we will consider is the workload (defined below) for which it is well known that the details of the scheduling policy are unimportant for us, as long as the policy is work-conserving.

The routing of jobs after service completions is determined using a routing matrix P, which is an J by J 0, 1 matrix $P = (P_{i,j}, 1 \leq i, j \leq J)$. It is assumed that $\sum_j P_{i,j} \leq 1$ for each i. (Namely, the sum is either 1 or 0). Upon service completion in buffer B_i, the job of class i is routed to buffer j if $P_{i,j} = 1$. Otherwise, if $\sum_j P_{i,j} = 0$, the jobs of class i leave the network. It is assumed that $P^n = 0$ for some positive integer n. It is easy to see that this condition is equivalent to saying that all jobs eventually leave the network.

It is known [CY01] that the traffic equation $\bar{\lambda}_i = \lambda_i + \sum_{1 \leq j \leq J} \bar{\lambda}_j P_{j,i}$ has a unique solution $\bar{\lambda} = (\bar{\lambda}_j)$ given simply as $\bar{\lambda} = [I - P^T]^{-1}\lambda$, where I is the J by J identity matrix. Let $\bar{\lambda}_{\max} = \max_j(\bar{\lambda}_j)$ (observe that $\lambda_j \leq \bar{\lambda}_j$ for every j and hence $\bar{\lambda}_{\max} \geq \lambda_{\max}$). Let $\bar{A}(t) = (\bar{A}_j(t))$ denote the vector of number of arrivals by time t that will eventually route to server j: $\bar{A}_j(t) = e_j^T(I + (P^T)^1 + (P^T)^2 + \ldots)A(t) =$

$e_j^T[I - P^T]^{-1}A(t)$ and e_j denotes the $j-th$ unit vector.

The traffic intensity vector is defined to be $\bar{\rho} = M^{-1}\bar{\lambda} = M^{-1}[I - P^T]^{-1}\lambda$. The traffic intensity of the entire server is $\rho = e^T\bar{\rho}$, where e is the J vector of ones. Let $Q_j(t)$ denote the queue length in buffer j at time t, let $Q(t) = (Q_j(t))$. We assume that $Q(0) = 1$. As for the case of TSC model, our results can be extended to the case $Q(0) \geq 0$, but for the results regarding steady-state behavior, the initialization of queues is irrelevant. Denote by W_k^j the waiting time of the k-th job arriving into buffer j. We let W_t denote the workload at time t. Namely, W_t is the time required to process all the jobs present in the system at time t, in the absence of the future arrivals. Note that W_t is also the virtual waiting time at time t when the scheduling policy is FIFO. Observe that if t_0 marks the beginning of a busy period and t_1 belongs to the same busy period (namely, the server was working continuously during the time interval $[t_0, t_1]$), then almost surely

$$W_{t_1} = \sum_{i=\bar{A}_1(t_0)}^{\bar{A}_1(t_1)} V_i^1 + \ldots + \sum_{i=\bar{A}_J(t_0)}^{\bar{A}_J(t_1)} V_i^J - (t_1 - t_0). \tag{2}$$

The model described above is denoted by MCSS(St) (Multiclass Single Server Stochastic) for short. It is known [Dai95] that if $\rho < 1$, and some additional technical assumption on interarrival and service time distributions hold then MCSS(St) is stable and enters the steady state in the same sense as described for the tandem queueing network. While in this case the steady-state distribution of many performance measures usually depends on the details of work-conserving policy used, the steady-state distribution of the workload does not depend on the policy, as we have discussed above. Let W_∞ denote the workload in steady state, and let B_∞ and I_∞ denote the steady-state duration of the busy and idle periods, respectively. Additionally, denote by $I_0, B_1, I_1, B_2, I_2, \ldots$ the alternating sequence of the lengths of the busy and idle periods of the MCSS(St) system, assuming that time zero initiates a busy period. Under

the same technical assumptions as above the following ergodic properties hold almost surely:

$$\lim_{t\to\infty} \frac{\int_0^t W_s ds}{t} = \mathbb{E}[W_\infty], \tag{3}$$

$$\lim_{n\to\infty} \frac{\sum_{1\le i\le n} B_i}{n} = \mathbb{E}[B_\infty], \tag{4}$$

$$\lim_{n\to\infty} \frac{\sum_{1\le i\le n} I_i}{n} = \mathbb{E}[I_\infty], \tag{5}$$

$$\lim_{n\to\infty} \frac{\sum_{1\le i\le n} B_i^2}{n} = \mathbb{E}[B_\infty^2]. \tag{6}$$

We denote by $n(t)$ the number of busy periods that have been initiated up to time t. Mathematically, we define $n(t)$ to satisfy $\sum_{1\le i\le n(t)-1}(B_i + I_i) < t \le \sum_{1\le i\le n(t)}(B_i + I_i)$. When $t \in [\sum_{1\le i\le n(t)-1}(B_i + I_i), \sum_{1\le i\le n(t)-1}(B_i + I_i) + B_{n(t)}]$, t falls on a busy period and using the definition of $n(t)$, we have $W(t) \le B_{n(t)}$. When $t \in [\sum_{1\le i\le n(t)-1}(B_i + I_i) + B_{n(t)}, \sum_{1\le i\le n(t)}(B_i + I_i)]$, t falls on idle period $I_{n(t)}$ and hence $W(t) = 0$. We let τ_i denote the beginning of the i-th busy period. This implies

$$\frac{\int_0^t W(s)ds}{t} = \frac{\sum_{i=1}^{n(t)} \int_{\tau_i}^{\min(\tau_i+B_i,t)} W(s)ds}{t} \le \frac{\sum_{1\le i\le n(t)} B_i^2}{\sum_{1\le i\le n(t)-1}(B_i + I_i)}$$

If (3),(4),(5) and (6) hold, then we also obtain

$$\mathbb{E}[W_\infty] \le \frac{\mathbb{E}[B_\infty^2]}{\mathbb{E}[B_\infty] + \mathbb{E}[I_\infty]} \le \frac{\mathbb{E}[B_\infty^2]}{\mathbb{E}[B_\infty]}. \tag{7}$$

This bound will turn useful when we apply our results for robust optimization models to the underlying stochastic model. As for the TSC case, we assume from now on $\rho < 1$. Rather than listing the assumptions leading to ergodic properties (3),(4),(5) and (6) we assume when stating our results, that the stochastic process W_t enters the steady-state as $t \to \infty$ and that the properties (3),(4),(5) and (6) holds almost surely.

2.3 Robust optimization type queueing systems

We now describe deterministic robust optimization type counterparts of the two stochastic queueing models described in the previous subsections.

We begin with TSC model and describe the corresponding model which we denote by TSC(RO) (Tandem Single Class Robust Optimization). The description of the network topology is the same as for TSC(St). However, it is not assumed that U_k, V_k^j and, as a result $Q(t), W_k^j, W_k$ are random variables. Rather we assume that these quantities are *arbitrary* subject to certain linear constraints detailed below. Additionally, we assume that the system starts empty $Q(0) = 0$ and only n jobs go through the system.

Specifically, consider a sequence of non-negative deterministic interarrival and service times $(U_k, 1 \le k \le n), (V_k^j, 1 \le k \le n), 1 \le j \le J$. Let

$$\phi(x) = \begin{cases} \sqrt{x \ln \ln x}, & x \ge e^e; \\ 1, & x < e^e. \end{cases} \tag{8}$$

We assume that there exist λ, Γ_a and $\mu_j, \Gamma_{s,j} \ge 0, 1 \le j \le J$ such that

$$\Big| \sum_{k+1 \le i \le n} U_k - \lambda^{-1}(n-k) \Big| \le \Gamma_a \phi(n-k), \qquad k = 0, 1, \ldots, n-1, \tag{9}$$

$$\Big| \sum_{k+1 \le i \le n} V_i^j - \mu_j^{-1}(n-k) \Big| \le \Gamma_{s,j} \phi(n-k), \qquad k = 0, 1, \ldots, n-1,\ j = 1, 2, \ldots, J. \tag{10}$$

It is because we need to consider tail summation $\sum_{k+1 \le i \le n}$ we assume that only n jobs going through the system, though we will be able to apply our results in the stochastic setting where infinite number of jobs pass through the system. Let $\Gamma = \max(\Gamma_a, \Gamma_{s,j})$. Borrowing from the robust optimization literature terminology ([BS04]), the parameters $\Gamma_a, \Gamma_{s,j}, \Gamma$ are called *budgets of uncertainty*. Note, that the values $U_k, V_k^j, k \ge 1$ uniquely define the corresponding performance measures $Q_j(t), W_k^j, W_k, k = 1, \ldots, n$. There is no notion of steady state quantities $Q_j(\infty), W_\infty$ for the model TSC(RO). The motivation for constraints (9) and (10) comes from the Law of the Iterated Logarithm, and we discuss the connection in a separate subsection.

