Exponential Tail Bounds on Queues: A Confluence of Non-Asymptotic Heavy Traffic and Large Deviations

Prakirt Raj Jhunjhunwala[1], Daniela Hurtado-Lange[2], Siva Theja Maguluri[3]
prj2122@columbia.edu, daniela.hurtado@kellogg.northwestern.edu, siva.theja@gatech.edu
[1]Columbia University, [2] Northwestern University, [3]Georgia Institute of Technology *

ABSTRACT

In general, obtaining the exact steady-state distribution of queue lengths is not feasible. Therefore, we focus on establishing bounds for the tail probabilities of queue lengths. We examine queueing systems under Heavy Traffic (HT) conditions and provide exponentially decaying bounds for the probability $\mathbb{P}(\epsilon q > x)$, where ϵ is the HT parameter denoting how far the load is from the maximum allowed load. Our bounds are not limited to asymptotic cases and are applicable even for finite values of ϵ, and they get sharper as $\epsilon \to 0$. Consequently, we derive non-asymptotic convergence rates for the tail probabilities. Furthermore, our results offer bounds on the exponential rate of decay of the tail, given by $-\frac{1}{x}\log \mathbb{P}(\epsilon q > x)$ for any finite value of x. These can be interpreted as non-asymptotic versions of Large Deviation (LD) results. To obtain our results, we use an exponential Lyapunov function to bind the moment-generating function of queue lengths and apply Markov's inequality. We demonstrate our approach by presenting tail bounds for a continuous time Join-the-shortest queue (JSQ) system.

1. INTRODUCTION

Queueing models are used to study performance of many systems such as cloud computing, data centers, ride hailing, call centers etc. In general, obtaining the complete distribution of queue lengths in these systems is intractable. Therefore, a common approach is to study asymptotic regimes. Recently, the Many-Server Heavy-Traffic (Many-Server-HT) regime has gained more popularity, where the system is loaded to maximum capacity while simultaneously increasing the number of servers. The system's behavior varies greatly depending on how quickly the load increases relative to the number of servers. As such one employs very different analysis techniques to study queueing systems in different regimes.

In the study of HT asymptotics, one typically scales the queue lengths using a parameter that represents the system's load. By denoting the load as $1-\epsilon$, where ϵ is the HT parameter, the HT limit is achieved when ϵ approaches zero. For a load balancing system in Heavy Traffic (HT), which when satisfies the so-called Complete Resource Pooling (CRP) condition, it is well-known that the scaled queue length follows an exponential distribution in the HT limit, which gives the tail probabilities of the limiting system. However, the rate of convergence of the tail probabilities (of the pre-limit system) to the corresponding HT value remains unknown.

Most real world systems involve Service Level Agreements (SLA), where customers are promised a specific level of service, including the maximum delay they can expect. Motivated by this, in this paper, we focus on establishing sharp bounds on the tail probabilities of scaled queue length of the pre-limit system, i.e., for $\epsilon > 0$. In particular, we get non-asymptotic bounds of the form $\mathbb{P}(\epsilon q > x) \leq \kappa(\epsilon, x)e^{-\theta(\epsilon)x}$, where q represents the total queue length in steady state. Here, $\theta(\epsilon)$ gives the decay rate of the tail probability of the pre-limit system, and $\theta(\epsilon)$ converges to the correct HT value as $\epsilon \to 0$. Recent results show the rate of convergence to HT in terms of the mean, moments, or Wasserstein's distance. These methods focus on the entire distribution of the queue lengths and drown the tail. For example, consider the second moment, and suppose ϵq converges in distribution to the random variable Υ. Then, from existing results, one obtains that $|\mathbb{E}[\epsilon^2 q^2] - \mathbb{E}[\Upsilon^2]|$ is $O(\epsilon)$, which gives a valid bound. From these results, one can obtain bounds in terms of tail probability of the form $|\mathbb{P}(\epsilon q > x) - \mathbb{P}(\Upsilon > x)| \leq O(\epsilon)$. However, these are not very informative as the tail probability itself can be much smaller than $O(\epsilon)$. Therefore, the rate of convergence of tail probabilities cannot be obtained using the existing methodologies. In this work, we correctly characterize $\theta(\epsilon)$ to obtain the rate of convergence of tail probability to the corresponding HT value. Our results are non-asymptotic in the sense that they are valid whenever ϵ is small, and not just when $\epsilon \to 0$. Also, our results are precise when ϵ gets closer to 0, recovering the HT results.

