M/G/k with Exponential Setup

Anshul Gandhi* **Mor Harchol-Balter***

September 2009
CMU-CS-09-166

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

*School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

Research supported by an NSF CCR-0615262 grant.

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of NSF.

Keywords: Queueing theory, M/G/k

Abstract

In this paper, we consider the M/G/k queueing system with setup times. This particular queueing model is common in manufacturing systems, where idle machines are turned off to save on operating costs, as well as in server farms, where idle servers are turned off to conserve power. While recent literature has analyzed the M/M/k system with exponential setup times, no closed-form solutions were obtained. We provide the first analytical closed form expressions for the mean response time, limiting distribution of the system states, as well as the z-transform for the number of jobs in system for the M/M/k system with exponential setup times. In particular, we prove the following decomposition property: the mean response time of the M/M/k system with exponential setup times differs from the mean response time of an M/M/k system without setup times, by a constant factor, which is the mean of the exponential setup time. Using matrix analytic methods and simulations, we show that the above decomposition property may also hold for the M/G/k system with exponential setup times.

1 Introduction

1.1 High-level motivation

We consider an M/G/k system, where there is a setup cost (time cost) required to turn on a server which is currently not in use. Setup costs are common in a wide variety of systems. In manufacturing systems there is often a "warmup" time needed to get a machine running, or a "transport time" needed when calling a staff member into work. In data centers, compute servers are often left off to save on power, but then there is a "reboot" time required to turn a server on when it is needed.

The particular model in this paper is motivated by our work on power management in data centers, where the response time (sojourn time) of jobs is a concern, but the overall power usage is a concern as well. In such settings, the setup cost is wasteful in two ways: (i) the setup cost creates a time delay, increasing overall mean response time, (ii) the setup cost wastes power, since full power is used during the duration of the setup cost, although no work is being done. It is therefore important in power management to understand, analytically, the effect of the setup cost, first on mean response time and then on power usage.

1.2 Specific model

Our model assumes that at any point in time, each of the k servers is in one of three states: OFF , ON (being used to serve a job), or SETUP (undergoing the setup cost so that it can be used to run a job). When servers are not in use, they are immediately switched to the OFF state. We assume that *at most one server at a time can be in the SETUP state* (this is common for limiting power usage). When a new job arrives, if there is already a server in the SETUP state, then the job simply joins the queue, otherwise the job picks an OFF server (assuming there is one) and switches it into the SETUP state. When a job completes service at a server, j, the job at the head of the queue is moved to server j, without the need for SETUP, since server j is already ON. Note that even if the job at the head of the queue was already waiting on another server i in SETUP mode, the job at the head of the queue is still directed to server j; server i is then turned off.

1.3 Results

We use the random variable I to denote the setup cost. For our M/M/k setup model, we find a peculiar and beautiful *decomposition property* in the case where I is exponentially-distributed. Letting T denote response time, we find that:

$$\mathbb{E}\left[T^{M/M/k/setup}\right] = \mathbb{E}\left[T^{M/M/k}\right] + \mathbb{E}[I] \qquad (1)$$

where $M/M/k/setup$ denotes an M/M/k system, where there is a setup cost, I, for turning servers on. That is, the mean response time in our setup model is just the mean response time for an M/M/k without setup, plus the mean setup time. To the best of our knowledge, this result is not known. The above result, as well as several other related results for the M/M/k, such as the transform of the number of jobs, is derived in Section 3.

The exponential assumptions in the above result make it analytically verifiable. In Section 4 we attempt to examine how far this result generalizes. Using matrix analytic methods, we are able to show that the above results extends to the case of a 2 server system, where the job size distribution is a 2-phase hyperexponential, H_2, and I is still distributed exponentially:

$$\mathbb{E}\left[T^{M/H_2/2/setup}\right] = \mathbb{E}\left[T^{M/H_2/2}\right] + \mathbb{E}[I]$$

Since the M/G/k is not analytically tractable, we resort to careful simulation experiments to determine (i) whether the result extends to an M/G/k as well, and (ii) whether the result requires that I be exponentially-distributed. We find that Eq. (1) extends to the M/G/k with *exponential setup costs*, when the job size distribution, G, is hyper-exponential, deterministic, bounded exponential or Bounded Pareto. However, if I is not exponentially-distributed, then Eq. (1) may not hold, not even in the case of an M/M/1.

Finally, in Section 5, we apply our results to analytically model power usage and response times in a server farm with setup costs. Our simple closed-form results enable us to solve for server farm parameters which minimize a weighted sum of mean response time and mean power usage.

2 Prior Work

While there has been a great amount of work on single-server queues with setup costs, including the M/G/1 queue with setup costs [15, 13, 5, 8, 6, 7] and the M/G/1/N finite buffer queue with setup costs [12, 14, 10], there has been very little work on multi-server queues with setup costs [1, 3]. We describe the above papers in more detail below.

2.1 M/G/1

As early as 1964, [15] showed that the M/G/1 with setup times has the following mean response time:

$$\begin{aligned} \mathbb{E}\left[T^{M/G/1/setup}\right] &= \mathbb{E}[S] + \frac{\lambda\mathbb{E}[S^2]}{2(1-\mathbb{E}[S])} + \frac{2\mathbb{E}[I] + \lambda\mathbb{E}[I^2]}{2(1+\lambda\mathbb{E}[I])} \qquad (2) \\ &= \mathbb{E}\left[T^{M/G/1}\right] + \frac{2\mathbb{E}[I] + \lambda\mathbb{E}[I^2]}{2(1+\lambda\mathbb{E}[I])} \qquad (3) \end{aligned}$$

In [13], the author considers a multi-class M/G/1 queue with setup times and a variety of queueing disciplines including FCFS and LCFS, and derives the Laplace - Stieltjes transforms of the waiting times for each class. In [5], the M/G/1 queue with generally distributed setup times and vacations is analyzed for the first two moments of the queue length, among other things. The author also considers various service disciplines that govern the vacation process, including the exhaustive service discipline, where the server goes on a vacation only when it has no outstanding jobs waiting in the queue. [8] consider a variant of the M/G/1 with setup times model, wherein the server is turned on only when some N ($N \geq 1$) jobs are present in the system. This particular policy of turning the server on only after N jobs have accumulated in the system is referred to as the N-policy in literature. The authors analyze the queue length distribution under the N-policy, and derive the optimal value of N which minimizes the total operation cost of the system. For the case of a batch arrival process, [6] analyzes a M^x/M/1 queue with random setup times for the queue size distribution. The author later extends his work in [7] to take into account a M^x/G/1 queue with random setup times.

2.2 M/G/1/N

For the case of a single server with a finite buffer, [14] considers a M/G/1/K queue with general service times under the N-policy. The queue length distribution and mean response time are derived under various service disciplines, including the exhaustive service discipline. In [10], the authors numerically analyze a M(n)/G/1/N queue with general setup times and state dependent arrival rates, for the queue length distribution. For the case of a batch arrival process, [12] analyze the queue length distribution of a single server vacation queue with both, general setup times and close-down times and a batch Markovian arrival process.

2.3 M/M/k

For the case of multiple servers with setup times, [3] consider an M/M/k queueing system with exponential service times. The authors solve the steady state equations for the associated Markov chain, using a combination of difference equations and matrix analytic methods. The recursive nature of the difference equations does not yield a closed-form solution, but can be solved numerically. The authors provide numerical results for specific instances of the waiting time, busy periods and queue lengths.

The difference equations method used by [3] was previously used in [1], where the authors consider a Markov chain similar to the M/M/k with exponential setup times. Again, the authors provide recursive formulations for various performance measures, which are then numerically solved for various examples.

The above approach differs from ours in that the above papers do not determine closed-form solutions for the limiting probabilities or the mean response time. In particular, while [3] assume a M/M/k setup model which is identical to ours, they do not derive Eq. (1), showing that the setup time is decomposably additive, nor do they observe this decomposition property in their graphs. They furthermore do not consider job size distributions other than the Exponential job size distribution. In particular, they don't look at phase-type distributions (eg. H_2) or bounded distributions.

3 M/M/k with Setup

We start, in Section 3.1, by presenting, in more detail, the M/M/k setup model that we will use throughout the paper. Then, we consider the case of $k = 1$, and analyze the mean response time for an M/M/1 with exponential setup times, in Section 3.2. We use results from [15] to show that the decomposition property (or Eq. (1)) holds true for the M/M/1 with exponential setup times. Later, in Appendix A, we derive the mean response time using two different techniques: (i) A tagged-job approach, and (ii) A Pre-emptive Last Come First Serve argument. Next, we consider the M/M/k model with exponential setup times. In Section 3.3, we derive the steady state limiting probabilities for this model, and in Section 3.4, we derive simple closed-form expressions for the mean response time, $\mathbb{E}[T]$, the z-transform of the number of jobs in the system, $\hat{N}(z)$, and other interesting performance measures. Finally, in Section 3.5, we consider an alternative version of the M/M/k model with setup costs, where we allow multiple servers to be in the SETUP mode simultaneously. Interestingly, Eq. (1) no longer holds for this particular model.

3.1 Model

We consider a multi server system with k homogenous servers, each with mean service rate $\mu = \frac{1}{\mathbb{E}[X]}$, where X denotes the job size. Unless stated otherwise, we assume that X is exponentially distributed. Jobs arrive into the system according to a Poisson process with rate λ. For stability, we assume that $k \cdot \mu > \lambda$. We define the load in the system as $\rho = \lambda \cdot \mathbb{E}[X] = \frac{\lambda}{\mu}$. Note that $0 \leq \rho < k$.

Each of the k servers is in one of three states: OFF, ON (being used to serve a job), or SETUP (undergoing the setup cost so that it can be used to run a job). When servers are not in use, they are immediately switched to the OFF state. When a new job arrives, if there is already a server in the SETUP state, then the job simply joins the queue, otherwise the job picks an OFF server (assuming there is one) and switches it into the SETUP state. We use I to denote the setup times, with $\mathbb{E}[I] = \frac{1}{\alpha}$. Unless stated otherwise, we assume that the setup times are exponentially distributed, with rate α. When a job completes service at a server, j, the job at the head of the queue is moved to server j, without the need for SETUP, since server j is already ON. Note that even if the job at the head of the queue was already waiting on another server i in SETUP mode, the job at the head of the queue is still directed to server j; server i is then turned off.

In the model just described, we allow at most one server to be turned on at any given time. In Section 3.5, we relax this condition, and consider an alternative version of the M/M/k with setup times, where we allow multiple servers to be in the SETUP mode simultaneously.

3.2 M/M/1

Theorem 1. *For an M/M/1 with exponentially distributed setup times, I, we have:*

$$\mathbb{E}\left[T^{M/M/1/setup}\right] = \mathbb{E}\left[T^{M/M/1}\right] + \mathbb{E}[I]$$

Proof

We use Eq. (2) from [15] to derive $\mathbb{E}\left[T^{M/M/1/setup}\right]$. Since I is Exponentially distributed, we have $\mathbb{E}[I^2] = 2\mathbb{E}^2[I]$. Thus, from Eq. (3), we have:

$$\begin{aligned}\mathbb{E}\big[T^{M/M/1/setup}\big] &= \mathbb{E}\big[T^{M/M/1}\big] + \frac{2\mathbb{E}[I] + \lambda\mathbb{E}[I^2]}{2(1+\lambda\mathbb{E}[I])} \\ &= \mathbb{E}\big[T^{M/M/1}\big] + \frac{2\mathbb{E}[I] + 2\lambda\mathbb{E}^2[I]}{2(1+\lambda\mathbb{E}[I])} \\ &= \mathbb{E}\big[T^{M/M/1}\big] + \mathbb{E}[I]\end{aligned}$$

Note that the expression for $\mathbb{E}\big[T^{M/M/1/setup}\big]$ satisfies Eq. (1) with $k = 1$. Thus, the decomposition property holds for an M/M/1 with exponential setup times. While in this section we have relied on the M/G/1 result from [15], the decomposition property can also be derived directly for the M/M/1, based on its memoryless property. Examples of two such direct proofs are provided in Appendix A.

3.3 M/M/k: Limiting probabilities

In this section, we analyze and solve the M/M/k with exponential setup, for the limiting probabilities. The method of solving the Markov chain for limiting probabilities is similar to Ivo's technique in [1].

Fig. 1 shows the Markov chain for our M/M/k with exponential setup costs. The states in the Markov chain are denoted as (a, b), where a represents the number of servers that are turned on and ready to serve (active), and b represents the number of jobs in the system. The Markov chain consists of $k + 1$ rows. The first row (from the top) consists of states where we have no active servers, the second row consists of states where we have exactly one active server, and so on. For the setup cost, recall that *only one server can be turned on at one time*. Thus, the rate of going from state (i, j) to state $(i + 1, j)$ is α for any $0 \leq i < k$ and $i < j$.

We'll now solve the Markov chain shown in Fig. 1 for the limiting probabilities of being in any state. We first find the limiting probabilities for the states in the 1st row, in terms of $\pi_{0,0}$. Next, we solve for the limiting probabilities of being in the states of the 2nd row, in terms of the solution for the 1st row, which in turn is expressed in terms of $\pi_{0,0}$. Continuing in this way, we can solve for the limiting probabilities of all the states of the Markov chain in terms of $\pi_{0,0}$. We'll then solve for $\pi_{0,0}$ using the equation $\sum_{i,j} \pi_{i,j} = 1$. This will give us the limiting probabilities for all the states in the Markov chain.

Theorem 2. *The limiting probabilities for the M/M/k with exponential setup times (whose Markov chain is as shown in Fig. 1) are given in terms of* $\pi_{0,0}$ *by:*

$$\begin{aligned}\pi_{i,j} &= \frac{\pi_{0,0}\cdot\gamma^i}{i!}\beta^j \quad \textit{for } 0 \leq i < k \textit{ and } j > i-1 \\ \pi_{k,j} &= \frac{\pi_{0,0}\gamma^k k\mu}{k!\cdot(k\mu - (\lambda+\alpha))}\beta^j - \frac{\pi_{0,0}k^k(\lambda+\alpha)}{k!\cdot(k\mu-(\lambda+\alpha))}\left(\frac{\rho}{k}\right)^j \quad \textit{for } j > k-1\end{aligned}$$

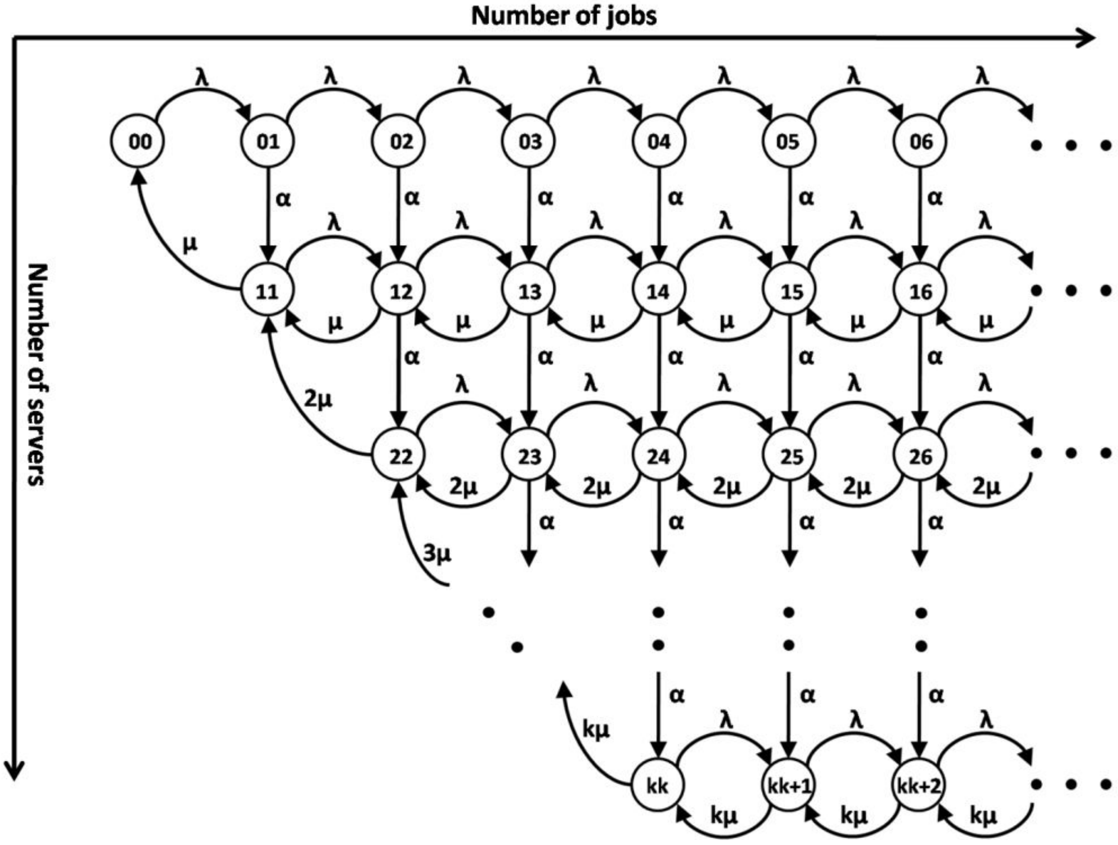


Figure 1: Markov chain for the M/M/k with exponential setup times.

where $\alpha = \frac{1}{\mathbb{E}[I]}$, $\beta = \frac{\lambda}{\lambda+\alpha}$ *and* $\gamma = \frac{\lambda+\alpha}{\mu}$.

