

Investigating the impact of inter-arrival time and service time on the performance of $H_2/H_2/k$

Ankur Sharma, Shayak Ghosal, Anshul Gandhi*

ankursharma@cs.stonybrook.edu, shayak.ghosal@stonybrook.edu, anshul@cs.stonybrook.edu

PACE Lab, Stony Brook University, USA

Abstract—An important objective in queueing theory is to understand the impact of inter-arrival times (IAT) and service times (ST) on the performance of a system. Many models and techniques have been developed towards this end. Consider the $G/G/1$ model that allows for arbitrary IAT and ST distributions. While exact performance analysis is difficult, several approximations exist to aid practitioners. For example, the Kingman’s formula approximates mean response time of a $G/G/1$ model in terms of the first and second moments of the IAT and ST distributions. However, such approximations are only accurate for very high load, and only consider the impact of the first two moments of IAT and ST. As such, a gap exists for the accurate analysis for arbitrary load that also considers the impact of higher moments of IAT and ST.

The goal of this paper is to address this gap. We leverage the fact that IAT and ST processes in computer systems can be well modeled by the flexible 2-phase Hyper-exponential distribution, H_2 , as shown by recent work. Based on the properties of H_2 , we set up the Markov chain for the $H_2/H_2/k$ model and numerically solve for performance metrics of interest. We then conduct an extensive evaluation to study the impact of various central moments of IAT and ST, including the squared coefficient of variation (SCV) and kurtosis, on the mean response time $E[T]$ of $H_2/H_2/k$ models. Our evaluation reveals several interesting results: $E[T]$ can depend superlinearly on arrival SCV at moderate load and non-monotonically on service SCV in multi-server systems, with kurtosis providing secondary modulation.

I. INTRODUCTION

Queueing models are widely employed to analyze and estimate the performance of computer systems. Common performance metrics of interest include response time, T , and number of jobs in system, N . Typically, the performance is modeled as a function of (at least) two input random variables: (i) the inter-arrival time, IAT, and (ii) the service time, ST. The IAT distribution represents the time between successive job arrivals to a system, whereas the ST distribution represents the time required to complete a job once it enters service. As such, an important question in queueing analysis is *how do the IAT and ST distributions impact response time?*

Ideally, this question should be posed in the context of the $G/G/k$ queueing model that allows for a general (or arbitrary) IAT distribution and a general ST distribution. While readily applicable in practice, the analysis of this model, even under $k = 1$ and just for the mean response time, is quite challenging, with practitioners often resorting to approximations (e.g., Kingman’s formula [1]) or bounds [2]. For the general $G/G/k$ model, the performance analysis is even more difficult,

with very few known results. On the other hand, there exists the classical $M/M/1$ queueing model where the IAT and ST distributions are Exponential. In this case, closed-form results are available not just for mean response time, $E[T]$, but also for the distribution of response times [3]. Several results also exist for the $M/M/k$ model, the k -server version of the $M/M/1$ model. However, real-world computer systems rarely have Exponential IAT or ST distributions [4]–[7]; the $M/M/1$ model is thus not very useful in practice. What is needed is a flexible distribution that is applicable in practice and lends itself to a tractable queueing model.

Prior work has shown that the *Hyper-exponential distribution* (a probabilistic mixture of Exponential distributions), even with just 2 phases, provides a good fit for the IAT and ST distribution of computer systems [8]–[10]; see Section III for more details. This 2-phase Hyper-exponential, H_2 , is flexible enough to mimic several distributions, including heavy-tailed distributions. Further, the $H_2/H_2/k$ queueing models are tractable, though requiring some effort, and are thus amenable to performance analysis. We thus consider the Hyper-exponential distribution in our investigation.

In this paper, we numerically investigate how the IAT and ST distributions affect the performance of $H_2/H_2/k$ systems. Specifically, *we study how the mean, squared coefficient of variation (SCV), and kurtosis of the IAT and ST distributions affect the mean response time of $H_2/H_2/k$ queues*. This moment-based parametrization lets us separate scale, second-order variability, and higher-order tail effects within the H_2 family. We also compare the marginal impact of IAT and ST distributions to ascertain which of these two input variables has a larger impact on performance. We use matrix analytic methods [11] to solve for the performance metrics of the $H_2/H_2/k$ models; this technique also allows us to analyze other interesting metrics, such as the distribution of number of jobs in the system and the probability of queueing.

Our numerical analysis provides several insights. First, both arrival and service SCV affect mean response time, but in different ways. Increasing the arrival SCV, C_a^2 , produces super-linear increases in $E[T]$, particularly at moderate load, whereas the service SCV, C_s^2 , has a comparatively more gradual effect, that may become non-monotone for larger server counts; see Section IV-B for details. Second, kurtosis has a secondary but visible effect. Increasing the arrival kurtosis, κ_a , tends to reduce $E[T]$ when C_a^2 is large, while its effect is negligible when C_a^2 is small; the effect of service kurtosis, κ_s , is almost negligible at high load and weak at moderate load. Finally,

*This work was supported by the National Science Foundation (NSF) under grants CCF-2324859, CNS-2214980, and CNS-2106434.

tail probabilities show broadly consistent trends: increasing C_a^2 produces the most pronounced increase in tail heaviness in our experiments, C_s^2 plays an analogous but less uniform role, and kurtosis matters mainly when the SCV is large.

The above findings cannot be immediately obtained by existing approximations (e.g., Kingman's [1]) that are only applicable in heavy traffic, often only include lower moments of IAT and ST, and are typically limited to a single server.

II. PRELIMINARIES AND RELATED WORK

In this paper, we consider a first-come-first-serve (FCFS) service discipline and independent and identically distributed (i.i.d.) IATs and STs. Let $\lambda = 1/E[IAT]$ be the mean request rate and let $\mu = 1/E[ST]$ be the mean service rate of the system; let k be the number of servers. Then, $\rho = \lambda/k \cdot \mu$ denotes the load of the system. We are interested in the response time, T , of the system; often, we consider the mean response time, $E[T]$, or the mean waiting time $E[W_q] = E[T] - E[ST]$.

G/G/k models: The $G/G/1$ model has been widely studied given the fact that it allows for arbitrary IAT and ST distributions. However, exact analysis or results for this model remain elusive. Lindley's recursive integral equation [12] gives a relationship satisfied by the stationary waiting time distribution of the $G/G/1$ model; however, solving the associated differential equations is challenging as it requires complex analysis and transform inversion. Instead, Kingman's formula [1] is often used in practice to approximate the mean waiting time, $E[W_q]$, as:

$$E[W_q] \approx \frac{\rho}{1-\rho} \cdot \frac{C_a^2 + C_s^2}{2} \cdot \frac{1}{\mu}, \quad (1)$$

where $C_a^2 = \frac{Var(IAT)}{(E[IAT])^2}$ (variance of IAT normalized by square of mean IAT, or square of the standard deviation of IAT normalized by mean IAT) and $C_s^2 = \frac{Var(ST)}{(E[ST])^2}$. Kingman's formula is known to provide accurate results in heavy traffic, that is, when $\rho \rightarrow 1$. Alternatively, diffusion approximation results also exist for the $G/G/1$ model [13].

