Fundamental Characteristics of Queues with Fluctuating Load

Varun Gupta* **Mor Harchol-Balter*†**
Alan Scheller-Wolf‡ **Uri Yechiali§**

August 2006
CMU-CS-06-117

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

*School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
‡Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA, USA
§Department of Statistics and Operations Research, School of Mathematical Sciences, Tel Aviv University, Tel Aviv, Israel
†Supported by NSF Career Grant CCR-0133077, NSF Theory CCR-0311383, NSF ITR CCR-0313148.

Keywords: Fluctuating load; MAP; MMPP; non-stationary arrivals/service; Ross's conjecture; stochastic ordering

Abstract

Systems whose arrival or service rates fluctuate over time are very common, but are still not well understood analytically. Stationary formulas are poor predictors of systems with fluctuating load. When the arrival and service processes fluctuate in a Markovian manner, computational methods, such as Matrix-analytic and spectral analysis, have been instrumental in the numerical evaluation of quantities like mean response time. However, such computational tools provide only limited insight into the *functional behavior* of the system with respect to its primitive input parameters: the arrival rates, service rates, and rate of fluctuation.

For example, the shape of the function that maps rate of fluctuation to mean response time is not well understood, even for an M/M/1 system. Is this function increasing, decreasing, monotonic? How is its shape affected by the primitive input parameters? Is there a simple closed-form approximation for the shape of this curve? Turning to user experience: How is the performance experienced by a user arriving into a "high load" period different from that of a user arriving into a "low load" period, or simply a random user. Are there stochastic relations between these? In this work, we provide the first answers to these fundamental questions.

"Characteristics of queues with non-stationary input streams are difficult to evaluate, therefore their bounds are of importance."

-TOMASZ ROLSKI [27]

1 Introduction

Motivation and model

The vast majority of queueing models assume a stationary process in order to derive performance characteristics, such as mean response time or mean number in system. In reality, computer systems have arrival rates which fluctuate over time. Furthermore, when the arrival rate is high, it is common to try to compensate by increasing the service rate, possibly by adding additional servers.

System designers often try to use standard queueing theorems, such as the stationary M/M/1 formulas, to predict the performance of their system. However, when the load fluctuates over time, it is not clear which stationary formula to use. One can try to average the load in some way over time, and use a stationary M/M/1 with the "average load," to predict system performance. However, as many system designers know, this is a very poor estimation of mean behavior. Furthermore, it completely ignores the differences in user perceived performance depending on whether the user arrives into a high-load or low-load state.

As people have become aware of the effects of fluctuating load, mathematical tools have been developed, such as matrix analytic methods and spectral analysis, which allow one to numerically evaluate systems in which the arrival rate and/or service rate change over time according to a Markovian process. While such tools provide numerical values for time-average behavior, they provide only limited insight into the functional behavior of the system with respect to the input parameters. These methods don't tell us how the mean response time is affected by the rate of fluctuation between high and low load, whether this is increasing or decreasing, whether it is monotonic, etc. These methods don't give us a complete sense of how the results vary as a function of the other input primitives, such as the arrival rate and service rate, or which parameters are most important.

In order to consider such questions, we evaluate a specific model for fluctuating load, shown in Figure 1.1. The system alternates between a "high" state and a "low" state, according to a Markovian process, where the system is in "high" for an exponentially-distributed time with rate α^H and in the "low" state for an exponentially-distributed time with rate α^L. While in the high state (respectively low state), arrivals occur according to a Poisson Process with rate λ^H (respectively, λ^L). Also while in the high state (respectively low state), services complete with exponential rate μ^H (respectively, μ^L). We define $\rho^H = \frac{\lambda^H}{\mu^H}$ and $\rho^L = \frac{\lambda^L}{\mu^L}$ and assume throughout that $\rho^H \geq \rho^L$ (but we *do not* assume any relationship between λ^H and λ^L or between μ^H and μ^L). We allow $\rho^H > 1$, provided that the system is still stable, as defined in Section 2. Note that the above model encompasses as special cases models with ON/OFF arrival processes (where $\lambda^L = 0$) and/or breakdowns ($\mu^H = 0$, in this case we define $\rho^H = \infty$). Even our simple Markovian model generates non-obvious and counter-intuitive behavior, and provides insight for more general models. In Section 3, we will consider a more general variant of our model where we allow for a burst of arrivals at each arrival instant, where the burst size can have an arbitrary distribution.

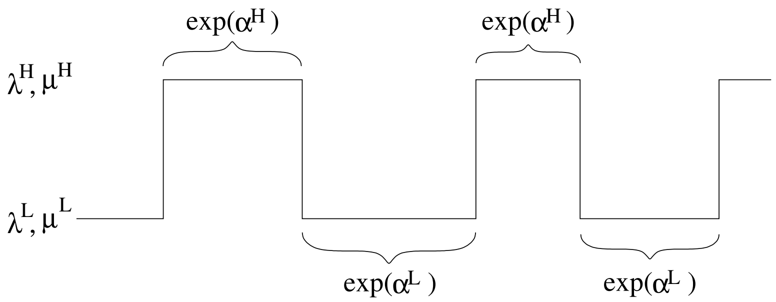


Figure 1.1: Alternating Load Model

Prior Work

Time-varying models, have been very widely studied since the earliest work in the 50's, continuing unabated to the present. (In the interest of brevity, we focus on models having non-deterministic switching behavior.) The earliest reference of this type is Clarke [9], who used generating functions to derive expressions for the number in queue. Soon thereafter other researchers applied transforms and generating functions to related models: Neuts [20], Çinlar [7, 6], Arjas [3]. Yechiali and Naor [34] used generating functions to reduce the solution of our model to that of obtaining the roots of a cubic equation. Using similar techniques, de Smit [10] obtained a Wiener-Hopf factorization for systems with MAP arrivals and general service; Sengupta [29] analyzed a system with Markovian arrival and service distributions and service interruptions; Takine and Sengupta [32] generalized [29] to MAP arrivals and general service; Adan and Kulkarni [2] allowed dependencies between successive arrivals and services in a MAP/G/1 framework; and finally Harrison and Zatschler [12] numerically derived the entire sojourn time distribution for very general Markovian systems which they call G-Queues.

A second class of highly effective analytical tools for time-varying models are the Matrix Analytical and related techniques. Neuts [21] used Matrix Analytical techniques to obtain numerical results for our model, observing that its behavior could be qualitatively different from the stationary $M/M/1$; Ramaswami [25] allowed general service times and Markovian Arrival Processes (MAP); Lucantoni, Meier-Hellstern and Neuts [15] modeled phase-type service and added server vacations; Sengupta [30] allowed dependencies between semi-Markov interarrival and semi-Markov service times; Takine et al. [31] combined Matrix Analytical techniques with generating functions to allow multiple customer classes and priorities; Lucantoni and Neuts [16] allowed batch MAP arrivals; Mitrani and Chakka numerically compared Matrix Analytic and spectral expansion techniques [18]; and finally Asmussen and Møller [4] solved matrix equations to get the entire waiting time distribution for a queue with MAP arrivals, phase-type service and multiple servers.

It is thus clear that researchers have been highly effective at developing methods to obtain *numerical results*, but what about basic properties, intuitive insights and simple approximations? Researchers have been at work in these directions as well. One of the classic conjectures in queueing theory was posed by Ross [28], who conjectured that increasing variability (fluctuation rate) in a Poisson arrival process would (weakly) increase mean customer delay, when the service rate does not vary. Rolski [27] confirmed this conjecture, and more recently Miyoshi and Rolski [19] extended the proof of Ross's conjecture to more general queueing models. Heyman [13] provided a contrasting insight – he gave an example of a deterministically varying arrival function that performs no worse than the stationary version. We continue this tradition in our current work – generalizing [13] by finding simple conditions which guarantee that a stationary system and its time-varying analog perform identically in our Markovian setting.

Another way to garner intuition for time-varying systems is to analyze limiting regimes. Very early on, Newell [22, 23, 24] used diffusion approximations for time-varying $M/M/1$ queues. Later, Massey [17] used uniform acceleration to derive the transient behaviors; Abate, Choudhury and Whitt [1] derived tail asymptotics for the waiting time and workload in MAP/GI/1 and MAP/MAP/1 queues; and Rider [26], Gelenbe and Rosenberg [11], Choudhury et al. [8], and Yang and Knessl [33] evaluated the special case when transitions happen much more slowly than arrivals or departures. Finally, Knessl and Yang [14] restricted themselves to a case in which the traffic intensity takes a very specific form, with the aim of generating insights for more general cases.

Our Goals

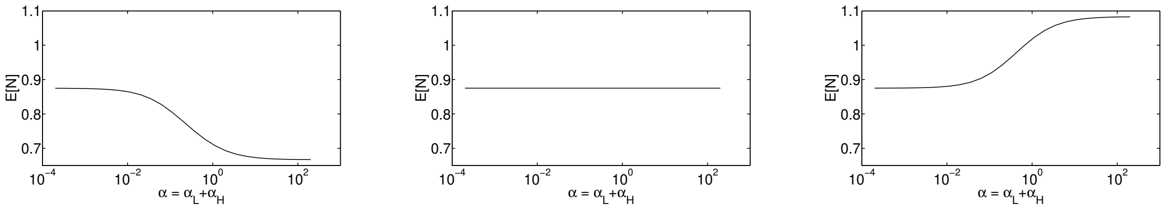
As we saw above, the prior work is very effective at producing computational results for our, and even more complex models. However, it is more limited at providing intuition. Part of the problem is that all these methods (generating functions, Matrix Analytical, Spectral Expansion) involve calculating the root of a cubic equation. While in theory a cubic polynomial can be solved analytically, in practice the solution is so cumbersome (dozens of lines in *Mathematica*) that there is no way to get a sense of the effect of the input parameters on the system performance. For example, the prior work does not provide a sense of the shape of the response time curve, nor how response time relates to the input primitives, such as the α^H and α^L parameters or the λ^H, λ^L, μ^H, μ^L parameters. Our goal in this work is to get this type of intuition.

One of the simplest/most fundamental questions is what happens when the rate of fluctuation (the α's) either approach zero or approach infinity. The prior work has not yet provided answers to even the very basic question of whether fast or slow fluctuations lead to higher mean response times. Ross [28] conjectured, and Rolski confirmed [27], that fluctuation leads to higher mean response time for the case where the mean service rate is a constant ($\mu^H = \mu^L = \mu$). In our more general model, however, where the service rate changes ($\mu^H \neq \mu^L$), we find in Section 2 that lower rate of fluctuation does *not* always lead to higher mean response time. There are cases where the response time is insensitive to the rate of fluctuation, or can even drop as the rate of fluctuation decreases. We derive a criterion, based on the notion of "slack," (s^H and s^L) where $s^H = \mu^H - \lambda^H$ and $s^L = \mu^L - \lambda^L$, which determines whether faster or slower rates of fluctuation result in better system performance.

Another fundamental question in the same vein is whether response time is always bounded by the two asymptotes, the case of high fluctuation rate and low fluctuation rate. Specifically, does a system with a "medium" fluctuation rate always have mean response time in between those two extreme cases? And if so, does mean response time change monotonically between those two extremes? To answer these questions,we start by deriving the transform for the number of jobs in our model (Section 3), and then we analyze a certain root of the denominator of this transform which allows us to answer these questions affirmatively in Section 4.

Our work also produces simple and accurate approximations for the mean number of jobs in the system, see Section 5. We do this by again starting with the transform derived in Section 3, but deriving approximations for its roots. We provide both a simple closed-form approximation which holds for all fluctuation rates (α's), as well as even simpler approximations which specialize for the case of only "high" or "low" α. While computational methods exist for obtaining the exact mean response time, our simple and accurate approximations have advantages over the exact results. From a computational perspective, the fact that our approximations are closed-form solutions means that they can easily be computed on any spread-sheet. More importantly our approximations provide the first results about the *shape* of the mean response time curve as a function of the fluctuation rate, α. In particular, they provide a simple and accurate approximation for the curve's functional form. The advantage of the simple functional form is that it shows which primitives are most important in determining mean response time, and allows for further sensitivity analysis. We also derive the closed-form fluid approximation for the mean number of jobs in the system for the case $\lambda^H > \mu^H$, and compare it numerically with the exact value.

In Section 6, we use our analysis to provide some insights into the behavior of the fluctuating load queue. We first ask the question: How does the mean number of jobs in the system vary as we scale the arrival and service rates? We find that the answer in this case is different from that in a $GI/GI/1$ queue, and, in fact, varies depending on the "slacks". Next we look at the effect of scaling the switching rates on the mean number of jobs, and identify the regimes where this scaling has a more pronounced effect, and the regimes where there is negligible effect on the mean number of jobs. Finally, we consider the problem of optimal capacity provisioning in a queue with fluctuating arrival rates, but a given total average service capacity. We provide a simple expression for near-optimal capacity splitting (irrespective of the switching rates). Further, our findings prove that under scenarios where μ^H and μ^L are under the control of a system designer, optimal capacity provisioning with a fluctuating arrival process leads to a smaller mean number of jobs, when compared to a system with a constant mean arrival rate. Thus, a fluctuating arrival



(a) $s^L > s^H$ (b) $s^L = s^H$ (c) $s^L < s^H$

Figure 2.2: Illustration of the behavior of E[N] as a function of α, described in Theorem 2.2. For all three figures, we fix $\rho^L = 0.2$ and $\rho^H = 0.6$, and $\alpha^L = \alpha^H$. The rest of the parameters are as follows: (a) $\mu^L = 1$, $\lambda^L = 0.2$, $\mu^H = 1$, $\lambda^H = 0.6$; (b) $\mu^L = 1$, $\lambda^L = 0.2$, $\mu^H = 2$, $\lambda^H = 1.2$; (c) $\mu^L = 1$, $\lambda^L = 0.2$, $\mu^H = 4$, $\lambda^H = 2.4$.

process is desirable as it can lead to more "efficient" resource provisioning.

Finally, while our results thus far have dealt with the overall time-average mean performance behavior, it is also of practical importance to understand how this time-average mean compares to the experience of a customer arriving into a "high" (H) period or a customer arriving into a "low" (L) period. Once again computational results can be used to evaluate specific instances, however we seek a qualitative ordering. We answer this question in Section 7, comparing three quantities: the number in system witnessed by an arrival into an H period, the number in system witnessed by an arrival into an L period, and the number in system witnessed by an arrival into a stationary system whose arrival rate is the weighted average of the two arrival rates and whose service rate is the weighted average of the two service rates. We find that a *stochastic dominance* relationship does exist. However, counter to intuition we find that while the number of jobs seen by an arrival into the 'average' system and the number of jobs seen by an arrival into an L phase are both stochastically dominated by the number of jobs seen by an arrival into an H phase, the number of jobs seen by an arrival into an L phase is *not* stochastically dominated by the number of jobs in the average system.

Throughout the majority of the report we investigate the characteristics of the mean number in system, $E[N]$, as through application of *Little's Law* (using the time-average arrival rate) this yields results for mean response time.