We denote the robust optimization counterpart of the MCSS(St) model by MCSS(RO). In this case it turns out to be convenient to consider infinite sequence of jobs. Thus consider infinite sequences of deterministic non-negative values $(U_k^j, k \ge 1), (V_k^j, k \ge 1), 1 \le j \le J$. It is assumed that values

$\lambda_j, \mu_j, \Gamma_{a,j}, \Gamma_{s,j} \geq 0,\ 1 \leq j \leq J$ exist such that

$$\Big| \sum_{1 \leq i \leq k} U_k^j - \lambda_j^{-1} k \Big| \leq \Gamma_{a,j} \phi(k), \quad k = 1, 2, \ldots,\ j = 1, 2, \ldots, J, \tag{11}$$

$$\Big| \sum_{1 \leq i \leq k} V_i^j - \mu_j^{-1} k \Big| \leq \Gamma_{s,j} \phi(k), \quad k = 1, 2, \ldots,\ j = 1, 2, \ldots, J. \tag{12}$$

For convenience we assume that at time zero the system begins with exactly one job in every class $j = 1, \ldots, J$: $Q_j(0) = 1$. Then the first after time zero external arrival into buffer j occurs at time U_1^j. As before, we let $\Gamma = \max(\Gamma_{a,j}, \Gamma_{s,j})$.

For technical reasons, we also assume that Γ in TSC(RO), MCSS(RO) constraints satisfies

$$\lambda \Gamma \geq e^{2e} \quad \text{and} \quad \min_j \lambda_j \Gamma \geq e^{2e}, \quad \text{respectively.} \tag{13}$$

2.4 The Law of the Iterated Logarithm

One of the cornerstones of the probability theory is the Law of the Iterated Logarithm (LIL) [Chu01], which states that given a i.i.d. sequence of random variables $X_1, \ldots, X_n, \ldots$ with zero mean and finite variance σ, the following holds almost surely,

$$\limsup_{n \to \infty} \frac{\sum_{1 \leq k \leq n} X_k}{\sigma \sqrt{2n \ln \ln n}} = 1, \quad \liminf_{n \to \infty} \frac{\sum_{1 \leq k \leq n} X_k}{\sigma \sqrt{2n \ln \ln n}} = -1.$$

The LIL extends immediately to non-zero mean i.i.d. sequences by subtracting $n\mathbb{E}[X_1]$ from $\sum_{1 \leq k \leq n} X_k$. Furthermore, LIL implies (in the case of zero-mean variables) that

$$\Gamma_{\text{LIL}} \triangleq \sup_{n \geq 1} \frac{|\sum_{1 \leq k \leq n} X_k|}{\sigma \sqrt{2} \phi(n)} < \infty, \tag{14}$$

where ϕ is defined in (8). Note that Γ_{LIL} is a random variable. Thus when we consider stochastic queueing models such as TSC(St) or MCSS(St), the constraints (9),(10),(11),(12) hold with probability one, with $\Gamma = \sqrt{2}\Gamma_{\text{LIL}}\sigma$, where Γ_{LIL} is defined in (14) for the corresponding random sequence. Specifically, let $\Gamma_a = \Gamma_{a,\text{LIL}} = \Gamma_{\text{LIL}}$ and $\sigma = \sigma_a$, when $X_k = U_{n-k} - \lambda^{-1}, 0 \leq k \leq n-1$ and U_k

is the sequence of interarrival times in the TSC(St) model. Similarly define $\Gamma_{s,j} = \Gamma_{s,j,\mathrm{LIL}}$ when $X_k = V^j_{n-k} - \mu_j^{-1}, 0 \le k \le n-1, 1 \le j \le J$. Observe, that for $\Gamma_a, \Gamma_{s,j}$ thus defined, the constraints (9),(10) hold for an infinite sequences of jobs (that is jobs which would have indices $-1, -2, \ldots$), even though we need it only for the first n jobs. For the MCSS(St) model define $\Gamma_{a,j} = \Gamma_{a,j,\mathrm{LIL}}, \Gamma_{s,j} = \Gamma_{s,j,\mathrm{LIL}}$ corresponding to the sequences $U^j_k - \lambda_j^{-1}, V^j_k - \lambda_j^{-1}, k \ge 1$, respectively. We obtain

Proposition 1. *Constraints (9),(10),(11),(12) hold with probability one for* $\Gamma_a = \sqrt{2}\Gamma_{a,LIL}\sigma_a$, $\Gamma_{s,j} = \sqrt{2}\Gamma_{s,j,LIL}\sigma_{s,j}$, $\Gamma_{a,j} = \sqrt{2}\Gamma_{a,j,LIL}\sigma_{a,j}$, *and* $\Gamma_{s,j} = \sqrt{2}\Gamma_{s,j,LIL}\sigma_{s,j}$, *respectively, where* $\Gamma_{\cdot,\cdot,LIL}$ *is defined in (14) for the corresponding sequence.*

As a conclusion, for *every* property derivable on the basis of these constraints in our deterministic robust optimization queueing network models, such as, for example, bounds on the sojourn time of the n-th job in TSC, the *same property* applies with probability one for the underlying stochastic network. This observation underlies the main idea of the paper.

3 Main results

In this section we state our main results on the performance bounds for robust optimization type queueing networks TSC(RO) and MCSS(RO), and the implications of our results for their stochastic counterparts TSC(St) and MCSS(St). We begin with TSC(RO) with the goal of obtaining a bound on the sojourn time.

Theorem 1. *The sojourn time of the n-th job in the TSC(RO) queueing system with constraints (9),(10) satisfies*

$$W_n \le \frac{7J^2\Gamma^2\lambda}{1-\rho^*} \ln\ln \frac{J\lambda\Gamma}{1-\rho^*} + J\lambda^{-1}. \tag{15}$$

Observe that the bound on the sojourn time is explicit. It is expressed directly in terms of the primitives of the queueing system such as arrival and service rates. Observe also that the upper bound

is independent from n. One can think of this bound as a "steady-state" bound on the sojourn time in the robust optimization model of the TSC system. Additionally, the constant Γ^2 is related to the "variances" of interarrival and service times viz a vi the LIL (14). It is known that in the stochastic GI/GI/1 queueing system the expected waiting time in steady state is approximately $(\sigma_a^2+\sigma_s^2)/(2\lambda(1-\rho))$, when the system is in heavy traffic, namely $\rho \to 1$. Namely, the expected waiting time depends linearly on the variances of interarrival and service time. Our bound (15) is thus consistent with this type of dependence. On the other hand, unfortunately, our bound depends quadratically on the number of servers J, whereas the correct dependence is known to be linear, at least in some special cases [Rei84],[GZ06].

The bound above does not have a correct $O((1-\rho^*)^{-1})$ scaling, which is known to be correct from the heavy-traffic theory perspective [Rei84],[GZ06]. However, the correction factor is a very slowly growing function $\ln\ln$. The upshot is that we can use this bound to obtain a bound on W_n and W_∞ in the underlying stochastic system. This is what we do next.

Corollary 1. *For every $n \geq 1$ the sojourn time of the n-th job in the TSC(St) queueing network satisfies*

$$\mathbb{E}[W_n] \leq \mathbb{E}\Big[\frac{7J^2\Gamma^2\lambda}{1-\rho^*}\ln\ln\frac{J\lambda\Gamma}{1-\rho^*}\Big] + J\lambda^{-1}. \tag{16}$$

where $\Gamma = \max_j(\sqrt{2}\sigma_a\Gamma_{a,LIL}, \sqrt{2}\sigma_{s,j}\Gamma_{s,j,LIL}, e^{2e}\lambda^{-1})$. If in addition the assumption (1) holds then

$$\mathbb{E}[W_\infty] \leq \mathbb{E}\Big[\frac{7J^2\Gamma^2\lambda}{1-\rho^*}\ln\ln\frac{J\lambda\Gamma}{1-\rho^*}\Big] + J\lambda^{-1}. \tag{17}$$

Proof. We first assume Theorem 1 is established. Note, in the context of the stochastic system, both W_n and Γ in Theorem 1 are random variables. We take $\Gamma = \max_j(\sqrt{2}\sigma_a\Gamma_{a,\mathrm{LIL}}, \sqrt{2}\sigma_{s,j}\Gamma_{s,j,\mathrm{LIL}}, e^{2e}\lambda^{-1})$ to satisfy (13), where $\Gamma_{\cdot,\cdot,\mathrm{LIL}}$ is defined in (14) for the corresponding sequence. Applying Proposition 1 we have that (15) holds with probability one for the underlying stochastic network. The bound (16) now follows from taking expectations of both sides of (15). The bound (17) follows from applying (1) to (16). □

We now turn our attention to the MCSS queueing model. Our approach for deriving a bound on the workload is based on first obtaining an upper bound on the duration of the busy period. Thus, we first give a bound on the duration of the busy period and then turn to the workload. Recall our assumption $Q(0) = 1$, though our results can readily be extended to the general case of $Q(0) \geq 0$. Thus, time $t = 0$ marks the beginning of a busy period.