Proof

Step 1: Solving the 1st row

The relevant balance equations for the 1st row are given by:

$$\begin{aligned} \pi_{0,j} \cdot (\lambda + \alpha) &= \pi_{0,j-1} \cdot \lambda \quad \text{for } j > 0. \\ \implies \pi_{0,j} &= \pi_{0,j-1} \cdot \left(\frac{\lambda}{\lambda + \alpha} \right) \quad \text{for } j > 0. \\ \implies \pi_{0,j} &= \pi_{0,j-1} \cdot \beta \quad \text{for } j > 0\text{, and } \beta = \tfrac{\lambda}{\lambda+\alpha}. \\ \implies \pi_{0,j} &= \pi_{0,0} \cdot \beta^j \quad \text{for } j > 0. \end{aligned} \tag{4}$$

We reserve the balance equation for $\pi_{0,0}$ for the 2nd row.

Step 2: Solving the 2nd row

The relevant balance equations for the 2nd row are given by:

$$\pi_{1,j} \cdot (\lambda + \alpha + \mu) = \pi_{1,j-1} \cdot \lambda + \pi_{0,j} \cdot \alpha + \pi_{1,j+1} \cdot \mu \quad \text{for } j > 1. \tag{5}$$

The RHS of Eq. (5) above consists of states of the 2nd row as well as states of the 1st row. Thus, this equation is inhomogeneous, with $\pi_{0,j}$ being the inhomogeneous part. Solutions for such equations are given by:

$$\pi_{1,j} = A_{1,1}x_1^j + A_{1,2}\beta^j \quad \text{for } j > 1\text{ ,} \tag{6}$$

where x_1 is a solution of the homogeneous equation:

$$x_1 \cdot (\lambda + \alpha + \mu) = \lambda + x_1^2 \cdot \mu \tag{7}$$

As we'll see, $A_{1,1}$ turns out to be 0, thus we thankfully don't need to compute x_1.

Plugging in Eqs. (4) and (6) into Eq. (5) for $j > 2$, we have:

$$\begin{aligned} A_{1,1}x_1^j \cdot (\lambda + \alpha + \mu) + A_{1,2}\beta^j \cdot (\lambda + \alpha + \mu) &= A_{1,1}x_1^{j-1} \cdot \lambda + A_{1,2}\beta^{j-1} \cdot \lambda + \pi_{0,0} \cdot \beta^j \cdot \alpha \\ &\quad + A_{1,1}x_1^{j+1} \cdot \mu + A_{1,2}\beta^{j+1} \cdot \mu \\ \implies A_{1,2} \cdot \left\{ \beta(\lambda + \alpha + \mu) - \lambda - \beta^2\mu \right\} &= \pi_{0,0} \cdot \beta \cdot \alpha \quad \text{(from Eq. (7))} \\ \implies A_{1,2}\frac{\lambda\mu\alpha}{(\lambda + \alpha)^2} &= \pi_{0,0} \cdot \frac{\lambda\alpha}{\lambda + \alpha} \\ \implies A_{1,2} &= \pi_{0,0} \cdot \frac{\lambda + \alpha}{\mu} \\ \implies A_{1,2} &= \pi_{0,0} \cdot \gamma \quad \text{where } \gamma = \tfrac{\lambda+\alpha}{\mu} \end{aligned} \tag{8}$$

To get $A_{1,1}$, we want to use the boundary condition for the 2nd row: use the balance equation for $\pi_{1,2}$, which will contain $\pi_{1,1}$. However, we first need to evaluate $\pi_{1,1}$. This requires us to use the balance equation for $\pi_{0,0}$, which we had intentionally left out in Step 1. Using the balance equation for $\pi_{0,0}$:

$$\begin{aligned}
\pi_{0,0} \cdot \lambda &= \pi_{1,1} \cdot \mu \\
\Longrightarrow \pi_{1,1} &= \pi_{0,0} \cdot \rho \qquad (9)
\end{aligned}$$

We now use Eqs. (4) and (6) in Eq. (5) for $j = 2$:

$$A_{1,1}x_1^2 \cdot (\lambda + \alpha + \mu) + A_{1,2}\beta^2 \cdot (\lambda + \alpha + \mu) = \pi_{1,1} \cdot \lambda + \pi_{0,0} \cdot \beta^2 \cdot \alpha + A_{1,1}x_1^3 \cdot \mu + A_{1,2}\beta^3 \cdot \mu$$

$$\begin{aligned}
\Longrightarrow A_{1,1} \cdot (\ldots) &= \pi_{0,0}\rho\lambda + \pi_{0,0} \cdot \beta^2 \cdot \alpha + A_{1,2}\beta^2(\beta\mu - \lambda - \alpha - \mu) \quad \text{(from Eq. (9))} \\
\Longrightarrow A_{1,1} \cdot (\ldots) &= \pi_{0,0}\lambda^2 \left(\frac{1}{\mu} + \frac{\alpha}{(\lambda + \alpha)^2}\right) - A_{1,2}\beta^2 \left(\frac{\lambda^2 + 2\lambda\alpha + \alpha^2 + \alpha\mu}{\lambda + \alpha}\right) \\
\Longrightarrow A_{1,1} \cdot (\ldots) &= \pi_{0,0}\frac{\beta^2}{\mu}\left\{\alpha\mu + (\lambda + \alpha)^2\right\} - \pi_{0,0}\beta^2\gamma \left(\frac{\lambda^2 + 2\lambda\alpha + \alpha^2 + \alpha\mu}{\lambda + \alpha}\right) \\
& \text{(from Eq. (8))} \\
\Longrightarrow A_{1,1} \cdot (\ldots) &= \cancel{\pi_{0,0}\frac{\beta^2}{\mu}\left\{\alpha\mu + (\lambda + \alpha)^2\right\}} - \cancel{\pi_{0,0}\frac{\beta^2}{\mu}\left(\lambda^2 + 2\lambda\alpha + \alpha^2 + \alpha\mu\right)}
\end{aligned}$$

$$\Longrightarrow A_{1,1} = 0 \qquad (10)$$

Thus, we have from Eqs. (6), (8) and (10):

$$\pi_{1,j} = \pi_{0,0} \cdot \beta^j \cdot \gamma \quad \text{for } j > 0 \qquad (11)$$

Note that the above equation is valid for $j = 1$ since $\beta \cdot \gamma = \rho$.

Step 3: Solving the 3rd row

Step 3 is similar to Step 2 above since the format of the balance equations remain the same. This is also true for Step 4, Step 5, ..., Step (k-1). For Step 3, we get:

$$\pi_{2,j} = \frac{\pi_{0,0} \cdot \gamma^2}{2!}\beta^j \quad \text{for } j > 1 \qquad (12)$$

$\vdots$

Step k: Solving the kth row

$$\pi_{k-1,j} = \frac{\pi_{0,0} \cdot \gamma^{k-1}}{(k-1)!}\beta^j \quad \text{for } j > k-2 \qquad (13)$$

Thus, we can combine the above results to say that:

$$\pi_{i,j} = \frac{\pi_{0,0} \cdot \gamma^i}{i!}\beta^j \quad \text{for } 0 \leq i < k \text{ and } j > i-1 \qquad (14)$$

Step k+1: Solving the (k+1)th row

The relevant balance equations for the (k+1)th row are given by:

$$\pi_{k,j} \cdot (\lambda + k\mu) \;=\; \pi_{k,j-1} \cdot \lambda + \pi_{k-1,j} \cdot \alpha + \pi_{k,j+1} \cdot k\mu \quad \text{for } j > k. \tag{15}$$

As before, the solution for such equations is given by:

$$\pi_{k,j} \;=\; A_{k,1} x_k^j + A_{k,2}\beta^j \quad \text{for } j > k \;, \tag{16}$$

where x_k is a solution of the homogeneous equation:

$$x_k \cdot (\lambda + k\mu) \;=\; \lambda + x_k^2 \cdot k\mu \tag{17}$$

This time, $A_{k,1}$ will not be zero. Thus, we need to solve the above equation for x_k. Solving Eq. (17) for x_k, we find $x_k = \frac{\rho}{k}$ (the other solution $x_k = 1$ is trivially discarded). Thus, we have:

$$\pi_{k,j} \;=\; A_{k,1}\left(\frac{\rho}{k}\right)^j + A_{k,2}\beta^j \quad \text{for } j > k \;, \tag{18}$$

Plugging in Eqs. (13) and (16) into Eq. (15) for $j > k+1$, we have:

$$\begin{aligned}
A_{k,1}x_k^j \cdot (\lambda + k\mu) + A_{k,2}\beta^j \cdot (\lambda + k\mu) \;&=\; A_{k,1}x_k^{j-1} \cdot \lambda + A_{k,2}\beta^{j-1} \cdot \lambda + \frac{\pi_{0,0} \cdot \gamma^{k-1}}{(k-1)!}\beta^j \cdot \alpha \\
&\quad + A_{k,1}x_k^{j+1} \cdot \mu + A_{k,2}\beta^{j+1} \cdot \mu \\
\implies A_{k,2} \cdot \left\{\beta(\lambda + k\mu) - \lambda - \beta^2 k\mu\right\} \;&=\; \frac{\pi_{0,0} \cdot \gamma^{k-1}}{(k-1)!} \cdot \beta \cdot \alpha \quad \text{(from Eq. (17))} \\
\implies A_{k,2} \cdot \frac{\beta\alpha}{\lambda + \alpha}\left\{k\mu - (\lambda + \alpha)\right\} \;&=\; \frac{\pi_{0,0} \cdot \gamma^{k-1}}{(k-1)!} \cdot \beta \cdot \alpha \\
\implies A_{k,2} \;&=\; \frac{\pi_{0,0} \cdot \gamma^{k-1}}{(k-1)!} \cdot \frac{\lambda + \alpha}{k\mu - (\lambda + \alpha)} \\
\implies A_{k,2} \;&=\; \frac{\pi_{0,0} \cdot \gamma^k \cdot \mu}{(k-1)! \cdot \{k\mu - (\lambda + \alpha)\}} \qquad (19)
\end{aligned}$$

To get $A_{k,1}$, we want to use a boundary condition for the (k+1)th row: use the balance equation for $\pi_{k,k+1}$, which will contain $\pi_{k,k}$. However, we first need to evaluate $\pi_{k,k}$. This requires us to use the balance equation for $\pi_{k-1,k-1}$. After a few steps of algebra, we find that:

$$\pi_{k,k} \;=\; \frac{\pi_{0,0} \cdot \rho^k}{k!} \tag{20}$$

We now use Eqs. (13) and (18) in Eq. (15) for $j = k+1$:

$$\begin{aligned}
A_{k,1}x_k^{k+1} \cdot (\lambda + k\mu) + A_{k,2}\beta^{k+1} \cdot (\lambda + k\mu) \;&=\; \pi_{k,k} \cdot \lambda + \frac{\pi_{0,0} \cdot \gamma^{k-1}}{(k-1)!}\beta^{k+1} \cdot \alpha + A_{k,1}x_k^{k+2} \cdot k\mu \\
&\quad + A_{k,2}\beta^{k+2} \cdot k\mu
\end{aligned}$$

$$\Longrightarrow A_{k,1} \cdot (\cancel{\lambda} + k\mu - \cancel{x_k k\mu}) \cdot x_k^{k+1} = \frac{\pi_{0,0} \cdot \rho^k}{k!}\lambda + \frac{\pi_{0,0} \cdot \gamma^{k-1}}{(k-1)!}\beta^{k+1} \cdot \alpha + A_{k,2}\beta^{k+1}(\beta k\mu - \lambda - k\mu)$$

(from Eq. (20) and since $x_k = \frac{\rho}{k}$)

$$\Longrightarrow A_{k,1} \cdot k\mu \cdot x_k^{k+1} = \frac{\pi_{0,0} \cdot \rho^{k-1}}{(k-1)!} \cdot \left(\frac{\lambda^2}{k\mu} + \frac{\alpha\lambda^2}{(\lambda+\alpha)^2}\right) - A_{k,2}\beta^{k+1} \cdot \left(\frac{\alpha k\mu}{\lambda+\alpha} + \lambda\right)$$

(since $\beta \cdot \gamma = \rho$)

$$\Longrightarrow A_{k,1} \cdot k\mu \cdot x_k^{k+1} = \frac{\pi_{0,0} \cdot \rho^{k-1} \cdot \lambda^2}{(k-1)!} \cdot \left(\frac{1}{k\mu} + \frac{\alpha}{(\lambda+\alpha)^2} - \frac{\alpha k\mu + \lambda(\lambda+\alpha)}{(k\mu - (\lambda+\alpha))(\lambda+\alpha)^2}\right)$$

(from Eq. (19) and since $\beta \cdot \gamma = \rho$)

$$\Longrightarrow A_{k,1} \cdot k\mu \cdot x_k^{k+1} = \frac{\pi_{0,0} \cdot \rho^{k-1} \cdot \lambda^2}{(k-1)!} \cdot \left(\frac{(k\mu - (\lambda+\alpha)) \cdot (\lambda+\alpha)^2 - k\mu(\lambda+\alpha)^2}{k\mu \cdot (k\mu - (\lambda+\alpha)) \cdot (\lambda+\alpha)^2}\right)$$

$$\Longrightarrow A_{k,1} \cdot k\mu \cdot x_k^{k+1} = \frac{\pi_{0,0} \cdot \rho^{k-1} \cdot \lambda^2}{(k-1)!} \cdot \left(\frac{-(\lambda+\alpha)^3}{k\mu \cdot (k\mu - (\lambda+\alpha)) \cdot (\lambda+\alpha)^2}\right)$$

$$\Longrightarrow A_{k,1} \cdot \cancel{k}\cancel{\mu} \cdot \frac{\cancel{\rho^{k+1}}}{k^{k+\cancel{1}}} = \frac{\pi_{0,0} \cdot \cancel{\rho^{k+1}} \cdot \cancel{\mu}}{k!} \cdot \left(\frac{-(\lambda+\alpha)}{(k\mu - (\lambda+\alpha))}\right)$$

$$\Longrightarrow A_{k,1} = -\frac{\pi_{0,0} \cdot k^{k-1} \cdot (\lambda+\alpha)}{((k-1)!)(k\mu - (\lambda+\alpha))} \tag{21}$$

Thus, we have from Eqs. (18), (19) and (21):

$$\pi_{k,j} = \frac{\pi_{0,0}\gamma^k k\mu}{k! \cdot (k\mu - (\lambda+\alpha))}\beta^j - \frac{\pi_{0,0}k^k(\lambda+\alpha)}{k! \cdot (k\mu - (\lambda+\alpha))}\left(\frac{\rho}{k}\right)^j \quad \text{for } j > k-1 \tag{22}$$

We have now solved for the limiting probabilities of all the states of the Markov chain in terms of $\pi_{0,0}$. We'll next solve for $\pi_{0,0}$ using the equation $\sum_{i,j} \pi_{i,j} = 1$:

Theorem 3. *For the M/M/k with exponential setup times,*

$$\pi_{0,0} = \left(1 - \frac{\lambda}{\lambda+\alpha}\right) \cdot \left\{\sum_{0 \le i < k} \frac{\rho^i}{i!} + \frac{\rho^k \mu}{(k-1)! \cdot (k\mu - \lambda)}\right\}^{-1} \tag{23}$$

where $\alpha = \frac{1}{\mathbb{E}[I]}$.