For the $G/G/k$ model, very few general results are known. Instead, approximate results under specific traffic conditions are available. For example, in heavy traffic, Eq. (1) has the following simple multi-server analogue [14]:

$$E[W] \approx \frac{\rho}{1-\rho} \cdot \frac{C_a^2 + C_s^2}{2k} \cdot \frac{1}{\mu} \quad (2)$$

Diffusion approximations [15] and other approximations [16] and numerical procedures also exist for analyzing $G/G/k$ models. Nonetheless, they remain approximations.

Almost all approximation results for $G/G/k$ models use at most the first two moments of the IAT and ST distributions. As a result, the effect of distributional shape beyond the mean and SCV has received comparatively little attention, especially for open queueing networks. Some prior work has analyzed the effect of kurtosis on the number of jobs in the queue, but not on response time [17]. Given the significant contribution of the second moment of IAT and ST in existing approximation

results for $G/G/k$ queues (see, for example, the C^2 terms in Eq. (2)), an important question to consider is **how do higher moments of the IAT and ST distributions impact performance?** Our numerical analysis in Section IV is guided by this question.

$H_2/H_2/k$ models: The $H_2/H_2/k$ model falls under the family of the $G/G/k$ model, but allows for feasible analysis given that a specific distribution (H_2 , in this case) is being considered for IAT and ST. For the case of $k = 1$, closed-form results for mean response time (in terms of the roots of a cubic equation) are available using transforms and spectral analysis [18]. For $k > 1$, no closed-form results are published for the $H_2/H_2/k$ model. Nonetheless, as we discuss next, it is possible to set up the Markov chain for the $H_2/H_2/k$ model and solve it for $E[T]$ via matrix analytic methods.

III. UNDERSTANDING THE H_2 DISTRIBUTION AND SOLVING THE $H_2/H_2/k$ MODEL

A. The Hyper-exponential distribution

The Hyper-exponential distribution with k phases, denoted as H_k , is a mixture of k Exponential distributions:

$$H_k = \begin{cases} \text{Exp}(\lambda_1) & \text{with probability } p_1 \\ \text{Exp}(\lambda_2) & \text{with probability } p_2 \\ \vdots & \\ \text{Exp}(\lambda_k) & \text{with probability } p_k, \end{cases} \quad (3)$$

where $p_1 + p_2 + \dots + p_k = 1$. Since the H_k is a probabilistic mixture of Exponentials, its moments can be easily derived in closed-form. For example, the mean (first moment) of H_k is $\sum_{i=1}^k p_i/\lambda_i$. A key advantage of the H_k distribution is the fact that it is **flexible** and can be tuned (via its λ and p parameters) to model most heavy-tailed distributions [19]. As such, the Hyper-exponential distribution, though not as popularly employed as the Exponential, has been used for modeling various parameters of interest in manufacturing systems [20] and computer science [10].

The H_2 distribution and its central moments: Recent work has shown that the Hyper-exponential distribution with just 2 phases (H_2) is flexible enough to accurately model the IAT and ST behavior of storage and internet traces [8], [9]. Further, with just 2 phases, the H_2 distribution is easy to analyze and lends itself to tractable queueing analysis, as we discuss next.

The H_2 distribution has three parameters and can be expressed as

$$H_2 = \begin{cases} \text{Exp}(\lambda_1) & \text{with probability } p, \\ \text{Exp}(\lambda_2) & \text{with probability } 1-p. \end{cases} \quad (4)$$

Given the above definition, and using the moments of the exponential distribution, the moments of an H_2 distribution can be written in closed form. If $X \sim H_2(p, \lambda_1, \lambda_2)$, then

$$E[X] = \frac{p}{\lambda_1} + \frac{1-p}{\lambda_2}, \quad (5)$$

$$Var(X) = \frac{2p}{\lambda_1^2} + \frac{2(1-p)}{\lambda_2^2} - (E[X])^2. \quad (6)$$

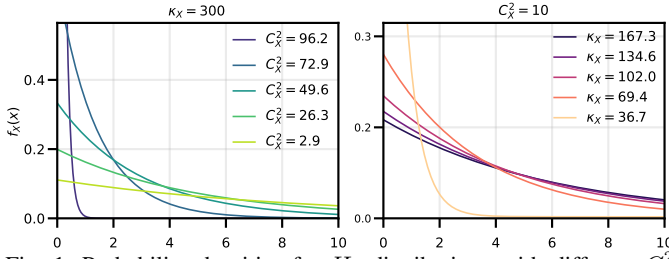


Fig. 1: Probability densities for H_2 distributions with different C^2 and κ values. The left plot shows distributions with varying C^2 for a fixed κ and the right one varies κ for a fixed C^2 .

To capture variance in a scale-independent manner, queueing analysis often employs the squared coefficient of variation (SCV), denoted by C_X^2 :

$$C_X^2 = \frac{\text{Var}(X)}{(E[X])^2}. \quad (7)$$

For an exponential distribution, $C^2 = 1$. For a proper non-degenerate hyperexponential distribution, $C^2 > 1$, because mixing exponentials with different rates introduces additional variability beyond that of a single exponential distribution.

Much of the existing approximation work on $G/G/k$ queues uses only the first two moments; for example, Kingman's formula in Eq. (1) depends on the mean and SCV of the inter-arrival and service-time distributions. To investigate the effect of higher moments, one can also consider the third and fourth central moments, namely skewness and kurtosis. We focus on kurtosis because it captures fourth-order shape information not determined by the mean and SCV, and is commonly used as a measure of tail heaviness [21]. Skewness, on the other hand, measures asymmetry. Kurtosis, denoted by κ , is defined as

$$\kappa = \text{Kurt}[X] = \frac{E[(X - E[X])^4]}{(\text{Var}(X))^2}. \quad (8)$$

A normal distribution has $\kappa = 3$, while an exponential distribution has $\kappa = 9$.

The SCV and kurtosis of a hyperexponential distribution are not independent. Using the standard moment formula for mixtures of Exponentials (see, e.g., Whitt [22]) and log-convexity of moments, it can be shown that:

$$\kappa_X \geq 3(C_X^2 + 1 + 1/C_X^2). \quad (9)$$

Since hyperexponential distributions have $C^2 > 1$, by Eq. (9), their $\kappa > 9$. In the H_2 family, changing κ while holding the mean and SCV fixed changes the balance between probability mass near short values and rare long values. Therefore, the three quantities $E[X]$, C^2 , and κ provide a useful parametrization for separating scale, second-order variability, and higher-order tail/shape effects in the queueing experiments. We use this moment-based parametrization to characterize the H_2 inter-arrival-time and service-time distributions throughout this work. We acknowledge that the broader concept of variability in a distribution may not fully be captured by moments alone [23]; however, it is still instructive to evaluate how performance is affected by such moments.