Theorem 2. *Given a MCSS(RO) queueing system with constraints (11),(12), let B be the duration of the busy period initiated at time 0. Then*

$$B \leq \frac{5(4J+3)^2\bar{\lambda}_{\max}^3\Gamma^4}{(1-\rho)^2} \ln\ln \frac{2(4J+3)\bar{\lambda}_{\max}^2\Gamma^2}{1-\rho}, \tag{18}$$

$$\textit{and} \quad \sup_{0\leq t\leq B} W(t) \leq \frac{2(4J+3)^2\bar{\lambda}_{\max}^3\Gamma^4}{1-\rho} \ln\ln \frac{(4J+3)\bar{\lambda}_{\max}^2\Gamma^2}{1-\rho} + \Gamma + 3\bar{\lambda}_{\max}^2\Gamma^3. \tag{19}$$

While the bound (19) corresponds to the maximum workload during a given busy period, the actual value of the bound does not depend on the busy period length explicitly. As it will become apparent from the proof, we use the same technique for obtaining a bound simultaneously on the duration of the busy period and maximum workload during the busy period. Let us now discuss the implications of these bounds for the underlying stochastic model MCSS(St).

Corollary 2. *Given a MCSS(St) model, suppose the relations (3),(4),(5) and (6) hold. Then*

$$\mathbb{E}[B_\infty] \leq \mathbb{E}\Big[\frac{5(4J+3)^2\bar{\lambda}_{\max}^3\Gamma^4}{(1-\rho)^2} \ln\ln \frac{2(4J+3)\bar{\lambda}_{\max}^2\Gamma^2}{1-\rho}\Big], \tag{20}$$

$$\mathbb{E}[W_\infty] \leq \mathbb{E}\Big[\frac{25(4J+3)^4\bar{\lambda}_{\max}^6\mu_{\max}\Gamma^8}{(1-\rho)^4}\Big(\ln\ln \frac{2(4J+3)\bar{\lambda}_{\max}^2\Gamma^2}{1-\rho}\Big)^2\Big], \tag{21}$$

where $\Gamma = \max_j(\sqrt{2}\sigma_{a,j}\Gamma_{a,j,LIL}, \sqrt{2}\sigma_{s,j}\Gamma_{s,j,LIL}, e^{2e}\lambda_{\min}^{-1})$.

Unfortunately, in this case the scaling of our bounds as $\rho \to 1$ deviates significantly from the correct behavior. From the heavy traffic theory [DK95], the correct behavior for the steady-state workload should be $O((1-\rho)^{-1})$. As for the steady-state busy period, the theory of M/G/1 queueing

system [Kle75] suggests the behavior $O((1-\rho)^{-\frac{3}{2}})$ as opposed to $O((1-\rho)^{-2}\ln\ln(1-\rho)^{-1})$ which we obtain. On the positive side, however, we managed to obtain explicit bounds on the performance measures which are expressed directly in terms of the stochastic primitives of the model, which we do not believe was possible using prior methods. We leave it as an interesting open problem to derive the performance bounds based on the robust optimization technique, which lead to the correct scaling behavior as $\rho \to 1$.

While the proofs of our main results are technically involved, conceptually they are not complicated. Before we turn to formal proofs, in order to help the reader, we outline below informally some of the key proof steps for our results.

For the TSC queueing network we first replace the constraints (9),(10) with more general constraints, see (22) and (23) below. Our results for the TSC network rely mostly on the Lindley's type recursion which in a single server queueing system recursively represents in the waiting time of the n-th job in terms of the interarrival and service times of the first n jobs. It is classical result of the queueing theory that this waiting time can be thought of as maximum of a random walk, with steps equalling in distribution to the difference between the interarrival and service times. We derive a similar relation in the form of a bound on the sojourn time of the n-th job in the TSC network. This bound is given in Theorem 3. Then we view this bound as an optimization problem and obtain a bound on the objective value by proving the concavity of the objective function and substituting explicit bounds from constraints (9),(10).

Our proofs for the MCSS queueing system rely on the relation (2). Namely, we take advantage of the fact that the workload is depleted with the unit rate during the busy period. Then we take advantage of the constraints (11),(12) to show that in the MCSS(RO) system the workload at time t during the busy period can be upper bounded by an expression of the form $-at+b\sqrt{t\ln\ln t}+c$ with strictly positive a, b. It is then not hard to obtain an explicit estimated t_0 such that this expression is negative for $t > t_0$.

Since this expression is an upper bound on a non-negative quantity (workload), then the duration of the busy period cannot be larger than t_0. This leads to an upper bound on the duration of the busy period in the MCSS(RO) system. In order to obtain a bound on the workload, we again take advantage of (2) and further obtain explicit upper bounds on the terms involving the sums of service times. We show that the workload at time t is at most $-at + b\sqrt{t \ln \ln t} + c$. We then obtain an upper bound on the workload during the busy period by obtaining explicit bounds on $\max_{t \geq 0} -at + b\sqrt{t \ln \ln t} + c$.

Our derivation of the bounds for the stochastic model MCSS(St) relies on the ergodic representation (3). We consider a modified system in which each busy period is initiated with simultaneous arrival of one job into *every* buffer j. This leads to a alternating renewal process with alternating i.i.d. busy and idle periods. We then obtain a bound on the steady-state workload in terms of the second moment of the busy period in the modified queueing system, using the renewal theory type arguments. It is this necessity to look at the second moment of the busy period which leads to a conservative scaling $O\Big((1-\rho)^{-4}(\ln\ln(1-\rho)^{-1})^2\Big)$ in our bound (21) on the steady-state workload.

4 Tandem single class queueing system analysis: proof of Theorem 1

In order to prove Theorem 1 we first generalize constraints (9),(10) and obtain a method for bounding W_n under more general uncertainty assumptions.

4.1 General upper bound on the sojourn times

Given a sequence of non-negative real values $\Gamma^j_{\min}(k), \Gamma^j_{\max}(k)$ $1 \leq j \leq J, 1 \leq k \leq n$, $\Gamma_{\min}(k), \Gamma_{\max}(k)$ $1 \leq k \leq n$, we consider the set of all sequences of service times and interarrival times $(V_i^j), (U_i)$

$j = 1, \ldots, J$, $i = 1, \ldots, n$ satisfying for all $k = 1, \ldots, n$

$$\Gamma^j_{\min}(k) \le \sum_{i=k}^{n} V_i^j \le \Gamma^j_{\max}(k), \tag{22}$$

$$\Gamma_{\min}(k) \le \sum_{i=k}^{n} U_i \le \Gamma_{\max}(k), \tag{23}$$

$$V_i^j, U_i \ge 0.$$

In the next theorem we obtain a bound on the sojourn time of the n-th job in TSC(RO) system in terms of values $\Gamma^j_{\min}(k), \Gamma^j_{\max}(k), \Gamma_{\min}(k), \Gamma_{\max}(k)$.

Theorem 3. *Suppose the relations (22) and (23) hold. Then*

$$W_n \le \max_{n \ge k_J \ge \ldots \ge k_1 \ge 1} \sum_{j=1}^{J-1} \left(\Gamma^j_{\max}(k_j) - \Gamma^j_{\min}(k_{j+1}+1) \right) + \Gamma^J_{\max}(k_J) - \Gamma_{\min}(k_1+1) \tag{24}$$

We now show how Theorem 3 implies our main result Theorem 1.

Proof of Theorem 1. The proof consists of two steps: the first step uses Theorem 3 to bound W_n with uncertainty sets (9),(10). The second step involves solving some associated maximization problem.