Proof

$$\begin{aligned}
1 &= \sum_{i,j} \pi_{i,j} \\
\implies 1 &= \sum_{\substack{i,j\\ i<k}} \pi_{i,j} + \sum_{j} \pi_{k,j} \\
\implies 1 &= \sum_{\substack{i,j\\ i<k}} \frac{\pi_{0,0}\gamma^i}{i!}\beta^j + \sum_{j} \pi_{0,0} \cdot \left\{ \frac{\gamma^k k\mu}{k! \cdot (k\mu - (\lambda+\alpha))}\beta^j - \frac{k^k(\lambda+\alpha)}{k! \cdot (k\mu - (\lambda+\alpha))}\left(\frac{\rho}{k}\right)^j \right\} \\
\implies \pi_{0,0}^{-1} &= \sum_{0\le i<k} \frac{\gamma^i}{i!}\left(\sum_{i\le j}\beta^j\right) + \frac{\gamma^k k\mu}{k! \cdot (k\mu - (\lambda+\alpha))}\sum_{k\le j}\beta^j - \frac{k^k(\lambda+\alpha)}{k! \cdot (k\mu - (\lambda+\alpha))}\sum_{k\le j}\left(\frac{\rho}{k}\right)^j \\
\implies \pi_{0,0}^{-1} &= \sum_{0\le i<k} \frac{\gamma^i}{i!} \cdot \frac{\beta^i}{(1-\beta)} + \frac{\gamma^k k\mu}{k! \cdot (k\mu - (\lambda+\alpha))} \cdot \frac{\beta^k}{(1-\beta)} - \frac{\lambda+\alpha}{k! \cdot (k\mu - (\lambda+\alpha))} \cdot \frac{\rho^k \cdot k\mu}{(k\mu-\lambda)} \\
\implies \pi_{0,0}^{-1} &= \frac{1}{1-\beta}\sum_{0\le i<k}\frac{\rho^i}{i!} + \frac{\rho^k\mu}{(k-1)! \cdot (k\mu - (\lambda+\alpha))} \cdot \frac{1}{(1-\beta)} - \frac{\lambda+\alpha}{(k-1)! \cdot (k\mu - (\lambda+\alpha))} \cdot \frac{\rho^k \cdot \mu}{(k\mu-\lambda)} \\
\implies \pi_{0,0}^{-1} &= \frac{1}{1-\beta}\sum_{0\le i<k}\frac{\rho^i}{i!} + \frac{\rho^k\mu}{(k-1)! \cdot (k\mu - (\lambda+\alpha))} \cdot \frac{1}{(1-\beta)} - \frac{\rho^k \cdot \mu}{(k-1)! \cdot (k\mu - (\lambda+\alpha))} \cdot \frac{\alpha}{(k\mu-\lambda)} \cdot \frac{1}{1-\beta} \\
\implies \pi_{0,0}^{-1} &= \frac{1}{1-\beta}\sum_{0\le i<k}\frac{\rho^i}{i!} + \frac{\rho^k\mu}{(k-1)! \cdot (1-\beta) \cdot (k\mu-\lambda)} \\
\implies \pi_{0,0} &= (1 - \frac{\lambda}{\lambda+\alpha}) \cdot \left\{ \sum_{0\le i<k}\frac{\rho^i}{i!} + \frac{\rho^k\mu}{(k-1)! \cdot (k\mu-\lambda)} \right\}^{-1} \qquad (24)
\end{aligned}$$

Surprisingly, we have $\pi_{0,0} = (1 - \frac{\lambda}{\lambda+\alpha}) \cdot \pi'_0$, where π'_0 is the limiting probability of having 0 jobs in an M/M/k system without setup.

3.4 M/M/k/setup: Performance measures

Now that we have all the limiting probabilities, we will proceed to calculate $\mathbb{E}\big[N^{M/M/k/setup}\big]$, the mean number of jobs in the M/M/k with exponential setup costs system. This will allow us to derive $\mathbb{E}\big[T^{M/M/k/setup}\big]$, the mean response time. We also derive $\hat{N}^{M/M/k/setup}(z)$, the z-transform of the number of jobs in the system, $\mathrm{Var}\big(N^{M/M/k/setup}\big)$, the variance of the number of jobs in system, and $\mathbb{E}\Big[K_{busy}^{M/M/k/setup}\Big]$, the expected number of servers ON or in SETUP.

Theorem 4. *The mean number of jobs,* $\mathbb{E}\big[N^{M/M/k/setup}\big]$*, for an M/M/k with exponential setup times is given by:*

$$\mathbb{E}\big[N^{M/M/k/setup}\big] = \rho + \frac{\beta}{1-\beta} + \frac{\pi_{0,0}\rho^k k\mu\lambda}{k!(1-\beta)(k\mu-\lambda)^2} \qquad (25)$$

where $\alpha = \frac{1}{\mathbb{E}[I]}$, $\beta = \frac{\lambda}{\lambda+\alpha}$ *and* $\pi_{0,0}$ *is as given by Eqn. (23).*

Proof

$$\begin{aligned}
\mathbb{E}\Big[N^{M/M/k/setup}\Big] &= \sum_{\substack{0\le i\le k\\ i\le j}} j\cdot\pi_{i,j}\\
\implies \mathbb{E}[N] &= \sum_{0\le i\le k}\left(\sum_{i\le j} j\cdot\pi_{i,j}\right)\\
&= \sum_{0\le i< k}\left(\sum_{i\le j} j\cdot\pi_{i,j}\right) + \sum_{k\le j} j\cdot\pi_{k,j}\\
&= \sum_{0\le i<k}\frac{\pi_{0,0}\gamma^i}{i!}\cdot\left(\sum_{i\le j} j\cdot\beta^j\right) + \sum_{k\le j} j\cdot\left\{\frac{\pi_{0,0}\gamma^k k\mu}{k!\cdot(k\mu-(\lambda+\alpha))}\beta^j - \frac{\pi_{0,0}k^k(\lambda+\alpha)}{k!\cdot(k\mu-(\lambda+\alpha))}\left(\frac{\rho}{k}\right)^j\right\}\\
&\qquad\text{(from Eqs. (14) and (22))}
\end{aligned}$$

$$\begin{aligned}
\implies \mathbb{E}[N] &= \sum_{0\le i<k}\frac{\pi_{0,0}\gamma^i}{i!}\cdot\left(\sum_{i\le j} j\cdot\beta^j\right) + \frac{\pi_{0,0}\gamma^k k\mu}{k!\cdot(k\mu-(\lambda+\alpha))}\cdot\left(\sum_{k\le j} j\cdot\beta^j\right)\\
&\quad -\frac{\pi_{0,0}k^k(\lambda+\alpha)}{k!\cdot(k\mu-(\lambda+\alpha))}\cdot\left(\sum_{k\le j} j\cdot\left(\frac{\rho}{k}\right)^j\right) \qquad (26)
\end{aligned}$$

We'll now find a closed form expression for $\sum_{i\le j} j\cdot x^j$, which will be useful for simplifying Eq. (26).

$$\begin{aligned}
Z &= i\cdot x^i + (i+1)\cdot x^{i+1} + (i+2)\cdot x^{i+2} + (i+3)\cdot x^{i+3} + \ldots\\
Z\cdot x &= 0\cdot x^i + (i+0)\cdot x^{i+1} + (i+1)\cdot x^{i+2} + (i+2)\cdot x^{i+3} + \ldots
\end{aligned}$$

$$Z\cdot(1-x) = i\cdot x^i + 1\cdot x^{i+1} + 1\cdot x^{i+2} + 1\cdot x^{i+3} + \ldots$$

$$\begin{aligned}
\implies Z &= \frac{(i-1)\cdot x^i}{1-x} + \frac{x^i}{(1-x)^2}\\
\implies \sum_{i\le j} j\cdot x^j &= \frac{(i-1)\cdot x^i}{1-x} + \frac{x^i}{(1-x)^2} \qquad (27)
\end{aligned}$$

Thus, using Eq. (27) in Eq. (26), we have:

$$\begin{aligned}
&= \sum_{0\le i<k} \frac{\pi_{0,0}\gamma^i\beta^i}{i!}\cdot\left(\frac{i-1}{(1-\beta)}+\frac{1}{(1-\beta)^2}\right)+\frac{\pi_{0,0}\gamma^k\beta^k k\mu}{k!\cdot(k\mu-(\lambda+\alpha))}\cdot\left(\frac{k-1}{(1-\beta)}+\frac{1}{(1-\beta)^2}\right)\\
&\quad -\frac{\pi_{0,0}\cancel{k^k}(\lambda+\alpha)}{k!\cdot(k\mu-(\lambda+\alpha))}\cdot\frac{\rho^k}{\cancel{k^k}}\cdot\left(\frac{k\mu\cdot(k-1)}{(k\mu-\lambda)}+\frac{k^2\mu^2}{(k\mu-\lambda)^2}\right)\\
&= \sum_{0\le i<k} \frac{\pi_{0,0}\rho^i}{i!}\cdot\left(\frac{i-1}{(1-\beta)}+\frac{1}{(1-\beta)^2}\right)+\frac{\pi_{0,0}\rho^k k\mu}{k!\cdot(k\mu-(\lambda+\alpha))}\cdot\left(\frac{k-1}{(1-\beta)}+\frac{1}{(1-\beta)^2}\right)\\
&\quad -\frac{\pi_{0,0}\rho^k(\lambda+\alpha)}{k!\cdot(k\mu-(\lambda+\alpha))}\cdot\left(\frac{k\mu\cdot(k-1)}{(k\mu-\lambda)}+\frac{k^2\mu^2}{(k\mu-\lambda)^2}\right)\\
&= \sum_{0\le i<k} \frac{\pi_{0,0}\rho^i}{i!}\cdot\left(\frac{i-1}{(1-\beta)}+\frac{1}{(1-\beta)^2}\right)+\frac{\pi_{0,0}\rho^k k\mu}{k!\cdot(k\mu-(\lambda+\alpha))}\cdot\left(\frac{k-1}{(1-\beta)}+\frac{1}{(1-\beta)^2}\right)\\
&\quad -\frac{\pi_{0,0}\rho^k(\lambda+\alpha)}{k!\cdot(k\mu-(\lambda+\alpha))}\cdot\left\{\left(\frac{k-1}{(1-\beta)}+\frac{1}{(1-\beta)^2}\right)-\left(\frac{k-1}{(1-\beta)}+\frac{1}{(1-\beta)^2}\right)\right\}\\
&\quad -\frac{\pi_{0,0}\rho^k(\lambda+\alpha)}{k!\cdot(k\mu-(\lambda+\alpha))}\cdot\left(\frac{k\mu\cdot(k-1)}{(k\mu-\lambda)}+\frac{k^2\mu^2}{(k\mu-\lambda)^2}\right)\\
&= \sum_{0\le i\le k} \frac{\pi_{0,0}\rho^i}{i!}\cdot\left(\frac{i-1}{(1-\beta)}+\frac{1}{(1-\beta)^2}\right)\\
&\quad +\frac{\pi_{0,0}\rho^k(\lambda+\alpha)}{k!\cdot(k\mu-(\lambda+\alpha))}\cdot\left\{\left(\frac{k-1}{(1-\beta)}+\frac{1}{(1-\beta)^2}\right)-\left(\frac{k\mu\cdot(k-1)}{(k\mu-\lambda)}+\frac{k^2\mu^2}{(k\mu-\lambda)^2}\right)\right\}\\
&= \sum_{0\le i\le k} \frac{\pi_{0,0}\rho^i}{i!}\cdot\left(\frac{i}{(1-\beta)}+\frac{\beta}{(1-\beta)^2}\right)\\
&\quad +\frac{\pi_{0,0}\rho^k(\lambda+\alpha)}{k!\cdot(k\mu-(\lambda+\alpha))}\cdot\left\{\left(\frac{k-1}{(1-\beta)}+\frac{1}{(1-\beta)^2}\right)-\left(\frac{k\mu\cdot(k-1)}{(k\mu-\lambda)}+\frac{k^2\mu^2}{(k\mu-\lambda)^2}\right)\right\}\\
&= \frac{\rho\cdot\pi_{0,0}}{1-\beta}\cdot\left\{\frac{1-\beta}{\pi_{0,0}}-\frac{\rho^k\mu}{(k-1)!\cdot(k\mu-\lambda)}\right\}+\frac{\beta\pi_{0,0}}{(1-\beta)^2}\cdot\left\{\frac{1-\beta}{\pi_{0,0}}-\frac{\rho^k\lambda}{k!\cdot(k\mu-\lambda)}\right\}\\
&\quad +\frac{\pi_{0,0}\rho^k(\lambda+\alpha)}{k!\cdot(k\mu-(\lambda+\alpha))}\cdot\left\{\left(\frac{(k-1)\cdot(k\mu\beta-\lambda)}{(1-\beta)(k\mu-\lambda)}\right)+\left(\frac{(k\mu\beta-\lambda)(2k\mu-k\mu\beta-\lambda)}{(1-\beta)^2(k\mu-\lambda)^2}\right)\right\}
\end{aligned}$$

(using Eq. (24))

$$\begin{aligned}
&= \rho+\frac{\beta}{1-\beta}-\frac{\pi_{0,0}\rho^k}{k!(1-\beta)(k\mu-\lambda)}\cdot\left(k\mu\rho+\frac{\beta\lambda}{1-\beta}\right)\\
&\quad +\frac{\pi_{0,0}\rho^k(\lambda+\alpha)}{k!\cdot(k\mu-(\lambda+\alpha))}\cdot\frac{k\mu\beta-\lambda}{(1-\beta)^2(k\mu-\lambda)^2}\cdot\{(k-1)(k\mu-\lambda)(1-\beta)+(2k\mu-k\mu\beta-\lambda)\}
\end{aligned}$$

$$
\begin{aligned}
\implies \mathbb{E}[N] &= \rho + \frac{\beta}{1-\beta} - \frac{\pi_{0,0}\rho^k\lambda}{k!(1-\beta)(k\mu-\lambda)} \cdot \left(k + \frac{\lambda}{\alpha}\right) \\
&\quad + \frac{\pi_{0,0}\rho^k(\lambda+\alpha)}{k!} \cdot \frac{\beta}{(1-\beta)^2(k\mu-\lambda)^2} \cdot \frac{(k\alpha+\lambda)(k\mu-\lambda)+k\mu\alpha}{\lambda+\alpha} \\
&= -\frac{\pi_{0,0}\rho^k\lambda}{k!(1-\beta)^2(k\mu-\lambda)^2} \cdot \left\{ \frac{(k\alpha+\lambda)(1-\beta)(k\mu-\lambda)}{\alpha} - \frac{(k\alpha+\lambda)(k\mu-\lambda)+k\mu\alpha}{\lambda+\alpha} \right\} \\
&\quad + \rho + \frac{\beta}{1-\beta} \\
\implies \mathbb{E}\left[N^{M/M/k/setup}\right] &= \rho + \frac{\beta}{1-\beta} + \frac{\pi_{0,0}\rho^k k\mu\lambda}{k!(1-\beta)(k\mu-\lambda)^2}
\end{aligned}
$$

Theorem 5. *The mean response time, $\mathbb{E}\left[T^{M/M/k/setup}\right]$, for an M/M/k with exponential setup times is given by:*

$$
\mathbb{E}\left[T^{M/M/k/setup}\right] = \frac{1}{\alpha} + \frac{\pi_{0,0}\rho^k k\mu}{k!(1-\beta)(k\mu-\lambda)^2} + \frac{1}{\mu} \tag{28}
$$

where $\alpha = \frac{1}{\mathbb{E}[I]}$, $\beta = \frac{\lambda}{\lambda+\alpha}$ and $\pi_{0,0}$ is as given by Eqn. (23).

Proof

We have $\mathbb{E}\left[N^{M/M/k/setup}\right]$ from Thm 4. We now invoke Little's law [11] to derive $\mathbb{E}\left[T^{M/M/k/setup}\right]$ from $\mathbb{E}\left[N^{M/M/k/setup}\right]$.

$$
\begin{aligned}
\mathbb{E}\left[T^{M/M/k/setup}\right] &= \frac{\mathbb{E}\left[N^{M/M/k/setup}\right]}{\lambda} \\
\implies \mathbb{E}\left[T^{M/M/k/setup}\right] &= \frac{1}{\mu} + \frac{1}{\alpha} + \frac{\pi_{0,0}\rho^k k\mu}{k!(1-\beta)(k\mu-\lambda)^2} \\
\implies \mathbb{E}\left[T^{M/M/k/setup}\right] &= \frac{1}{\alpha} + \frac{\pi_{0,0}\rho^k k\mu}{k!(1-\beta)(k\mu-\lambda)^2} + \frac{1}{\mu}
\end{aligned}
$$

This brings us to the decomposition property.

Corollary

$$
\mathbb{E}\left[T^{M/M/k/setup}\right] = \mathbb{E}\left[T^{M/M/k}\right] + \mathbb{E}[I]
$$

Proof

Follows from the fact that $\mathbb{E}\left[T^{M/M/k}\right] = \frac{\pi_{0,0}\rho^k k\mu}{k!(1-\beta)(k\mu-\lambda)^2} + \frac{1}{\mu}$.

We now derive the z-transform for the number of jobs in the system. The z-transform, for a random variable X, is defined as $\hat{X}(z) = \mathbb{E}\left[z^X\right]$.

Theorem 6. *The z-transform for the number of jobs in the system, $\hat{N}^{M/M/k/setup}(z)$, is given by:*

$$\hat{N}^{M/M/k/setup}(z) \;=\; \frac{\pi_{0,0}\cdot k\mu\rho^k}{(k!)(k\mu-(\lambda+\alpha))}\cdot\frac{z^k}{(1-\beta z)} - \frac{\pi_{0,0}\cdot\rho^k(\lambda+\alpha)}{(k!)(k\mu-(\lambda+\alpha))}\cdot\frac{z^k}{(1-\frac{\rho z}{k})} + \frac{\pi_{0,0}}{1-\beta z}\sum_{0\le i<k}\frac{(\rho z)^i}{i!}$$

where $\alpha=\frac{1}{\mathbb{E}[I]}$, $\beta=\frac{\lambda}{\lambda+\alpha}$, $\gamma=\frac{\lambda+\alpha}{\mu}$ and $\pi_{0,0}$ is as given by Eqn. (23).