To visually illustrate these shape effects, Figure 1 presents two sets of H_2 probability density functions. The left subfigure demonstrates the effect of varying the squared coefficient of variation while holding kurtosis constant. As C_X^2 increases, the distribution becomes significantly peakier and more concentrated near the origin. Because the kurtosis is fixed, this increased variability is driven by an abundance of very small and moderate values, rather than increasingly extreme rare-tail events. The second panel shows the reverse scenario: holding total variability constant ($C_X^2 = 10$) while varying kurtosis. In this case, a higher κ_X yields a flatter, fatter-tailed distribution. Conversely, a lower κ_X produces a distribution with a sharper peak near the origin and a much weaker far tail, meaning extreme values are less likely.

B. Matrix analytic method solution for $H_2/H_2/k$ queues

The H_2 distribution, being a probabilistic mixture of two exponentials, admits a Markovian representation; consequently, the $H_2/H_2/k$ queue can be modelled as a continuous-time Markov chain. Figure 2 shows the components of this Markov chain where the IAT distribution is $H_2(p, \lambda_1, \lambda_2)$ and the ST distribution is $H_2(q, \mu_1, \mu_2)$. The state space is represented as (n, a, b) where $n \geq 0$ is the total number of jobs in the system (including those being served and those waiting, if any); $a \leq \min(n, k)$ is the number of jobs being served whose service time is in phase 1 of the ST distribution; and $b \in \{1, 2\}$ is the phase of the arrival time of the next job about to arrive to the system. The blue and red arrows show transitions corresponding to arrival and completion of a job, respectively, with their transition rates denoted above them.

On an arrival from phase b (with rate λ_b), the number of jobs increases by one, the next inter-arrival phase is chosen as Bernoulli(p), and if a server is idle the new job immediately enters service with service phase chosen as Bernoulli(q). On a service completion, the number of jobs decreases by one; if the system was not queued, the number of phase-1 jobs (i.e., a in the state tuple) changes only according to the phase of the departing job. The departure rate aggregates over all busy servers: phase-1 servers each contribute μ_1 and phase-2 servers each contribute μ_2 , giving a total rate of $a\mu_1 + (\min(n, k) - a)\mu_2$. When the system is queued, a departure is immediately replaced by a waiting job, so the transition also accounts for the service phase of the departing job and the newly entering job. This is an infinite Markov Chain with an initial portion when $n \leq k$ (first two subfigures of Figure 2), and a repeating portion for $n > k$ (third subfigure).

Given the repeating structure of the Markov chain, the $H_2/H_2/k$ model can be solved via the iterative matrix analytic methods to obtain the stationary probabilities $\pi_{n,a,b}$ for the Markov Chain states. The steady-state distribution of the number of jobs (N) in the system, can then be obtained as $P(N = n) = \sum_{a,b} \pi_{n,a,b}$. This steady state distribution of N can be used to derive the mean response time, $E[T]$.

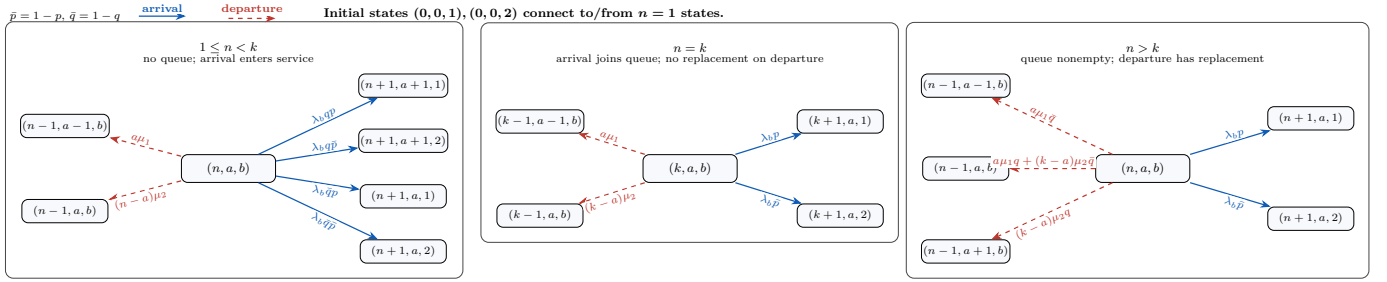


Fig. 2: Illustrative states of the $H_2/H_2/k$ Markov Chain with $IAT \sim H_2(p, \lambda_1, \lambda_2)$ and the $ST \sim H_2(q, \mu_1, \mu_2)$.

IV. EVALUATION RESULTS

This section presents our evaluation results and insights for our $H_2/H_2/k$ model where the IAT is distributed as $H_2(p, \lambda_1, \lambda_2)$ and ST is distributed as $H_2(q, \mu_1, \mu_2)$. For our evaluation, we sweep through different ranges for the mean, C^2 , and κ for the IAT and ST distributions. As discussed earlier in Section III-A, this moment-based parametrization lets us vary scale, variability, and higher-order tail/shape effects separately. The sweep is done by solving for the H_2 parameters (e.g., p, λ_1, λ_2 for IAT) that result in the desired mean, C^2 , and κ values via Eqs. (5), (7), and (8). The resulting system consists of coupled polynomial equations, which we symbolically solved using Mathematica. This resulted in closed-form (yet tedious) expressions, which translated to a unique feasible solution in each case. Note, per Eq. (9), that the values of C^2 and κ cannot be chosen independently; we thus exhaustively choose relevant values satisfying this constraint.

We first present our results for the impact of the first moments of IAT and ST on $E[T]$ for $H_2/H_2/k$ with $k = 1, 2, 5, 10$, corresponding to varying number of servers. We then analyze the impact of C_a^2 (the squared coefficient of variation of IAT) and C_s^2 (the squared coefficient of variation

of ST), and finally, the impact of κ_a (the kurtosis of IAT) and κ_s (the kurtosis of ST). For the C^2 and κ experiments, we consider two representative load values, $\rho = \frac{\lambda}{k \cdot \mu} = 0.3$ (representative of realistic loads reported by cloud computing providers [24], [25]) and $\rho = 0.9$ (representative of high load). Unless mentioned otherwise, we fix $E[ST] = 1$ for ease of presentation, following the approach in prior works [1].

A. Impact of mean of IAT and ST

Figure 3 shows $E[T]$ on a log scale as a function of $E[IAT]$ for $k = 1, 2, 5, 10$. We fix $E[ST] = 1$, $\kappa_a = 100$, and $\kappa_s = 80$, and consider four representative pairs of squared coefficients of variation: $(C_a^2, C_s^2) = (4, 4), (4, 25), (25, 4), (25, 25)$. As expected, mean response time decreases steeply with $E[IAT]$ at low values, then gradually flattens. Increasing the number of servers lowers the absolute response times; by $k = 10$, all curves approach the service time floor $E[ST] = 1$ once $E[IAT]$ is sufficiently far from the high-load region.