2 Anomalous Behavior of Fluctuating Load Queue

We start our work by asking the most basic of questions: How does the mean number of jobs in the system, $E[N]$, compare in the case when the load fluctuates slowly (low α), as compared with the case where the load fluctuates quickly (high α)? For all the work that has been done on numerically evaluating instances of our model, the question of whether $E[N]$ is higher under low

α or high α has not been addressed. Although intuition would tell us that low α should lead to higher $E[N]$ because there is seemingly more variability in the load in this case, this fact has not been proven. In this section we prove that lower α does *not* always lead to higher $E[N]$, and we derive a criterion that tells us when $E[N]$ increases for low α and when it *decreases* for low α. Before we can state our theorem, we need to define a quantity which we call *slack* and which we use throughout the paper.

Definition 2.1 *The* slack *during the low load period is defined as $s^L \equiv \mu^L - \lambda^L$. The slack during the high load period is defined as $s^H \equiv \mu^H - \lambda^H$.*

Recall that we make no assumptions about μ^L, μ^H, λ^L, or λ^H, except to assume that $\rho^H \equiv \frac{\lambda^H}{\mu^H} > \rho^L \equiv \frac{\lambda^L}{\mu^L}$. We allow $\rho^H > 1$, so long as stability is met. The remainder of the section will be spent proving Theorem 2.2 below; providing a condition for stability; and discussing the nebulous concept of "load," in a load-fluctuating system.

Theorem 2.2 *Let $\alpha = \alpha^L + \alpha^H$.*
If $s^L < s^H$, then $E[N^{\alpha\to 0}] < E[N^{\alpha\to\infty}]$.
If $s^L > s^H$, then $E[N^{\alpha\to 0}] > E[N^{\alpha\to\infty}]$.
If $s^L = s^H$, then $E[N^{\alpha\to 0}] = E[N^{\alpha\to\infty}]$.

Corollary 2.3 *If $\mu^H = \mu^L$, then $E[N^{\alpha\to 0}] \geq E[N^{\alpha\to\infty}]$ for all settings. This confirms Ross's Conjecture.*

We start with a discussion of the two extreme values of $E[N]$ when $\rho^H < 1$; the case where α^L and α^H are both very low, and the case where α^L and α^H are both very high. When the α's are very low, $E[N]$ can be shown to be a weighted mixture of the mean numbers of jobs under two stationary M/M/1 queues: one with load ρ^L and the other with load ρ^H. This may seem obvious, but it will be formally verified via our analysis in Section 3. Specifically, we have:

$$E[N^{\alpha\to 0}] = \frac{\frac{\rho^L}{1-\rho^L}\cdot\frac{1}{\alpha^L} + \frac{\rho^H}{1-\rho^H}\cdot\frac{1}{\alpha^H}}{\frac{1}{\alpha^L} + \frac{1}{\alpha^H}}$$

By contrast, when α^L and α^H are very high, fluctuations are very rapid. In this case, our analysis in Section 3 will show that the system converges to a single M/M/1 queue with load ρ^A:

$$\rho^A = \frac{\lambda^A}{\mu^A} = \frac{\frac{\lambda^H}{\alpha^H} + \frac{\lambda^L}{\alpha^L}}{\frac{\mu^H}{\alpha^H} + \frac{\mu^L}{\alpha^L}}$$

where μ^A and λ^A are the average service and arrival rates,

$$\mu^A = \frac{\frac{\mu^H}{\alpha^H} + \frac{\mu^L}{\alpha^L}}{\frac{1}{\alpha^H} + \frac{1}{\alpha^L}}, \quad \lambda^A = \frac{\frac{\lambda^H}{\alpha^H} + \frac{\lambda^L}{\alpha^L}}{\frac{1}{\alpha^H} + \frac{1}{\alpha^L}}$$

That is,

$$E[N^{\alpha\to\infty}] = \frac{\rho^A}{1-\rho^A}$$

Observation 2.4 *We observe that ρ^A as defined above serves as a* stability criterion *for the system under all α^H and α^L values, since $\rho^A < 1$ is equivalent to saying that the time-average arrival rate is less than the time-average service rate. However, ρ^A* does not *represent the* true load. *Specifically*

$$\rho^A \neq 1 - \pi_0$$

where π_0 represents the fraction of time that the system is idle. In fact, we conjecture that determining π_0 is as hard a problem as determining $E[N]$. (These last two observations were also made by Yechiali and Naor [34].)

We now prove Theorem 2.2.

Proof: The necessary and sufficient condition for
$E[N^{\alpha\to 0}] < E[N^{\alpha\to\infty}]$ is:

$$\frac{\frac{\rho^L}{(1-\rho^L)\alpha^L} + \frac{\rho^H}{(1-\rho^H)\alpha^H}}{\frac{1}{\alpha^L} + \frac{1}{\alpha^H}} < \frac{\rho^A}{1-\rho^A} \tag{2.1}$$

which reduces to

$$\lambda^H(1-c) + \lambda^L\left(1 - \frac{1}{c}\right) > 0$$

where $c = \frac{\mu^L - \lambda^L}{\mu^H - \lambda^H}$. Or,

$$\lambda^H c^2 - c(\lambda^H + \lambda^L) + \lambda^L < 0$$

The solution to the above inequality is $c \in \left(\frac{\lambda^L}{\lambda^H}, 1\right)$. Note that $c > \frac{\lambda^L}{\lambda^H}$ is equivalent to $\rho^H > \rho^L$, which is trivially true. Therefore the only other condition is $c < 1$, or $(\mu^L - \lambda^L) < (\mu^H - \lambda^H)$. Note, we are assuming $\rho^H < 1$, otherwise this behavior is not possible. The remaining cases in the theorem are proven analogously. ■

The behavior of the fluctuating-load queue is illustrated in Figure 2.2.

Intuition for Theorem 2.2

While the proof of Theorem 2.2 was purely algebraic, we can provide some intuition for the observed behavior. Recall that in an $M/M/1$ queue with arrival rate λ and service rate μ, the mean response time is given by $\frac{1}{\mu-\lambda}$. That is, the mean response time of an $M/M/1$ queue is the inverse of the slack between the arrival and service rates. Thus if we compared two $M/M/1$ queues, one operating at rates λ^H and μ^H and the other at rates λ^L and μ^L, when $(\mu^H - \lambda^H) > (\mu^L - \lambda^L)$, the former system exhibits a lower mean response time.

Now let us compare the fraction of customers departing during the H phase in the fluctuating load queue when $\rho^H < 1$. As $\alpha \to 0$, almost all customers arriving during the H phase depart during the same H phase. Therefore, $\frac{\lambda^H/\alpha^H}{\lambda^H/\alpha^H+\lambda^L/\alpha^L}$ fraction of customers depart during the H phase as $\alpha \to 0$. When $\alpha \to \infty$, the fraction of customers departing during the H phases is just the fraction of service capacity offered during the H phase, that is, $\frac{\mu^H/\alpha^H}{\mu^H/\alpha^H+\mu^L/\alpha^L}$.

As can easily be seen, the fraction of customers departing during H phase increases by a factor of ρ^H/ρ^A as the switching rates decrease from ∞ to 0. Thus when $s^H > s^L$, as switching rates decrease, an increasing fraction of customers experience lower mean response times due to lower slack offered in the H phases causing a lowering of overall mean response time and hence mean number of jobs in the system.

3 Analysis

We first define the following quantities.

Definition 3.1 *N^L is defined as the random variable for the number of jobs at the instants when the system switches from a Low (L) to a High (H) phase. The z-transform of N^L is denoted by $\widehat{\Pi^L}(z)$. Similarly, N^H represents the random variable for the number of jobs at the end of H phases and $\widehat{\Pi^H}(z)$ denotes the z-transform of N^H.*

N^L and N^H are illustrated in Figure 3.3. Our approach is based on deriving the expressions for $\widehat{\Pi^L}(z)$ and $\widehat{\Pi^H}(z)$. In Section 3.1 we prove that knowledge of the *distributions at switching points* suffices to determine the distribution of the number of jobs in the system at a randomly sampled point in time. To derive $\widehat{\Pi^L}(z)$ and $\widehat{\Pi^H}(z)$, we will first obtain a *transition function* which maps the distribution of number of jobs at a switching point to the distribution at the next switching point (Section 3.2, equation (3.6)). This transition function will then allow us to express $\widehat{\Pi^L}(z)$ in terms of $\widehat{\Pi^H}(z)$, and vice-versa (see Section 3.3, equations (3.9)-(3.10)). Finally we solve these to get expressions for $\widehat{\Pi^L}(z)$ and $\widehat{\Pi^H}(z)$ in terms of π_0 only (see Section 3.3, equation (3.8)). All the transform derivations described above will assume a more general model than we have considered so far, where we allow for a *burst of arrivals* at each arrival instant, where the burst size can be arbitrary.

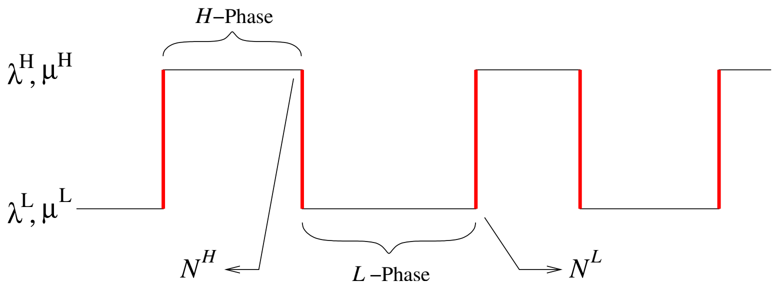


Figure 3.3: Switching Points used in Analysis

3.1 Conditional PASTA

Let N denote the random variable for the number of jobs in the system at a randomly sampled point. The following theorem relates the distribution of N with those of N^L and N^H.

Theorem 3.2 *N has the same distribution as $\mathscr{N}$, where*

$$\mathscr{N} = \begin{cases} N^L & w.p. \ \frac{\alpha^H}{\alpha^L+\alpha^H} \\ N^H & w.p. \ \frac{\alpha^L}{\alpha^L+\alpha^H} \end{cases}$$

Proof: Let $\widehat{\Pi}(z)$ be the z-transform of N. Proving the above theorem is equivalent to proving

$$\widehat{\Pi}(z) = \frac{\frac{\widehat{\Pi^L}(z)}{\alpha^L} + \frac{\widehat{\Pi^H}(z)}{\alpha^H}}{\frac{1}{\alpha^L} + \frac{1}{\alpha^H}} \tag{3.2}$$

Let $\widehat{\Pi^L}(z,t)$ be the z-transform for the number of jobs in the system t units of time after the start of the L phase, conditioned on the phase being longer than t, and let $\widehat{\Pi^H}(z,t)$ be the corresponding quantity for H phase. Note that $\widehat{\Pi^L}(z,0) = \widehat{\Pi^H}(z)$. Also by conditioning on the length of an L phase,

$$\widehat{\Pi^L}(z) = \int_{u=0}^{\infty} \widehat{\Pi^L}(z,u)\alpha^L e^{-\alpha^L u} du \tag{3.3}$$

We will use renewal-reward theory to prove equation (3.2). The renewal cycles consist of a single L phase followed by a single H phase. The instantaneous reward earned at time t is given by $r(t) = z^{n(t)}$ where $n(t)$ is the number of jobs in the system at time t. Clearly, $\widehat{\Pi}(z)$ is the long run average rate at which reward is earned. Therefore,

$$
\begin{aligned}
\widehat{\Pi}(z) &= \frac{E[\text{reward in } L \text{ phase}] + E[\text{reward in } H \text{ phase}]}{E[\text{length of } L \text{ phase}] + E[\text{length of } H \text{ phase}]} \\
&= \left(\frac{1}{\alpha^L} + \frac{1}{\alpha^H}\right)^{-1} \left[\int_{t=0}^{\infty}\int_{u=0}^{t} \widehat{\Pi^L}(z,u)du\, \alpha^L e^{-\alpha^L t}dt + \int_{t=0}^{\infty}\int_{u=0}^{t} \widehat{\Pi^H}(z,u)du\, \alpha^H e^{-\alpha^H t}dt\right] \\
&= \left(\frac{1}{\alpha^L} + \frac{1}{\alpha^H}\right)^{-1} \left[\int_{u=0}^{\infty}\int_{t=u}^{\infty} \widehat{\Pi^L}(z,u)du\, \alpha^L e^{-\alpha^L t}dt + \int_{u=0}^{\infty}\int_{t=u}^{\infty} \widehat{\Pi^H}(z,u)du\, \alpha^H e^{-\alpha^H t}dt\right] \\
&= \left(\frac{1}{\alpha^L} + \frac{1}{\alpha^H}\right)^{-1} \left[\frac{\int_{u=0}^{\infty} \widehat{\Pi^L}(z,u)\, \alpha^L e^{-\alpha^L u}du}{\alpha^L} + \frac{\int_{u=0}^{\infty} \widehat{\Pi^H}(z,u)\, \alpha^H e^{-\alpha^H u}du}{\alpha^H}\right] && (3.4) \\
&= \frac{\frac{\widehat{\Pi^L}(z)}{\alpha^L} + \frac{\widehat{\Pi^H}(z)}{\alpha^H}}{\frac{1}{\alpha^L} + \frac{1}{\alpha^H}} && (3.5)
\end{aligned}
$$

■

The intuition behind Theorem 3.2 is the PASTA (Poisson Arrivals See Time Averages) property exhibited by the Markovian switching process, as we prove next.

Theorem 3.3 *The time average distribution of number of jobs in the system during L phases (respectively H phases) is the same as the distribution of N^L (respectively N^H).*

Proof: Consider a slight modification of our system where whenever the system switches from H to L, we restart the system with an initial number of jobs sampled from the distribution of N^H. It is obvious that the time average distribution of number of jobs during the L phases in our original system is the same as the time average distribution of the number of jobs during the L phases in this modified system. Now consider another queueing system where we set off timers according to a Poisson process with rate α^L. Whenever a timer expires, we restart the system with some number of jobs sampled from the distribution of N^H. This can be visualized as seeing only the L phases of our modified queueing system stitched together. Since the timer events are a Poisson process, by PASTA, the distribution of number of jobs at these event instants is the same as the time average distribution, which is the time average distribution of number of jobs during the L phases in the modified system and hence the same as the time average distribution of jobs during L phases in the original system.
To further justify the use of PASTA, the time average distribution of number of jobs in the final system is the distribution at a randomly sampled point in time. Since the timer events are Poisson, the distribution of elapsed time since the immediately preceding timer expiration and a random time instant is also exponential with mean $\frac{1}{\alpha^L}$. Therefore, the distribution at such a ramdom time is the distribution an $exp(\alpha^L)$ time after the start of an L phase, precisely N^L by definition. ■

Now, since the long term fraction of time spent in L phases is $\frac{\alpha^H}{\alpha^L+\alpha^H}$ and in H is $\frac{\alpha^L}{\alpha^L+\alpha^H}$, the linear combination of Theorem 3.2 follows. Although we have proved the above result for only one observable quantity, the number of jobs in the system, the result holds for any observable quantity e.g. square of number of jobs in system, age of the job in service, z-transform of the number of jobs in service.

To summarize, although we defined N^L and N^H to be the distributions of number of jobs at switching points, they are the same as the distributions for number of jobs seen by an arbitrary arrival during *the L or H phase, respectively.*

3.2 Derivation of Transition functions

Our goal in this section is to derive a transition function which maps the distribution of the number of jobs at a switching point to the distribution at the next switching point. To do this, we first need to return to a simple $M/M/1$ queue (without fluctuating load), and consider its transient behavior with respect to the number of jobs at time $T \sim exp(\alpha)$, given a distribution on the number of jobs at time 0.