Proof

The z-transform, for a random variable X, is defined as $\hat{X}(z)=\mathbb{E}\big[z^X\big]$. We now derive $\hat{N}^{M/M/k/setup}(z)$, the z-transform of the number of jobs in the system.

$$\begin{aligned}\hat{N}^{M/M/k/setup}(z) &= \sum_{i=0}^{k} Pr[N=i]\cdot z^i\\ &= \left(\sum_{i=0}^{k-1}\sum_{j=i}^{\infty}\pi_{i,j}z^j\right)+\sum_{j=k}^{\infty}\pi_{k,j}z^j\\ &= \left(\sum_{i=0}^{k-1}\sum_{j=i}^{\infty}\frac{\pi_{0,0}\gamma^i\beta^j}{i!}z^j\right)+\sum_{j=k}^{\infty}\left(\frac{\pi_{0,0}\gamma^k k\mu}{k!\cdot(k\mu-(\lambda+\alpha))}\beta^j-\frac{\pi_{0,0}k^k(\lambda+\alpha)}{k!\cdot(k\mu-(\lambda+\alpha))}\left(\frac{\rho}{k}\right)^j\right)\cdot z^j\\ &= \frac{\pi_{0,0}}{1-\beta z}\sum_{0\le i<k}\frac{(\rho z)^i}{i!}+\frac{\pi_{0,0}\cdot k\mu\rho^k}{(k!)(k\mu-(\lambda+\alpha))}\cdot\frac{z^k}{(1-\beta z)}-\frac{\pi_{0,0}\cdot\rho^k(\lambda+\alpha)}{(k!)(k\mu-(\lambda+\alpha))}\cdot\frac{z^k}{(1-\frac{\rho z}{k})}\end{aligned}$$

Using the z-transform from Thm. 6, we can derive $\mathrm{Var}\big(N^{M/M/k/setup}\big)$, the variance of the number of jobs in the system.

Theorem 7. *The variance of the number of jobs in the system, $\mathrm{Var}\big(N^{M/M/k/setup}\big)$, is given by:*

$$\mathrm{Var}\big(N^{M/M/k/setup}\big) \;=\; \mathrm{Var}\big(N^{M/M/k}\big)+\frac{\beta}{1-\beta}\left(1+\frac{\beta}{1-\beta}\right)+\frac{2\pi_{0,0}\cdot\rho^{k+2}\cdot\beta^2}{k\cdot(k-\rho)\cdot(1-\beta)^3\cdot(k!)}$$

where $\alpha=\frac{1}{\mathbb{E}[I]}$, $\beta=\frac{\lambda}{\lambda+\alpha}$ and $\pi_{0,0}$ is as given by Eqn. (23) and

$$\mathrm{Var}\big(N^{M/M/k}\big)=\rho(1+\rho)+\frac{k\pi_{0,0}\cdot\rho^{k+1}\cdot(k+\rho)\cdot(1+k-\rho)}{(k!)(1-\beta)(k-\rho)^3}-\left(\rho+\frac{k\pi_{0,0}\cdot\rho^{k+1}}{(k!)(1-\beta)(k-\rho)^2}\right)^2$$

Proof

We break the proof down into two steps. In Step 1, we derive the variance of the number of jobs for an M/M/k without setup times, $\mathrm{Var}\big(N^{M/M/k}\big)$. Then, in Step 2, we use the z-transform from Thm. 6 to derive $\mathrm{Var}\big(N^{M/M/k/setup}\big)$.

Step 1: $\mathrm{Var}\big(N^{M/M/k}\big)$

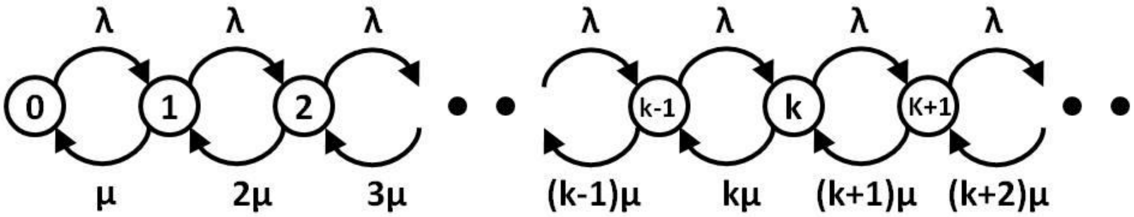


Figure 2: Markov chain for the M/M/k without setup times.

The Markov chain for an M/M/k without setup times is shown in Fig. 2. Based on the chain, we can easily calculate the steady-state limiting probabilities of the system:

$$\pi_i \quad = \quad \begin{cases} \frac{\pi_0 \rho^i}{i!} & \text{if } 1 < i \leq k \\ \pi_k \cdot \left(\frac{\rho}{k}\right)^{i-k} & \text{if } i > k \end{cases} \tag{29}$$

where

$$\pi_0 \quad = \quad \left\{ \sum_{0 \leq i < k} \frac{\rho^i}{i!} + \frac{\rho^k \mu}{(k-1)! \cdot (k\mu - \lambda)} \right\}^{-1} \tag{30}$$

We now calculate the mean number of jobs in the M/M/k system, $\mathbb{E}\left[N^{M/M/k}\right]$:

$$\begin{aligned} \mathbb{E}\left[N^{M/M/k}\right] &= \sum_{i=1}^{\infty} i \cdot \pi_i \\ &= \pi_0 \left(\sum_{i=1}^{k} i \cdot \frac{\rho^i}{i!} \right) + \pi_k \left(\sum_{i=k+1}^{\infty} i \cdot \left(\frac{\rho}{k}\right)^{i-k} \right) \qquad \text{(from Eq. 29)} \\ &= \pi_0 \left(\rho \sum_{i=0}^{k-1} \frac{\rho^i}{i!} \right) + \pi_k \left(\sum_{i=1}^{\infty} (k+i) \cdot \left(\frac{\rho}{k}\right)^i \right) \\ &= \rho - \frac{\pi_0 \rho^{k+1}}{(k!)(1-\frac{\rho}{k})} + \frac{\pi_0 \rho^{k+1}}{k \cdot (k!)(1-\frac{\rho}{k})} \left(k + \frac{1}{1-\frac{\rho}{k}} \right) \\ &= \rho + \frac{\pi_0 \rho^{k+1}}{k \cdot (k!) \cdot (1-\frac{\rho}{k})^2} \end{aligned} \tag{31}$$

In a similar manner, we can calculate $\mathbb{E}\Big[N^{2^{M/M/k}}\Big]$. We find that:

$$\mathbb{E}\Big[N^{2^{M/M/k}}\Big] \;=\; \rho(1+\rho) + \frac{k\pi_0 \cdot \rho^{k+1} \cdot (k+\rho) \cdot (1+k-\rho)}{(k!) \cdot (k-\rho)^3} \tag{32}$$

We can now calculate $\mathrm{Var}\big(N^{M/M/k}\big) = \mathbb{E}\Big[N^{2^{M/M/k}}\Big] - \mathbb{E}^2\big[N^{M/M/k}\big]$ from Eqs. (32) and (31):

$$\begin{aligned}
\mathrm{Var}\big(N^{M/M/k}\big) &= \mathbb{E}\Big[N^{2^{M/M/k}}\Big] - \mathbb{E}^2\big[N^{M/M/k}\big] \\
&= \rho(1+\rho) + \frac{k\pi_0 \cdot \rho^{k+1} \cdot (k+\rho) \cdot (1+k-\rho)}{(k!) \cdot (k-\rho)^3} - \left(\rho + \frac{\pi_0 \rho^{k+1}}{k \cdot (k!) \cdot (1-\frac{\rho}{k})^2}\right)^2 \\
&= \rho(1+\rho) + \frac{k\pi_{0,0} \cdot \rho^{k+1} \cdot (k+\rho) \cdot (1+k-\rho)}{(k!)(1-\beta)(k-\rho)^3} \\
&\quad - \left(\rho + \frac{k\pi_{0,0} \cdot \rho^{k+1}}{(k!)(1-\beta)(k-\rho)^2}\right)^2 \quad \text{(from Eq. (24))}
\end{aligned} \tag{33}$$

Step 2: $\mathrm{Var}\big(N^{M/M/k/setup}\big)$

We use the z-transform from Thm. 6 to derive $\mathrm{Var}\big(N^{M/M/k/setup}\big)$. $\mathrm{Var}\big(N^{M/M/k/setup}\big)$ is defined as:

$$Var(N) \;=\; \mathbb{E}\big[N^2\big] - \mathbb{E}^2[N] \tag{34}$$

From the theory of z-transforms, we know that:

$$\frac{d^2}{dz^2}\hat{N}^{M/M/k/setup}(z)\Big|_{z=1} = \mathbb{E}\Big[N^{2^{M/M/k/setup}}\Big] - \mathbb{E}^2\big[N^{M/M/k/setup}\big] \tag{35}$$

Thus, we can calculate $\mathbb{E}\Big[N^{2^{M/M/k/setup}}\Big]$, since we already know $\mathbb{E}\big[N^{M/M/k/setup}\big]$ from Thm. 4.

$$\hat{N}^{M/M/k/setup}(z) \;=\; \frac{\pi_{0,0}}{1-\beta z}\sum_{0\le i<k}\frac{(\rho z)^i}{i!} + \frac{\pi_{0,0} \cdot k\mu\rho^k}{(k!)(k\mu-(\lambda+\alpha))} \cdot \frac{z^k}{(1-\beta z)} - \frac{\pi_{0,0} \cdot \rho^k(\lambda+\alpha)}{(k!)(k\mu-(\lambda+\alpha))} \cdot \frac{z^k}{(1-\frac{\rho z}{k})}$$

We now differentiate $\hat{N}^{M/M/k/setup}(z)$ twice, and set $z=1$ to get $\frac{d^2}{dz^2}\hat{N}^{M/M/k/setup}(z)\big|_{z=1}$. Due to the complexity of the equations, we skip a few steps of algebra when deriving $\frac{d^2}{dz^2}\hat{N}^{M/M/k/setup}\left(z\right)\Big|_{z=1}$.

$$\begin{aligned}
\frac{d^2}{dz^2}\hat{N}^{M/M/k/setup}(z)\Big|_{z=1} &= \frac{\pi_{0,0}\rho^{k+1}}{k!}\Bigg(\frac{k\cdot(k+\rho)}{(1-\beta)\cdot(k-\rho)^2}+\frac{2k\beta\cdot(k-3\rho)}{(1-\beta)^2\cdot(k-\rho)^3} \\
&\quad +\frac{2\rho\cdot(k^2+\beta^2\rho^2-2k\cdot\rho\cdot\beta^2)}{k\cdot(k-\rho)^3\cdot(1-\beta)^3}\Bigg)+\left(\rho+\frac{\beta}{1-\beta}\right)^2+\left(\frac{\beta}{1-\beta}\right)^2
\end{aligned}$$

$$\begin{aligned}
\implies \mathbb{E}\Big[N^{2^{M/M/k/setup}}\Big] &= \frac{\pi_{0,0}\rho^{k+1}}{k!}\Bigg(\frac{k\cdot(k+\rho)}{(1-\beta)\cdot(k-\rho)^2}+\frac{2k\beta\cdot(k-3\rho)}{(1-\beta)^2\cdot(k-\rho)^3} \\
&\quad +\frac{2\rho\cdot(k^2+\beta^2\rho^2-2k\cdot\rho\cdot\beta^2)}{k\cdot(k-\rho)^3\cdot(1-\beta)^3}\Bigg)+\left(\rho+\frac{\beta}{1-\beta}\right)^2+\left(\frac{\beta}{1-\beta}\right)^2 \\
&\quad +\rho+\frac{\beta}{1-\beta}+\frac{k\cdot\pi_{0,0}\rho^{k+1}}{(k!)\cdot(1-\beta)\cdot(k-\rho)^2}\ \text{(from Eq. (35) and Thm. 4)}
\end{aligned}$$

$$\begin{aligned}
\implies \mathrm{Var}\Big(N^{M/M/k/setup}\Big) &= \frac{\pi_{0,0}\rho^{k+1}}{k!}\Bigg(\frac{k\cdot(k+\rho)}{(1-\beta)\cdot(k-\rho)^2}+\frac{2k\beta\cdot(k-3\rho)}{(1-\beta)^2\cdot(k-\rho)^3}+\frac{2\rho\cdot(k^2+\beta^2\rho^2-2k\cdot\rho\cdot\beta^2)}{k\cdot(k-\rho)^3\cdot(1-\beta)^3}\Bigg) \\
&\quad +\left(\rho+\frac{\beta}{1-\beta}\right)^2+\left(\frac{\beta}{1-\beta}\right)^2+\rho+\frac{\beta}{1-\beta}+\frac{k\cdot\pi_{0,0}\rho^{k+1}}{(k!)\cdot(1-\beta)\cdot(k-\rho)^2} \\
&\quad -\left(\rho+\frac{\beta}{1-\beta}+\frac{k\cdot\pi_{0,0}\rho^{k+1}}{(k!)\cdot(1-\beta)\cdot(k-\rho)^2}\right)^2\ \text{(from Thm. 4)} \\
&= \frac{\pi_{0,0}\rho^{k+1}}{k!}\Bigg(\frac{k\cdot(k+\rho+1)}{(1-\beta)\cdot(k-\rho)^2}+\frac{2k\beta\cdot(k-3\rho)}{(1-\beta)^2\cdot(k-\rho)^3}+\frac{2\rho\cdot(k^2+\beta^2\rho^2-2k\cdot\rho\cdot\beta^2)}{k\cdot(k-\rho)^3\cdot(1-\beta)^3}\Bigg) \\
&\quad +\mathrm{Var}\Big(N^{M/M/k}\Big)+\frac{\beta}{1-\beta}\left(1+\frac{\beta}{1-\beta}\right)-\frac{2k\beta\cdot\pi_{0,0}\cdot\rho^{k+1}}{(k!)\cdot(k-\rho)^2\cdot(1-\beta)^2} \\
&\quad -\frac{k\cdot\pi_{0,0}\cdot\rho^{k+1}\cdot(k+\rho)\cdot(1+k-\rho)}{(k!)\cdot(k-\rho)^3\cdot(1-\beta)}\ \text{(from Eq. (33))} \\
&= \frac{\pi_{0,0}\rho^{k+1}}{k!}(-\frac{2k\rho}{(1-\beta)\cdot(k-\rho)^3}-\frac{4k\beta\cdot\rho}{(1-\beta)^2\cdot(k-\rho)^3} \\
&\quad +\frac{2\rho\cdot(k^2+\beta^2\rho^2-2k\cdot\rho\cdot\beta^2)}{k\cdot(k-\rho)^3\cdot(1-\beta)^3})+Var(N)^{M/M/k}+\frac{\beta}{1-\beta}\left(1+\frac{\beta}{1-\beta}\right)
\end{aligned}$$

$$= \frac{\pi_{0,0}\rho^{k+1}}{k!} \cdot \frac{2\rho\beta^2}{k \cdot (1-\beta)^3 \cdot (k-\rho)} + Var(N)^{M/M/k} + \frac{\beta}{1-\beta}\left(1 + \frac{\beta}{1-\beta}\right)$$

$$= Var(N)^{M/M/k} + \frac{\beta}{1-\beta}\left(1 + \frac{\beta}{1-\beta}\right) + \frac{2\pi_{0,0} \cdot \rho^{k+2} \cdot \beta^2}{k \cdot (k-\rho) \cdot (1-\beta)^3 \cdot (k!)}$$

Finally, we derive $\mathbb{E}\left[K_{busy}^{M/M/k/setup}\right]$, the expected number of servers either ON or in SETUP.