The figure also shows that variability in both the arrival and service distributions matters, but not equally. The curve for $(C_a^2, C_s^2) = (25, 25)$ is consistently the highest, while $(4, 4)$ is the lowest. More importantly, the two curves with $C_a^2 = 4$,

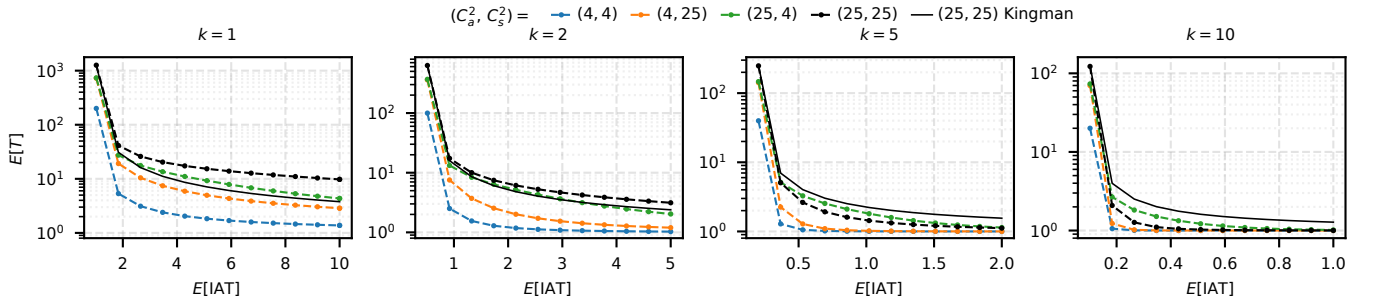


Fig. 3: Impact of $E[IAT]$ on mean response time of $H_2/H_2/k$ for $k = 1, 2, 5, 10$ and different settings of (C_a^2, C_s^2) . Here $E[ST] = 1$, $\kappa_a = 100$, and $\kappa_s = 80$. The y-axis is on log scale. Kingman's approximation (black, solid line) is shown for $(C_a^2, C_s^2) = (25, 25)$.

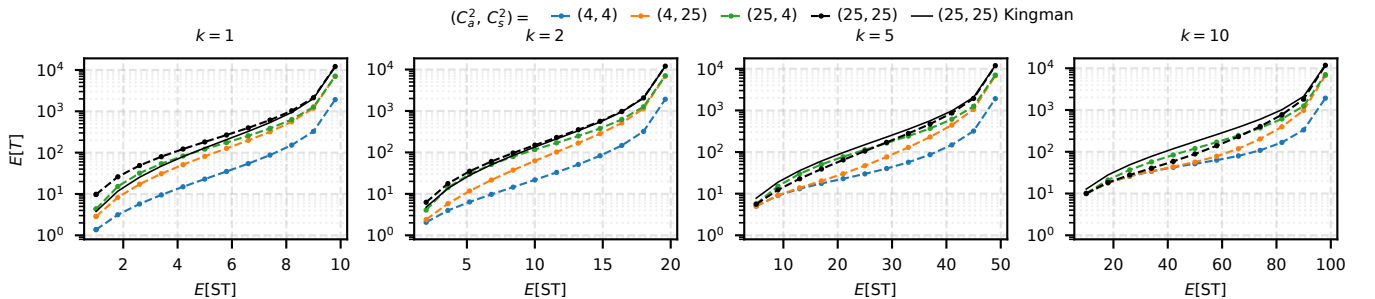


Fig. 4: Impact of $E[ST]$ on mean response time of $H_2/H_2/k$ for $k = 1, 2, 5, 10$ and different settings of (C_a^2, C_s^2) . Here $E[IAT] = 10$, $\kappa_a = 100$, and $\kappa_s = 80$. The y-axis is on log scale. Kingman's approximation (black, solid line) is shown for $(C_a^2, C_s^2) = (25, 25)$.

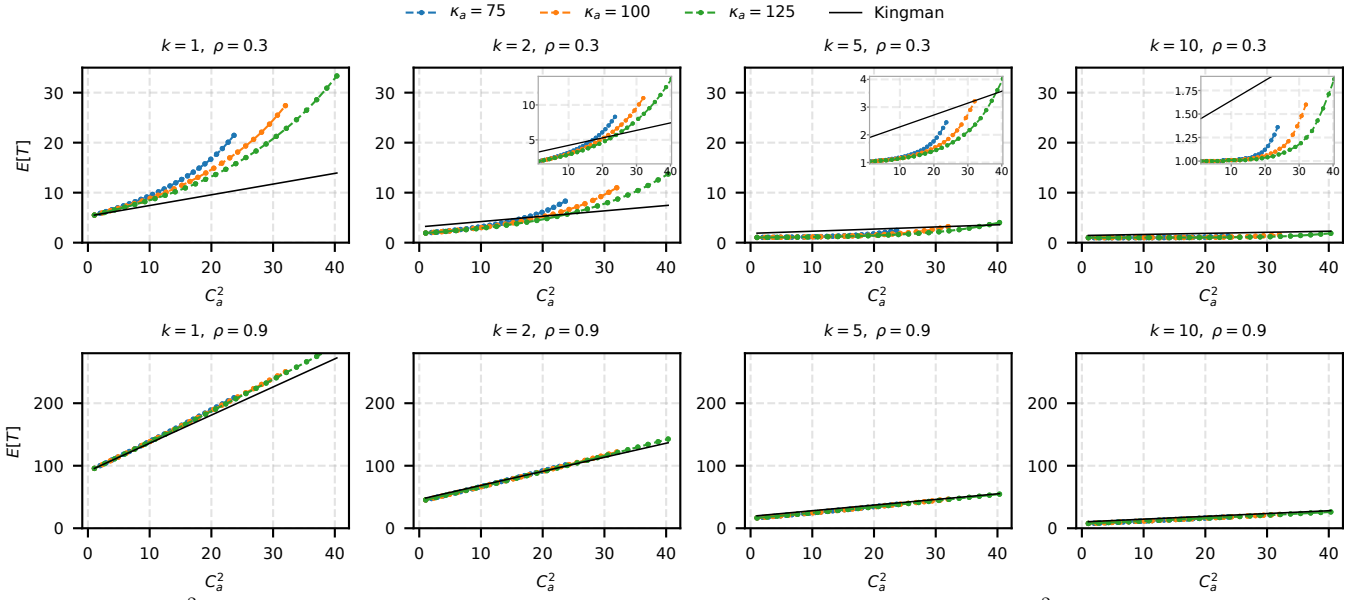


Fig. 5: Impact of C_a^2 on mean response time of $H_2/H_2/k$ for $k = 1, 2, 5, 10$ and $\rho \in 0.3, 0.9$. We fix $C_s^2 = 20$ and $\kappa_s = 100$, and vary $\kappa_a \in 75, 100, 125$. Kingman's approximation is shown as the black, solid line.

namely (4, 4) and (4, 25), remain below the two curves with $C_a^2 = 25$, namely (25, 4) and (25, 25). **Increasing the arrival SCV from 4 to 25 significantly impacts $E[T]$** , while the impact on $E[T]$ of **increasing C_s^2 from 4 to 25 is less significant**.