We set $\Gamma_{\min}(k) = \lambda^{-1}(n+1-k) - \Gamma_a \phi(n+1-k), \Gamma_{\max}(k) = \lambda^{-1}(n+1-k) + \Gamma_a \phi(n+1-k), \Gamma^j_{\min}(k) = \mu_j^{-1}(n+1-k) - \Gamma_{s,j}\phi(n+1-k), \Gamma^j_{\max}(k) = \mu_j^{-1}(n+1-k) + \Gamma_{s,j}\phi(n+1-k)$, where ϕ is defined by (8). From Theorem 3 we obtain:

$$\begin{aligned} W_n \le & \max_{n \ge k_J \ge \ldots \ge k_1 \ge 1} \sum_{j=1}^{J-1} \left(\mu_j^{-1}(n+1-k_j) + \Gamma_{s,j}\phi(n+1-k_j) \right) \\ & - \sum_{j=1}^{J-1} \left(\mu_j^{-1}(n+1-k_{j+1}-1) - \Gamma_{s,j}\phi(n+1-k_{j+1}-1) \right) \\ & + \left(\mu_J^{-1}(n+1-k_J) + \Gamma_{s,J}\phi(n+1-k_J) \right) \; - \; \left(\lambda^{-1}(n+1-k_1-1) - \Gamma_a \phi(n+1-k_1-1) \right) \end{aligned}$$

Since $n \ge k_{j+1} \ge k_j \quad \forall j$, we can replace μ_j^{-1} by $\mu_{\min}^{-1} = \max(\mu_1^{-1}, \mu_2^{-1} \ldots, \mu_J^{-1}) < \lambda^{-1}$ and preserve

inequality. Similarly, we can replace $\Gamma_{s,1}, \Gamma_{s,2}, \ldots, \Gamma_{s,J}, \Gamma_a$ by Γ. We obtain:

$$\begin{aligned}
W_n &\leq \max_{n \geq k_J \geq \ldots \geq k_1 \geq 1} \sum_{j=1}^{J-1} \Big[\mu_{\min}^{-1}\big(k_{j+1} + 1 - k_j\big) + \Gamma\big(\phi(n+1-k_j) + \phi(n-k_{j+1})\big)\Big] \\
&\quad + \big(\mu_{\min}^{-1}(n+1-k_J) + \Gamma\phi(n+1-k_J)\big) - \big(\lambda^{-1}(n-k_1) - \Gamma\phi(n-k_1)\big) \\
&\leq \max_{n \geq k_1 \geq 1} \mu_{\min}^{-1}(n-k_1) + 2J\Gamma\phi(n+1-k_1) \\
&\quad + J\mu_{\min}^{-1} - \lambda^{-1}(n-k_1) \quad \text{where we used } k_1 \leq k_2 \leq \ldots \leq k_J \text{ to combine } \Gamma \text{ terms} \\
&= \max_{n \geq k_1 \geq 1} (n+1-k_1)(\mu_{\min}^{-1} - \lambda^{-1}) + 2J\Gamma\phi(n+1-k_1) + (J-1)\mu_{\min}^{-1} + \lambda^{-1} \\
&\leq \max_{n \geq k_1 \geq 1} (n+1-k_1)(\mu_{\min}^{-1} - \lambda^{-1}) + 2J\Gamma\phi(n+1-k_1) + J\lambda^{-1} \quad \text{since } \lambda^{-1} > \mu_{\min}^{-1}
\end{aligned}$$

We let $x = n + 1 - k_1$. Since $1 \leq k_1 \leq n$ we have that $1 \leq x \leq n$ and obtain:

$$\begin{aligned}
W_n &\leq \max_{n \geq x \geq 1} x(\mu_{\min}^{-1} - \lambda^{-1}) + 2J\Gamma\phi(x) + J\lambda^{-1} \\
&\leq \max_{x \geq 1} x(\mu_{\min}^{-1} - \lambda^{-1}) + 2J\Gamma\phi(x) + J\lambda^{-1} \qquad (25)
\end{aligned}$$

Putting $a = \lambda^{-1} - \mu_{\min}^{-1}, b = J\Gamma, c = J\lambda^{-1}$, and using the assumption (13), we have $b/a = \lambda J\Gamma/(1-\rho^*) \geq e^{2e}$, namely, the condition (34) is satisfied. Applying Proposition 2 from Appendix we obtain

$$W_n \leq \frac{7\lambda J^2\Gamma^2}{1-\rho^*} \ln\ln \frac{\lambda J\Gamma}{1-\rho^*} + J\lambda^{-1}.$$

This completes the proof of the theorem. □

4.2 Proof of Theorem 3

Job 1 enters the system first, followed by jobs $2, 3, \ldots, n$. Let U_i^j be the time between the arrival of job i and job $i-1$ into server j for $i = 2, \ldots, n$ and $j = 1, \ldots, J$. Specifically, $U_i^1 = U_i$, and we define

$U_1^j = V_1^{j-1}$ for $j = 2, \ldots, J$. The following relations are well known in the queueing theory [Kle75].

$$W_i^j = \max(W_{i-1}^j + V_{i-1}^j - U_i^j, 0) \qquad \forall\ i = 2, \ldots, n,\ j = 1, \ldots, J, \tag{26}$$

$$U_i^j = V_i^{j-1} + I_i^{j-1} \qquad \forall\ i = 2, \ldots, n,\ j = 2, \ldots, J, \tag{27}$$

$$W_i^j = \max\Big\{ \max_{1 \le k \le i-1} \sum_{l=k}^{i-1} \big(V_l^j - U_{l+1}^j\big), 0 \Big\} \qquad \forall\ i = 2, \ldots, n,\ j = 1, \ldots, J, \tag{28}$$

$$W_{i-1}^j = W_i^j - I_i^j - (V_{i-1}^j - U_i^j) \qquad \forall\ i = 2, \ldots, n,\ j = 1, \ldots, J. \tag{29}$$

We now prove some more detailed results regarding the dynamics of our queueing system.

Corollary 3. *The following relations hold for $k = 2, \ldots, n-1$:*

$$\sum_{i=k+1}^{n} U_i^2 = \sum_{i=k+1}^{n} (V_i^1 + I_i^1) = W_n^1 - W_k^1 + \sum_{i=k+1}^{n} U_i^1 + V_n^1 - V_k^1.$$

Proof. The first equality follows from (27). To prove the second equality we use (29) to obtain

$$\sum_{i=k+1}^{n} (V_i^1 + I_i^1) = \sum_{i=k+1}^{n} (W_i^1 - W_{i-1}^1 + U_i^1) + V_n^1 - V_k^1,$$

and the result follows. □

Lemma 1.

$$W_n = \max_{n \ge k_J \ge \ldots \ge k_1 \ge 1} \sum_{i=k_1}^{k_2} V_i^1 + \sum_{i=k_2}^{k_3} V_i^2 + \ldots + \sum_{i=k_{J-1}}^{k_J} V_i^{J-1} + \sum_{i=k_J}^{n} V_i^J - \sum_{i=k_1+1}^{n} U_i^1. \tag{30}$$

Proof. We prove Lemma 1 by induction. We let $W_i^{j,S} = W_i^j + V_i^j$ denote the sojourn time of customer i in server j.

Case $J = 1$: We first define $\sum_{i=j+1}^{j} \equiv 0$ for all j. Using (28) and $V_i^j \ge 0$ we have for any $n = 2, \ldots, n$:

$$\begin{aligned} W_n^{1,S} &= \max\Big(\max_{n-1 \ge k_1 \ge 1} \sum_{i=k}^{n-1} \big(V_i^1 - U_{i+1}^1\big), 0 \Big) + V_n^1 \\ &= \max\Big(\max_{n \ge k_1 \ge 1} \sum_{i=k_1}^{n} V_i^1 - \sum_{i=k_1+1}^{n} U_i^1, V_n^1 \Big) \\ &= \max_{n \ge k_1 \ge 1} \Big(\sum_{i=k_1}^{n} V_i^1 - \sum_{i=k_1+1}^{n} U_i^1 \Big) \quad \text{and this completes case } J = 1. \end{aligned}$$

Case $J > 1$: Note that $W_n = W_n^{1,S} + (W_n^{2,S} + \ldots + W_n^{J,S})$ and denotes the sojourn time of job n in J-server system. We suppose that the result holds for a $J-1$ tandem system and proceed by induction:

$$
\begin{aligned}
& \max_{n\geq k_J\geq\ldots\geq k_1\geq 1} \sum_{i=k_1}^{k_2} V_i^1 + \sum_{i=k_2}^{k_3} V_i^2 + \ldots + \sum_{i=k_{J-1}}^{k_J} V_i^{J-1} + \sum_{i=k_J}^{n} V_i^J - \sum_{i=k_1+1}^{n} U_i^1 \\
= & \max_{n\geq k_J\geq\ldots\geq k_1\geq 1} \Big(\sum_{i=k_1}^{k_2} V_i^1 - \sum_{i=k_1+1}^{k_2} U_i^1 \Big) - \sum_{i=k_2+1}^{n} U_i^1 + \sum_{i=k_2}^{k_3} V_i^2 + \ldots + \sum_{i=k_{J-1}}^{k_J} V_i^{J-1} + \sum_{i=k_J}^{n} V_i^J \\
= & \max_{n\geq k_J\geq\ldots\geq k_2\geq 1} \Big[\max_{k_1:k_2\geq k_1\geq 1} \Big(\sum_{i=k_1}^{k_2} V_i^1 - \sum_{i=k_1+1}^{k_2} U_i^1 \Big) - \sum_{i=k_2+1}^{n} U_i^1 + \sum_{i=k_2}^{k_3} V_i^2 + \ldots + \sum_{i=k_{J-1}}^{k_J} V_i^{J-1} + \sum_{i=k_J}^{n} V_i^J \Big] \\
= & \max_{n\geq k_J\geq\ldots\geq k_2\geq 1} W_{k_2}^{1,S} - \sum_{i=k_2+1}^{n} U_i^1 + \sum_{i=k_2}^{k_3} V_i^2 + \ldots + \sum_{i=k_{J-1}}^{k_J} V_i^{J-1} + \sum_{i=k_J}^{n} V_i^J \quad \text{the base case } J=1 \text{ is used} \\
= & \max_{n\geq k_J\geq\ldots\geq k_2\geq 1} \Big(W_{k_2}^{1,S} - \sum_{i=k_2+1}^{n} U_i^1 \Big) + \sum_{i=k_2}^{k_3} V_i^2 + \ldots + \sum_{i=k_{J-1}}^{k_J} V_i^{J-1} + \sum_{i=k_J}^{n} V_i^J \\
= & \max_{n\geq k_2\geq\ldots\geq k_J\geq 1} \Big(W_{n}^{1,S} - \sum_{i=k_2+1}^{n} U_i^2 \Big) + \sum_{i=k_2}^{k_3} V_i^2 + \ldots + \sum_{i=k_{J-1}}^{k_J} V_i^{J-1} + \sum_{i=k_J}^{n} V_i^J \\
& \text{we used Corollary 3 and } W_{k_2}^{1,S} = W_{k_2}^1 + V_{k_2}^1 \\
= & W_n^{1,S} + \max_{n\geq k_J\geq\ldots\geq k_2\geq 1} \sum_{i=k_2}^{k_3} V_i^2 + \ldots + \sum_{i=k_{J-1}}^{k_J} V_i^{J-1} + \sum_{i=k_J}^{n} V_i^J - \sum_{i=k_2+1}^{n} U_i^2 \\
= & W_1^{1,S} + (W_n^{2,S} + \ldots + W_n^{J,S}) \quad \text{by inductive assumption on } J-1 \text{ server system}
\end{aligned}
$$

and the proof follows from definition of sojourn time W_n. □

Proof of Theorem 3. The result follows immediately from Lemma 1. □

5 Multiclass single server analysis: proofs of main results

5.1 Proof of Theorem 2

Lemma 2. *For every t satisfying*

$$t \geq \max_j(\lambda_j^{-1}e^e, \lambda_j^{-1} + 3\lambda_j^{-1}\lambda_{\max}^2\Gamma^2), \tag{31}$$

the following holds: $A_j(t) \leq t\lambda_j + 3\lambda_j^2\Gamma^2\phi(t\lambda_j)$.

Proof. Assume first $A_j(t) < e^e$. Then applying (11) corresponding to the case $A_j(t) < e^e$, we obtain $A_j(t)\lambda_j^{-1} - \Gamma_{a,j} \leq t$, namely $A_j(t) \leq \lambda_j t + \lambda_j \Gamma_{a,j} \leq \lambda_j t + \lambda_j \Gamma$. Since $\lambda_j \Gamma, \phi(t\lambda_j) \geq 1$ from (13) and (8), the desired result is obtained. For the rest of the proof assume $A_j(t) \geq e^e$. Applying (11), we obtain $A_j(t)\lambda_j^{-1} - \Gamma_{a,j}\sqrt{A_j(t)\ln\ln A_j(t)} \leq t$. Which gives

$$\frac{A_j(t) - t\lambda_j}{\sqrt{A_j(t)\ln\ln A_j(t)}} \quad \leq \lambda_j \Gamma_{a,j} \quad \leq \lambda_j \Gamma. \tag{32}$$

Define b_j by: $b_j = t\lambda_j + 3\lambda_j^2\Gamma^2\sqrt{t\lambda_j \ln\ln t\lambda_j}$. Observe that:

$$\begin{aligned}
\frac{b_j - t\lambda_j}{\sqrt{b_j \ln\ln b_j}} &= \frac{3\lambda_j^2\Gamma^2\sqrt{t\lambda_j \ln\ln t\lambda_j}}{\Big((t\lambda_j + 3\lambda_j^2\Gamma^2\sqrt{t\lambda_j \ln\ln t\lambda_j})\ln\ln(t\lambda_j + 3\lambda_j^2\Gamma^2\sqrt{t\lambda_j \ln\ln t\lambda_j})\Big)^{\frac{1}{2}}} \\
&\geq \frac{3\lambda_j^2\Gamma^2\sqrt{t\lambda_j \ln\ln t\lambda_j}}{\Big((t\lambda_j + 3\lambda_j^2\Gamma^2\sqrt{t^2\lambda_j^2})\ln\ln(t\lambda_j + 3\lambda_j^2\Gamma^2\sqrt{t^2\lambda_j^2})\Big)^{\frac{1}{2}}} \quad \text{since } t\lambda_j \geq \ln\ln t\lambda_j \text{ for } t\lambda_j \geq e^e \text{ from (31)} \\
&= \frac{3\lambda_j^2\Gamma^2\sqrt{t\lambda_j \ln\ln t\lambda_j}}{\Big((t\lambda_j)(1 + 3\lambda_j^2\Gamma^2)\ln\ln(t\lambda_j)(1 + 3\lambda_j^2\Gamma^2)\Big)^{\frac{1}{2}}} \\
&\geq \frac{3\lambda_j^2\Gamma^2\sqrt{t\lambda_j \ln\ln t\lambda_j}}{\Big((t\lambda_j)(1 + 3\lambda_j^2\Gamma^2)\ln\ln(t\lambda_j)^2\Big)^{\frac{1}{2}}} \quad \text{since } t\lambda_j > 1 + 3\lambda_j^2\Gamma^2 \text{ from (31)} \\
&\geq \frac{3\lambda_j^2\Gamma^2\sqrt{\ln\ln t\lambda_j}}{\sqrt{(4\lambda_j^2\Gamma^2)(2\ln\ln t\lambda_j)}} \quad \text{since } 2\ln\ln t\lambda_j > \ln\ln(t\lambda_j)^2 \text{ for } t\lambda_j \geq e^e \text{ and } \lambda_j\Gamma \geq 1 \\
&\geq \lambda_j\Gamma \quad \text{by simplifying above expression.}
\end{aligned}$$

Since $\frac{x - t\lambda_j}{\sqrt{x\ln\ln x}}$ is an increasing function for $x \geq e^e$ and from (32), we have that $b_j \geq A_j(t)$ and the result is obtained. □

We now obtain an upper bound on the cumulative arrival processes $\bar{A}_j(t), 1 \leq j \leq J$.

Lemma 3. *For every t satisfying (31), the following holds*

$$\phi(\bar{A}_j(t)) \leq \big((2 + 6\lambda_{\max}^2\Gamma^2)\big)^{\frac{1}{2}}\phi(\bar{\lambda}_j t)$$

Proof. Consider first the case $\bar{A}_j(t) < e^e$. From (8), we have that $\phi(\bar{A}_j(t)) = 1$ and applying (31), the lemma follows. Now we consider the case $\bar{A}_j(t) \geq e^e$. Recall that $\bar{A}_j(t) = e_j^T[I - P^T]^{-1}A(t)$. Applying

Lemma 2

$$\bar{A}_j(t) \le e_j^T[I-P^T]^{-1}\lambda t + e_j^T[I-P^T]^{-1}\begin{bmatrix} 3\lambda_1^2\Gamma^2\phi(t\lambda_1) \\ 3\lambda_2^2\Gamma^2\phi(t\lambda_2) \\ \vdots \\ 3\lambda_J^2\Gamma^2\phi(t\lambda_J) \end{bmatrix}$$

$$\le e_j^T[I-P^T]^{-1}\lambda t + 3\lambda_{\max}^2\Gamma^2 e_j^T[I-P^T]^{-1}\lambda t \qquad \text{applying (31) and } x \ge \phi(x) \text{ for } x \ge e^e$$

$$= \bar{\lambda}_j t(1+3\lambda_{\max}^2\Gamma^2), \qquad \text{applying the definition of } \bar{\lambda}_j.$$

Applying this bound we also obtain

$$\ln\ln \bar{A}_j(t) \le \ln\ln(\bar{\lambda}_j t(1+3\lambda_{\max}^2\Gamma^2))$$

$$\le \ln\ln(\bar{\lambda}_j t)^2 \qquad \text{using assumption (31)}$$

$$= \ln\ln\bar{\lambda}_j t + \ln 2$$

$$\le 2\ln\ln\bar{\lambda}_j t, \qquad \text{using } \bar{\lambda}_j t \ge \lambda_j t \ge e^e \text{ from (31)}.$$

Combining the previous bounds with definition of $\phi(x)$, the lemma follows. □

Lemma 4. *For every t satisfying (31), we have: $\bar{m}^T\bar{A}(t) - t \le (\rho-1)t + 3\lambda_{\max}\Gamma^2\phi(\lambda_{\max}t)$.*