Theorem 8. *The expected number of servers either ON or in SETUP, is given by:*

$$\mathbb{E}\left[K_{busy}^{M/M/k/setup}\right] = \beta + \rho - \frac{\pi_{0,0}\rho^k\beta}{k!(1-\beta)(1-\frac{\rho}{k})}$$

where $\alpha = \frac{1}{\mathbb{E}[I]}$, $\beta = \frac{\lambda}{\lambda+\alpha}$ *and* $\pi_{0,0}$ *is as given by Eqn. (23).*

Proof

From Section 3.1, we know that a server can be in any of the following three states: (i) OFF, (ii) ON or (iii) SETUP. We are interested in the expected number of servers not in the OFF state. We can assign $K_{busy}^{M/M/k/setup}$ values to each of the states in the M/M/k Markov chain shown in Fig 1. For example, in state $(0,0)$, $K_{busy}^{M/M/k/setup}(0,0) = 0$. In the state $(0,1)$, $K_{busy}^{M/M/k/setup}(0,1) = 1$. For states $(0,j)$, where $j > 1$, $K_{busy}^{M/M/k/setup}(0,j) = 1$ again, since we only allow one server to be in the SETUP mode at any time. In general, for state (i,j), we have:

$$K_{busy}^{M/M/k/setup}(i,j) = \begin{cases} i & \text{if } i = j \\ i+1 & \text{otherwise} \end{cases} \qquad (36)$$

Thus, the expected number of servers either ON or in SETUP, $\mathbb{E}\left[K_{busy}^{M/M/k/setup}\right]$, is given by:

$$\mathbb{E}\left[K_{busy}^{M/M/k/setup}\right] = \sum_{i,j} \pi_{i,j} \cdot K_{busy}^{M/M/k/setup}(i,j) \qquad (37)$$

Using Eqs. (36), (14) and (22) in the above equation, we get:

$$
\begin{aligned}
\mathbb{E}\left[K_{busy}^{M/M/k/setup}\right] &= \left(\sum_{i=0}^{k-1}\sum_{j=i+1}^{\infty}\frac{\pi_{0,0}\cdot\gamma^i\cdot\beta^j}{i!}\cdot(i+1)\right)+\left(\sum_{i=1}^{k-1}\frac{\pi_{0,0}\cdot\rho^i}{i!}\cdot i\right)\\
&\quad +k\cdot\sum_{i=k}^{\infty}\left(\frac{\pi_{0,0}\gamma^k k\mu}{k!\cdot(k\mu-(\lambda+\alpha))}\beta^i-\frac{\pi_{0,0}k^k(\lambda+\alpha)}{k!\cdot(k\mu-(\lambda+\alpha))}\left(\frac{\rho}{k}\right)^i\right)\\
&= \pi_{0,0}\cdot\frac{\beta}{1-\beta}\sum_{i=0}^{k-1}\left(\frac{\rho^i}{(i-1)!}+\frac{\rho^i}{i!}\right)+\pi_{0,0}\sum_{i=1}^{k-1}\left(\frac{\rho^i}{(i-1)!}\right)\\
&\quad +\frac{k\pi_{0,0}\rho^k}{(k!)(1-\beta)(1-\frac{\rho}{k})}\\
&= \pi_{0,0}\cdot\frac{\beta}{1-\beta}\left(\rho\cdot\left(\frac{1-\beta}{\pi_{0,0}}-\frac{\rho^{k-1}}{(k-1)!}-\frac{\rho^k}{(k!)(1-\frac{\rho}{k})}\right)+\frac{1-\beta}{\pi_{0,0}}-\frac{\rho^k}{(k!)(1-\frac{\rho}{k})}\right)\\
&\quad +\pi_{0,0}\rho\cdot\left(\frac{1-\beta}{\pi_{0,0}}-\frac{\rho^{k-1}}{(k-1)!}-\frac{\rho^k}{(k!)(1-\frac{\rho}{k})}\right)+\frac{k\pi_{0,0}\rho^k}{(k!)(1-\beta)(1-\frac{\rho}{k})}\text{ (by Eq. (24))}\\
&= \beta+\rho-\frac{\pi_{0,0}\rho^k\beta}{k!(1-\beta)(1-\frac{\rho}{k})}
\end{aligned}
$$

3.5 M/M/k with multiple SETUP servers

Throughout this paper, we assume that only one server is allowed to be in the SETUP mode at a time. However, it is interesting to ask what happens when multiple servers are allowed to be in the SETUP mode simultaneously. While the Markov chain for the case of multiple SETUP servers is quite similar to that for a single SETUP server, it does not lend itself to a pretty analysis. In Appendix B, we attempt to analyze the M/M/2 where multiple servers can be in the SETUP mode simultaneously (in this case, at most 2). The limiting probabilities are far more complex, making it unlikely to obtain a closed-form expression for the mean response time. It is also not obvious that any decomposition property exists.

4 M/G/k with Setup

So far we have looked at exponential job sizes and exponential setup costs. In this section, we examine whether or not, the decomposition property (given by Eq. (1)) extends to other job size distributions and other setup time distributions. First, in Section 4.1, we consider the M/H_2/2 with exponential setup costs. This particular model can be analyzed via matrix analytic methods (see [9] for an excellent reference on matrix analytic methods). We find that the decomposition property holds for the M/H_2/2 with exponential setup costs. Then, in Section 4.2, we consider the M/G/k with exponential setup costs. Since this model cannot be analyzed via matrix analytic

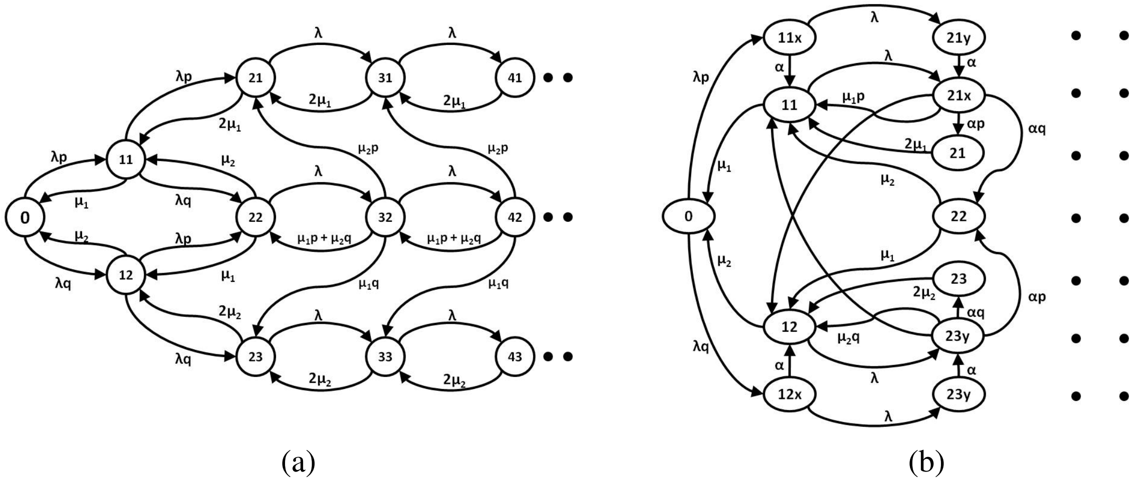


Figure 3: Markov chain for the M/H$_2$/2 (a) without setup times and (b) with exponential setup times.

State	Description of state
0	No jobs in the system.
11	1 job in the system, and it is of Type 1.
12	1 job in the system, and it is of Type 2.
21	2 jobs in the system, and both are of Type 1.
22	2 jobs in the system, with one of Type 1 and the other of Type 2.
23	2 jobs in the system, and both are of Type 2.
31	3 jobs in the system, and the 2 jobs at the head of the queue are both of Type 1.
32	3 jobs in the system, and the 2 jobs at the head of the queue are of Types 1 and 2 respectively.
33	3 jobs in the system, and the 2 jobs at the head of the queue are both of Type 2.

Table 1: State space description for the Markov chain in Fig. 3 (a).

methods, we resort to careful simulations to determine whether Eq. (1) holds for general job size distributions. Surprisingly, we find that the decomposition property appears to hold for the M/G/k with exponential setup costs. Finally, in Section 4.3, we consider the M/G/k with general setup costs and show that Eq. (1) might not hold in this case.

4.1 M/H$_2$/2 with Exponential setup

The M/H$_2$/2 with exponential setup times can be analyzed numerically, via matrix analytic methods. First, consider the simple M/H$_2$/2 without setup times. Again, the average arrival rate is λ. The job size is denoted by X, where, with probability p, X is exponentially distributed with rate μ_1, and with probability $(1-p)$, X is exponentially distributed with rate μ_2. Under this notation, the Markov chain for the M/H$_2$/2 can be represented as shown in Fig. 3 (a). A description of some of the states in the Markov chain is given in Table 1.

Next, we analyze the M/H$_2$/2 with exponential setup times system. Let I denote the setup times

State	Description of state
0	No jobs in the system.
11x	1 job in the system, and it is of Type 1. 1 server in setup mode.
11	1 job in the system, and it is of Type 1. 1 server active.
12x	1 job in the system, and it is of Type 2. 1 server in setup mode.
12	1 job in the system, and it is of Type 2. 1 server active.
21y	2 job in the system, and the one at the head of the queue is of Type 1. 1 server in setup mode.
21x	2 jobs in the system, and the one at the head of the queue is of Type 1. 1 server active, and serving job of Type 1. 1 server in setup mode.
21	2 jobs in the system, and both are of Type 1. 2 servers active, both serving jobs of Type 1.
22	2 jobs in the system, with one of Type 1 and the other of Type 2.. 2 servers active, one serving job of Type 1, other serving job of Type 2.
23y	2 job in the system, and the one at the head of the queue is of Type 2. 1 server in setup mode.
23x	2 jobs in the system, and the one at the head of the queue is of Type 2. 1 server active, and serving job of Type 2. 1 server in setup mode.
23	2 jobs in the system, and both are of Type 2. 2 servers active, both serving jobs of Type 2.

Table 2: State space description for the Markov chain in Fig. 3 (b).

and let $\alpha = \frac{1}{\mathbb{E}[I]}$. The Markov chain for this particular model is given in Fig. 3 (b). Due to the complexity of the state space, we only show a part of the chain. A description of some of the states in the Markov chain is given in Table 2. While the Markov chain is complex, it is tractable via matrix analytic methods due to its regular repeating structure.

Our results for the M/H$_2$/2 with exponential setup costs are shown in Fig. 4. In our results, $\rho = \lambda \cdot \mathbb{E}[X]$, where $\mathbb{E}[X] = \frac{p}{\mu_1} + \frac{1-p}{\mu_2}$ is the mean job size. Also, $C^2 = \frac{Var(X)}{\mathbb{E}^2[X]}$, denotes the squared coefficient of variation. Thus, from Fig. 4, it appears that the M/H$_2$/2 with exponential setup costs satisfies Eq. (1) for different values of ρ and C^2.

4.2 M/G/k with Exponential setup

In this section, we explore whether Eq. (1) extends to an M/G/k with exponential setup costs. We use simulation results to determine the mean response time of an M/G/k system with, and without setup times, and check whether Eq. (1) holds. The job size distributions we try out are listed in Table 3, along with their mean ($\mathbb{E}[X]$) and squared co-efficient of variation (C^2).

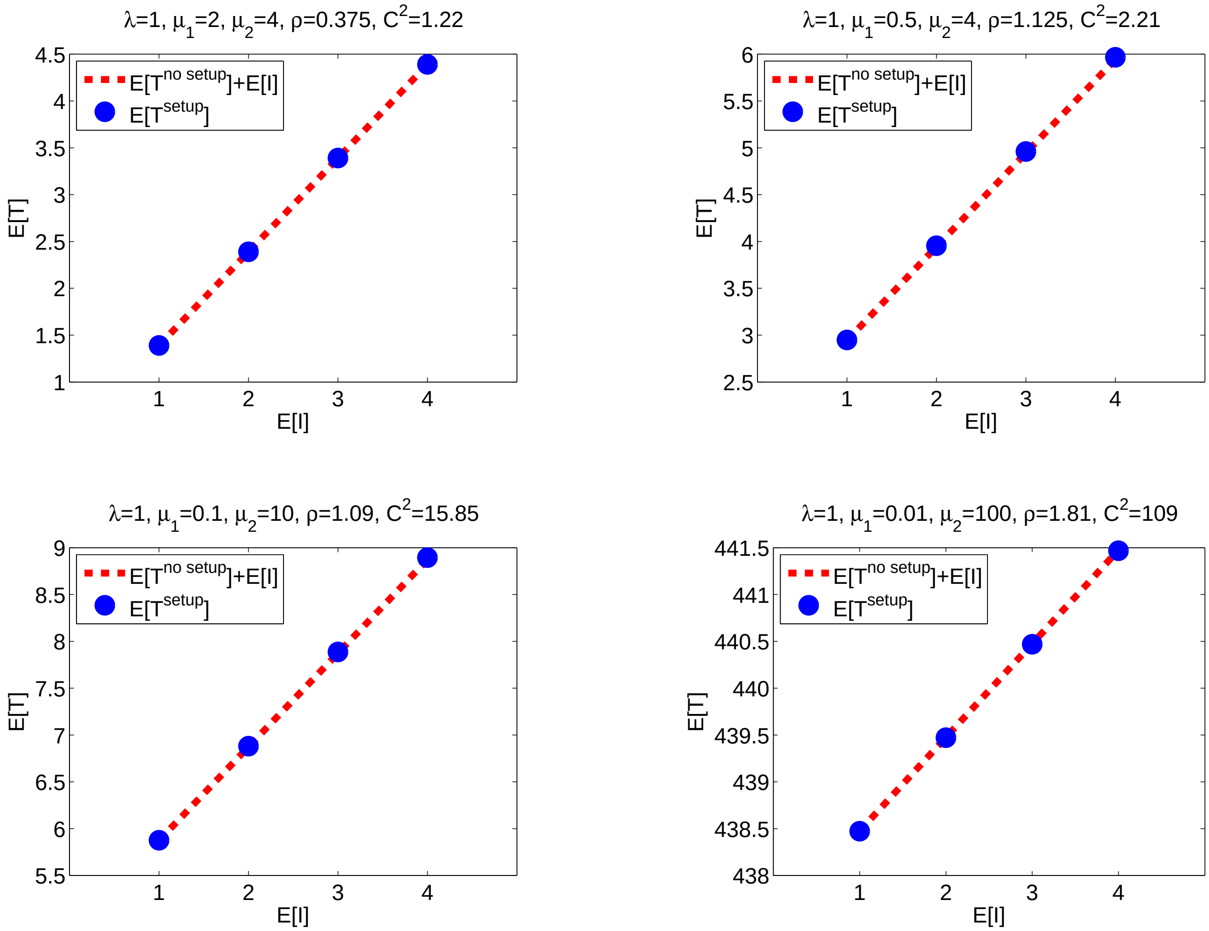


Figure 4: Matrix-analytic results for an M/H$_2$/2 with exponential setup costs. Results show that $\mathbb{E}\left[T^{M/H_2/2/setup}\right] = \mathbb{E}\left[T^{M/H_2/2}\right] + \mathbb{E}[I]$.

Distribution	$\mathbb{E}[X]$ **(Mean)**	C^2 **(Squared co-eff. of variation)**
Exponential (rate=2)	0.5	1
Hyper-exponential ($\mu_1 = 2$, $\mu_2 = 4$, p=0.5)	0.38	1.22
Uniform (0.25, 0.75)	0.5	0.08
Uniform (0.1, 1)	0.55	0.22
Bounded Exponential (min=0, max=2)	0.46	0.81
Bounded Exponential (min=0.2, max=2.2)	0.66	0.4
Deterministic (mean=0.5)	0.5	0
Bounded Pareto (min=1, max=2, α=1)	1.39	0.04
Bounded Pareto (min=1, max=10, α=1)	2.56	0.53
Bounded Pareto (min=1, max=100, α=1)	4.65	3.62
Bounded Pareto (min=1, max=1000, α=1)	6.91	19.92

Table 3: The different job size distributions used in our simulations.

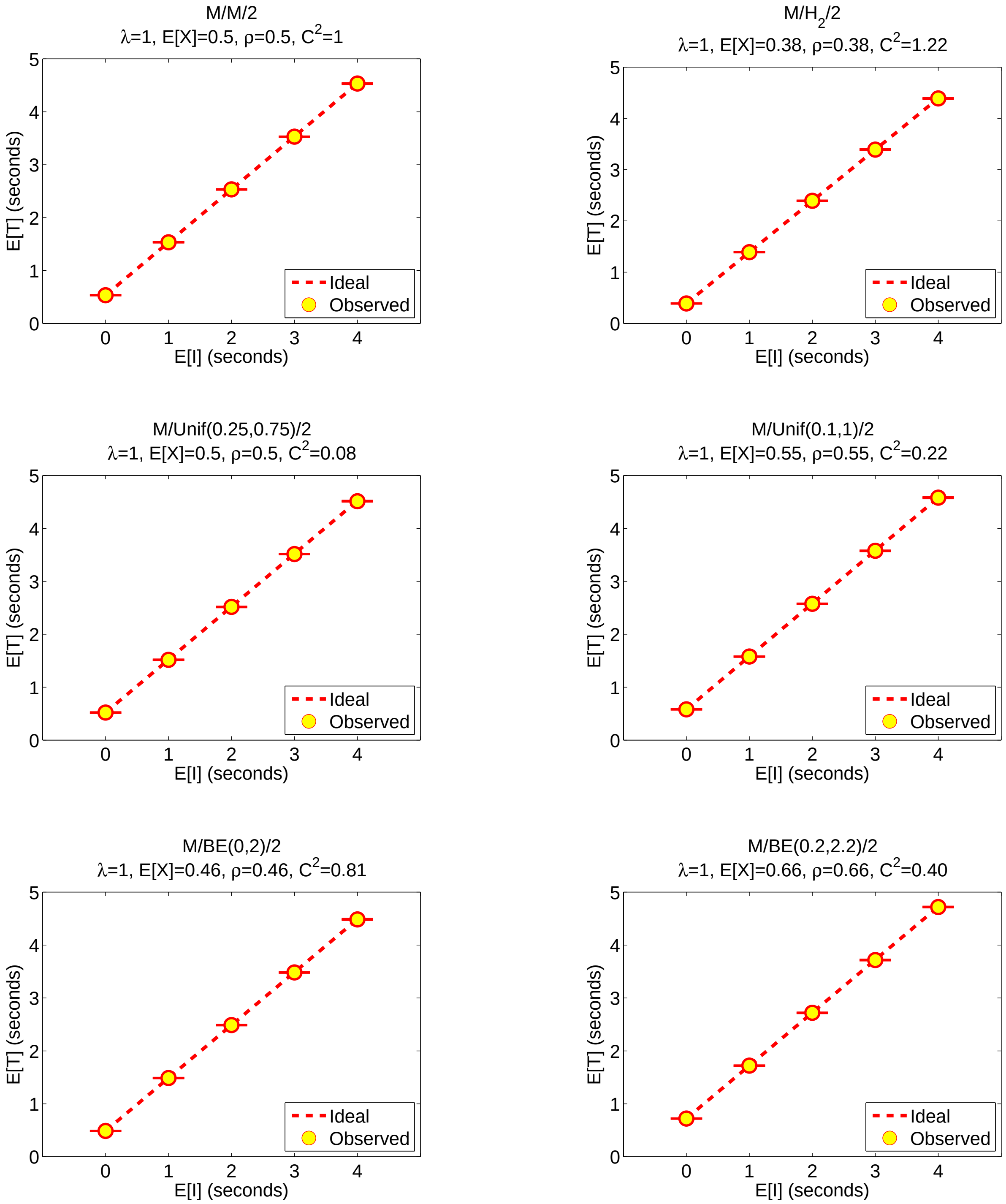


Figure 5: Simulation results for exponential (M), hyper-exponential (H_2), uniform (Unif) and Bounded Exponential (BE) job size distributions for an M/G/2 with exponential setup times. The parameters on the BE distribution denote the min and max values respectively. Confidence intervals (shown as horizontal lines) indicate good agreement with Eq. (1).