Figure 4 fixes $E[IAT] = 10$ and varies $E[ST]$. As $E[ST]$ increases, $E[T]$ increases in each subfigure (corresponding to different k settings). The **same ordering of the C^2 cases is again visible**: (4, 4) gives the smallest response times and (25, 25) gives the largest, while the two cases with $C_a^2 = 4$ remain below the two cases with $C_a^2 = 25$. Thus, the separation seen in Figure 3 persists when the service mean is varied instead of the inter-arrival mean.

In addition to the exact $H_2/H_2/k$ results, Figures 3 and 4 also show the Kingman approximation for the single-server case and its $1/k$ -scaled version for the multi-server cases. We plot this approximation only for the high-variability case $C_a^2 = C_s^2 = 25$. As expected, the Kingman approximation is close to the exact curve at high load (high $E[ST]$ or low $E[IAT]$), while the difference becomes larger at low load; Kingman underestimates for $k = 1, 2$ and overestimates for $k = 5, 10$.

B. Impact of squared coefficient of variation of IAT and ST

We now study in detail the impact of C_a^2 and C_s^2 on $E[T]$. We fix $E[ST] = 1$ and choose $E[IAT]$ so that $\rho = \frac{E[ST]}{k \cdot E[IAT]}$ equals either 0.3 or 0.9. Figure 5 shows $E[T]$ as a function of C_a^2 while fixing $C_s^2 = 20$ and $\kappa_s = 100$. The different curves in each subfigure correspond to different κ_a values, and for each κ_a , C_a^2 goes up to its maximum permissible value subject to Eq. (9). For all server counts and for both load values, **$E[T]$ increases with C_a^2** . For the moderate load setting of $\rho = 0.3$, **the increase is visibly convex and superlinear**. This behavior is consistent with the shapes of the H_2 arrival distribution seen in Figure 1: when κ_a is fixed, increasing C_a^2 places more mass near very small inter-arrival times, producing burstier arrivals.

The pattern is consistent across k ; as k increases, absolute response times and their sensitivity to C_a^2 diminish. The inset zoomed figures (shown within subfigures for $k = 2, 5, 10$ and $\rho = 0.3$) confirm that the qualitative behavior remains similar.

Observation 1: $E[T]$ increases superlinearly with C_a^2 across all loads and server counts, with a large overall range of increase.

In the moderate-load regime, **Kingman's approximation captures the increasing trend with C_a^2 but diverges from the true curves**. Because the actual curves exhibit a superlinear dependence on C_a^2 , this discrepancy widens as C_a^2 increases.

At $\rho = 0.9$, the absolute response times are much larger, and the curves in Figure 5 are close to linear in C_a^2 . In this high-load regime, the curves are also close to the Kingman-type heavy-load approximation, and the different κ_a curves are nearly indistinguishable. Thus, at high load, the dominant effects are the load and the SCV, while the additional effect of changing κ_a is small. At $\rho = 0.3$, the curves separate more clearly. **For a fixed value of C_a^2 , larger κ_a generally gives a somewhat smaller $E[T]$, especially when C_a^2 is large**. This effect is most visible for smaller k ; as k increases, the absolute response times decrease and the curves move closer together.

Figure 6 shows $E[T]$ as a function of C_s^2 , with $C_a^2 = 20$ and $\kappa_a = 100$ fixed. For $k = 1, 2$, $E[T]$ increases with C_s^2 at both load values. At $\rho = 0.9$, the increase is again nearly linear, the different κ_s curves are very close to each other, and the Kingman approximation fairly captures the trend.

For the moderate-load case ($\rho = 0.3$), **the behavior depends on k** . For $k = 1$ and $k = 2$, $E[T]$ increases roughly linearly in C_s^2 , with a slope close to Kingman's prediction (though the absolute values deviate). **This is in contrast to the superlinear dependence on C_a^2 : the sensitivity of $E[T]$ to C_s^2 remains moderate and does not amplify at large C_s^2** .

As k increases, the behavior departs from linearity and, **surprisingly, even from monotonicity**. For $k = 5$ and $k = 10$, the curves appear nearly flat on the common y -axis scale, but

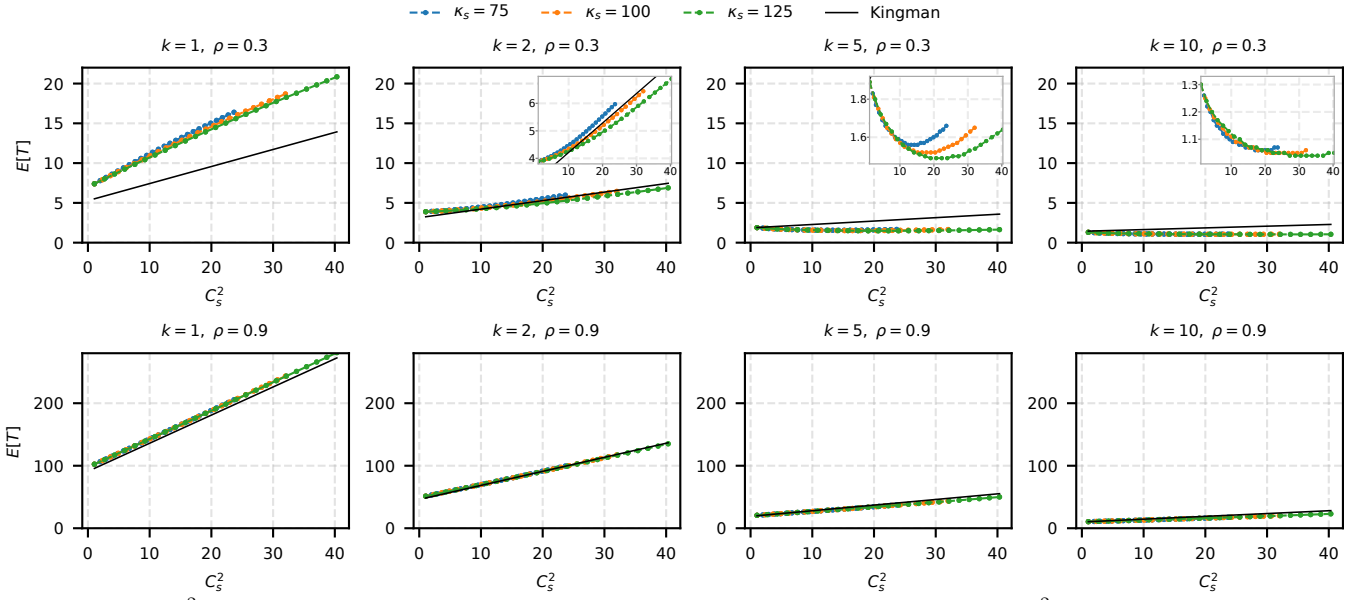


Fig. 6: Impact of C_s^2 on mean response time of $H_2/H_2/k$ for $k = 1, 2, 5, 10$ and $\rho \in 0.3, 0.9$. We fix $C_a^2 = 20$ and $\kappa_a = 100$, and vary $\kappa_s \in 75, 100, 125$. Kingman's approximation is shown as the black, solid line.