Proof. Applying definition of $\bar{A}_j(t)$, we have

$$\bar{m}^T\bar{A}(t) - t = \bar{m}^T[I-P^T]^{-1}A(t) - t$$

$$\le m^T[I-P^T]^{-1}\big(\lambda t + 3\lambda_{\max}\Gamma^2\phi(\lambda_{\max}t)\lambda\big) - t \qquad \text{from Lemma 2}$$

$$= \sum_j m_j\bar{\lambda}_j t + 3\lambda_{\max}\Gamma^2\phi(\lambda_{\max}t)\sum_j m_j\bar{\lambda}_j - t \qquad \text{applying the definition of } \bar{\lambda}_j$$

$$= (\rho-1)t + 3\lambda_{\max}\Gamma^2\phi(\lambda_{\max}t)\rho$$

and the lemma follows from applying the condition $\rho < 1$ to the second term. □

We now obtain an upper bound in the duration of the busy period. Recall the identity (2). Since the busy period begins at time zero its duration is upper bounded by the first time t such that

$$\sum_{i=1}^{\bar{A}_1(t)} V_i^1 + \ldots + \sum_{i=1}^{\bar{A}_J(t)} V_i^J - t < 0. \tag{33}$$

Consider any t satisfying the lower bound (31). We have

$$\begin{aligned}
&\sum_{i=1}^{\bar{A}_1(t)} V_i^1 + \ldots + \sum_{i=1}^{\bar{A}_J(t)} V_i^J - t \\
&\qquad \le \sum_{j=1}^{J} \mu_j^{-1} \bar{A}_j(t) + \sum_{j=1}^{J} \Gamma_{a,j} \phi(\bar{A}_j(t)) - t \quad \text{applying (11),(12)} \\
&\qquad \le \bar{m}^T \bar{A}(t) - t + \sum_{j=1}^{J} \Gamma_{a,j} \big((2 + 6\lambda_{\max}^2 \Gamma^2)\big)^{\frac{1}{2}} \phi(\bar{\lambda}_j t) \quad \text{applying Lemma 3} \\
&\qquad \le t(\rho - 1) + 3\lambda_{\max} \Gamma^2 \phi(\lambda_{\max} t) + \sum_{j=1}^{J} \Gamma (2 + 6\lambda_{\max}^2 \Gamma^2)^{\frac{1}{2}} \phi(\bar{\lambda}_j t) \quad \text{applying Lemma 4} \\
&\qquad \le t(\rho - 1) + (4J + 3) \bar{\lambda}_{\max} \Gamma^2 \phi(\bar{\lambda}_{\max} t),
\end{aligned}$$

where we have used a crude estimate $2+6\lambda_{\max}^2\Gamma^2 < 16\lambda_{\max}^2\Gamma^2$, justified by (13). We now apply Lemma 6 with $x = \bar{\lambda}_{\max} t, a = \bar{\lambda}_{\max}^{-1}(1-\rho), b = (4J+3)\bar{\lambda}_{\max}\Gamma^2/2$ and $c = 0$. The condition (34) is implied by assumption (13), and the second condition of Lemma 6 is satisfied since $c = 0$. We obtain that (33) holds for all t satisfying (31) and

$$\begin{aligned}
t &\ge \frac{18(4J+3)^2 \bar{\lambda}_{\max}^2 \Gamma^4}{4\bar{\lambda}_{\max} \bar{\lambda}_{\max}^{-2} (1-\rho)^2} \ln\ln \frac{3(4J+3)\bar{\lambda}_{\max}\Gamma^2}{2\bar{\lambda}_{\max}^{-1}(1-\rho)} \\
&\ge \frac{5(4J+3)^2 \bar{\lambda}_{\max}^3 \Gamma^4}{(1-\rho)^2} \ln\ln \frac{2(4J+3)\bar{\lambda}_{\max}^2\Gamma^2}{1-\rho}.
\end{aligned}$$

Observe using (13) that the right-hand side of the last expression is larger than the right-hand side of (31). Combining two cases we obtain (18).

We now turn to (19). First suppose t does not satisfy (31). Denote the right-hand side of (31) by C. That is $t < C$. Observe that $W(t) \le (C - t) + W(C) \le C + W(C)$ as the workload at time C corresponds in addition to arrivals during $[t, C]$. So now we focus on the case when t satisfies (31). We

use Proposition 2 from Appendix and obtain

$$\sup_{C\le t\le B} W(t) \le \frac{7(4J+3)^2\bar{\lambda}_{\max}^2\Gamma^4}{4\bar{\lambda}_{\max}^{-1}(1-\rho)}\ln\ln\frac{(4J+3)\bar{\lambda}_{\max}\Gamma^2}{2\bar{\lambda}_{\max}^{-1}(1-\rho)}$$
$$\le \frac{2(4J+3)^2\bar{\lambda}_{\max}^3\Gamma^4}{1-\rho}\ln\ln\frac{(4J+3)\bar{\lambda}_{\max}^2\Gamma^2}{1-\rho}.$$

From (13), we have $\Gamma \ge \lambda_{\min}^{-1}$. We conclude that

$$\sup_{0\le t\le B} W(t) \le \frac{2(4J+3)^2\bar{\lambda}_{\max}^3\Gamma^4}{1-\rho}\ln\ln\frac{(4J+3)\bar{\lambda}_{\max}^2\Gamma^2}{1-\rho} + \Gamma + 3\bar{\lambda}_{\max}^2\Gamma^3.$$

This completes the proof of the theorem.

5.2 Proof of Corollary 2

First we establish bound (20). Let $t = 0$ mark the beginning of a busy period with (random) length B_∞ in steady state. This means that there is an arrival into one of the classes j_0 at time 0. Consider a modified system where the first arrivals into classes $j \ne j_0, \lambda_j > 0$ after time 0 are artificially pushed down to exactly time 0. Namely, now at time zero there is an arrival into every class j with $\lambda_j > 0$. The subsequent arrivals into these classes are also pushed earlier by the same amount, thus creating an i.i.d. renewal process initiated at time 0. Let $\hat{B}$ be the busy period initiated in the modified system at time 0. It is easy to see that almost surely $\hat{B} \ge B_\infty$. However, now that we have arrivals in every class at time zero, applying Proposition 1 and our result for the robust optimization counterpart queueing system, namely applying part (18) of Theorem 2, we obtain the required bound by taking the expected values of both sides of (18). This establishes part (20).

In order to prove (21), we use a bound (7). Using our earlier argument for the proof of (20) but applying it to the second moment of $\hat{B}$ we obtain

$$\mathbb{E}[B_\infty^2] \le \mathbb{E}[\hat{B}^2] \le \mathbb{E}\Big[\frac{25(4J+3)^4\bar{\lambda}_{\max}^6\Gamma^8}{(1-\rho)^4}\Big(\ln\ln\frac{2(4J+3)\bar{\lambda}_{\max}^2\Gamma^2}{1-\rho}\Big)^2\Big].$$

On the other hand, we trivially have have $\mathbb{E}[B_\infty] \ge \min_{1\le j\le J} m_j = 1/\mu_{\max}$, since every busy period involves at least one service completion. The result then follows.

6 Conclusion

Using ideas from the robust optimization theory we have developed a new method for conducting performance analysis of queueing networks. The essence of our approach is replacing stochastic primitives of the underlying queueing system with deterministic quantities which satisfy the implications of some probability laws. These implications take the form of linear constraints and for the case of two queueing systems, namely Tandem Single Class queueing networks and Multiclass Single Server queueing system, we have managed to derive explicit upper bounds on some performance measures such as sojourn times and workloads. Then we showed that the bounds implied by the Law of the Iterated Logarithm are applicable for the underlying stochastic queueing system leading to explicit and non-asymptotic performance bounds on the same performance measures. We are not aware of any other method of performance analysis which can provide similar performance bounds in queueing model of similar generality.

We have just scratched the surface of possibilities in this paper and we certainly expect that our approach can be strengthened and extended in multiple directions, some of which we outline below. First we expect that our approach extends to even more general models, such as, for example multiclass queueing networks or more general processing networks [Har00]. The performance bounds can be obtained perhaps again by introducing linear constraints implied by probability laws and using some sort of a Lyapunov function for obtaining bounds in the resulting robust optimization type queueing model. Another important direction is identifying new probability laws which lead to tighter constraints than the ones implied by the LIL. Ideally, one would like to be able to obtain bounds which faithfully represent the scaling behavior of the performance measures of interest in the heavy traffic regime as the (bottleneck) traffic intensity ρ converges to the unity. Further, it would be interesting to obtain performance bounds on the tail probability of the performance measure of interest, perhaps by constructing constraints implied by bounds on the tail probabilities of the underlying stochastic processes. For example, perhaps

one can obtain large deviations type bounds by considering the linear constraints implied by the large deviations bounds on the underlying stochastic processes. Deeper connection between the results of this paper and the results in the adversarial queueing theory and the related queueing literature is worth investigating as well.