Consider an $M/M/1$ queue with service rate μ where with rate λ arrivals occur (possibly, more than 1). Let $N(t)$ be the number of jobs in the system at time t and $\widehat{\Pi}(z,t)$ be the z-transform of $N(t)$. Let T be an exponentially distributed random variable with mean $\frac{1}{\alpha}$. We represent $\widehat{\Pi}(z,T)$, the z-transform of $N(T)$, by $\widehat{\Pi}_\alpha(z)$. The following Theorem expresses $\widehat{\Pi}_\alpha(z)$ as a function of $\widehat{\Pi}(z,0)$.

Theorem 3.4

$$\widehat{\Pi}_\alpha(z) = \frac{\alpha z\widehat{\Pi}(z,0) - \mu(1-z)\pi_\alpha}{\alpha z - \mu(1-z) + \lambda z(1-\widehat{A}(z))} \tag{3.6}$$

where $\widehat{A}(z)$ is the z-transform of the burst size distribution and if we let ξ denote the root of denominator of (3.6) *in the interval* $(0,1)$*, then,*

$$\pi_\alpha = \frac{\alpha\xi\widehat{\Pi}(\xi,0)}{\mu(1-\xi)} \tag{3.7}$$

The constant π_α is equal to the idle probability at T.

Proof: The proof of the above theorem is a trivial extension of Bailey's [5] work on transient analysis of $M/M/1$ queues to incorporate bursts. We mention it here for completeness. Let a_j be the probability that the burst size is j (wlog, $a_0 = 0$). Also, let $p_i(t)$ be the probability that there are i jobs in the system at time t. We can now write the differential equations for this system:

$$\begin{aligned}
\frac{dp_i(t)}{dt} &= \lambda\sum_{j=1}^{i} a_j p_{i-j}(t) - (\lambda+\mu)p_i(t) + \mu p_{i+1}(t) \\
\frac{p_0(t)}{dt} &= -\lambda p_0(t) + \mu p_1(t)
\end{aligned}$$

which gives,

$$z\frac{\partial\widehat{\Pi}(z,t)}{\partial t} = \widehat{\Pi}(z,t)\left\{\mu(1-z) - \lambda z(1-\widehat{A}(z))\right\} - \mu(1-z)p_0(t)$$

Integrating by parts, we get the expression for $\widehat{\Pi}_\alpha(z)$ as:

$$\begin{aligned}\widehat{\Pi}_\alpha(z) &= \int_0^\infty \widehat{\Pi}(z,t)\alpha e^{-\alpha t}dt \\ &= \frac{\alpha z\widehat{\Pi}(z,0) - \mu(1-z)p_0(T)}{\alpha z - \mu(1-z) + \lambda z(1-\widehat{A}(z))}\end{aligned}$$

To complete the solution we need to find $p_0(T)$ $(= \pi_\alpha)$. The denominator in the expression of $\widehat{\Pi}_\alpha(z)$ has the value $-\mu < 0$ at $z = 0$ and $\alpha > 0$ at $z = 1$ and therefore, a root $\xi \in (0, 1)$. For $\widehat{\Pi}_\alpha(z)$ to converge inside the unit disk $|z| < 1$, ξ must also be a root of the numerator. Hence,

$$p_0(T) = \frac{\alpha\xi\widehat{\Pi}(\xi,0)}{\mu(1-\xi)}$$

■

The transition functions mapping the distribution at the start of an L or an H phase to the end of the phase is obtained by specifying μ, λ, α and $\widehat{A}(z)$ in (3.6).

3.3 Distributions at Switching Points

The transition function we have derived will now allow us to write fixed point equations, the solution to which will be the desired expressions for $\widehat{\Pi^L}(z)$ and $\widehat{\Pi^H}(z)$. To get there, let $\widehat{A^L}(z)$ (respectively $\widehat{A^H}(z)$) be the z-transform for the distribution of the burst sizes for arrivals during L (respectively H) phases. Let $\overline{A^L}$ and $\overline{A^H}$ be the mean burst sizes during L and H phases. Theorem 3.5, gives the expressions for $\widehat{\Pi^L}(z)$ and $\widehat{\Pi^H}(z)$.

Theorem 3.5

$$\widehat{\Pi^L}(z) = \frac{z\alpha^H\sigma^L + z\alpha^L\sigma^H - (1-z)\sigma_z^H(\mu^L\pi_0^L)}{z\alpha^H\sigma_z^L + z\alpha^L\sigma_z^H - (1-z)\sigma_z^H\sigma_z^L} \tag{3.8}$$

where,

$$\sigma_z^L = \mu^L - \lambda^L z \frac{1-\widehat{A^L}(z)}{1-z}$$
$$\sigma_z^H = \mu^H - \lambda^H z \frac{1-\widehat{A^H}(z)}{1-z}$$
$$\sigma^L = \sigma_z^L\Big|_{z=1} = \mu_L - \lambda_L \overline{A^L}$$
$$\sigma^H = \sigma_z^H\Big|_{z=1} = \mu_H - \lambda_H \overline{A^H}$$
$$\pi_0^L = \mathit{Pr}\{N^L = 0\}$$

The expression for $\widehat{\Pi^H}(z)$ *is completely symmetric to* (3.8)

Proof: From (3.6), we can write the following relations

$$\widehat{\Pi^L}(z) = \frac{\alpha^L z \widehat{\Pi^H}(z) - \mu^L(1-z)\pi_0^L}{\alpha^L z - \mu^L(1-z) + \lambda^L z(1-\widehat{A^L}(z))} \tag{3.9}$$
$$\widehat{\Pi^H}(z) = \frac{\alpha^H z \widehat{\Pi^L}(z) - \mu^H(1-z)\pi_0^H}{\alpha^H z - \mu^H(1-z) + \lambda^H z(1-\widehat{A^H}(z))} \tag{3.10}$$

where $\pi_0^L = Pr\{N^L = 0\}$ and $\pi_0^H = Pr\{N^H = 0\}$ are unknowns. We can solve (3.9)-(3.10) for $\widehat{\Pi^L}(z)$ to get

$$\widehat{\Pi^L}(z) = \frac{z\alpha^L\mu^H\pi_0^H + z\alpha^H\mu^L\pi_0^L - (1-z)\sigma_z^H\mu^L\pi_0^L}{z\alpha^H\sigma_z^L + z\alpha^L\sigma_z^H - (1-z)\sigma_z^H\sigma_z^L} \tag{3.11}$$

By substituting $z = 1$ in the above equation we get one equation relating π_0^L and π_0^H:

$$\frac{\frac{\mu^L}{\alpha^L}\pi_0^L + \frac{\mu^H}{\alpha^H}\pi_0^H}{\frac{\mu^L}{\alpha^L} + \frac{\mu^H}{\alpha^H}} = 1 - \frac{\frac{\lambda^L\overline{A^L}}{\alpha^L} + \frac{\lambda^H\overline{A^H}}{\alpha^H}}{\frac{\mu^L}{\alpha^L} + \frac{\mu^H}{\alpha^H}} \tag{3.12}$$
$$= 1 - \rho^A \tag{3.13}$$

It turns out that there is a very simple explanation for (3.12) based on *Little's Law*. Imagine stretching the L periods by a factor of μ^L and H periods by μ^H. Now in this transformed time, the service rate is a constant, 1. The switching times are distributed as $exp(\alpha^L/\mu^L)$ and $exp(\alpha^H/\mu^H)$. The time average arrival rate scales by a factor of $\left(\frac{1}{\alpha^L} + \frac{1}{\alpha^H}\right)$ divided by $\left(\frac{\mu^L}{\alpha^L} + \frac{\mu^H}{\alpha^H}\right)$, because the arrivals that were earlier occurring in $\left(\frac{1}{\alpha^L} + \frac{1}{\alpha^H}\right)$ now occur in $\left(\frac{\mu^L}{\alpha^L} + \frac{\mu^H}{\alpha^H}\right)$ amount of time. Using similar reasoning, the idle probability of this system is the expression on the LHS of (3.12). Applying Little's Law at the server in this scaled system gives (3.12).

To complete the solution, we need one more equation relating π_0^L and π_0^H. Using PASTA we know that

$$\frac{\frac{\pi_0^L}{\alpha^L} + \frac{\pi_0^H}{\alpha^H}}{\frac{1}{\alpha^L} + \frac{1}{\alpha^H}} = \pi_0$$

where π_0 is the long term fraction of the time when system is idle. However, π_0 is also unknown and is *not* equal to $1 - \rho^A$. But, by using (3.12), (3.11) simplifies to (3.8). ∎

Corollary 3.6 *When $\widehat{A^L}(z) = \widehat{A^H}(z) = z$ (burst size $\equiv 1$), $\widehat{\Pi^L}(z)$ and $\widehat{\Pi^H}(z)$ become:*

$$\widehat{\Pi^H}(z) = \frac{z\alpha^H(\mu^L - \lambda^L) + z\alpha^L(\mu^H - \lambda^H) - (1-z)(\mu^H - \lambda^H z)(\mu^H \pi_0^H)}{z\alpha^H(\mu^L - \lambda^L z) + z\alpha^L(\mu^H - \lambda^H z) - (1-z)(\mu^H - \lambda^H z)(\mu^L - \lambda^L z)} \tag{3.14}$$

$$\widehat{\Pi^L}(z) = \frac{z\alpha^H(\mu^L - \lambda^L) + z\alpha^L(\mu^H - \lambda^H) - (1-z)(\mu^H - \lambda^H z)(\mu^L \pi_0^L)}{z\alpha^H(\mu^L - \lambda^L z) + z\alpha^L(\mu^H - \lambda^H z) - (1-z)(\mu^H - \lambda^H z)(\mu^L - \lambda^L z)} \tag{3.15}$$

Expressions (3.14)-(3.15) agree with those derived by Yechiali and Naor [34]. In the rest of the paper we will analyze the special case mentioned in Corollary 3.6. Again, we can solve for π_0^L and π_0^H by noticing that the cubic polynomial in the denominators of these expressions has a root in $(0, 1)$ where both numerators must also be 0. Numerically solving for this root, however, does not achieve our goals of getting simple and intuitive expressions.

4 Results - Monotonicity of E[N]

In Section 2, we analyzed how asymptotic behavior of our 2-phase fluctuating load queue compares for the cases $\alpha \to 0$ and $\alpha \to \infty$. We are now interested in determining the behavior in the mid-α range. To be precise, we want to answer questions like: If we fix the values of $\mu^L, \mu^H, \lambda^L, \lambda^H$ and the ratio $\frac{\alpha^L}{\alpha^H}$, but start increasing α ($\alpha = \alpha^H + \alpha^L$) from 0 to ∞, how do $E[N], E[N^L]$ and $E[N^H]$ behave? Are they always between the asymptotes of Section 2? Do they increase or decrease monotonically with α? What is the asymptotic behavior when $\rho^H > 1$?

We will first derive expressions for π_0^L and π_0^H as linear combinations of the asymptotes, and show that these are monotonic, in Section 4.2. Then, in Section 4.3, we will introduce the parameter r, which can be thought of as denoting the *ratio*, or the mixture, of the two asymptotes at the given parameter setting. We will use this parameterization for our ultimate expressions for $E[N]$, recasting the expressions in Section 4.2 in terms of r. These expressions will be used to prove the desired monotonicity results for $E[N]$, first for $\rho^H < 1$, in Section 4.4, and then in Section 4.5 for $\rho^H > 1$.

Throughout, we disregard the singular case $\rho^H = 1$.

4.1 Definitions

To clean up our derivations we start with some definitions.

Definition 4.1 *The normalized switching rate, Δ, is defined as*

$$\Delta = \frac{\alpha^L}{\mu^L} + \frac{\alpha^H}{\mu^H} \tag{4.16}$$

The intuition behind using Δ instead of α is that Δ can be thought of as absorbing the scale of the problem by stretching the L periods by a factor of μ^L and the H phases by a factor of μ^H. This leads to identical queueing behavior (at least at the switching points) but further analysis only depends on ρ^A, ρ^L and ρ^H as will be seen later since the service rates are now a constant 1 (we will assume $\mu^L, \mu^H \neq 0$ [1]). Using Δ, Equations (3.14)-(3.15), can be rewritten as follows:

$$\widehat{\Pi^H}(z) = \frac{z\Delta(1-\rho^A) - (1-z)(1-\rho^L z)\pi_0^H}{z\Delta(1-\rho^A z) - (1-z)(1-\rho^L z)(1-\rho^H z)} \tag{4.17}$$

$$\widehat{\Pi^L}(z) = \frac{z\Delta(1-\rho^A) - (1-z)(1-\rho^H z)\pi_0^L}{z\Delta(1-\rho^A z) - (1-z)(1-\rho^L z)(1-\rho^H z)} \tag{4.18}$$

Definition 4.2 *Let $F(z)$ denote the quadratic in the denominators of* (4.17) *and* (4.18)*:*

$$F(z) = z\Delta(1-\rho^A z) - (1-z)(1-\rho^L z)(1-\rho^H z). \tag{4.19}$$

The roots of $F(z)$ [2] will play an important part in our analysis.

Definition 4.3 *We define χ to be the root of $F(z)$ that lies in the interval $(0,1)$. Further, let*

$$\theta = \left(\frac{1-\rho^A}{1-\rho^A\chi}\right)$$

To convince ourselves that $F(z)$ has exactly *one root in $(0,1)$, note that $F(0) < 0$, $F(1) > 0$ and $F(1/\rho^L) \leq 0$.*

4.2 Monotonicity of the π_0's

The following two theorems establish the desired monotonicity property.

Theorem 4.4 *For $\rho^H \geq 1$, π_0^H, π_0^L and π_0 decrease monotonically as switching rates decrease.*

[1] When either of these is 0, the solution reduces to solving a quadratic.

[2] It is interesting to note that if $Q(x)$ represents the characteristic matrix polynomial obtained during the spectral expansion solution of our fluctuating load queue, then $\frac{\det[Q(x)]}{x-1} = Kx^3F(x^{-1})$ for a constant K.

Theorem 4.5 *for $\rho^H < 1$*

(i) π_0^H decreases monotonically as switching rates decrease.

(ii) π_0^L increases monotonically as switching rates decrease.

(iii) π_0 decreases monotonically as switching rates decrease if and only if $\mu^L > \mu^H$.