zooming reveals **a shallow non-monotone dependence: $E[T]$ first decreases, reaches a minimum, and then increases**. The effect is small in absolute terms, with $E[T]$ ranging approximately 1.5–1.9 for $k = 5$ and 1.04–1.30 for $k = 10$; however, **it establishes that, for multi-server queues at moderate load, the response time can depend non-monotonically on C_s^2 . The Kingman approximation, being an increasing straight line, gets very inaccurate for these cases.**

Observation 2: $E[T]$ increases with C_s^2 more moderately than C_a^2 , becoming non-monotone at moderate loads with larger k .

To rule out the possibility that this non-monotonic behavior is a numerical artifact, we independently simulated the $H_2/H_2/k$ queue using Ciw, a Python-based discrete-event simulator for queueing networks [26]. The simulated mean response times closely matched those from our matrix analytical method. For example, for $C_s^2 = 3$ and 18 in the $k = 5$, $\rho = 0.3$, $\kappa_s = 125$ case in Fig. 6, the analytical values are 1.78 and 1.50, respectively. For each point, we ran 100 independent one-million-second simulations (about 1.5 million arrivals per run), with appropriate warm-up and cool-down periods. The corresponding simulated means were 1.780 and 1.495, with standard deviations within 1.2 percent. These results support the observed non-monotonic behavior.

This non-monotone behavior can be explained by separating the mean waiting time $E[W_q]$ into the frequency of waiting and the severity of waiting. Since $E[ST]$ is fixed in this experiment, the behavior of $E[T] = E[W_q] + E[ST]$ is determined by $E[W_q]$. We use the decomposition:

$$E[W_q] = P(W_q > 0) \cdot E[W_q | W_q > 0]. \quad (10)$$

The first factor measures how often an arriving job finds all servers busy, while the second factor measures the mean wait time conditional on this event. As discussed in Section III-B, our matrix-analytic solution gives the stationary probabilities

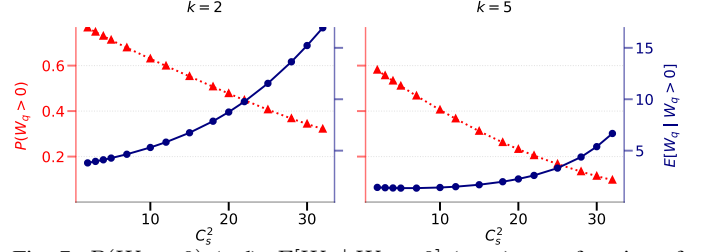


Fig. 7: $P(W_q > 0)$ (red), $E[W_q | W_q > 0]$ (navy), as a function of C_s^2 for $\rho = 0.3$, $\kappa_s = 125$, and $k = 2$ and $k = 5$.

$\pi_{n,a,b}$ of the states of the system, where n is the total number of jobs in the system, a is the number of phase-1 services in progress, and $b \in \{1, 2\}$ is the arrival phase. From these the probability that an arrival has to wait can be computed directly. If arrivals occur from phase b at rate λ_b , then:

$$P(W_q > 0) = \frac{\sum_{n \geq k, a, b} \pi_{n,a,b} \lambda_b}{\sum_{n,a,b} \pi_{n,a,b} \lambda_b}. \quad (11)$$

The weighting by λ_b is needed since the distribution seen by arrivals is not, in general, the same as the time-average.

Figure 7 shows $P(W_q > 0)$ and $E[W_q | W_q > 0]$ as functions of C_s^2 for $\kappa_s = 125$, $k \in \{2, 5\}$, and $\rho = 0.3$ (corresponding to the top row, middle two subfigures in Figure 6). This plot shows that **increasing C_s^2 decreases $P(W_q > 0)$ for both $k = 2$ and $k = 5$** . This is consistent with the behavior of the H_2 service distribution as seen in Figure 1: when κ_s is fixed, increasing C_s^2 makes the distribution more concentrated near zero, so many services are completed very quickly. In a multi-server system, these short services help keep servers available and reduce the frequency with which an arrival sees all servers busy. At the same time, $E[W_q | W_q > 0]$ increases with C_s^2 , because the same increase in variability also produces longer service times, which make waiting episodes more severe once queueing occurs. The difference between $k = 2$ and $k = 5$ is explained by the balance of these

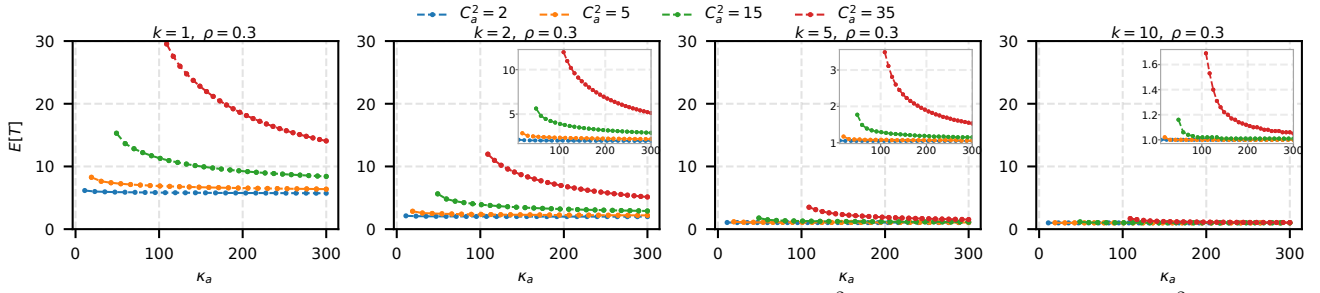


Fig. 8: Impact of κ_a on $E[T]$ of $H_2/H_2/k$ for $k = 1, 2, 5, 10$ and $\rho = 0.3$. We fix $C_s^2 = 20$ and $\kappa_s = 100$, and vary $C_a^2 \in 2, 5, 15, 35$.