Finally, we expect that the philosophy of replacing the *probability model* with *implications of the probability model* will prove useful in non-queueing contexts as well, whenever one has to deal with the issues of stochastic analysis of complicated functionals of stochastic primitives.

Acknowledgements

The authors would like to thank Dmitriy Katz for stimulating discussions and the anonymous reviewers for providing constructive feedback. Research partially supported by NSF grants DMI-0556106 and CMMI-0726733.

References

[AAF+96] M. Andrews, B. Awerbuch, A. Fernandez, Jon Kleinberg, T. Leighton, and Z. Liu, *Universal stability results for greedy contention-resolution protocols*, Proc. 27th IEEE Symposium on Foundations of Computer Science (1996), 380–389.

[BGT96] D. Bertsimas, D. Gamarnik, and J. Tsitsiklis, *Stability conditions for multiclass fluid queueing networks*, IEEE Trans. Automat. Control **41** (1996), 1618–1631.

[BGT01] ———, *Performance of multiclass Markovian queueing networks via piecewise linear Lyapunov functions*, Ann. of Appl. Prob. **11** (2001), no. 4, 1384–1428.

[BKR+01] A. Borodin, J. Kleinberg, P. Raghavan, M. Sudan, and D. Williamson, *Adversarial queueing theory*, Journal of ACM **48** (2001), 13–38.

[BNM99] D. Bertsimas and J. Nino-Mora, *Optimization of multiclass queueing networks with changeover times via the achievable region method: Part II, the multi-station case*, Mathematics of Operations Research **24** (1999), 331–361.

[BPT94] D. Bertsimas, I. Paschalidis, and J. Tsitsiklis, *Optimization of multiclass queueing networks: Polyhedral and nonlinear characterization of achievable performance*, The Annals of Applied Probability **4** (1994), 43–75.

[BS03] D. Bertsimas and M. Sim, *Robust Discrete Optimization and Network Flows*, Mathematical Programming Series B **98** (2003), 49–71.

[BS04] D. Bertsimas and M. Sim, *The price of robustness*, Oper. Res. **52** (2004), 3553.

[BTN98] A. Ben-Tal and A. Nemirovski, *Robust convex optimization*, Math. Oper. Res. **23** (1998), 769–805.

[BTN99] ———, *Robust solutions of uncertain linear programs*, Oper. Research Lett. **25** (1999), 1–13.

[Chu01] K.L. Chung, *A course in probability theory*, third ed., Academic Press, 2001.

[Cru91a] R. Cruz, *A calculus for network delay, part I: network elements in isolation*, IEEE Trans. Information Theory **37** (1991), no. 1, 114–131.

[Cru91b] ———, *A calculus for network delay, part II: network analysis*, IEEE Trans. Information Theory **37** (1991), no. 1, 132–141.

[CY01] H. Chen and D. Yao, *Fundamentals of queueing networks: Performance, asymptotics and optimization*, Springer-Verlag, 2001.

[Dai95] J. G. Dai, *On the positive Harris recurrence for multiclass queueing networks: A unified approach via fluid models*, Ann. Appl. Probab. **5** (1995), 49–77.

[DK95] J. G. Dai and T. G. Kurtz, *A multiclass station with Markovian feedback in heavy traffic*, Math. Oper. Res. **20** (1995), no. 3, 721–742.

[DM95] J. G. Dai and S. P. Meyn, *Stability and convergence of moments for multiclass queueing networks via fluid limit models*, IEEE Transcation on Automatic Controls **40** (1995), 1889–1904.

[Gam00] D. Gamarnik, *Using fluid models to prove stability of adversarial queueing networks*, IEEE Transactions on Automatic Control. (Conference version in FOCS98.) **4** (2000), 741–747.

[Gam03] ———, *Stability of adaptive and non-adaptive packet routing policies in adversarial queueing networks*, SIAM Journal on Computing. (Conference version in STOC99.) (2003), 371–385.

[Goe99] A. Goel, *Stability of networks and protocols in the adversarial queueing model for packet routing*, Proc. 10th ACM-SIAM Symposium on Discrete Algorithms (1999).

[GOW04] A. Ganesh, N. O'Connell, and D. Wischik, *Big queues*, Springer-Verlag, Lecture Notes in Mathematics, Vol. 1838., 2004.

[GP93] G. Gallager and A. Parekh, *A generalized processor sharing approach to flow control in integrated services networks: the single node case*, IEEE/ACM Transactions on Networking **1** (1993), no. 3, 344–357.

[GP94] ———, *A generalized processor sharing approach to flow control in integrated services networks: the multiple node case*, IEEE/ACM Transactions on Networking **2** (1994), no. 2, 137–150.

[GZ06] D. Gamarnik and A. Zeevi, *Validity of heavy traffic steady-state approximations in open queueing networks*, Ann. Appl. Prob. **16** (2006), no. 1, 56–90.

[Har90] J. M. Harrison, *Brownian motion and stochastic flow systems*, Krieger Publishing Company, 1990.

[Har00] ———, *Stochastic networks and activity analysis*, Ann. Appl. Probab. **10** (2000), 75–103.

[JOK97] H. Jin, J. Ou, and P. R. Kumar, *The throughput of irreducible closed Markovian queueing networks: functional bounds, asymptotic loss, efficiency, and the Harrison-Wein conjectures.*, Mathematics of Operations Research **22** (1997), 886–920.

[KK94] S. Kumar and P. R. Kumar, *Performance bounds for queueing networks and scheduling policies*, IEEE Transactions on Automatic Control **8** (1994), 1600–1611.

[Kle75] L. Kleinrock, *Queueing systems*, John Wiley and Sons, Inc., 1975.

[KM04] P. R. Kumar and J. Morrison, *New linear program performance bounds for queueing networks*, Journal of Optimization Theory and Applications (2004), 575–597.

[LR87] G. Latouche and V. Ramaswami, *Introduction to matrix analytic methods in stochastic modeling*, Society for Industrial Mathematics, 1987.

[MT93] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Springer-Verlag. London, UK., 1993.

[Rei84] M. I. Reiman, *Open queueing networks in heavy traffic*, Mathematics of Operations Research **9** (1984), 441–458.

[Sig90] K. Sigman, *The stability of open queueing networks, stochastic processes and their applications*, Stochastic Processes and their Applications **35** (1990), 11–25.

[Soy73] A. L. Soyster, *Convex programming with set-inclusive constraints and applications to inexact linear programming*, Oper. Res. **21** (1973), 1154–1157.

[SW95] A. Shwartz and A. Weiss, *Large deviations for performance analysis*, Chapman and Hall, 1995.

[Tur86] J. Turner, *New directions in communications, or which way to the information age?*, IEEE Commun. Mag. **24** (1986), 8–15.

[Whi02] W. Whitt, *Stochastic-process limits*, Springer, 2002.

Appendix. Preliminary technical results

In this section we establish some preliminary technical results. Using ϕ as defined by (8), we let $U(x) = -ax + 2b\phi(x) + c$ for some positive constants a, b, c satisfying

$$\frac{b}{a} \geq e^{2e}. \tag{34}$$

Lemma 5. *$U(x)$ is strictly concave for $x \geq e^e$.*

Proof.

$$\begin{aligned}
\frac{\partial U(x)}{\partial x} &= -a + b\sqrt{\frac{\ln\ln x}{x}} + \frac{b}{\ln x}\frac{1}{\sqrt{x\ln\ln x}} \\
\frac{\partial^2 U(x)}{\partial x^2} &= b\Big(x^{-\frac{1}{2}}\frac{1}{2}(\ln\ln x)^{-\frac{1}{2}}\frac{1}{\ln x}\frac{1}{x} + (\ln\ln x)^{\frac{1}{2}}(-\frac{1}{2}x^{-\frac{3}{2}})\Big) \\
&\quad + b\Big(-(\ln x)^{-2}\frac{1}{x}(x\ln\ln x)^{-\frac{1}{2}} + (\ln x)^{-1}(-\frac{1}{2})(x\ln\ln x)^{-\frac{3}{2}}(\frac{1}{\ln x} + \ln\ln x)\Big) \\
&= bx^{-\frac{3}{2}}(\frac{1}{2})(\ln\ln x)^{-\frac{1}{2}}\Big(\frac{1}{\ln x} - (\ln\ln x)\Big) \\
&\quad + b\Big(-(\ln x)^{-2}\frac{1}{x}(x\ln\ln x)^{-\frac{1}{2}}\Big) + b\Big((\ln x)^{-1}(-\frac{1}{2})(x\ln\ln x)^{-\frac{3}{2}}(\frac{1}{\ln x} + \ln\ln x)\Big) \\
&< 0 \quad \text{since all three terms on RHS above are negative for } x \geq e^e
\end{aligned}$$