Proof: (Theorems 4.4 and 4.5). Our first step is to derive expressions for π_0^L and π_0^H. As mentioned previously, since $F(\chi) = 0$ and $\chi < 1$, the numerators of $\widehat{\Pi^H}(z)$ and $\widehat{\Pi^L}(z)$ must be zero at $z = \chi$ for these z-transforms to converge inside the unit disk $|z| < 1$. Therefore,

$$\begin{aligned}\pi_0^H &= \frac{\chi}{(1-\chi)(1-\rho^L\chi)}\Delta(1-\rho^A)\\ \pi_0^L &= \frac{\chi}{(1-\chi)(1-\rho^H\chi)}\Delta(1-\rho^A)\end{aligned}$$

yielding

$$\pi_0^H(1-\rho^L\chi) = \pi_0^L(1-\rho^H\chi)$$

Combining this with (3.12) and using the definition of θ gives simpler expressions for π_0^H and π_0^L:

$$\begin{aligned}\pi_0^H &= (\rho^A)^{-1}[(1-\rho^A)\rho^H - \theta(\rho^H - \rho^A)]\\ \pi_0^L &= (\rho^A)^{-1}[(1-\rho^A)\rho^L + \theta(\rho^A - \rho^L)]\end{aligned}$$

We can write the above as follows:

$$\pi_0^H = (1-\rho^A)\left[\frac{1-\theta}{1-(1-\rho^A)}\right] + (1-\rho^H)\left[\frac{\theta-(1-\rho^A)}{1-(1-\rho^A)}\right] \tag{4.20}$$

$$\pi_0^L = (1-\rho^A)\left[\frac{1-\theta}{1-(1-\rho^A)}\right] + (1-\rho^L)\left[\frac{\theta-(1-\rho^A)}{1-(1-\rho^A)}\right] \tag{4.21}$$

or equivalently as

$$\pi_0^H = (1-\rho^A)\left[\frac{\omega-\theta}{\omega-(1-\rho^A)}\right] + 0\left[\frac{\theta-(1-\rho^A)}{\omega-(1-\rho^A)}\right] \tag{4.22}$$

$$\pi_0^L = (1-\rho^A)\left[\frac{\omega-\theta}{\omega-(1-\rho^A)}\right] + \left(\frac{(\rho^H-\rho^L)(1-\rho^A)}{\rho^H-\rho^A}\right)\left[\frac{\theta-(1-\rho^A)}{\omega-(1-\rho^A)}\right] \tag{4.23}$$

where $\omega = \frac{\rho^H(1-\rho^A)}{(\rho^H-\rho^A)}$. By observing that $\lim_{\Delta\to\infty}\theta = (1-\rho^A)$ and $\lim_{\Delta\to 0}\theta = \min(1,\omega)$, equations (4.20)-(4.21) can be seen as expressing π_0^H and π_0^L as a convex combination of the limiting cases when $\rho^H < 1$. Similarly, equations (4.22)-(4.23) express π_0^H and π_0^L as a convex combination of the limiting cases when $\rho^H > 1$.

Since χ and hence θ are monotonic in Δ [3], this proves the monotonicity in π_0^H and π_0^L. Since π_0 is a linear combination of π_0^L and π_0^H, the behavior of π_0 will also be monotonic between its limiting values $\lim_{\alpha\to 0} \pi_0$ and $\lim_{\alpha\to\infty} \pi_0$. For $\rho^H \geq 1$, it is easy to see that

$$\lim_{\alpha\to\infty} \pi_0 < \lim_{\alpha\to 0} \pi_0$$

For $\rho^H < 1$ we have

$$\lim_{\alpha\to\infty} \pi_0 > \lim_{\alpha\to 0} \pi_0 \iff \rho^H(\mu^L - \mu^H) > \rho^L(\mu^L - \mu^H)$$

Since $\rho^L \leq \rho^H$, the last part of the theorem follows. ■

4.3 Parameterization in Terms of r

While parameterization in terms of θ is sufficient to show monotonicity of π_0^L and π_0^H, parameterization by θ does *not* yield the simplest expressions for $E[N]$, which is our ultimate aim. We identify a new parameter r which will allow us to express $E[N]$ as a convex combination of two limits, similar to what we did above for π_0^H and π_0^L. We will show how to express $E[N^L]$ and $E[N^H]$ as convex combinations of their limiting curves. Similar expression for $E[N]$ can be obtained using the following fact proved in Section 3.1:

$$E[N] = \frac{\frac{E[N^L]}{\alpha^L} + \frac{E[N^H]}{\alpha^H}}{\frac{1}{\alpha^L} + \frac{1}{\alpha^H}}$$

We start by defining the parameter r for the cases $\rho^H < 1$ and $\rho^H \geq 1$.

Definition 4.6 *For $\rho^H < 1$, we define $r_{(\rho^H<1)}$ as:*

$$r_{(\rho^H<1)} \equiv \frac{\ell_{(\rho^H<1)} - \theta}{\Delta \ell'_{(\rho^H<1)}} \tag{4.24}$$

where

$$\ell_{(\rho^H<1)} \equiv \lim_{\Delta\to 0} \theta = 1 \tag{4.25}$$

$$\ell'_{(\rho^H<1)} \equiv \lim_{\Delta\to 0} \frac{\ell_{(\rho^H<1)} - \theta}{\Delta} = \frac{\rho^A}{(1-\rho^L)(1-\rho^H)} \tag{4.26}$$

$$\tag{4.27}$$

[3] χ is the root of a polynomial that is the sum of a cubic with roots $1, \frac{1}{\rho^L}, \frac{1}{\rho^H}$ and a quadratic that is positive in $\left(0, \frac{1}{\rho^A}\right)$ and increases uniformly with α. Therefore, $F(z)$ increases uniformly in $\left(0, \frac{1}{\rho^A}\right)$ as α increases, hence proving monotonicity of χ.

Definition 4.7 *For $\rho^H > 1$, we define $r_{(\rho^H>1)}$ as:*

$$r_{(\rho^H>1)} \equiv \frac{\ell_{(\rho^H>1)} - \theta}{\Delta \ell'_{(\rho^H>1)}} \tag{4.28}$$

where

$$\ell_{(\rho^H>1)} \equiv \lim_{\Delta\to 0} \theta = \frac{\rho^H(1-\rho^A)}{\rho^H - \rho^A} \tag{4.29}$$

$$\ell'_{(\rho^H>1)} \equiv \lim_{\Delta\to 0} \frac{\ell_{(\rho^H>1)} - \theta}{\Delta} = \frac{\rho^A\rho^H(1-\rho^A)}{(\rho^H-\rho^L)(\rho^H-1)(\rho^H-\rho^A)} \tag{4.30}$$

Whenever unambiguous, we will suppress the subscripts on r. By the way we have defined r, $\lim_{\alpha\to 0} r = 1$ and $\lim_{\alpha\to\infty} r = 0$. To obtain the limits mentioned in the above definitions, we first substitute $z = \left(\frac{\theta-(1-\rho^A)}{\rho^A\theta}\right)$ in (4.19) to obtain the following cubic polynomial:

$$\begin{aligned} g(\theta) &= (\rho^A)^2\Delta[\theta-(1-\rho^A)]\theta \\ &\quad - [\theta-1][\theta(\rho^A-\rho^L)+\rho^L(1-\rho^A)][\theta(\rho^H-\rho^A)-\rho^H(1-\rho^A)] \end{aligned} \tag{4.31}$$

Then, θ is the root of the above polynomial lying in the interval $(1-\rho^A, 1)$.

We now present one of our main results, expressing $E[N^L]$ and $E[N^H]$ as convex combinations of the limiting curves in terms of r.

Theorem 4.8 *a. For $\rho^H < 1$*

$$E[N^H] = \left[\frac{\rho^A}{1-\rho^A}\right] + \left[\frac{\rho^H}{1-\rho^H} - \frac{\rho^A}{1-\rho^A}\right] r_{(\rho^H<1)} \tag{4.32}$$

$$E[N^L] = \left[\frac{\rho^A}{1-\rho^A}\right] + \left[\frac{\rho^L}{1-\rho^L} - \frac{\rho^A}{1-\rho^A}\right] r_{(\rho^H<1)} \tag{4.33}$$

b. For $\rho^H > 1$

$$E[N^H] = \left[\frac{\rho^A}{1-\rho^A}\right] + \left[\frac{\rho^A}{1-\rho^A} + \frac{\rho^H}{\rho^H-1}\right]\left[\frac{1-\ell_{(\rho^H>1)}}{-\ell'_{(\rho^H<1)}\Delta} + \left(\frac{\ell'_{(\rho^H>1)}}{-\ell'_{(\rho^H<1)}}\right) r_{(\rho^H>1)}\right] \tag{4.34}$$

$$E[N^L] = \left[\frac{\rho^A}{1-\rho^A}\right] + \left[\frac{\rho^A}{1-\rho^A} - \frac{\rho^L}{1-\rho^L}\right]\left[\frac{1-\ell_{(\rho^H>1)}}{-\ell'_{(\rho^H<1)}\Delta} + \left(\frac{\ell'_{(\rho^H>1)}}{-\ell'_{(\rho^H<1)}}\right) r_{(\rho^H>1)}\right] \tag{4.35}$$

Proof: We start by differentiating our transforms from (4.17)-(4.18) and set $z = 1$, which results in:

$$E[N^H] = \frac{\rho^A}{1-\rho^A} - \frac{(1-\rho^L)(1-\rho^H-\pi_0^H)}{\Delta(1-\rho^A)} \tag{4.36}$$

$$E[N^L] = \frac{\rho^A}{1-\rho^A} - \frac{(1-\rho^H)(1-\rho^L-\pi_0^L)}{\Delta(1-\rho^A)} \tag{4.37}$$

By substituting π_0^H and π_0^L in terms of θ from (4.20)-(4.21) into (4.36)-(4.37) and collecting the terms dependent on Δ, we get the expressions in (4.32)-(4.33). Similarly, substituting π_0^H and π_0^L from (4.22)-(4.23) into (4.36)-(4.37) results in (4.34)-(4.35). ■

Note that for the case $\rho^H < 1$, we express $E[N^H]$ and $E[N^L]$ in (4.32)-(4.33) as the convex combination of two constants. This is not possible when $\rho^H > 1$ because when $\Delta \to 0$, the mean number of jobs becomes unbounded. However for this case, we can write $E[N^H]$ (and respectively $E[N^L]$) as a convex combination of two asymptotic functions of the form $a + \frac{b}{\Delta}$ which are only separated by a constant.

4.4 Monotonicity of Number in System, $\rho^H < 1$

The following theorem proves monotonicity of $E[N]$ for $\rho^H < 1$:

Theorem 4.9 *For the case $\rho^H < 1$*

(i) *$E[N^L]$ decreases and $E[N^H]$ increases monotonically as the switching rates decrease.*

(ii) *If $s^L < s^H$ then $E[N]$ decreases monotonically as switching rates decrease, otherwise it increases monotonically.*

Proof: For succinctness, we will define the following quantities

$$N_\infty \equiv \frac{\rho^A}{1-\rho^A}$$

$$N_0 \equiv \frac{\frac{1}{\alpha^H}\frac{\rho^H}{1-\rho^H} + \frac{1}{\alpha^L}\frac{\rho^L}{1-\rho^L}}{\frac{1}{\alpha^H} + \frac{1}{\alpha_L}}$$

From (4.32)-(4.33), we can write the expectations of number of jobs in system as

$$E[N^H] = \left[\frac{\rho^A}{1-\rho^A}\right] + \left[\frac{\rho^H}{1-\rho^H} - \frac{\rho^A}{1-\rho^A}\right] r \tag{4.38}$$

$$E[N^L] = \left[\frac{\rho^A}{1-\rho^A}\right] + \left[\frac{\rho^L}{1-\rho^L} - \frac{\rho^A}{1-\rho^A}\right] r \tag{4.39}$$

$$E[N] = N_\infty + [N_0 - N_\infty]\, r \tag{4.40}$$

From the limits proved in (4.25)-(4.26), we know that for the case $\rho^H < 1$, $\lim_{\Delta\to\infty} r = 0$ and $\lim_{\Delta\to 0} r = 1$. We will now show that r is monotonic in Δ. This will imply that $E[N^H]$ increases and $E[N^L]$ decreases monotonically as switching rates decrease and $E[N]$ will increase or decrease depending on whether s^L is larger or smaller than s^H, respectively.

By substituting $\theta = 1 - r\Delta\ell'_{(\rho^H<1)}$ in $g(\theta)$ from (4.31), we get the following polynomial relating Δ and r:

$$\begin{aligned} h(\Delta, r) =& r^2\Delta^2(\ell'_{(\rho^H<1)})^2[\rho^A(1-\rho^L)(1-\rho^H) + r(\rho^A-\rho^L)(\rho^H-\rho^A)] \\ & - r\Delta\ell'_{(\rho^H<1)}\rho^A[(1+\rho^A)(1-\rho^L)(1-\rho^H) + r\{(\rho^A-\rho^L)(\rho^H-1) + (1-\rho^L)(\rho^H-\rho^A)\}] \\ & + (\rho^A)^2(1-\rho^L)(1-\rho^H)(1-r) \end{aligned} \tag{4.41}$$

The above may be viewed as a cubic for r in terms of Δ, or, alternatively, as a quadratic for Δ in terms of r. Therefore, for any r there can be at most two values of Δ. Since r is a continuous function of Δ, $h(\Delta, r)$ must cross $r = c$ with $c \in (0, 1)$ an odd number of times and with $c \in (-\infty, 0) \cup (1, \infty)$ an even number of times. For $r > 1$, the product of the two roots of the above quadratic is negative and hence does not have two positive roots. The case $r < 0$ cannot arise because as mentioned earlier $\chi \in (0, 1)$ and hence $\theta = \left(\frac{1-\rho^A}{1-\rho^A\chi}\right) \in (1-\rho^A, 1)$. Therefore r decreases monotonically from 1 to 0 as Δ increases. ■

4.5 Monotonicity of Number in System, $\rho^H > 1$

Complementary to Theorem 4.9, the following Theorem proves monotonicity of $E[N]$ for $\rho^H > 1$:

Theorem 4.10 *When $\rho^H > 1$, $E[N], E[N^L]$ and $E[N^H]$ increase monotonically as the switching rates decrease.*

Proof: From (4.34)-(4.35), we can write the expression for $E[N]$ as

$$E[N] = N_\infty + [N_\infty - N_0]\left(\frac{1-\ell_{(\rho^H>1)}}{-\ell'_{(\rho^H<1)}\Delta} + \frac{\ell'_{(\rho^H>1)}}{\ell'_{(\rho^H<1)}}r\right) \tag{4.42}$$

From the limits proved in (4.29)-(4.30), we know that for $\rho^H > 1$, $\lim_{\Delta\to\infty} r = 0$ and $\lim_{\Delta\to 0} r = 1$. We will now show that r is monotonic in Δ. From equations (4.34)-(4.35) and (4.42), this will imply that $E[N^H]$,$E[N^L]$ and $E[N]$ all increase monotonically as switching rates decrease. Moreover $E[N]$ is bounded between two curves which are separated by the constant $\mathcal{C}$ given by,

$$\mathcal{C} = \left[\frac{\rho^A}{1-\rho^A} + \frac{\frac{1}{\alpha^H}\frac{\rho^H}{\rho^H-1} - \frac{1}{\alpha^L}\frac{\rho^L}{1-\rho^L}}{\frac{1}{\alpha^H} + \frac{1}{\alpha_L}}\right]\left(\frac{\ell'_{(\rho^H>1)}}{-\ell'_{(\rho^H<1)}}\right)$$

The proof of monotonicity of r for this case will be along the same lines as the $\rho^H < 1$ case. By substituting $\theta = \ell_{(\rho^H>1)} - r\Delta\ell'_{(\rho^H>1)}$ in (4.31), we get the following polynomial relating Δ and r:

$$\begin{aligned} h(\Delta, r) =& r^2(\ell'_{(\rho^H>1)})^2\Delta^2\left[(\rho^A)^2 + r\ell'_{(\rho^H>1)}(\rho^A-\rho^L)(\rho^H-\rho^A)\right] \\ &- r\ell'_{(\rho^H>1)}\Delta\rho^A\left[\frac{\rho^A(1-\rho^A)(\rho^H+\rho^A)}{\rho^H-\rho^A} + r\ell'_{(\rho^H>1)}\left((1-\rho^A)(\rho^H-\rho^L)-(\rho^H-1)(\rho^A-\rho^L)\right)\right] \\ &+ \rho^A\rho^H\left(\frac{\rho^A(1-\rho^A)}{\rho^H-\rho^A}\right)^2(1-r) \end{aligned} \tag{4.43}$$

As before, the case $r < 0$ cannot arise and for $r > 1$ the product of the roots is negative. Following the same arguments as in Section 4.4, we have the desired results. ■

5 Results - Simple Approximations

Having established the monotonicity property of $E[N]$ with respect to Δ (and hence α), we now turn to the question of obtaining tight approximations for the $E[N]$ versus Δ curve that are simple and can be easily analyzed. Evaluating these approximations yields insights into the behavior of $E[N]$ versus the system primitives, in particular α. In Sections 4.4 and 4.5, we expressed $E[N]$ as a function of r. To recapitulate, for $\rho^H < 1$ we have from (4.40)

$$E[N] = a + br_{(\rho^H<1)}$$

for some constants a, b and for $\rho^H > 1$ we have from (4.42)

$$E[N] = a' + b'r_{(\rho^H>1)} + \frac{c'}{\Delta}$$

for some different constants a', b', c'. The aim of this section is to get simple approximations for r. We handle the cases $\rho^H < 1$ and $\rho^H > 1$ together by defining the following quantities.