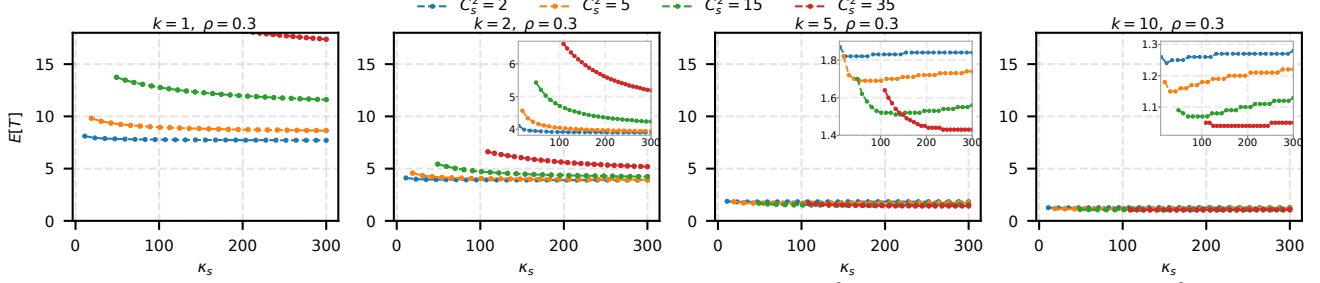


Fig. 9: Impact of κ_s on $E[T]$ of $H_2/H_2/k$ for $k = 1, 2, 5, 10$ and $\rho = 0.3$. We fix $C_a^2 = 20$ and $\kappa_a = 100$, and vary $C_s^2 \in 2, 5, 15, 35$.

two effects. For $k = 2$, the increase in conditional severity dominates the decrease in waiting probability, whereas for $k = 5$, the decrease in $P(W_q > 0)$ initially dominates, so $E[T]$ decreases; only for larger C_s^2 does the conditional severity become large enough to reverse the trend.

C. Impact of kurtosis of IAT and ST

We now analyze the impact of the kurtosis of IAT and ST on $E[T]$. We keep the SCV fixed and vary the kurtosis parameter of the distributions. Figure 8 shows $E[T]$ as a function of the arrival kurtosis κ_a , with $C_s^2 = 20$ and $\kappa_s = 100$. The **effect of κ_a is most pronounced when the arrival SCV C_a^2 is large**. For $C_a^2 = 2$, the curves are nearly flat across all subfigures. For $C_a^2 = 15$, and especially for $C_a^2 = 35$, **increasing κ_a reduces $E[T]$** . This **reduction is substantial at moderate load**. For example, at $\rho = 0.3$ and $k = 1$, the $C_a^2 = 35$ curve drops from about 30 at $\kappa_a = 100$ to roughly 14 near $\kappa_a = 300$. For $k = 5$ and $k = 10$, the absolute response times are smaller, but the same decreasing trend remains visible for large C_a^2 , though with a smaller magnitude. **At high load, the dependence on κ_a is much weaker**. For $\rho = 0.9$ (figure omitted due to lack of space), changing κ_a produces a minimal decrease in $E[T]$.

The decreasing trend with κ_a can be understood from the structure of the H_2 family. For fixed mean and SCV, increasing arrival kurtosis makes very large interarrival times more prominent in the tail. These long gaps can allow a built-up queue to drain, thereby reducing the mean response time. Lower κ_a , in contrast, places relatively more mass on short-to-moderate interarrival times, producing more bursty arrivals and hence larger congestion. This mechanism is noticeable only when arrival variability is already large enough to create substantial queue build-up, which explains why the effect is weak for $C_a^2 = 2$ but strong for $C_a^2 = 35$.

Figure 9 shows the corresponding results for the service kurtosis κ_s , with $C_a^2 = 20$ and $\kappa_a = 100$ fixed. The **effect**

of κ_s is much weaker. For $k = 1$ and $k = 2$ at $\rho = 0.3$, increasing κ_s reduces $E[T]$ modestly, especially for the larger values of C_s^2 , while the lower- C_s^2 curves are nearly flat. At $\rho = 0.9$ (figure omitted), all curves are almost flat.

The $k = 5$ and $k = 10$ subfigures also show nearly flat curves. However, at $\rho = 0.3$, zoomed-in views reveal a very slight non-monotonicity. As illustrated by the service-time shapes in Figure 1, and as in the case of C_s^2 , this dependence should be viewed as the net effect of several interacting features: mass near small service times, tail behavior, queue formation, server occupancy, and conditional waiting.

Observation 3: Kurtosis provides secondary modulation, mainly when the corresponding SCV is large, with κ_a having a more visible effect than κ_s .

D. Extension to other performance metrics and tails

Our matrix-analytic solution of the $H_2/H_2/k$ Markov chain allows us to analyze metrics beyond mean response time. By solving for the steady-state distribution, we can recover the distribution of the number of jobs in the system, making it possible to study tail probabilities, such as $P(N > n)$.

Figure 10 shows the tail probability $P(N > n)$ when the ST distribution is fixed and the IAT distribution is varied. The two subfigures correspond to $k = 1$ and $k = 5$, with the load fixed at $\rho = 0.3$. The dominant effect in Figure 10 is the SCV of the IAT distribution, C_a^2 . For both $k = 1$ and $k = 5$, curves with $C_a^2 = 20$ exhibit substantially heavier tails than those with $C_a^2 = 4$. Thus, increasing arrival variability leads not only to a larger mean response time, as observed earlier, but also to a higher probability of large numbers of jobs accumulating in the system. The effect of κ_a is visible only at larger values of C_a^2 . For $C_a^2 = 4$ the two κ_a curves (green and blue) almost coincide; for $C_a^2 = 20$, the smaller value $\kappa_a = 70$ produces the heavier tail, while $\kappa_a = 150$ leads to a faster decay. This is consistent with the earlier observation on mean response time.

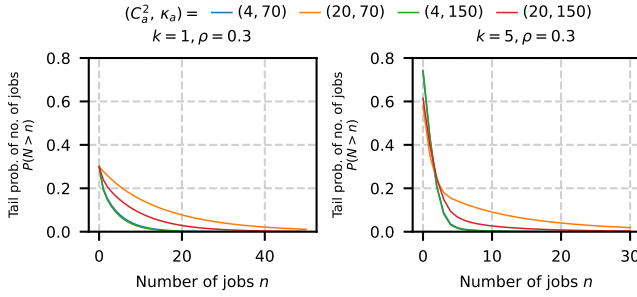


Fig. 10: Tail probability of jobs in the system, $P(N > n)$, with fixed ST distribution ($C_s^2 = 4$, $\kappa_s = 25$) and varying IAT parameters.

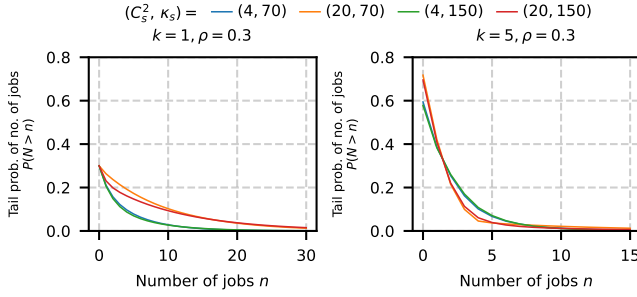


Fig. 11: Tail probability of jobs in the system, $P(N > n)$, with fixed IAT distribution ($C_a^2 = 4$, $\kappa_a = 25$) and varying ST parameters.