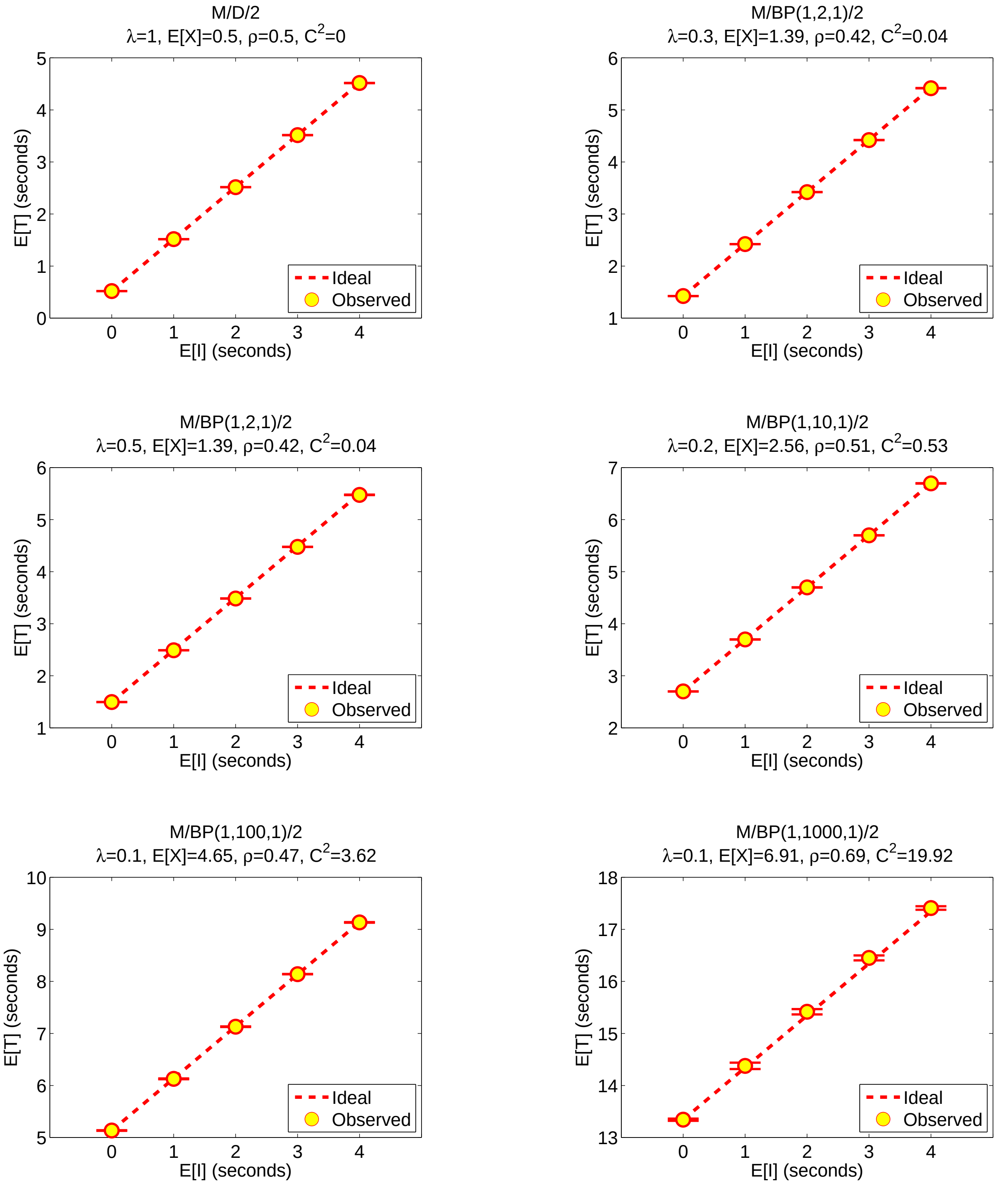


Figure 6: Simulation results for deterministic (D) and Bounded Pareto (BP) job size distributions for an M/G/2 with exponential setup times. The parameters on the BP distribution denote the min, max and α values respectively. Confidence intervals (shown as horizontal lines) indicate good agreement with Eq. (1).

Each simulation consists of 10^7 arrivals, and we average our results over multiple runs for each job size distribution. We also report the 95% confidence intervals in each case. In all simulations, $k = 2$. Due to lack of space, we omit simulation results for $k > 2$ in this paper, however, we have verified our results for $k = 4$ as well.

Figs. 5 and 6 show our simulation results for a range of job size distributions, including Exponential (M/M/2), Hyper-exponential (M/H_2/2), Uniform (M/Unif/2), Bounded Exponential (M/BE/2), Deterministic (M/D/2) and Bounded Pareto (M/BP/2). In all cases, we see that the M/G/k with exponential setup times appears to satisfy Eq. (1).Tight confidence intervals (shown as horizontal lines) indicate good agreement between the simulation results and Eq. (1).

4.3 M/G/k with General setup

We now consider the case of an M/G/k with general setup times. As we will show, the decomposition property is not satisfied by most setup time distributions that are not exponential. Consider, for example, the M/G/1 with general setup time, I.

From Eq. (3), we have:

$$\mathbb{E}\big[T^{M/G/1/setup}\big] = \mathbb{E}\big[T^{M/G/1}\big] + \frac{2\mathbb{E}[I] + \lambda\mathbb{E}[I^2]}{2(1+\lambda\mathbb{E}[I])}$$

The decomposition property is satisfied if and only if:

$$\begin{aligned} \frac{2\mathbb{E}[I] + \lambda\mathbb{E}[I^2]}{2(1+\lambda\mathbb{E}[I])} &= \mathbb{E}[I] \\ \Longrightarrow \mathbb{E}\big[I^2\big] &= 2\mathbb{E}^2[I] \\ \Longrightarrow C_I^2 &= \frac{\mathbb{E}[I^2] - \mathbb{E}^2[I]}{\mathbb{E}^2[I]} = 1 \qquad (38) \end{aligned}$$

where C_I^2 is the squared co-efficient of variation for the setup times. While Eq. (38) is clearly satisfied by exponentially distributed I, it is also satisfied by a few other distribution, for example:

$$I = \begin{cases} 0 & \text{with probability } \frac{1}{2} \\ 1 & \text{with probability } \frac{1}{2} \end{cases} \qquad (39)$$

but clearly not by most distributions considered in this paper.

In general, the mean response time for an M/M/k with non-exponentially distributed setup times can differ significantly from the mean response time for an M/M/k with exponentially distributed setup times. For example, Fig. 7 shows our simulation results for an M/M/2 with deterministic setup times. Clearly, the mean response time for an M/M/2 with deterministic setup times does not satisfy Eq. (1).

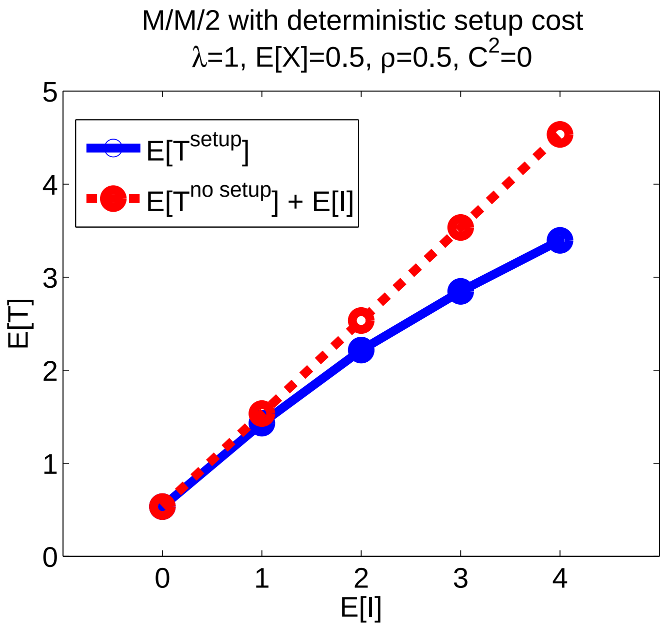


Figure 7: Mean response time vs. mean setup time for an M/M/2 with deterministic setup times. Clearly, Eq. (1) is no longer valid in this case.

5 Application

As stated earlier, our M/M/k with setup costs model is motivated by the problem of power management in server farms. Most server farm operators today are interested in minimizing both power usage and mean response time. While minimizing mean response time points to turning on many servers, minimizing power usage points to turning on few servers. To save on power usage, it is customary to turn servers OFF when they are not in use. However, this means that a SETUP cost is required every time a server needs to be turned on. This SETUP cost is both wasteful with respect to mean response time and with respect to power usage, since power is needed during the whole SETUP period. Since the SETUP cost can have a big effect on both power and response time, it is important that we take it into account when optimizing the configuration of our server farm.

The performance metric we consider in this section is a weighted sum of the mean response time, $\mathbb{E}[T]$, and the mean power consumption, $\mathbb{E}\left[Power^{M/M/k/setup}\right]$, given by

$$PERF = \mathbb{E}\left[Power^{M/M/k/setup}\right] + c \cdot \mathbb{E}\left[T^{M/M/k/setup}\right]$$

The weight parameter c has units of Watts/sec, and can be thought of as the price required, in Watts, to lower the mean response time of the server farm by 1 second. A similar weighted linear combination has previously been used in literature [2, 4, 16].

We assume the same algorithm as is assumed throughout the paper: We have a k server system. When servers are not in use they are switched to OFF. At most one server can be in the SETUP state. When a new job arrives, if there is already a server in the SETUP state, it will queue up;

otherwise, it will move one of the OFF servers (assuming there is one) to the SETUP state. When a job completes, the job at the head of the queue is moved to that server, without the need for SETUP.

Our goal is to determine the optimal number of servers $k = k^*$ that should be used to minimize PERF. The answer is not obvious. Since the servers are turned on only when needed, one might assume that having an infinite number of servers can't be that bad, since the servers that are turned OFF don't cost us power. However, by limiting the total number of servers, k, we can force the jobs to queue up. This can help us avoid the SETUP cost of turning on a new server every time a new job comes in. Thus a higher SETUP cost should point to a lower value of k^*, for optimizing PERF.

In Section 5.1 we derive a closed-form expression for $\mathbb{E}\left[Power^{M/M/k/setup}\right]$ and for PERF. We then use this expression in Section 5.2 to determine k^* for the above problem under a variety of load settings, SETUP costs, and values for the weight parameter, c.

5.1 $\mathbb{E}\left[Power^{M/M/k/setup}\right]$

From Section 3.1, we know that a server can be in any of the following three states: (i) OFF, (ii) ON or (iii) SETUP. While in the OFF state, we assume that the power consumption of a server is 0. While in the ON or SETUP states, the power consumption of the server is assumed to be a constant, P_{max}. Given this information, we see that $\mathbb{E}\left[Power^{M/M/k/setup}\right] = P_{max} \cdot \mathbb{E}\left[K_{busy}^{M/M/k/setup}\right]$. Thus, using Thm. 8, we have:

Theorem 9. *For an M/M/k with exponential setup times, the mean power consumption is given by:*

$$\mathbb{E}\left[Power^{M/M/k/setup}\right] = P_{max}\left(\beta + \rho - \frac{\pi_{0,0}\rho^k\beta}{k!(1-\beta)(1-\frac{\rho}{k})}\right)$$

where $\alpha = \frac{1}{\mathbb{E}[I]}$, $\beta = \frac{\lambda}{\lambda+\alpha}$ *and* $\pi_{0,0}$ *is as given by Eqn. (23).*

Proof

Follows trivially from that fact that $\mathbb{E}\left[Power^{M/M/k/setup}\right] = P_{max} \cdot \mathbb{E}\left[K_{busy}^{M/M/k/setup}\right]$, and from Thm. 8.

Finally, combining Thms. 5 and 9, we have:

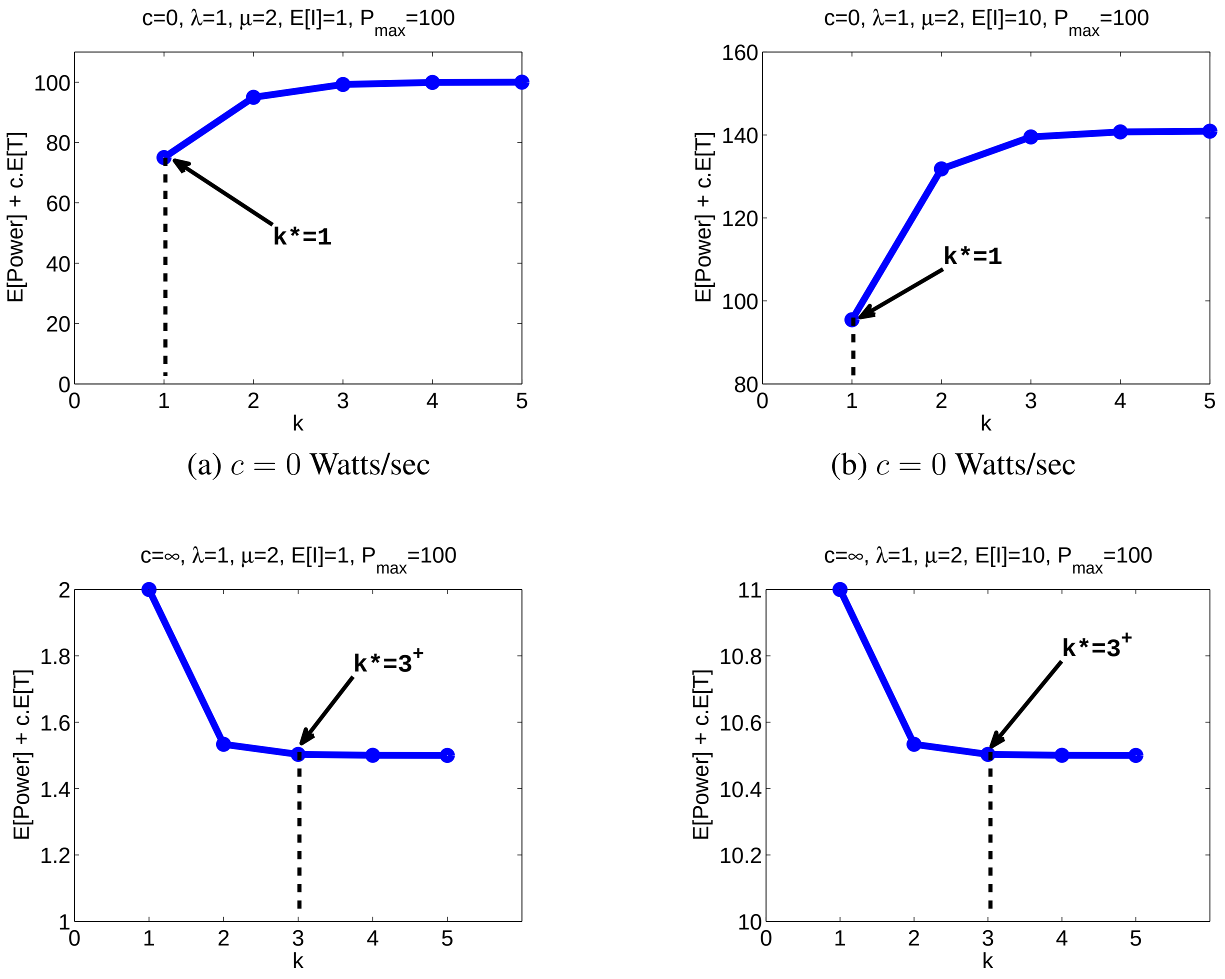
Theorem 10. *For an M/M/k with exponential setup times, the PERF value is given by:*

$$PERF = \left(\beta + \rho - \frac{\pi_{0,0}\rho^k\beta}{k!(1-\beta)(1-\frac{\rho}{k})}\right) \cdot P_{max} + c \cdot \left(\frac{1}{\alpha} + \frac{\pi_{0,0}\rho^k k\mu}{k!(1-\beta)(k\mu-\lambda)^2} + \frac{1}{\mu}\right)$$

where $\alpha = \frac{1}{\mathbb{E}[I]}$, $\beta = \frac{\lambda}{\lambda+\alpha}$ *and* $\pi_{0,0}$ *is as given by Eqn. (23).*

Proof

Follows trivially from that fact that $PERF = \mathbb{E}\left[Power^{M/M/k/setup}\right] + c \cdot \mathbb{E}\left[T^{M/M/k/setup}\right]$, and from Thms. 5 and 9.