Definition 5.1

$$u \stackrel{def}{=} \min(1, \rho^H), \qquad v \stackrel{def}{=} \max(1, \rho^H)$$

For succinctness, we also define the following constants

$$\begin{aligned} c_1 &= (v-u)(v-\rho^L) \\ c_2 &= (u-\rho^A)(v-\rho^L) - (v-u)(\rho^A-\rho^L) \end{aligned}$$

We first derive r^*, an approximation for r, shown in equation (5.44). The r^* approximation is highly accurate under all values of Δ (and hence α), and yet it is a closed-form expression, which does not require the solution of a cubic.

In Section 5.1 we go further by finding simpler approximations for r^* in the case where Δ is "low" and the case where Δ is "high" separately (we make low and high precise in the coming sections). Although these approximations $r^*_{\Delta:low}$ (equation (5.48)) and $r^*_{\Delta:high}$ (equation (5.47)) are intended to work only for "low" and "high" Δ, we will find (see Figure 5.4) that using just these two expressions gives us an excellent sense of the shape of the $E[N]$ curve as a function of Δ (and hence α). We also give another approximation, $r^*_{\Delta:med}$ (equation (5.49)), for the special case: $\rho^H \approx 1$.

Claim 5.2 *The r versus Δ curve is well-approximated by the r^* versus Δ curve where,*

$$r^* = \frac{2c_1}{c_1 + (v+\rho^A)\Delta + \sqrt{c_1^2 + 2\Delta((v+\rho^A)c_1 + 2vc_2) + (v-\rho^A)^2\Delta^2}} \tag{5.44}$$

Although, we don't have a formal proof of the above claim we will provide arguments in support of the same. The $r^3\Delta^2$ terms in (4.41) and (4.43) go to 0 as $\Delta \to 0$. Also as $\Delta \to \infty$, $r\Delta$ approaches a constant but $r^3\Delta^2$ again goes to 0. Therefore, by neglecting this term in $h(\Delta, r)$, we get the following quadratic equation in r,

$$\mathtt{h}^*(\Delta, r) = r^2 v\left[\Delta^2\rho^A - \Delta c_2\right] - rc_1\left[\Delta(v+\rho^A) + c_1\right] + c_1^2 \tag{5.45}$$

where we have used u and v to combine (4.41) and (4.43). The polynomial $\mathtt{h}^*(\Delta, r)$ gives a very good approximation to $h(\Delta, r)$ around the root of interest. The approximation r^* is obtained by taking the root of $\mathtt{h}^*(\Delta, r)$

$$\begin{aligned} r^* &= c_1\frac{\Delta(v+\rho^A) + c_1 - \sqrt{(\Delta(v+\rho^A)+c_1)^2 - 4v(\Delta^2\rho^A - \Delta c_2)}}{2v(\Delta^2\rho^A - \Delta c_2)} \\ &= \frac{2c_1}{c_1 + (v+\rho^A)\Delta + \sqrt{c_1^2 + 2\Delta((v+\rho^A)c_1 + 2vc_2) + (v-\rho^A)^2\Delta^2}} \end{aligned} \tag{5.46}$$

The sign of the discriminant in (5.46) has to be negative because

1. If the coefficient of r^2 is negative then the product of the roots is negative and minus sign will give the positive root.

2. If the coefficient of r^2 is positive then both roots are positive and minus sign will give the smaller of the roots.

5.1 Simpler Approximations

In this section, we start with the expression for r^* and simplify it further by looking at two different Δ regimes, high and low.

Case: High Δ

Claim 5.3 *When* $\Delta \gg 2\left|\frac{(v+\rho^A)c_1+2vc_2}{(v-\rho^A)^2}\right| \equiv t_h$, $r^*_{\Delta:high}$ *approximates* r *where*

$$r^*{}_{\Delta:high} = \frac{c_1/v}{\Delta + \frac{c_1+c_2}{(v-\rho^A)}} \tag{5.47}$$

Proof: Starting from (5.44) and making appropriate approximations, we get,

$$\begin{aligned}
r^* &= \frac{2\left(\frac{c_1}{\Delta}\right)}{(v+\rho^A)+\frac{c_1}{\Delta}+\sqrt{(v-\rho^A)^2+\left(\frac{c_1}{\Delta}\right)^2+2\left(\frac{(v+\rho^A)c_1+2vc_2}{\Delta}\right)}} \\
&= \frac{2\left(\frac{c_1}{\Delta}\right)}{(v+\rho^A)+\frac{c_1}{\Delta}+(v-\rho^A)\sqrt{1+\left(\frac{c_1}{\Delta(v-\rho^A)}\right)^2+2\left(\frac{(v+\rho^A)c_1+2vc_2}{\Delta(v-\rho^A)^2}\right)}} \\
&\approx \frac{2\left(\frac{c_1}{\Delta}\right)}{(v+\rho^A)+\frac{c_1}{\Delta}+(v-\rho^A)\left\{1+\frac{(v+\rho^A)c_1+2vc_2}{\Delta(v-\rho^A)^2}+\frac{1}{2}\left(\frac{c_1}{\Delta(v-\rho^A)}\right)^2\right\}} \\
&\approx \frac{2\left(\frac{c_1}{\Delta}\right)}{(v+\rho^A)+\frac{c_1}{\Delta}+\left\{(v-\rho^A)+\left(\frac{(v+\rho^A)c_1+2vc_2}{\Delta(v-\rho^A)}\right)\right\}} \\
&= \frac{\left(\frac{c_1}{\Delta}\right)}{v+v\left(\frac{c_1+c_2}{\Delta(v-\rho^A)}\right)} \\
&= \frac{c_1/v}{\Delta+\frac{c_1+c_2}{(v-\rho^A)}}
\end{aligned}$$

■

Case: Low Δ

Claim 5.4 *When* $\Delta \ll 2\left|\frac{c_1^2}{2(v+\rho^A)c_1+4vc_2}\right| \equiv t_l$, $r^*_{\Delta:low}$ *approximates* r *where*

$$r^*{}_{\Delta:low} = \frac{1}{1+\frac{(c_1(v+\rho^A)+vc_2)\Delta}{c_1^2}} \tag{5.48}$$

Proof: Starting from (5.44) and making appropriate approximations, we get,

$$\begin{aligned}
r^* &= \frac{2c_1}{c_1 + (v+\rho^A)\Delta + \sqrt{c_1^2 + ((v-\rho^A)\Delta)^2 + (2(v+\rho^A)c_1 + 4vc_2)\Delta}} \\
&= \frac{2c_1}{c_1 + (v+\rho^A)\Delta + c_1\sqrt{1 + \left(\frac{(v-\rho^A)\Delta}{c_1}\right)^2 + \frac{(2(v+\rho^A)c_1+4vc_2)\Delta}{c_1^2}}} \\
&\approx \frac{2c_1}{c_1 + (v+\rho^A)\Delta + c_1\left\{1 + \frac{1}{2}\left(\frac{(v-\rho^A)\Delta}{c_1}\right)^2 + \frac{((v+\rho^A)c_1+2vc_2)\Delta}{c_1^2}\right\}} \\
&= \frac{2c_1}{c_1 + (v+\rho^A)\Delta + \left\{c_1 + \frac{1}{2}\left(\frac{(v-\rho^A)^2\Delta^2}{c_1}\right) + \frac{((v+\rho^A)c_1+2vc_2)\Delta}{c_1}\right\}} \\
&= \frac{c_1}{c_1 + \frac{(c_1(v+\rho^A)+vc_2)\Delta}{c_1} + \frac{1}{4}\left(\frac{(v-\rho^A)^2\Delta^2}{v_1}\right)} \\
&= \frac{1}{1 + \frac{(c_1(v+\rho^A)+vc_2)\Delta}{c_1^2}}
\end{aligned}$$

■

We defined the "high" Δ regime as $\Delta \gg t_h$ and the "low" Δ regime as $\Delta \ll t_l$. We will now provide very simple bounds on these thresholds. First observe that

$$(v-u)(v+\rho^L) < \frac{(v+\rho^A)c_1 + 2vc_2}{v-\rho^A} < (v-\rho^L)(1+\rho^H)$$

Therefore,

$$\begin{aligned}
t_h &= 2\left|\frac{(v+\rho^A)c_1 + 2vc_2}{(v-\rho^A)^2}\right| \\
&< 2\frac{(v-\rho^L)(1+\rho^H)}{(v-\rho^A)} < \frac{2v(1+\rho^H)}{(v-\rho^A)}
\end{aligned}$$

and

$$\begin{aligned}
t_l &= \left|\frac{c_1^2}{2(v+\rho^A)c_1 + 4vc_2}\right| \\
&> \frac{(v-\rho^L)(1-\rho^H)^2}{2(v-\rho^A)(1+\rho^H)} > \frac{(1-\rho^H)^2}{2(1+\rho^H)}
\end{aligned}$$

Another salient question concerns the size of the area between the thresholds for the two regimes;

how wide $\frac{t_h}{t_l}$ is:

$$\begin{aligned}\frac{t_h}{t_l} &= \left(2\frac{(v+\rho^A)c_1+2vc_2}{(v-\rho^A)c_1}\right)^2\\ &< \left(2\frac{(v-\rho^L)(1+\rho^H)(v-\rho^A)}{(v-\rho^A)(v-\rho^L)(v-u)}\right)^2\\ &= \left(2\frac{(1+\rho^H)}{(v-u)}\right)^2\end{aligned}$$

Therefore as $\rho^H \to 1$, this gap increases (Figure 5.4 (a)-(d)). We handle this special case next.

Case: Intermediate Δ, $\rho^H \approx 1$
As we have noted above, when $\rho^H \to 1$, the gap where neither $r^*_{\Delta:low}$ nor $r^*_{\Delta:high}$ approximation is tight increases. The reason this happens is that the range of switching rates where the Δ term dominates the constant c_1^2 term and the Δ^2 term in the radical of Equation (5.44) increases. Therefore, for this case we give the following approximation, $r^*_{\Delta:med}$, obtained from (5.44) by just keeping the Δ term of the radical:

$$r^*_{\Delta:med} = \frac{2c_1}{c_1+(v+\rho^A)\Delta+\sqrt{2\Delta(c_1(v+\rho^A)+2vc_2)}} \tag{5.49}$$

The $r^*_{\Delta:med}$ approximation is illustrated in Figure 5.4(e)-(f). This approximation supplements $r^*_{\Delta:low}$ and $r^*_{\Delta:high}$ and depending on the switching rates and ρ^H, one should be chosen appropriately for observing the functional behavior.

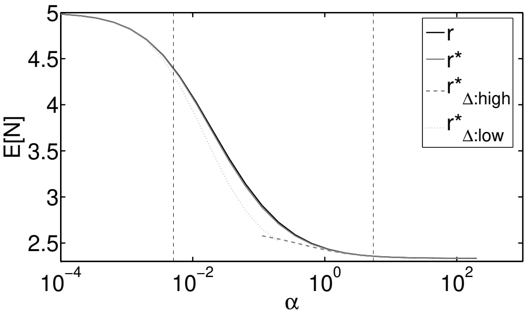
5.2 Fluid Limit for $\rho^H > 1$ case

When the arrival pattern causes transient overload during the H phase, fluid limit of the queue length process allows us to get a simple approximation for the queue length distribution. This approximation becomes tighter when $\alpha \ll \{\mu, \lambda\}$.