Our work bridges the gap between the Large Deviations (LD) and Many-Server-HT regimes. When one studies the LD regime, the goal is to find the exponential rate at which the tail probability decays, which is precisely given by $\theta(\epsilon)$. As such, our tail bounds can be used to recover the non-asymptotic LD results. Thus, our tail bounds are at a confluence of non-asymptotic HT and non-asymptotic LD. To the best of our knowledge, such comprehensive LD results have not been previously reported in the existing literature.

2. MODEL: JSQ SYSTEM

We consider a continuous-time queueing system consisting of n Single-Server Queues (SSQ) in parallel, each serving jobs according to first-come-first-serve. At any time t,

*This work was partially supported by NSF grants EPCN-2144316 and CMMI-2140534.

let $\mathbf{q}(t)$ denote the queue length vector, where $q_i(t)$ is the queue length of i^{th} queue. For the ease of notation, we use $\overline{q}(t)$ to denote the total queue length at time t, i.e., $\overline{q}(t) = \sum_{i=1}^{n} q_i(t)$. Jobs arrive to the system according to a Poisson process with rate λ_n, and service times are exponentially distributed with rate μ. When a job arrives, it is dispatched according to JSQ, that is, the job is sent to the queue with index $i^*(t) \in \arg\min_i q_i(t)$, breaking ties uniformly at random.

The system load is $\rho_n := \frac{\lambda_n}{n\mu} < 1$, and we define $\epsilon_n = 1 - \rho_n$. In the context of the JSQ system, we consider the Many-Server-HT regime, where the system size n grows to infinity while the HT parameter ϵ_n approaches zero. Specifically, we consider $\epsilon_n = n^{-\alpha}$ with $\alpha > 1$ constant, and take the limit as $n \to \infty$. We use $\mathbf{q}$ to denote the steady-state queue length vector, and $\overline{q} := \sum_{i=1}^{n} q_i$.

2.1 Results for JSQ system

It is well known that the HT distribution of the scaled steady-state total queue length $\epsilon\overline{q}$ converges to an exponential random variable as $\epsilon \to 0$ [1]. Further, as shown in [2], the result extends to Many-Server-HT, where $\epsilon_n\overline{q}$ converges to an exponential random variable in distribution if $\alpha > 2$, and it was conjectured that the result also holds for $\alpha \in (1, 2]$. In Theorem 1, we complete the result by demonstrating that $\epsilon_n\overline{q}$ converges in distribution to an exponential random variable $\alpha > 1$.

Theorem 1. *Suppose the system satisfies the condition $\lambda_n = n\mu(1 - n^{-\alpha})$, i.e., $\epsilon_n = n^{-\alpha}$, where $\alpha > 1$. Then, for any $\theta < 1$, we have $n^{1-\alpha}\mathbf{q} \xrightarrow{d} \Upsilon\mathbf{1}$ as $n \to \infty$, where Υ is an exponential random variable with mean 1.*

The result in Theorem 1 thus fills the gap in the literature and is in conjunction with all the prior work [2]. A crucial step to this end is in establishing State Space Collapse (SSC) for the JSQ system. In HT, that is as $\epsilon_n \downarrow 0$, we have $\epsilon_n q_i \approx \frac{\epsilon_n}{n}\sum_{i=1}^{n} q_i$ for all $i \in \{1, 2, \ldots, n\}$, that is, the n-dimensional queueing vector collapses to a one-dimensional subspace. This phenomenon, called SSC, is a key property of the JSQ system in heavy traffic. In order to prove Theorem 1, we prove that the JSQ system satisfies SSC for all value of $\alpha > 1$.

Theorem 2. *Suppose the JSQ system satisfies the condition $\lambda_n = n\mu(1 - \epsilon_n)$. Let ϵ_n is small enough such that $n\epsilon_n \log\left(\frac{1}{\epsilon_n}\right) < \kappa_1$, where κ_1 and κ_2 are positive constants. Suppose $\theta_n := \frac{1}{\epsilon_n}\log\frac{1}{1-\epsilon_n}$. Then, for all $x > 1 - \epsilon_n$ we have*

$$e^{-\theta_n x} \le \mathbb{P}\Big(\epsilon_n\overline{q} > x\Big) \le \Big[2ex\big(1 - \kappa_2 n\epsilon_n \log\epsilon_n\big)\Big]e^{-\theta_n x},$$

where the lower bound holds for any $n \ge 1$ and $\epsilon_n \in (0, 1)$. As a consequence, we have the following large deviation result,

$$\lim_{x\to\infty} -\frac{1}{x}\log\mathbb{P}\Big(\epsilon_n\overline{q} > x\Big) = \theta_n := \frac{1}{\epsilon_n}\log\frac{1}{1-\epsilon_n}. \quad (1)$$