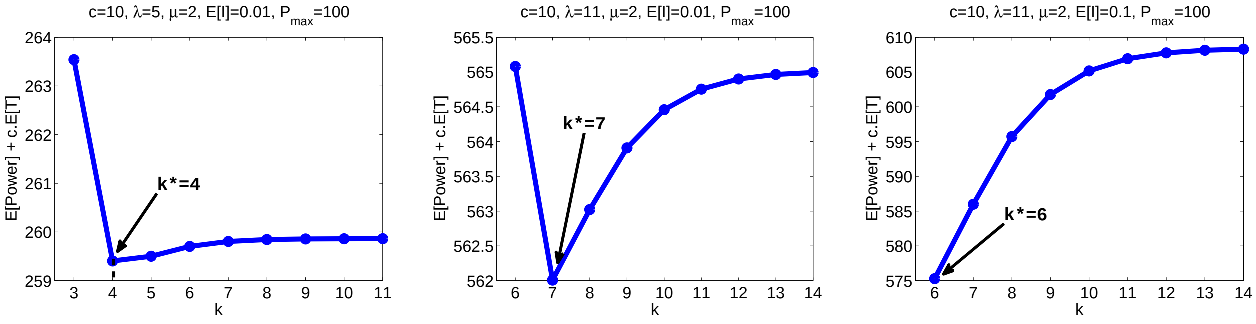
(a) $c = 0$ Watts/sec

(b) $c = 0$ Watts/sec

(c) $c = \infty$ Watts/sec

(d) $c = \infty$ Watts/sec

Figure 8: Results showing optimal k^* value for $c = 0$ Watts/sec ((a) and (b)) and $c = \infty$ Watts/sec ((c) and (d)).

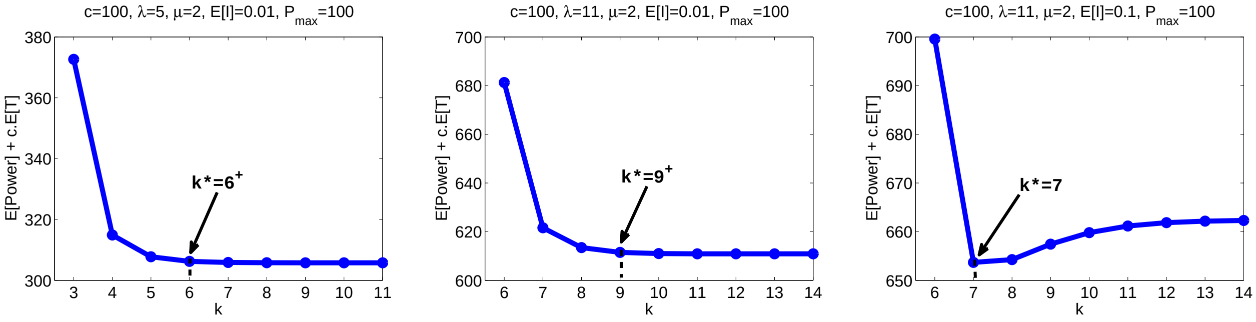


(a) $c = 10$ Watts/sec

(b) $c = 10$ Watts/sec
high ρ

(c) $c = 10$ Watts/sec
high ρ and high $\mathbb{E}[I]$

Figure 9: Results showing optimal k^* value for $c = 10$ Watts/sec.



(a) $c = 10$ Watts/sec

(b) $c = 10$ Watts/sec
high ρ

(c) $c = 10$ Watts/sec
high ρ and high $\mathbb{E}[I]$

Figure 10: Results showing optimal k^* value for $c = 100$ Watts/sec.

5.2 Optimization

In this section we deduce the optimal k value, k^*, which minimizes the PERF metric, under a variety of settings including different loads ρ, different weight-parameters c, and different SETUP values $\mathbb{E}[I]$.

We start with the case of $c = 0$. Here the goal is simply to minimize $\mathbb{E}\big[Power^{M/M/k/setup}\big]$. This is done by using the lowest possible number of servers, namely $k^* = 1$, regardless of ρ or $\mathbb{E}[I]$. Results are shown in Figs. 8 (a) and (b). Less trivial is the case of $c = \infty$. Here the goal is simply to minimize $\mathbb{E}\big[T^{M/M/k/setup}\big]$. This should point to maximizing the number of servers k. Observe that since power is not relevant when $c = \infty$, having a higher SETUP cost should not effect the k^* value, since as shown in Eqn. (1), the SETUP cost is additive. Figs. 8 (c) and (d) show our results for the value of $c = \infty$. Notice that after the k value gets sufficiently high, the mean response time remains constant, and hence we choose k^* as the lowest of these k values.

Fig. 9 shows our results when the weight parameter is $c = 10$ Watts/sec and Fig. 10 shows results for a weight parameter of $c = 100$ Watts/sec. In both figures, the (a) graph shows the case of low load, and low SETUP cost. The (b) graph shows the case of high load, but still low SETUP cost. Finally, the (c) graph shows the case of high load, and high SETUP cost.

The following trends are apparent: (i) As c is increased, $\mathbb{E}\big[T^{M/M/k/setup}\big]$ matters more, and, as expected k^* increases; (ii) As ρ is increased, as expected k^* needs to go up to keep response times from exploding; (iii) As $\mathbb{E}[I]$ is increased, k^* goes down, as expected because it is better to leave jobs waiting in the queue and avoid the SETUP cost of turning on a new server.

6 Conclusion

In this paper, we start by considering the M/M/k queueing system with setup times. We provide the first analytical closed form expressions for the mean response time, limiting distribution of the number of jobs in the system, and the z-transform for the number of jobs in system for the M/M/k system with exponential setup times. In particular, we prove the following *decomposition property*: the mean response time of the M/M/k system with exponential setup times differs from the mean response time of an M/M/k system without setup times, by an additive constant, which is the mean of the exponential setup time. Using matrix analytic methods and simulations, we show that the above decomposition property may also hold for the M/G/k system with exponential setup times. The fact that the setup time is exponentially-distributed is important: in fact we prove that the decomposition property cannot hold when the setup time distribution has squared coefficient of variation different from 1. Finally, we present a motivating application of our work: power management in server farms. Using our simple closed form expressions for the M/M/k with setup costs, we derive the optimal number of servers to be used in a server farm with setup time, so as to minimize a weighted sum of mean power and mean response time.

References

[1] I.J.B.F Adan and J. van der Wal. Combining make to order and make to stock. *OR Spektrum*, 20:73–81, 1998.

[2] Susanne Albers and Hiroshi Fujiwara. Energy-efficient algorithms for flow time minimization. *ACM Trans. Algorithms*, 3(4):49, 2007.

[3] J. R. Artalejo, A. Economou, and M. J. Lopez-Herrero. Analysis of a multiserver queue with setup times. *Queueing Syst. Theory Appl.*, 51(1-2):53–76, 2005.

[4] Nikhil Bansal, Ho-Leung Chan, and Kirk Pruhs. Speed scaling with an arbitrary power function. In *SODA '09: Proceedings of the Nineteenth Annual ACM -SIAM Symposium on Discrete Algorithms*, pages 693–701, Philadelphia, PA, USA, 2009. Society for Industrial and Applied Mathematics.

[5] Wolfgang Bischof. Analysis of $M/G/1$-queues with setup times and vacations under six different service disciplines. *Queueing Syst. Theory Appl.*, 39(4):265–301, 2001.

[6] Gautam Choudhury. On a batch arrival poisson queue with a random setup time and vacation period. *Comput. Oper. Res.*, 25(12):1013–1026, 1998.

[7] Gautam Choudhury. An $M^X/G/1$ queueing system with a setup period and a vacation period. *Queueing Syst. Theory Appl.*, 36(1/3):23–38, 2000.

[8] Sun Hur and Seung-Jin Paik. The effect of different arrival rates on the N-policy of $M/G/1$ with server setup. *Applied Mathematical Modelling*, 23(4):289 – 299, 1999.

[9] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods in Stochastic Modeling*. ASA-SIAM, Philadelphia, 1999.

[10] Huan Li, Yixin Zhu, and Ping Yang. Computational analysis of $M(n)/G/1/N$ queues with setup time. *Computers & Operations Research*, 22(8):829 – 840, 1995.

[11] J.D.C. Little. A proof of the queueing formula $l = \lambda w$. *Operations Research*, 9:383–387, 1961.

[12] Zhisheng Niu, Tao Shu, and Yoshitaka Takahashi. A vacation queue with setup and close-down times and batch markovian arrival processes. *Perform. Eval.*, 54(3):225–248, 2003.

[13] Hideaki Takagi. Priority queues with setup times. *Oper. Res.*, 38(4):667–677, 1990.

[14] Hideaki Takagi. $M/G/1/K$ queues with N-policy and setup times. *Queueing Systems Theory Appl.*, 14(1-2):79–98, 1993.

[15] P.D. Welch. On a generalized $M/G/1$ queueing process in which the first customer of each busy period receives exceptional service. *Operations Research*, 12:736–752, 1964.

[16] Adam Wierman, Lachlan L. H. Andrew, and Ao Tang. Power-aware speed scaling in processor sharing systems. *INFOCOM*, 2009.

A M/M/1 with exponential setup times

We are interested in analyzing mean response time in an M/M/1 with a setup costs, denoted as $\mathbb{E}\big[T^{M/M/1/setup}\big]$. We use random variable I to denote the setup cost. We use S to denote the service time of a job. The mean arrival rate into the system is λ. Both S and I are assumed to be exponentially-distributed.

A.1 M/M/1 with Exponential Setup – Tagged-job approach

We'll use a tagged-job approach to analyze the mean response time for a job in the M/M/1 with setup costs model. Using the PASTA (Poisson Arrivals See Time Averages) property, a job coming into the M/M/1 queue, sees N_Q number of jobs in the queue. Each of these N_Q jobs has size S. Then with probability $\rho = \frac{\lambda}{\mu}$, our tagged job sees the excess of a job in service, which is just S (due to Exponential job sizes), and with probability $(1-\rho)$, he sees a setup time, I, because there are no jobs in service.

Thus, the queueing time of the tagged job, T_Q^{setup}, can be expressed as:

$$\begin{aligned}
T_Q^{setup} &= \sum_{i=1}^{N_Q^{setup}} S_i + \rho\cdot S + (1-\rho)\cdot I \\
\implies \mathbb{E}\big[T_Q^{setup}\big] &= \mathbb{E}\big[N_Q^{setup}\big]\cdot\mathbb{E}[S] + \rho\cdot\mathbb{E}[S] + (1-\rho)\cdot\mathbb{E}[I] \\
\implies \mathbb{E}\big[T_Q^{setup}\big] &= \mathbb{E}\big[T_Q^{setup}\big]\cdot\rho + \rho\cdot\mathbb{E}[S] + (1-\rho)\cdot\mathbb{E}[I] \\
\implies \mathbb{E}\big[T_Q^{setup}\big] &= \frac{\rho}{1-\rho}\mathbb{E}[S] + \mathbb{E}[I] \\
\implies \mathbb{E}\big[T_Q^{setup}\big] &= \mathbb{E}\Big[T_Q^{M/M/1}\Big] + \mathbb{E}[I] \\
\implies \mathbb{E}\big[T_Q^{setup}\big] + \mathbb{E}[S] &= \mathbb{E}\Big[T_Q^{M/M/1}\Big] + \mathbb{E}[S] + \mathbb{E}[I] \\
\implies \mathbb{E}\big[T_Q^{setup}\big] &= \mathbb{E}\big[T^{M/M/1}\big] + \mathbb{E}[I] \\
\implies \mathbb{E}\big[T^{M/M/1/setup}\big] &= \mathbb{E}\big[T^{M/M/1}\big] + \mathbb{E}[I]
\end{aligned}$$

The last equation above is the same as Eq. (1) with $k = 1$.

A.2 M/M/1 with Exponential Setup – Proof via PLCFS

Let ρ denote the load in an M/M/1 without setup costs and ρ' denote the load an M/M/1 with setup cost I. Likewise B denotes the length of an M/M/1 busy period, while B' denotes the length of the busy period in an M/M/1 with setup cost I. Observe that ρ' and B are independent of the scheduling policy used, so long as the scheduling policy is work-conserving.

$$\begin{aligned}
\mathbb{E}[B] &= \frac{\mathbb{E}[S]}{1-\rho} \\
\mathbb{E}[B'] &= \frac{\mathbb{E}[S+I]}{1-\rho} \qquad (40) \\
\rho &= \frac{\mathbb{E}[B]}{\mathbb{E}[B]+\frac{1}{\lambda}} \\
\rho' &= \frac{\mathbb{E}[B']}{\mathbb{E}[B']+\frac{1}{\lambda}} = \frac{\frac{\mathbb{E}[S+I]}{1-\rho}}{\frac{\mathbb{E}[S+I]}{1-\rho}+\frac{1}{\lambda}} = \frac{\mathbb{E}[I]+\mathbb{E}[S]}{\mathbb{E}[I]+\frac{1}{\lambda}} \qquad (41)
\end{aligned}$$

Response time in an M/M/1 will be denoted by T, while that in an M/M/1 with setup cost I will be denoted by T^{setup}. To derive $\mathbb{E}[T^{setup}]$, we observe that $\mathbb{E}[T^{setup}]$ is the same whether we use FCFS scheduling or PLCFS (Preemptive Last Come First Served) scheduling. To see this, consider the continuous Markov chain depicting number of jobs in the system under each scheduling policy. The Markov chains are identical.

Thus it suffices to derive $\mathbb{E}[T^{setup}]$ assuming PLCFS scheduling ,which is what we will do. To derive $\mathbb{E}[T^{setup}]$, we will condition on what the arrival sees. There are 3 possibilities:

1. Arrival sees an idle system – with probability $1-\rho'$
2. Arrival sees a busy server, and the server is currently in setup mode – with probability $\rho' \cdot \frac{\mathbb{E}[I]}{\mathbb{E}[B']}$
3. Arrival sees a busy server, and the server is not in setup mode – with probability $\rho' \cdot \frac{\mathbb{E}[B']-\mathbb{E}[I]}{\mathbb{E}[B']}$

What's important here is that, if a new arrival comes in when the server is in setup mode, the server stays in setup mode, even while the new arrival is pushed to the top of the stack. This is because the setup must complete in full before any new work can begin.

$$\begin{aligned}
\mathbb{E}\big[T^{setup}\big] &= (1-\rho')\mathbb{E}[B'] + \rho' \cdot \frac{\mathbb{E}[I]}{\mathbb{E}[B']} \cdot \left(\frac{\mathbb{E}[I+S]}{1-\rho}\right) + \rho' \cdot \frac{\mathbb{E}[B'] - \mathbb{E}[I]}{\mathbb{E}[B']} \cdot \frac{\mathbb{E}[S]}{1-\rho} \\
&= (1-\rho')\mathbb{E}[B'] + \rho'\mathbb{E}[I] + \rho' \cdot \frac{\mathbb{E}[B'] - \mathbb{E}[I]}{\mathbb{E}[B']} \cdot \frac{\mathbb{E}[S]}{1-\rho} \\
&= (1-\rho')\mathbb{E}[B'] + \rho'\mathbb{E}[I] + \rho' \cdot \frac{\mathbb{E}[B'] - \mathbb{E}[I]}{\mathbb{E}[S+I]} \cdot \mathbb{E}[S] \quad \text{(Using Eq. (40))} \\
&= \frac{\frac{1}{\lambda} - \mathbb{E}[S]}{\mathbb{E}[I] + \frac{1}{\lambda}} \cdot \mathbb{E}[B'] + \frac{\mathbb{E}[I+S]}{\mathbb{E}[I] + \frac{1}{\lambda}} \cdot \mathbb{E}[I] + \frac{\mathbb{E}[B'] - \mathbb{E}[I]}{\mathbb{E}[I] + \frac{1}{\lambda}} \cdot \mathbb{E}[S] \quad \text{(Using Eq. (41))} \\
&= \frac{(\frac{1}{\lambda} - \mathbb{E}[S]) \cdot \mathbb{E}[B'] + (\mathbb{E}[I] + \mathbb{E}[S]) \cdot \mathbb{E}[I] + (\mathbb{E}[B'] - \mathbb{E}[I]) \cdot \mathbb{E}[S]}{\mathbb{E}[I] + \frac{1}{\lambda}} \\
&= \frac{\mathbb{E}[B'] + \lambda \cdot \mathbb{E}^2[I]}{\lambda \cdot \mathbb{E}[I] + 1} \\
&= \frac{\frac{\mathbb{E}[I+S]}{1-\rho} + \lambda \cdot \mathbb{E}^2[I]}{\lambda \cdot \mathbb{E}[I] + 1} \quad \text{(Using Eq. (40))} \\
&= \frac{\mathbb{E}[I] + \frac{\mathbb{E}[S] + \rho \cdot \mathbb{E}[I]}{1-\rho} + \lambda \cdot \mathbb{E}^2[I]}{\lambda \cdot \mathbb{E}[I] + 1} \\
&= \frac{\mathbb{E}[S]}{1-\rho} + \mathbb{E}[I] \\
&= \mathbb{E}[T] + \mathbb{E}[I] \\
\implies \mathbb{E}\big[T^{M/M/1/setup}\big] &= \mathbb{E}\big[T^{M/M/1}\big] + \mathbb{E}[I]
\end{aligned}$$

Again, the last equation above is the same as Eq. (1) with $k = 1$.