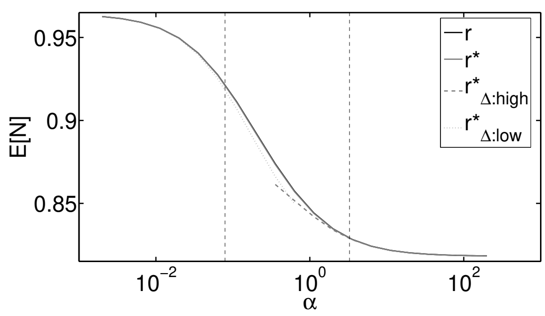
We will denote the stochastic process denoting the state of the environment as $E(t)$. In our case $E(t) \in \{H, L\}$. We define a fluid process $Y(t)$ in this reference environment process by the following differential equations:

$$\frac{dY(t)}{dt} = \begin{cases} r^{E(t)} & \text{if } Y(t) > 0 \\ \left(r^{E(t)}\right)^+ & \text{if } Y(t) = 0 \end{cases}$$

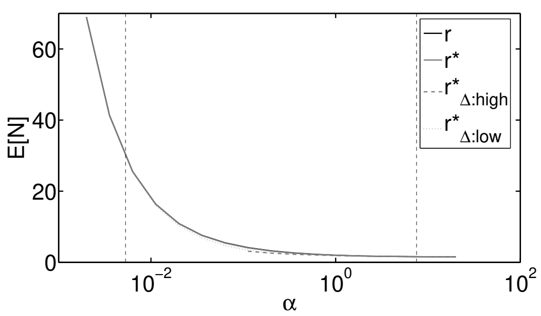
where $r^H = -s^H$ and $r^L = -s^L$ denote the fluid flow rates in that environment state. The fluid scaling theorem [8] states:



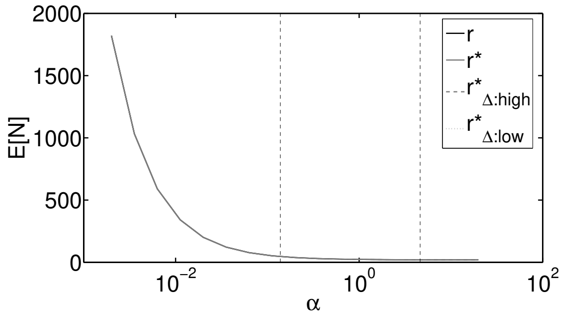
(a) $\rho^H = 0.9$, $\rho^L = 0.5$



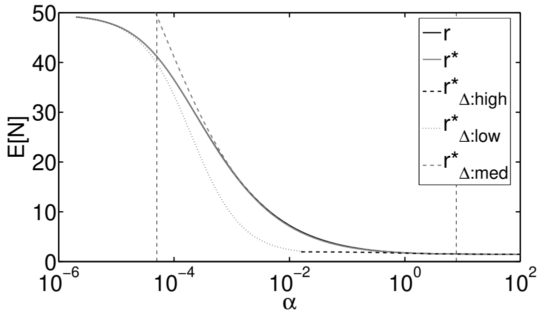
(b) $\rho^H = 0.6$, $\rho^L = 0.3$



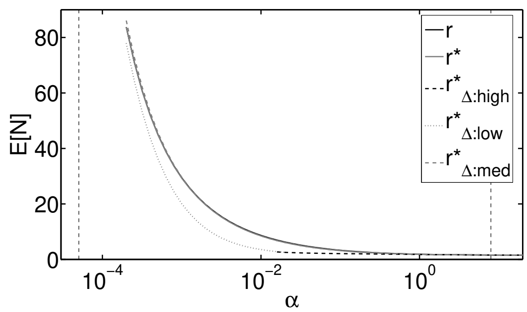
(c) $\rho^H = 1.1$, $\rho^L = 0.1$



(d) $\rho^H = 1.4$, $\rho^L = 0.5$



(e) $\rho^H = 0.99$, $\rho^L = 0.2$



(f) $\rho^H = 1.01$, $\rho^L = 0.2$

Figure 5.4: Illustration of E[N] as a function of α $(= \alpha^H + \alpha^L)$, using the exact r; our closed-form approximation r^*; and our very simple approximations $r^*_{\Delta:low}$ and $r^*_{\Delta:high}$. The top row shows examples where $\rho^H < 1$. The middle row shows examples where $\rho^H > 1$. The bottom row illustrates the approximation $r^*_{\Delta:med}$ when $\rho^H \approx 1$. The vertical lines in each plot indicate the thresholds for the low Δ and high Δ regimes. In all cases $\mu^H = \mu^L = 1$ and $\alpha^H = \alpha^L$.

Theorem 5.5 *Let N_ϵ be the queue length process of the fluctuating load queue run in environment* *process* $E_\epsilon \equiv \{E(\epsilon t)\ :\ t \geq 0\}$. *Let* $E_\epsilon(\cdot/\epsilon) \rightarrow E(\cdot)$ *in* $D[0,\infty)$ *w.p.* 1 *as* $\epsilon \downarrow 0$. *If* $\epsilon N_\epsilon(0) \rightarrow y$ *in* $\mathbb{R}$ *w.p.* 1 *as* $\epsilon \downarrow 0$, *then*

$$\epsilon N_\epsilon(\cdot/\epsilon) \rightarrow Y(\cdot) \textit{ in } D[0,\infty) \textit{ w.p. } 1 \textit{ as } \epsilon \downarrow 0, \tag{5.50}$$

where Y *is the stochastic fluid process with environment process* E*, deterministic flowrates* $r^E = -s^e$ *(*$E \in \{H, L\}$*) and initial content* $Y(0) = y$*.*

Since we are interested in the stationary distribution, we can interpret this theorem in the following simpler way,

Theorem 5.6 *Let* N_ϵ *denote the stationary distribution of the fluctuating load queue with switching rates* $\epsilon\alpha^H$ *and* $\epsilon\alpha^L$*. Then*

$$\epsilon N_\epsilon \xrightarrow{d} Y \textit{ in } \mathbb{R} \textit{ as } \epsilon \downarrow 0 \tag{5.51}$$

where Y *is the stationary distribution of the stochastic fluid process with switching rates* α^H *and* α^L *and deterministic flow rates* $r^E = -s^E$ *(*$E \in \{H, L\}$*).*

The following theorem gives the distribution of Y.

Theorem 5.7 *Let* Y *be the stationary stochastic fluid process defined in Theorem 5.6. Let* Y^H *and* Y^L *be the random variables for the time average fluid levels during* H *and* L *phases, respectively. The distributions of* Y*,* Y^L *and* Y^H *are given by*

$$Y \sim \begin{cases} 0 & w.p. \ \left(1 + \frac{\alpha^L/s^L}{\alpha^H/s^H}\right)\frac{\alpha^H}{\alpha^L+\alpha^H} \\ exp\left(-\frac{\alpha^H}{s^H} - \frac{\alpha^L}{s^L}\right) & w.p. \ \left(1 - \frac{s^H}{s^L}\right)\frac{\alpha^L}{\alpha^L+\alpha^H} \end{cases} \tag{5.52}$$

$$Y^H \sim exp\left(-\frac{\alpha^H}{s^H} - \frac{\alpha^L}{s^L}\right) \tag{5.53}$$

$$Y^L \sim \begin{cases} 0 & w.p. \ \left(1 + \frac{\alpha^L/s^L}{\alpha^H/s^H}\right) \\ exp\left(-\frac{\alpha^H}{s^H} - \frac{\alpha^L}{s^L}\right) & w.p. \ \left(-\frac{\alpha^L/s^L}{\alpha^H/s^H}\right) \end{cases} \tag{5.54}$$

Proof: We begin by noting that $Pr[Y^H = 0] = 0$. Let $p_0^L = Pr[Y^L = 0]$. Also let $f^H(y)$ and $f^L(y)$ be the density functions of Y^H and Y^L, respectively, with support on $(0, +\infty]$. Further, let $\widetilde{Y_+^H}(s)$ and $\widetilde{Y_+^L}(s)$ be the Laplace-Stieltjes transforms of f^H and f^L, respectively. Then,

$$\begin{aligned} p_0^L &= \int_{y=0}^{\infty} f^H(y) e^{-\left(\frac{\alpha^L y}{\mu^L - \lambda^L}\right)} dy \\ &= \widetilde{Y^H}\left(\frac{\alpha^L}{\mu^L - \lambda^L}\right) \end{aligned}$$

We also have the following expressions for the probability density functions (using Theorem 3.3)

$$f^H(y) = \left[p_0^L \alpha^H e^{-\left(\frac{\alpha^H y}{\lambda^H - \mu^H}\right)} + \int_{x=0^+}^{y} f^L(x) \alpha^H e^{-\left(\frac{\alpha^H (y-x)}{\lambda^H - \mu^H}\right)} dx \right] \frac{1}{\lambda^H - \mu^H}$$

$$f^L(y) = \left[\int_{x=y}^{\infty} f^H(x) \alpha^L e^{-\left(\frac{\alpha^L (x-y)}{\mu^L - \lambda^L}\right)} dx \right] \frac{1}{\mu^L - \lambda^L}$$

which gives the following relationships between $\widetilde{Y_+^H}(s)$ and $\widetilde{Y_+^L}(s)$:

$$\widetilde{Y_+^H}(s) = \left[\widetilde{Y_+^H}\left(\frac{\alpha^L}{\mu^L - \lambda^L}\right) + \widetilde{Y_+^L}(s) \right] \left(\frac{\frac{\alpha^H}{\lambda^H - \mu^H}}{\frac{\alpha^H}{\lambda^H - \mu^H} + s} \right)$$

$$\widetilde{Y_+^L}(s) = \left[\widetilde{Y_+^H}(s) - \widetilde{Y_+^H}\left(\frac{\alpha^L}{\mu^L - \lambda^L}\right) \right] \left(\frac{\frac{\alpha^L}{\mu^L - \lambda^L}}{\frac{\alpha^L}{\mu^L - \lambda^L} - s} \right)$$

The above equations solve to

$$\widetilde{Y^H}(s) = \frac{\frac{\alpha^H}{\lambda^H - \mu^H} - \frac{\alpha^L}{\mu^L - \lambda^L}}{\frac{\alpha^H}{\lambda^H - \mu^H} - \frac{\alpha^L}{\mu^L - \lambda^L} + s} \tag{5.55}$$

$$\widetilde{Y^L}(s) = \left(1 - \frac{\alpha^L/(\mu^L - \lambda^L)}{\alpha^H/(\lambda^H - \mu^H)}\right) + \left(\frac{\alpha^L/(\mu^L - \lambda^L)}{\alpha^H/(\lambda^H - \mu^H)}\right) \widetilde{Y^H}(s) \tag{5.56}$$

The distribution of Y^H is of an exponential random variable with rate $\left(\frac{\alpha^H}{\lambda^H - \mu^H} - \frac{\alpha^L}{\mu^L - \lambda^L}\right)$. Y^L is the combination of an atom at 0 and an exponential distribution with the same rate. Taking a weighted average of Y^H and Y^L gives the distribution of Y. ■

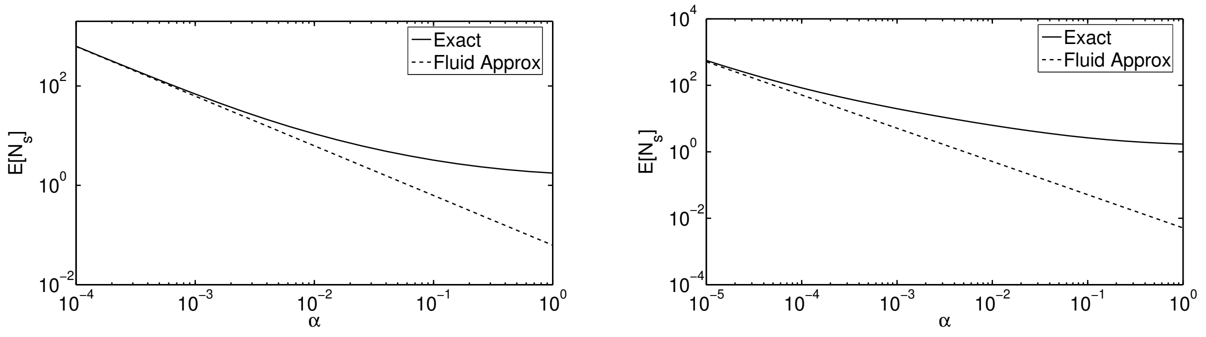
Corollary 5.8 *The mean fluid level is given by*

$$E[Y] = \left(\frac{s^H - s^L}{\alpha^H + \alpha^L}\right) \left(\frac{\alpha^L / s^L}{\alpha^H / s^H + \alpha^L / s^L}\right) \tag{5.57}$$

Figure 5.5 shows a comparison of the $E[N]$ vs. α curve with that obtained using the fluid limit of the system. The $E[N]$ axis is also plotted on $\log$ scale to illustrate the gap at higher values of switching rates.

6 Behavioral insights into the fluctuating load queue

Having established some fundamental properties of the fluctuating load queue, we further examine the behavior of this system. One question that we ask is, if we double $\lambda^{\{L,H\}}$ and $\mu^{\{L,H\}}$ but



(a) $\lambda^H = 1.1$, $\lambda^L = 0.1$ (b) $\lambda^H = 1.01$, $\lambda^L = 0.2$

Figure 5.5: Comparison of fluid approximation with actual mean queue length. In all cases $\mu^H = \mu^L = 1$ and $\alpha^H = \alpha^L$.

keep the switching rates the same, what happens to the mean response time? Does it decrease by twice (as in a $GI/GI/1$ system) or can it decrease by a smaller or a larger factor? If so, by how much? We address this question in Section 6.1. Another important question is, when is the effect of slowing the switching rate most felt on mean response time (Section 6.2)? We conclude this section with an application. In Section 6.3 we consider the question of how to optimally split a given average service capacity given a traffic arrival pattern (α, λ's).

6.1 Effect of scaling the arrival and service rates

It is well known that for a $GI/GI/1$ system, scaling both the arrival rates and service rates by a factor of k leads to identical queueing behavior as the original system but mean response times are scaled by a factor $\frac{1}{k}$. This is because the new system can be seen as a scaled version of the original system where time is sped up by a factor of k. For the same reason, if in our fluctuating load queue we scale the arrival, service and switching rates by a factor of k, the mean response time of the new system will be $\frac{1}{k}$ times that of the original system, as in a $GI/GI/1$ queue. But what happens if we only scale the arrival and service rates?
Let us represent the original system (with arrival rates $\lambda^{\{L,H\}}$, service rates $\mu^{\{L,H\}}$ and switching rates $\alpha^{\{L,H\}}$) by system A. Let system B be the same as system A except that the switching rates are halved, and let system C represent the system we want to compare with system A, that is with arrival rates $2\lambda^{\{L,H\}}$, service rates $2\mu^{\{L,H\}}$ and switching rates $\alpha^{\{L,H\}}$. Clearly, mean response time of system C, $E[T_C]$, is half the mean response time of system B, $E[T_B]$ using time-scaling argument. The problem now is to compare the response times of system A ($E[T_A]$) and system B. We consider several cases:
Case $s^L = s^H$: For this case, as noted before, the mean response time is invariant to the switching rates and therefore system A and B have same mean response times. Consequently, $E[N_C] =$

$\frac{1}{2}E[N_B] = \frac{1}{2}E[N_A]$.
Case $s^L > s^H$, $\rho^H < 1$: Here, moving from A to B leads to an increase in mean response times. However, we will show that the mean response time (equivalently, mean queue length) does not increase by more than twice in the process. To show this, it suffices to prove that $\left|\frac{d\log E[N]}{d\log\Delta}\right|$ is bounded by 1. But $E[N] = a + br$ for some positive constants a and b and hence,

$$\begin{aligned}\left|\frac{d\log E[N]}{d\log\Delta}\right| &= \left|\frac{d\log(a+br)}{d\log\Delta}\right| \\ &= \left|\frac{br}{a+br}\right| \cdot \left|\frac{d\log r}{d\log\Delta}\right| \\ &< \left|\frac{d\log r}{d\log\Delta}\right|\end{aligned}$$

We will find it easier to prove $\frac{d\log\Delta}{d\log r} \leq -1$. This would imply $-1 \leq \frac{d\log r}{d\log\Delta} \leq 0$ and hence $\left|\frac{d\log E[N]}{d\log\Delta}\right| \leq \left|\frac{d\log r}{d\log\Delta}\right| \leq 1$.

Lemma 6.1

$$\frac{d\log\Delta}{d\log r} \leq -1$$

Proof: See Appendix A.1. ■

Therefore the mean response time of system B is within a factor of 2 of response time of system A. Thus, $\frac{1}{2}E[T_A] \leq E[T_C] \leq E[T_A]$.
Case $s^L < s^H$, $\rho^H < 1$: The mean response time of system B is now less than that of system A. In this case we can express the mean number of jobs in system as $E[N] = a + b(1-r)$ for some positive constants a and b and hence,

$$\begin{aligned}\left|\frac{d\log E[N]}{d\log\Delta}\right| &= \left|\frac{d\log(a+b(1-r))}{d\log\Delta}\right| \\ &= \left|\frac{d\log(a+b(1-r))}{d\log r}\right| \cdot \left|\frac{d\log r}{d\log\Delta}\right| \\ &= \frac{br}{a+b(1-r)}\left|\frac{d\log r}{d\log\Delta}\right|\end{aligned}$$

As a consequence of the proof of Lemma 6.1, we actually obtain the stronger inequality $-(1-r) \leq \frac{d \log r}{d \log \Delta} \leq 0$. Therefore,

$$
\begin{aligned}
\left|\frac{d \log E[N]}{d \log \Delta}\right| &= \frac{br}{a+b(1-r)}\left|\frac{d \log r}{d \log \Delta}\right| \\
&\leq \frac{br(1-r)}{a+b(1-r)} \\
&\leq r \\
&\leq 1
\end{aligned}
$$

Therefore, $\frac{1}{2}E[N_A] \leq E[N_B] \leq E[N_A]$ and hence $\frac{1}{4}E[N_A] \leq E[N_C] \leq \frac{1}{2}E[N_A]$.