Figure 11 presents the complementary experiment, in which the IAT distribution is held fixed and the ST distribution is varied. The service-side effects in Figure 11, like the arrival-side effects in Figure 10, are driven more by SCV than by kurtosis. For $k = 1$, increasing C_s^2 from 4 to 20 produces a noticeably heavier tail. The effect of κ_s is comparatively small and becomes noticeable only when C_s^2 is large. For $k = 5$, the tail probabilities decay much more rapidly and the separation among curves for different service-time parameter settings is considerably smaller. The counter-intuitive phenomenon noted earlier is also apparent in the multi-server case: a lower value of C_s^2 yields a slightly heavier tail compared to the higher one.

Observation 4: Our results in Figures 10 and 11 show that C_a^2 is the primary driver of tail heaviness, with C_s^2 playing an analogous but weaker role; in both cases, κ provides secondary modulation that matters only at large SCV values.

V. CONCLUSIONS

This paper numerically evaluated how moment-based summaries of the IAT and ST distributions affect the performance of $H_2/H_2/k$ queues. Using the Markovian structure induced by the two-phase hyper-exponential representation, we solved the model by matrix-analytic methods and studied mean response time as well as tail probabilities of the number of jobs in the system. Our evaluation shows that C_a^2 produces superlinear increases in $E[T]$ at moderate load, while C_s^2 has a qualitatively different effect: a more gradual dependence for small k that becomes non-monotone in multi-server systems, explained by competing effects of waiting frequency and waiting severity. Kurtosis acts as a secondary modulator that matters mainly when the corresponding SCV is large, with a more pronounced effect at moderate load.

REFERENCES

- [1] J. F. C. Kingman, "The single server queue in heavy traffic," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 57, no. 4, pp. 902–904, 1961.
- [2] L. Kleinrock, *Queueing Systems, Volume 2*. Wiley-Interscience, 1976.
- [3] Kleinrock, L., *Queueing Systems, Volume 1*. Wiley-Interscience, 1975.
- [4] K. Keeton, A. Veitch, D. Obal, and J. Wilkes, "I/O Characterization of Commercial Workloads," in *3rd Workshop on Computer Architecture Evaluation using Commercial Workloads*, Toulouse, France, 2000.
- [5] R. Gracia-Tinedo, Y. Tian, J. Sampé, H. Harkous, J. Lenton, P. García-López, M. Sánchez-Artigas, and M. Vukolic, "Dissecting UbuntuOne: Autopsy of a Global-scale Personal Cloud Back-end," in *Proceedings of the 2015 Internet Measurement Conference*, Tokyo, Japan, pp. 155–168.
- [6] S. Kavalanekar, B. Worthington, Q. Zhang, and V. Sharda, "Characterization of storage workload traces from production Windows Servers," in *Proceedings of the 2008 IEEE International Symposium on Workload Characterization*, Seattle, WA, USA, 2008, pp. 119–128.
- [7] A. Verma, R. Koller, L. Useche, and R. Rangaswami, "SRCMap: Energy proportional storage using dynamic consolidation," in *Proceedings of the 8th USENIX Conference on File and Storage Technologies*, 2010.
- [8] M. Wajahat, A. Yele, T. Estro, A. Gandhi, and E. Zadok, "Analyzing the distribution fit for storage workload and internet traffic traces," *Performance Evaluation*, vol. 142, pp. 102–121, 2020.
- [9] Wajahat, Muhammad and Yele, Aditya and Estro, Tyler and Gandhi, Anshul and Zadok, Erez, "Distribution Fitting and Performance Modeling for Storage Traces," in *Proceedings of the 27th IEEE MASCOTS*, ser. MASCOTS '19, Rennes, France, 2019.
- [10] A. Feldmann and W. Whitt, "Fitting mixtures of exponentials to long-tail distributions to analyze network performance models," *Performance Evaluation*, vol. 31, no. 3-4, pp. 245–279, 1998.
- [11] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Philadelphia, PA, USA: ASA-SIAM, 1999.
- [12] D. Lindley, "The theory of queues with a single server," *Mathematical Proc. of the Cambridge Phil. Soc.*, vol. 48, no. 2, pp. 277–289, 1952.
- [13] P. G. Harrison and N. M. Patel, *Performance modelling of communication networks and computer architectures*. Addison-Wesley, 1993.
- [14] W. Whitt, "Approximations for the gi/g/m queue," *Production and Operations Management*, vol. 2, no. 2, pp. 114–161, 1993.
- [15] B. Halachimi and W. Franta, "A diffusion approximate solution to the g/g/k queueing system," *Computers and Operations Research*, vol. 4, no. 1, pp. 37–46, 1977.
- [16] W. Whitt, "Approximations for the gi/g/m queues," *Productions and Operations Management*, vol. 2, no. 2, pp. 114–161, 1993.
- [17] A. Brandwajn and T. Begin, "Higher-order distributional properties in closed queueing networks," *Perf. Eval.*, vol. 66 (11), pp. 607–620, 2009.
- [18] V. N. Tarasov, "Analysis of queues with hyperexponential arrival distributions," *Problems of Information Transmission*, vol. 52, no. 1, pp. 14–23, 2016.
- [19] D. R. Cox, "A use of complex probabilities in the theory of stochastic processes," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 51, no. 2, pp. 313–319, 1955.
- [20] H. Papadopoulos and C. Heavey, "Queueing theory in manufacturing systems analysis and design: A classification of models for production and transfer lines," *European journal of operational Research*, vol. 92, no. 1, pp. 1–27, 1996.
- [21] L. DeCarlo, "On the meaning and use of kurtosis," *Psychological Methods*, vol. 2, no. 3, pp. 292–307, 1997.
- [22] W. Whitt, "On approximations for queues, iii: Mixtures of exponential distributions," *AT&T Bell Labs Technical Journal*, vol. 63, no. 1, pp. 163–175, 1984.
- [23] V. Gupta, M. Harchol-Balter, J. G. Dai, and B. Zwart, "On the inapproximability of m/g/k: why two moments of job size distribution are not enough," *Queueing Systems*, vol. 64, no. 1, pp. 5–48, 2010.
- [24] H. Liu, "A Measurement Study of Server Utilization in Public Clouds," in *Proceedings of the 9th IEEE International Conference on Dependable, Autonomic and Secure Computing*, Sydney, 2011, pp. 435–442.
- [25] C. Lu, K. Ye, G. Xu, C. Z. Xu, and T. Bai, "Imbalance in the cloud: An analysis on Alibaba cluster trace," in *Proceedings of the 2017 IEEE International Conference on Big Data*, Boston, USA, pp. 2884–2892.
- [26] G. I. Palmer, V. A. Knight, P. R. Harper, and A. L. Hawa, "Ciw: An open-source discrete event simulation library," *Journal of Simulation*, vol. 13, no. 1, pp. 68–82, 2019.