B M/M/k with multiple SETUP servers

In this section, we attempt to analyze the Markov chain for the M/M/2 with exponential setup, under a different model, wherein multiple servers can be in the SETUP mode simultaneously (in this case, at most 2). The corresponding Markov chain is shown in Fig. 11.

We first find the limiting probabilities for the Markov chain states in the 1st row, in terms of $\pi_{0,0}$. Next, we solve for the limiting probabilities of being in the states of the 2nd row, in terms of the solution for the 1st row, which in turn is expressed in terms of $\pi_{0,0}$. At this point, we'll see that the limiting probabilities of the states in the 2nd row involve messy square roots. This makes it unlikely that a simple closed-form solution for mean response time exists.

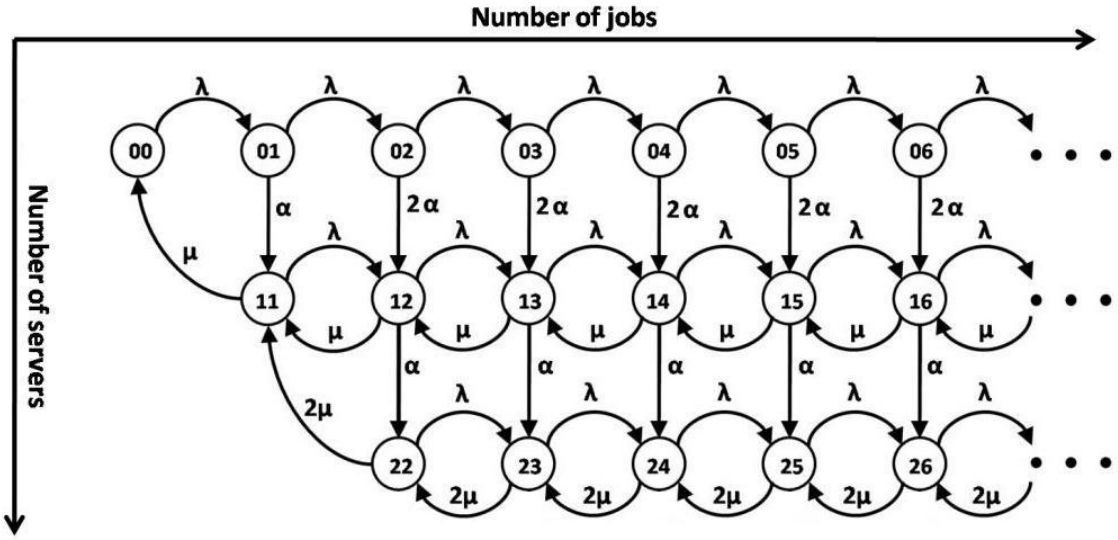


Figure 11: Markov chain for the M/M/2 with exponential setup times, under the assumption that multiple servers can be in the SETUP mode simultaneously.

Step 1: Solving the 1st row
The relevant balance equations for the 1st row are given by:

$$\begin{aligned}\pi_{0,1}\cdot(\lambda+\alpha) &= \pi_{0,0}\cdot\lambda \\ \Longrightarrow \pi_{0,1} &= \pi_{0,0}\cdot\beta \quad \text{where } \beta=\tfrac{\lambda}{\lambda+\alpha}. \qquad (42)\\ \pi_{0,j}\cdot(\lambda+2\alpha) &= \pi_{0,j-1}\cdot\lambda \quad \text{for } j>1. \\ \Longrightarrow \pi_{0,j} &= \pi_{0,j-1}\cdot\beta' \quad \text{for } j>1\text{, and } \beta'=\tfrac{\lambda}{\lambda+2\alpha}. \\ \Longrightarrow \pi_{0,j} &= \pi_{0,0}\cdot\beta\cdot\beta'^{j-1} \quad \text{for } j>1. \qquad (43)\end{aligned}$$

We reserve the balance equation for $\pi_{0,0}$ for the 2nd row.

Step 2: Solving the 2nd row
The relevant balance equations for the 2nd row are given by:

$$\pi_{1,j}\cdot(\lambda+\alpha+\mu) = \pi_{1,j-1}\cdot\lambda+\pi_{0,j}\cdot 2\alpha+\pi_{1,j+1}\cdot\mu \quad \text{for } j>1. \qquad (44)$$

The RHS of Eq. (44) above consists of states of the 2nd row as well as states of the 1st row. Thus, this equation is inhomogeneous, with $\pi_{0,j}$ being the inhomogeneous part. Solutions for such equations are given by:

$$\pi_{1,j} = A_{1,1}x_1^j + A_{1,2}\beta'^j \quad \text{for } j>1\ , \qquad (45)$$

where x_1 is a solution of the homogeneous equation:

$$
\begin{aligned}
x_1 \cdot (\lambda + \alpha + \mu) &= \lambda + x_1^2 \cdot \mu && (46) \\
\Longrightarrow x_1 &= \frac{(\lambda+\alpha+\mu) \pm \sqrt{(\lambda+\alpha+\mu)^2 - 4\lambda\mu}}{2\mu} && (47)
\end{aligned}
$$

Clearly, x_1 has a very messy form. Thus, we hope that it's coefficient, $A_{1,1}$, turns out to be zero. Unfortunately, we find that $A_{1,1}$ is in fact, non-zero.

Substituting Eqs. (43) and (45) into Eq. (44) for $j > 2$, we have:

$$
\begin{aligned}
A_{1,1}x_1^j \cdot (\lambda+\alpha+\mu) + A_{1,2}\beta'^j \cdot (\lambda+\alpha+\mu) &= A_{1,1}x_1^{j-1} \cdot \lambda + A_{1,2}\beta'^{j-1} \cdot \lambda + \pi_{0,0} \cdot \beta \cdot \beta'^{j-1} \cdot 2\alpha \\
&\quad + A_{1,1}x_1^{j+1} \cdot \mu + A_{1,2}\beta'^{j+1} \cdot \mu \\
\Longrightarrow A_{1,2} \cdot \left\{\beta'(\lambda+\alpha+\mu) - \lambda - \beta'^2\mu\right\} &= \pi_{0,0} \cdot \beta \cdot 2\alpha \quad \text{(from Eq. (46))} \\
\Longrightarrow A_{1,2}\frac{\lambda\alpha(2\mu - 2\alpha - \lambda)}{(\lambda+2\alpha)^2} &= \pi_{0,0} \cdot \frac{2\lambda\alpha}{\lambda+\alpha} \\
\Longrightarrow A_{1,2} &= \frac{2\pi_{0,0} \cdot (\lambda+2\alpha)^2}{(\lambda+\alpha)(2\mu-2\alpha-\lambda)} && (48)
\end{aligned}
$$

To get $A_{1,1}$, we want to use a boundary condition for the 2nd row: use the balance equation for $\pi_{1,2}$, which will contain $\pi_{1,1}$. However, we first need to evaluate $\pi_{1,1}$. This requires us to use the balance equation for $\pi_{0,0}$, which we had intentionally left out in Step 1. Using the balance equation for $\pi_{0,0}$:

$$
\begin{aligned}
\pi_{0,0} \cdot \lambda &= \pi_{1,1} \cdot \mu \\
\Longrightarrow \pi_{1,1} &= \pi_{0,0} \cdot \rho && (49)
\end{aligned}
$$

We now use Eqs. (43) and (45) in Eq. (44) for $j = 2$:

$$
\begin{aligned}
A_{1,1}x_1^2 \cdot (\lambda+\alpha+\mu) + A_{1,2}\beta'^2 \cdot (\lambda+\alpha+\mu) &= \pi_{1,1} \cdot \lambda + \pi_{0,0}\beta\beta' \cdot 2\alpha + A_{1,1}x_1^3 \cdot \mu + A_{1,2}\beta'^3 \cdot \mu \\
\Longrightarrow A_{1,1} \cdot \left(x_1^2 \cdot (\lambda+\alpha+\mu - x_1\mu)\right) &= \pi_{0,0}\rho\lambda + \pi_{0,0}\beta\beta' \cdot 2\alpha + A_{1,2}\beta'^2(\beta'\mu - \lambda - \alpha - \mu) \quad \text{(from Eq. (49))}
\end{aligned}
$$

$$\begin{aligned}
\implies A_{1,1} \cdot \lambda \cdot x_1 &= \pi_{0,0}\lambda^2 \left(\frac{1}{\mu} + \frac{2\alpha}{(\lambda+\alpha)(\lambda+2\alpha)} \right) - A_{1,2}\beta'^2 \left(\frac{\lambda^2 + 3\lambda\alpha + 2\alpha^2 + 2\alpha\mu}{\lambda + 2\alpha} \right) \\
\implies A_{1,1} \cdot \lambda \cdot x_1 &= \pi_{0,0}\lambda^2 \frac{(\lambda^2 + 3\lambda\alpha + 2\alpha^2 + 2\alpha\mu)}{\mu(\lambda+\alpha)(\lambda+2\alpha)} - \frac{(2\pi_{0,0}\lambda^2)(\lambda^2 + 3\lambda\alpha + 2\alpha^2 + 2\alpha\mu)}{(\lambda+\alpha)(2\mu - 2\alpha - \lambda)(\lambda+2\alpha)} \\
&\qquad \text{(from Eq. (48))} \\
\implies A_{1,1} \cdot \lambda \cdot x_1 &= \frac{-\pi_{0,0}\lambda^2 \cancel{(\lambda+2\alpha)}(\lambda^2 + 3\lambda\alpha + 2\alpha^2 + 2\alpha\mu)}{\mu(\lambda+\alpha)\cancel{(\lambda+2\alpha)}(2\mu - 2\alpha - \lambda)} \\
\implies A_{1,1} \cdot \lambda \cdot x_1 &= \frac{-\pi_{0,0}\lambda^2(\lambda^2 + 3\lambda\alpha + 2\alpha^2 + 2\alpha\mu)}{\mu(\lambda+\alpha)(2\mu - 2\alpha - \lambda)} \\
\implies A_{1,1} \cdot \lambda \cdot x_1 &\neq 0 \\
\implies A_{1,1} &\neq 0 \quad \text{(since } x_1 \neq 0 \text{ from Eq. (47))} \qquad (50)
\end{aligned}$$

Thus we see that the expression for limiting probabilities, given by Eq. (45) involves an x_1 term with messy square-roots, and a non-zero coefficient, $A_{1,1}$. While one can continue along these same lines to determine the $\pi_{2,j}$ limiting probabilities, the form of the $\pi_{1,j}$ limiting probabilities suggests that it is unlikely that we will find a simple closed-form expression for mean response time.

To better understand the mean response time for the M/M/k with multiple SETUP servers, we applied matrix analytic methods (see [9] for an excellent reference on matrix analytic methods) to analyze the Markov chain.

Fig. 12 shows our results for the mean response time versus mean setup time for an M/M/k system with exponential setup times, for both, the model in Section 3.3 (shown as dots) as well as the model in this section (shown as crosses). We consider different values of k and $\rho = \lambda \cdot \mathbb{E}[X]$ in our results. Clearly, the decomposition property of Eq. (1) is not satisfied by the M/M/k with multiple SETUP servers model, however, it is not clear whether some other decomposition property might exist.

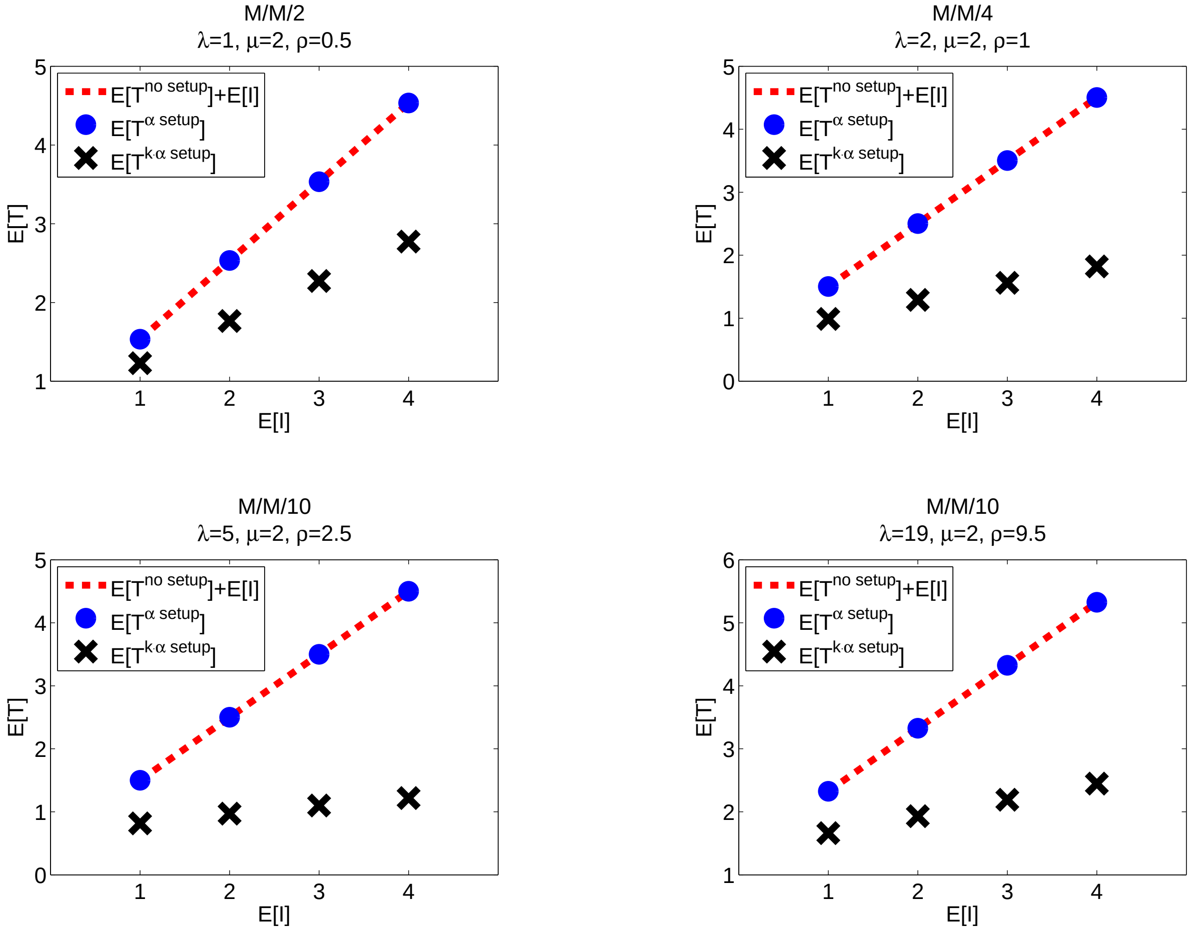


Figure 12: Matrix-analytic results for an M/M/k with exponential setup costs, under two different models: (i) The α model, where only one server can be in the SETUP mode at any time (represented by dots) and (ii) The $k\cdot\alpha$ model, where multiple servers can be in the SETUP mode simultaneously (represented by crosses). Results show that the $k \cdot \alpha$ model does not satisfy the decomposition property as expressed by Eq. (1), whereas the α model does satisfy this property.