Case $\rho^H > 1$**:** The expression for the mean response time of system A is $E[N_A] = a' + b'r + \frac{c'}{\Delta}$ for some positive constants a', b', c'. To prove $\left|\frac{d \log E[N]}{d \log \Delta}\right| < 1$, we begin with the following lemma:

Lemma 6.2 *For* $f(x), g(x) > 0$

$$
\frac{d \log(f(x)+g(x))}{dx} \leq \max\left\{\frac{d \log f(x)}{dx}, \frac{d \log g(x)}{dx}\right\}
$$

Proof:

$$
\begin{aligned}
\frac{d \log(f(x)+g(x))}{dx} &= \frac{\partial \log(f(x)+g(x))}{\partial \log f(x)} \frac{d \log f(x)}{x} + \frac{\partial \log(f(x)+g(x))}{\partial \log g(x)} \frac{d \log g(x)}{x} \\
&= \frac{f(x)}{f(x)+g(x)} \frac{d \log f(x)}{dx} + \frac{g(x)}{f(x)+g(x)} \frac{d \log g(x)}{dx} \\
&\leq \max\left\{\frac{d \log f(x)}{dx}, \frac{d \log g(x)}{dx}\right\}
\end{aligned}
$$

■

Corollary 6.3 *For* $f(x), g(x) > 0$

$$
\left|\frac{d \log(f(x)+g(x))}{dx}\right| \leq \max\left\{\left|\frac{d \log f(x)}{dx}\right|, \left|\frac{d \log g(x)}{dx}\right|\right\}
$$

Using Corollary 6.3 to bound $\left|\frac{d \log E[N]}{d \log \Delta}\right|$,

$$\begin{aligned}
\left|\frac{d\log E[N]}{d\log\Delta}\right| &= \left|\frac{d\log\left(a'+b'r+c'/\Delta\right)}{d\log\Delta}\right| \\
&\le \max\left\{\left|\frac{d}{d\log\Delta}\log a'\right|, \left|\frac{d}{d\log\Delta}\log\left(b'r\right)\right|, \left|\frac{d}{d\log\Delta}\log\left(\frac{c'}{\Delta}\right)\right|\right\} \\
&= \max\left\{0, \left|\frac{d\log r}{d\log\Delta}\right|, \left|\frac{d\log\Delta}{d\log\Delta}\right|\right\} \\
&\le \max\left\{1,1\right\} \\
&= 1
\end{aligned}$$

where we have used Lemma 6.1 to bound $\left|\frac{d\log r}{d\log\Delta}\right|$ by 1. Further, asymptotically as $\Delta \to 0$, $E[N_A] \approx \frac{c'}{\Delta}$. Therefore, when the switching rates are very small, $E[N_B] \approx 2E[N_A]$ implying $E[N_C] \approx E[N_A]$. To see why doubling the arrival and service rates has almost no effect on mean response times, observe that during the H phase the queues grow at a rate of $(\lambda^H - \mu^H)$. Doubling the arrival and switching rates causes the queues to build up twice as fast. Although each customer spends half the time at the server in system C, they have to wait for almost twice as many customers as in system A, nullifying the benefit of faster service rates.

6.2 Effect of switching rates on Mean Response Times

We have yet to directly address the question of when does changing the switching rates have a big effect on the mean response times and when does it have almost no effect. We will try and answer this question here.

Case $\rho^H > 1$: As we have mentioned previously, the $E[N]$ vs Δ curve for this case is bounded between two curves of the form $a + \frac{b}{\Delta}$, which indicates that the effect of switching rates keeps increasing as the switching rates become smaller.

Case $\rho^H < 1$: From Figure 5.4 one can see that for this case, there is a certain zone within which changes in Δ affect the mean response times a lot, and beyond which the curve asymptotes. We will use the approximations we have derived in Section 5 to find the boundaries of this zone, noting that $r^*_{\Delta:low}$ (5.48) is a tight approximation at the left boundary and $r^*_{\Delta:high}$ (5.47) is a tight approximation at the right boundary. Since the exact mathematical notion of boundary is very fuzzy, we will take them as the intersection of the tangent at the inflexion point of the $r - \log\Delta$ curve and $r = 1$ or $r = 0$. After going through the calculations, one gets the following expression for the left boundary

$$\Delta_L = \frac{(c_1/e)^2}{c_1(1+\rho^A)+c_2} \tag{6.58}$$

and

$$\Delta_H = \frac{(c_1 + c_2)e^2}{1 - \rho^A} \tag{6.59}$$

for the right boundary. Figure 6.6 shows an example of these boundaries.

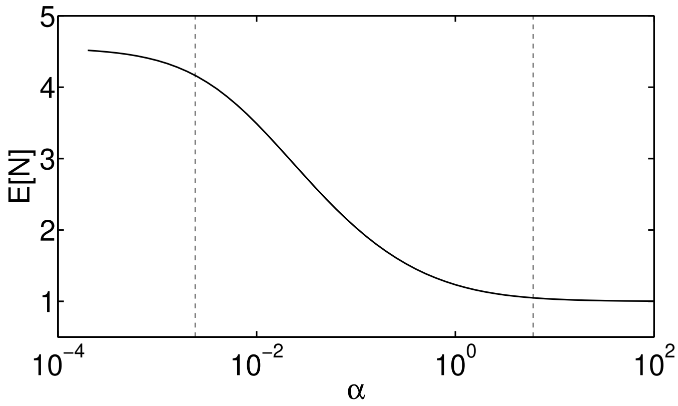


Figure 6.6: Illustration of the calculations in Section 6.2. The vertical lines are the zone boundaries. ($\mu^H = \mu^L = 1$, $\alpha^H = \alpha^L$, $\rho^H = 0.9$, $\rho^L = 0.1$).

6.3 Optimal Capacity Splitting

So far we have concerned ourselves with analysing a system with given service rates μ^L and μ^H. From a system designer's point of view, the question that is of more importance is: Given a traffic arrival pattern $(\lambda^L, \lambda^H, \alpha^L, \alpha^H)$ and a certain average service capacity μ^A, how should it be split over the L and H phases so as to minimise the mean response times? Is load balancing across the L and H phases a good policy?

Let

$$p^H = \frac{\alpha^L}{\alpha^L + \alpha^H}$$
$$p^L = \frac{\alpha^H}{\alpha^L + \alpha^H}$$

Instead of finding the optimal split of service capacity for some setting of α's, we will find that policy which minimises $E[N^{\alpha \to 0}]$. We hope this policy will provide near optimal response times for all scales of switching rates. Formally, the optimization problem for minimizing $E[N^{\alpha\to 0}]$ is:

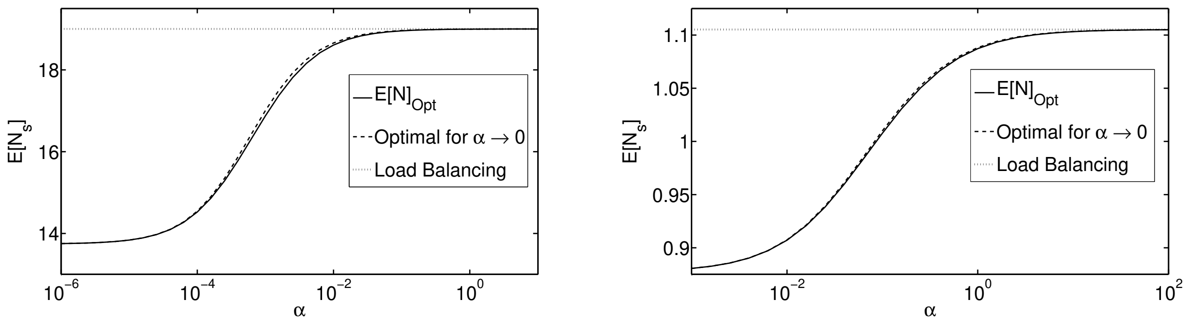
$$
\begin{aligned}
\text{minimize} \quad & p^H \frac{\lambda^H}{\mu^H - \lambda^H} + p^L \frac{\lambda^L}{\mu^L - \lambda^L} \\
\text{over} \quad & \mu^H, \mu^L \\
\text{subject to} \quad & p^H \mu^H + p^L \mu^L \ = \ \mu^A \\
& \mu^H, \mu^L \ > \ 0
\end{aligned}
$$

The solution to this optimization problem is straightforward and is given by,

$$\mu^{H*} = \lambda^H + (\mu^A - \lambda^A)\frac{\sqrt{\lambda^H}}{p^H\sqrt{\lambda^H} + p^L\sqrt{\lambda^L}} \tag{6.60}$$

$$\mu^{L*} = \lambda^L + (\mu^A - \lambda^A)\frac{\sqrt{\lambda^L}}{p^H\sqrt{\lambda^H} + p^L\sqrt{\lambda^L}} \tag{6.61}$$

For this solution, $\frac{\lambda^H}{\mu^{H*}} > \frac{\lambda^L}{\mu^{L*}}$ and $(\mu^{H*} - \lambda^H) > (\mu^{L*} - \lambda^L)$. Also note that load balancing $\left(\frac{\lambda^H}{\mu^H} = \frac{\lambda^L}{\mu^L} = \frac{\lambda^A}{\mu^A}\right)$ is *not* the optimal solution. Figure 6.7 shows how the above capacity provisioning policy performs in comparison with the optimal capacity splitting policy. As can be seen from the figure, minimising $E[N^{\alpha\to 0}]$ gives near-optimal mean response times for all scales of switching rates. An important implication of this fact is that it is not only sometimes good to have variability in the arrival process but in fact it is *desirable*. By splitting capacity intelligently over the high arrival rate and low arrival rate phases, one can get lower mean response times than if the system had a fixed arrival rate.



(a) $\lambda^H = 1.8$, $\lambda^L = 0.1$ (b) $\lambda^H = 0.95$, $\lambda^L = 0.1$

Figure 6.7: Illustration of the optimal $E[N]$ as a function of α ($= \alpha^H + \alpha^L$). The dashed curve represents $E[N]$ using the optimal splitting of service capacity and the solid curve represents the $E[N]$ curve obtained by setting $\mu^H = \mu^{H*}$ and $\mu^L = \mu^{L*}$. In all cases $\mu^A = 1$ and $\alpha^H = \alpha^L$.

7 Results – Stochastic Ordering

Most results in queueing theory describe the experience of an arbitrary arrival to a system. But, in a time-varying system, an arrival may know that she is not "arbitrary"; she may know whether she is arriving into a high load or a low load period. In this case the salient question, as far as the arrival is concerned, is not about an *arbitrary* arrival's experience, but rather about *her* experience (conditional on the type of period into which she arrives).

To explore this question we compare, N^H, N^L, and N^{ρ^A} stochastically, where the last term denotes the number in system seen by an arrival to a *stationary* queue with the same average load, ρ^A, as our time-varying system.

Note that the distribution of future service rates, and thus response time, is *completely determined* by the number in system seen upon arrival and the type of period arrived into. Moreover, if only arrival rates vary (i.e. if service rates are constant), stochastic orderings for number in system immediately translate into stochastic orderings for response times.

Intuition leads one to believe that an arrival into a high load state should see more customers than one arriving into a low load state in expectation, but whether there is a stochastic dominance between these, that is, $N^H \geq_{st} N^L$, is not obvious; we prove this to be true. Furthermore, one might also believe that an arrival during a high load state would see more customers than an arrival into the average system, and that an arrival into the average system would see more customer than an arrival during the low load state, $N^H \geq_{st} N^{\rho^A} \geq_{st} N^L$. Surprisingly, we find that this statement is only partially true: The first inequality holds but the second does not in general. Thus our system exhibits a striking lack of symmetry.

We start with a preliminary result:

Lemma 7.1 *Given an $M/M/1$ queue with load ρ and stationary distribution $\mathcal{X}$, if we start this system with an initial distribution $X(0)$, then*

$$X(t) \geq_{st} X(t+s) \geq_{st} \mathcal{X} \quad \forall s,t \geq 0$$
$$\Longleftrightarrow$$
$$Pr\{X(0)=j\} \geq \rho\, Pr\{X(0)=j-1\}$$

The directions of all the inequalities can be reversed to get the condition for a stochastically increasing system.

Proof: Define a discrete time process Y such that $Y(0) =_{st} X(0)$. For $i > 0$, $Y(i)$ evolves as:

$$Y(i) = \begin{cases} Y(i-1)+1 & w.p.\ \frac{\lambda}{\mu+\lambda} \\ (Y(i-1)-1)^+ & w.p.\ \frac{\mu}{\mu+\lambda} \end{cases}$$

We couple the processes X and Y as follows: since $X(0) =_{st} Y(0)$ we choose the same initial value for these. Set timers according to a Poisson process with rate $(\mu+\lambda)$ and at the i^{th} expiration of the timer (say at t_i) we set $X(t_i) = Y(i)$. Using this coupling,

$$X(t) \geq_{st} X(t+s) \geq_{st} \mathcal{X} \quad \forall s, t \geq 0$$
$$\Longleftrightarrow$$
$$Y(i) \geq_{st} Y(i+k) \geq_{st} \mathcal{X} \quad \forall i, k \geq 0.$$

We first examine the condition for $Y(i+1) \leq_{st} Y(i)$:

$$\text{Pr}\{Y(i+1) \geq j\} \leq \text{Pr}\{Y(i) \geq j\}$$
$$\Longleftrightarrow$$
$$\text{Pr}\{Y(i) \geq j-1\}\tfrac{\lambda}{\lambda+\mu} + \text{Pr}\{Y(i) \geq j+1\}\tfrac{\mu}{\lambda+\mu} \leq \text{Pr}\{Y(i) \geq j\}\left(\tfrac{\lambda}{\lambda+\mu} + \tfrac{\mu}{\lambda+\mu}\right)$$
$$\Longleftrightarrow$$
$$\left[\text{Pr}\{Y(i) \geq j-1\} - \text{Pr}\{Y(i) \geq j\}\right]\tfrac{\lambda}{\lambda+\mu} \leq \left[\text{Pr}\{Y(i) \geq j\} - \text{Pr}\{Y(i) \geq j+1\}\right]\tfrac{\mu}{\lambda+\mu}$$
$$\Longleftrightarrow$$
$$\text{Pr}\{Y(i) = j-1\}\rho \leq \text{Pr}\{Y(i) = j\}$$