Theorem 2 establishes the exponential decay of the tail of the total queue length for a JSQ system in the Many-Server-HT regimes. Further, the result in Theorem 2 is consistent with the fact that the distribution of the scaled steady-state total queue length, i.e., $\epsilon_n\overline{q}$, converges to an exponential random variable in distribution, as n grows to ∞.

In Theorem 2, we are able to characterize the exact tail decay rate of the continuous time JSQ system. Our result implies that, in Many-Server-HT with $\alpha > 1$ and when the term $n\epsilon_n \log\left(\frac{1}{\epsilon_n}\right)$ is small enough, the decay rate of the JSQ system exactly matches the tail decay rate of an SSQ. This is a significant advancement compared to existing literature. Previous work primarily focused on comparing the behavior of the JSQ system with an SSQ under the limiting conditions, specifically as $\epsilon_n \to 0$. In contrast, our work examines the behavior of a pre-limit JSQ system and directly compares it to the corresponding SSQ. Our bounds on tail probability on JSQ system, presented in Theorem 2, can be decomposed into terms as discussed below.

SSC violation: For the JSQ system, the SSC violation term is given by $\big(1 - \kappa_2 n\epsilon_n \log\epsilon_n\big)$. In non-asymptotic HT conditions (i.e., when $n < \infty$ in Many-Server-HT), the SSC property is not fully satisfied. This introduces an additional multiplicative term in the tail probability bound, which is captured by $1 - \kappa_2 n\epsilon_n \log\epsilon_n$, and reflects the extent to which SSC is violated. Further, in Theorem 2, we need the term $n\epsilon_n \log\frac{1}{\epsilon_n}$ to be small enough, to ensure that the SSC is not completely violated, and behaviour of the JSQ system is close to a corresponding SSQ.

Pre-limit tail: The pre-limit tail denotes the actual decay rate of the tail probability of $\epsilon_n\overline{q}$ under non-asymptotic HT condition, i.e., $n < \infty$. For the continuous-time JSQ system, we exactly characterize the pre-limit tail, which is given by θ_n. When $\epsilon_n = n^{-\alpha}$ with $\alpha > 1$, as $n \to \infty$, the tail of $\epsilon_n\overline{q}$ matches that of an exponential distribution with mean 1, as $\lim_{n\to\infty}\theta_n = 1$. Further, note that, the deviation of the pre-limit tail from the corresponding HT value is given by $|\theta_n - 1|$, which is of order $O(\epsilon_n)$.

Pre-exponent error: In the context of the JSQ system, the pre-exponent error is represented by the expression $2ex$. This error term arises from using Markov's Inequality to obtain tail-probability bounds using MGF. To clarify this error term, consider a random variable X that follows an exponential distribution with rate λ. In this case, the MGF of X is given by $\mathbb{E}[\exp(\theta X)] = \frac{1}{1-\theta/\lambda}$ for all $\theta < \lambda$. Applying Markov's Inequality to the MGF and optimizing over the value of θ, we obtain $\mathbb{P}(X > x) \le e\lambda x e^{-\lambda x}$. The upper bound differs from the actual tail of X by a multiplicative factor of $e\lambda x$, which arises from using Markov's Inequality. We acknowledge that it may be possible to eliminate the Markov-Inequality error by employing more complex techniques. However, we have chosen to rely solely on Markov's Inequality for our analysis to maintain simplicity.

For more details on this work, please refer to [3].

3. REFERENCES

[1] D. Hurtado-Lange and S. T. Maguluri. Transform methods for heavy-traffic analysis. *Stochastic Systems*, 10(4):275–309, 2020.

[2] D. Hurtado-Lange and S. T. Maguluri. A load balancing system in the many-server heavy-traffic asymptotics. *Queueing Systems*, 101(3-4):353–391, 2022.

[3] P. R. Jhunjhunwala, D. Hurtado-Lange, and S. T. Maguluri. Exponential tail bounds on queues: A confluence of non-asymptotic heavy traffic and large deviations, 2023.