□

Lemma 6. *Assuming (34) and* $e^e > (c/b)^2$,

$$U(x) < 0 \qquad \forall\ x > (18b^2/a^2)\ln\ln(3b/a).$$

Proof. Since $(18b^2/a^2)\ln\ln(3b/a) > e^e$, throughout the proof we restrict ourselves to the domain $x \geq e^e$. Since in addition $x > (c/b)^2$, we have $b\phi(x) \geq b\sqrt{x} > c$. In this range $-ax+2b\phi(x)+c \leq -ax+3b\phi(x) = -ax + 3b\sqrt{x\ln\ln x}$. This quantity is less than zero provided

$$\Big(\frac{x}{\ln\ln x}\Big)^{\frac{1}{2}} > \frac{3b}{a} \triangleq \alpha.$$

It is easy to check that $x/\ln\ln x$ is a strictly increasing function with $\lim_{x\to\infty}(x/\ln\ln x) = \infty$. Let x_0 be the unique solution of $x/\ln\ln x = \alpha^2$ on $x \geq e^e$. We claim that $x_0 \leq 2\alpha^2\ln\ln\alpha$. The assertion of the lemma follows from this bound. Let $A = 2\alpha^2\ln\ln\alpha$. Then

$$\begin{aligned}
\frac{A}{\ln\ln A} &= \frac{2\alpha^2\ln\ln\alpha}{\ln(2\ln\alpha + \ln^{(3)}\alpha + \ln 2)} \\
&\geq \frac{2\alpha^2\ln\ln\alpha}{\ln(4\ln\alpha)} \quad \text{since } \ln\alpha \geq \ln^{(3)}\alpha \text{ and } \ln\alpha > \ln 2 \\
&\geq \frac{2\alpha^2\ln\ln\alpha}{2\ln(\ln\alpha)} \quad \text{since } \ln\alpha > \ln(b/a) \geq 2e > 4. \\
&= \alpha^2.
\end{aligned}$$

This implies $x_0 \leq A$ and the proof is complete. □

Proposition 2. *Under the assumption (34)*

$$\sup_{x\geq 0} U(x) \leq 7(b^2/a)\ln\ln(b/a) + c.$$

Proof. Since $a > 0$, then the supremum in $\sup_{x\geq 0} U(x)$ is achieved. Let x^* be any value achieving $\max_{x\geq 0} U(x)$. First suppose $0 \leq x^* < e^e$. It follows from the definition of ϕ in (8) that $\phi(x^*) = 1$ and thus $U(x^*) = -ax^* + 2b + c$. Using $0 \leq x^* < e^e$ and assumption (34), it is straightforward to check that $U(x^*)$ is indeed upper bounded from above by $7(b^2/a)\ln\ln(b/a)+c$. Next, we consider the case $x^* = e^e$,

and using the fact that $a > 0$, we obtain $U(x^*) \leq 2b \cdot \sqrt{e^e \ln\ln(e^e)} + c$. It is again straightforward to check that the aforementioned bound is upper bounded from above by $7(b^2/a)\ln\ln(b/a) + c$.

We now consider the case $x^* > e^e$. By Lemma 5, x^* is the unique point satisfying $\frac{\partial U(x^*)}{\partial x^*} = 0$, if it exists. The remainder of the proof is devoted to the final case where we obtain

$$0 = \frac{\partial U(x^*)}{\partial x^*} = -a + \frac{b(\frac{1}{\ln x^*} + \ln\ln x^*)}{\sqrt{x^* \ln\ln x^*}} \tag{35}$$

Continuing further, (35) implies

$$\frac{\sqrt{x^* \ln\ln x^*}}{\ln\ln x^* + \frac{1}{\ln x^*}} = \frac{b}{a} \triangleq \alpha. \tag{36}$$

Note

$$\frac{x^*}{\ln\ln x^*} > \alpha^2$$
$$\frac{x^*}{2\ln\ln x^*} < \alpha^2 \text{ since } \ln\ln x^* > \frac{1}{\ln x^*} \text{ for } x \geq e^e$$

It is easy to check that $x/\ln\ln x$ is a strictly increasing function for $x \geq e^e$ and $\lim_{x\to\infty}(x/\ln\ln x) = \infty$. (34) implies that there exist unique $x_{\min}$ and $x_{\max}$ satisfying

$$\frac{x_{\min}}{\ln\ln x_{\min}} = \alpha^2 \qquad \frac{x_{\max}}{2\ln\ln x_{\max}} = \alpha^2$$

The monotonicity of $x/\ln\ln x$ implies $x_{\min} \leq x^* \leq x_{\max}$. In order to complete the proof of the proposition, we will first state and prove Lemmas 7 and 8.

Lemma 7. *$x_{\min} \geq \alpha^2 \ln\ln\alpha$ and $x_{\max} \leq 4\alpha^2 \ln\ln\alpha$.*

Proof. Let $B_1 = \alpha^2 \ln\ln\alpha$. Then

$$\begin{aligned} \frac{B_1}{\ln\ln B_1} &= \frac{\alpha^2 \ln\ln\alpha}{\ln\ln(\alpha^2 \ln\ln\alpha)} \\ &< \frac{\alpha^2 \ln\ln\alpha}{\ln\ln\alpha} \quad \text{since } \ln\ln\alpha \geq 1 \text{ for } \alpha \geq e^{2e} \\ &= \alpha^2. \end{aligned}$$

Thus since $\frac{x}{\ln\ln x}$ is increasing for $x \geq e^e$, we have $x_{\min} \geq B_1$ and the first assertion is established.

Let $B_2 = 4\alpha^2 \ln\ln\alpha$. Then

$$\begin{aligned}
\frac{B_2}{2\ln\ln B_2} &= \frac{4\alpha^2 \ln\ln\alpha}{2\ln\ln(4\alpha^2 \ln\ln\alpha)} \\
&= \frac{4\alpha^2 \ln\ln\alpha}{2\ln(2\ln\alpha + \ln^{(3)}\alpha + \ln 4)} \\
&\geq \frac{4\alpha^2 \ln\ln\alpha}{2\ln(4\ln\alpha)} \quad \text{since } \ln\alpha \geq \ln^{(3)}\alpha \text{ and } \ln\alpha > \ln 4 \\
&\geq \frac{4\alpha^2 \ln\ln\alpha}{4\ln(\ln\alpha)} \quad \text{since } \ln\alpha \geq 2e > 4. \\
&= \alpha^2.
\end{aligned}$$

Thus, again since $x/\ln\ln x$ is increasing for $x \geq e^e$, then the second assertion follows. □

Lemma 7 and $x_{\min} \leq x^* \leq x_{\max}$ imply

$$\alpha^2 \ln\ln\alpha \leq x^* \leq 4\alpha^2 \ln\ln\alpha. \tag{37}$$

Lemma 8. $\sqrt{x_{\max}\ln\ln x_{\max}} \leq 4\alpha \ln\ln\alpha$.

Proof.

$$\begin{aligned}
\sqrt{x_{\max}\ln\ln x_{\max}} &\leq \sqrt{\left(4\alpha^2\ln\ln\alpha\right)\ln\ln\left(4\alpha^2\ln\ln\alpha\right)} \text{ by Lemma 7} \\
&= \alpha\sqrt{4\ln\ln\alpha}\sqrt{\ln\left(2\ln\alpha + \ln^{(3)}\alpha + \ln 4\right)} \\
&\leq \alpha\sqrt{4\ln\ln\alpha}\sqrt{\ln\left(4\ln\alpha\right)} \text{ since } \ln\alpha \geq \ln^{(3)}\alpha \text{ and } \ln\alpha \geq \ln(e^{2e}) > \ln 4 \\
&\leq \alpha\sqrt{4\ln\ln\alpha}\sqrt{2\ln\ln\alpha} \text{ since } \ln\alpha > 4
\end{aligned}$$

and the lemma follows from the last step. □

We now complete the proof of Proposition 2. We have

$$\begin{aligned}
U(x^*) &\leq -ax^* + 2b\sqrt{x^* \ln\ln x^*} + c \\
&\leq -ax_{\min} + 2b\sqrt{x_{\max} \ln\ln x_{\max}} + c \quad \text{since } x_{\min} \leq x^* \leq x_{\max} \\
&\leq -ax_{\min} + 8b\alpha \ln\ln\alpha + c \ \text{ by Lemma 8} \\
&\leq -a\alpha^2 \ln\ln\alpha + 8b\alpha \ln\ln\alpha + c \ \text{ by Lemma 7} \\
&= 7(b^2/a)\ln\ln(b/a) + c.
\end{aligned}$$

□