Thus $Y(i+1) \leq_{st} Y(i)$ *iff* $\text{Pr}\{Y(i) = j\} \geq \rho\text{Pr}\{Y(i) = j-1\}\ \forall j \geq 1$, as required. Now, if $Y(i+1) \leq_{st} Y(i)\ \forall i \geq 0$ it immediately follows that $Y(i) \geq_{st} \mathcal{X},\ \forall i \geq 0$, because $Y(n)$ converge in distribution to $\mathcal{X}$ as $n \to \infty$. To complete the proof, we need to show that

$$\text{Pr}\{Y(i) = j-1\}\rho \leq \text{Pr}\{Y(i) = j\} \quad \forall j > 0$$

implies

$$\text{Pr}\{Y(i+1) = j-1\}\rho \leq \text{Pr}\{Y(i+1) = j\} \quad \forall j > 0$$

which follows for $j > 1$ since:

$$\begin{aligned}
&\text{Pr}\{Y(i+1) = j\} \\
&= \text{Pr}\{Y(i) = j-1\}\frac{\lambda}{\lambda+\mu} + \text{Pr}\{Y(i) = j+1\}\frac{\mu}{\lambda+\mu} \\
&\geq [\rho\text{Pr}\{Y(i) = j-2\}]\frac{\lambda}{\lambda+\mu} + [\rho\text{Pr}\{Y(i) = j\}]\frac{\mu}{\lambda+\mu} \\
&= \rho\text{Pr}\{Y(i+1) = j-1\}
\end{aligned}$$

For $j = 1$, we have

$$
\begin{aligned}
&\Pr\{Y(i+1) = 1\} \\
&= \Pr\{Y(i) = 0\}\frac{\lambda}{\lambda+\mu} + \Pr\{Y(i) = 2\}\frac{\mu}{\lambda+\mu} \\
&\geq \Pr\{Y(i) = 0\}\frac{\rho\mu}{\lambda+\mu} + [\rho\Pr\{Y(i) = 1\}]\frac{\mu}{\lambda+\mu} \\
&= \rho[\Pr\{Y(i) = 0\} + \Pr\{Y(i) = 1\}]\frac{\mu}{\lambda+\mu} \\
&= \rho\Pr\{Y(i+1) = 0\}
\end{aligned}
$$

For $\rho \geq 1$, the system cannot decrease stochastically because the stationary distribution does not exist. The condition for such a system to be stochastic increasing is the same as that for an $M/M/1$ with $\rho < 1$. ■

Theorem 7.2 *For our alternating load system,*

$$N^H \geq_{st} N^L$$

Proof: We will prove that starting an $M/M/1$ with load ρ^L and initial distribution as N^H satisfies the conditions of Lemma 7.1 and will result in a stochastically decreasing process. Then, since N^L is the random variable for the number of jobs at a time chosen from the distribution $exp(\alpha^L)$, it too will be stochastically smaller than the initial distribution, N^H.

By factoring the polynomials in the numerator and the denominator of (3.14) we can write $\widehat{\Pi^H}(z)$ as:

$$
\begin{aligned}
\widehat{\Pi^H}(z) &= \frac{\lambda^L\mu^H\pi_0^H}{\lambda^L\lambda^H}\frac{(\delta - z)(z-\chi)}{(z-a)(z-b)(z-\chi)} \\
&= \frac{\pi_0^H}{\rho^H}\frac{(\delta - z)}{(z-a)(z-b)}
\end{aligned}
\tag{7.62}
$$

where $0 \leq \chi < 1$, with say $a \leq b$. (As mentioned in Section 3, the denominator has a root χ in $(0, 1)$ and the numerator must also have a root equal to χ for the z-transform to converge in the unit disc $|z| < 1$.) Similarly

$$\widehat{\Pi^L}(z) = \frac{\pi_0^L}{\rho^L}\frac{(\gamma - z)}{(z-a)(z-b)} \tag{7.63}$$

The fact that a, b, χ, δ and γ are all real for $\rho^A < 1$ can be easily verified. Also, $\delta, \gamma > a$ since the z-transform is an increasing function of z, it must become negative via a discontinuity. Using $\widehat{\Pi^L}(0) = \pi_0^L$ and $\widehat{\Pi^H}(0) = \pi_0^H$:

$$\delta = \rho^H ab, \quad \gamma = \rho^L ab \tag{7.64}$$

Evaluating (4.19) for information about the roots a and b, we have $F(1/\rho^H) \geq 0$, $F(1) > 0$, $F(1/\rho^A) \leq 0$, $F(1/\rho^L) \leq 0$. Therefore: $\max\{1, \frac{1}{\rho^H}\} \leq a \leq \frac{1}{\rho^A} \leq \frac{1}{\rho^L} \leq b$. Combining these with (7.64): $1 < a \leq \gamma \leq b \leq \delta$.

Let $p_i^H = Pr\{N^H = i\}$; to derive p_i^H, we will expand (7.62).

$$\begin{aligned}
\widehat{\Pi^H}(z) &= \frac{\pi_0^H(\delta - z)}{\rho^H(b-a)} \left[\frac{1}{a}\left(\frac{1}{1-z/a}\right) - \frac{1}{b}\left(\frac{1}{1-z/b}\right)\right] \\
&= \frac{\pi_0^H}{\rho^H(b-a)} \left[\left\{\left(\frac{\delta}{a}-1\right) - \left(\frac{\delta}{b}-1\right)\right\}\right. \\
&\quad + z\left\{\left(\frac{\delta}{a}-1\right)\frac{1}{a} - \left(\frac{\delta}{b}-1\right)\frac{1}{b}\right\} \\
&\quad \left. + z^2\left\{\left(\frac{\delta}{a}-1\right)\frac{1}{a^2} - \left(\frac{\delta}{b}-1\right)\frac{1}{b^2}\right\} + \cdots\right]
\end{aligned} \tag{7.65}$$

Note that the last representation is what we would obtain by writing out the spectral expansion solution, with $\frac{1}{a}$ and $\frac{1}{b}$ as the two eigenvalues and the probability distribution as the sum of two geometric distributions.

Let $\nu_i^H = \frac{p_{i+1}^H}{p_i^H}$. From (7.65),

$$\begin{aligned}
\nu_i^H &= \frac{\left(\frac{\delta}{a}-1\right)\frac{1}{a^{i+1}} - \left(\frac{\delta}{b}-1\right)\frac{1}{b^{i+1}}}{\left(\frac{\delta}{a}-1\right)\frac{1}{a^{i}} - \left(\frac{\delta}{b}-1\right)\frac{1}{b^{i}}} \\
&= \frac{\zeta u^{i+1} - \eta v^{i+1}}{\zeta u^i - \eta v^i} \\
&= \left(\frac{\zeta u^{i+1} - \eta v^{i+1}}{\zeta u^i - \eta v^i}\right)\left(\frac{\zeta u^{i+1} - \eta v^{i+1}}{\zeta u^{i+1} - \eta v^{i+1}}\right) \\
&= \frac{(\zeta u^{i+1} - \eta v^{i+1})^2 - \zeta\eta u^i v^i(u^2+v^2) + \zeta\eta u^i v^i(u^2+v^2)}{(\zeta u^i - \eta v^i)(\zeta u^{i+1} - \eta v^{i+1})} \\
&= \frac{(\zeta u^i - \eta v^i)(\zeta u^{i+2} - \eta v^{i+2})}{(\zeta u^i - \eta v^i)(\zeta u^{i+1} - \eta v^{i+1})} + \frac{\zeta\eta u^i v^i(u^2+v^2-2uv)}{(\zeta u^i - \eta v^i)(\zeta u^{i+1} - \eta v^{i+1})} \\
&= \nu_{i+1}^H + \frac{\zeta\eta u^i v^i(u-v)^2}{(\zeta u^i - \eta v^i)(\zeta u^{i+1} - \eta v^{i+1})} \\
&\geq \nu_{i+1}^H
\end{aligned}$$

Since the p_i^H are a mixture of two geometrics, one decaying with rate $\frac{1}{a}$ and the other with $\frac{1}{b}$, and $\frac{1}{a} \geq \frac{1}{b}$, as i increases the first component dominates and the rate of decay effectively becomes $\frac{1}{a}$; or, $\lim_{i\to\infty} \nu_i^H = \frac{1}{a} \geq \rho^L$. Also because ν_i^H are decreasing, $\nu_i^H \geq \frac{1}{a} \geq \rho^L\ \forall i$. ∎

Theorem 7.3 *For our model:*
$N^H \geq_{st} N^{\rho^A}$ *but* $N^L \not\leq_{st} N^{\rho^A}$.

Proof: From the proof of Theorem 7.2, $\nu_i^H \geq \nu_{i+1}^H \; \forall i \geq 0$ and $\lim_{i\to\infty} \nu_i^H = \frac{1}{a} \geq \rho^A$. Therefore, $\nu_i^H \geq \rho^A \; \forall i \geq 0$. Then using Lemma 7.1, $N^H \geq_{st} N^{\rho^A}$.

Returning to the proof of Theorem 7.2, define $q_i^L = Pr\{N^L \geq i\}$. Using $\widehat{\Pi^L}(1) = 1$ in (7.63),

$$\frac{\pi_0^L}{\rho^L} = \frac{(a-1)(b-1)}{(\gamma-1)} \tag{7.66}$$

Thus, using the formula for p_i^L, derived from the expansion of $\widehat{\Pi^L}(z)$ analogous to (7.65), and using (7.66):

$$\begin{aligned}
q_i^L &= \sum_{j=i}^{\infty} p_i^L \\
&= \frac{\pi_0^L}{\rho^L(b-a)} \left[\left(\frac{\gamma}{a} - 1\right) \frac{1}{a^i} \left(\frac{1}{1-\frac{1}{a}}\right) - \left(\frac{\gamma}{b} - 1\right) \frac{1}{b^i} \left(\frac{1}{1-\frac{1}{b}}\right) \right] \\
&= \frac{(a-1)(b-1)}{(\gamma-1)(b-a)} \left[\left(\frac{\gamma-a}{a-1}\right) \frac{1}{a^i} + \left(\frac{b-\gamma}{b-1}\right) \frac{1}{b^i} \right]
\end{aligned}$$

Let $c = \frac{(b-1)(\gamma-a)}{(\gamma-1)(b-a)}$. Then $q_i^L = c\frac{1}{a^i} + (1-c)\frac{1}{b^i}$ and $0 \leq c \leq 1$. Also recall that $a \leq \frac{1}{\rho^A}$. Let $k = \lceil \log_{(a\rho^A)} c \rceil + 1$ so that $(a\rho^A)^k < c$. Now,

$$q_k^L = c\frac{1}{a^k} + (1-c)\frac{1}{b^k} \geq \frac{c}{a^k} > (\rho^A)^k = q_k^{\rho^A}$$

Clearly, $\forall j \geq k$, $q_j^L > q_j^{\rho^A}$ and hence $N^L \not\leq_{st} N^{\rho^A}$. In fact, $N^L =_{st} N^{\rho^A}$ if and only if $\rho^A = \rho^H = \rho^L$. ∎

Above we saw how N^H and N^L compare stochastically for a particular setting of arrival, service and switching rates. While such results seem theoretically appealing, these are of limited utility. In Section 4 we showed that $E[N^H]$, $E[N^L]$ and $E[N]$ are monotonic in α $(= \alpha^L + \alpha^H)$. It is interesting to ask the question: Do any of these monotonic behaviors extend to the stronger setting of stochastic monotonicity. We conjecture the following.

Conjecture 7.4 *For given values of $\mu^{\{L,H\}}$, $\lambda^{\{L,H\}}$ and the ratio α^L/α^H, N^H increases stochastically as switching rate α $(= \alpha^L + \alpha^H)$ decreases.*

8 Conclusion

In this paper we have considered very basic, yet open, questions regarding the response time of a queue with time-varying load. We have found that the response time can both increase or decrease when the load fluctuates more slowly, and we have derived a simple *slack criterion* to specify the behavior. We have also proven the first *monotonicity* results for systems with time-varying load, as well as the first *stochastic ordering* results for these systems. Finally we have provided the first results on the *shape* of the mean response time in a queue with fluctuating load, as a function of the rate of fluctuation and other input primitives. These latter results were enabled by the derivation of a series of approximations for the mean number of jobs in the system, which are accurate and also very simple and closed-form, telling us how the shape of the mean number of jobs is affected by the input primitives.

We hope that our research will stimulate others to ask further fundamental questions about time-varying systems. For example, we have seen that $E[N]$, $E[N^H]$ and $E[N^L]$ are all monotonic in α. Further, we conjecture that a stronger result may exist, namely that the random variable N^H is stochastically monotonic in α. However this is entirely non-obvious, particularly since N^L is not stochastically monotonic.

References

[1] J. Abate, G. Choudhury, and W. Whitt. Asymptotics for steady-state tail probabilities in structured Markov queueing models. *Commun. Statist.-Stoch. Mod.*, 10(1):99–143, 1994.

[2] I. J. B. F. Adan and V. G. Kulkarni. Single-server queue with Markov-dependent inter-arrival and service times. *QUESTA*, 45:113–134, 2003.

[3] E. Arjas. On the use of a fundamental identity in the theory of semi-Markov queues. *Adv. Appl. Prob.*, 4:271–284, 1972.

[4] S. Asmussen and J. Møller. Calculation of the steady state waiting time distribution in GI/PH/c and MAP/PH/c queues. *QUESTA*, 37:9–29, 2001.

[5] N. T. J. Bailey. A continuous time treatment of a simple queue using generating functions. *J. R. Statist. Soc., Series B*, 16(2):288–291, 1954.

[6] E. Çinlar. Queues with semi-Markov arrivals. *J. Appl. Prob.*, 4:365–379, 1967.

[7] E. Çinlar. Time dependence of queues with semi-Markov services. *J. Appl. Prob.*, 4:356–364, 1967.

[8] G. L. Choudhury, A. Mandelbaum, M. I. Reiman, and W. Whitt. Fluid and diffusion limits for queues in slowly changing environments. *Stoch. Mod.*, 13:121–146, 1997.

[9] A. B. Clark. A waiting line process of the Markov type. *Ann. Math. Statist.*, 27:452–459, 1956.

[10] J. H. A. de Smit. The single server semi-Markov queue. *Stoch. Proc. and Appl.*, 22:37–50, 1986.

[11] E. Gelenbe and C. Rosenberg. Queues with slowly varying arrival and service processes. *Man. Sci.*, 36(8):928–937, 1990.

[12] P. Harrison and H. Zatschler. Sojourn time distributions in modulated G-queues with batch processing. In *Quantitative Evaluation of Systems (QEST)*, pages 90–99, 2004.

[13] D. P. Heyman. On Ross's conjectures about queues with non-stationary Poisson arrivals. *J. Appl. Prob.*, 19:245–249, 1982.

[14] C. Knessl and Y. P. Yang. An exact solution for an $M(t)/M(t)/1$ queue with time-dependent arrivals and service. 40:233–245, 2002.

[15] D. M. Lucantoni, K. S. Meier-Hellstern, and M. Neuts. A single server queue with server vacations and a class of non-renewal arrival processes. *Adv. App. Prob.*, 22:676–705, 1990.

[16] D. M. Lucantoni and M. Neuts. Some steady-state distributions for the MAP/SM/1 queue. *Commun. Statist.-Stoch. Mod.*, 10(3):575–598, 1994.

[17] W. A. Massey. Asymptotic analysis of the time dependent $M/M/1$ queue. *Math. of OR*, 10(2):305–327, 1985.

[18] I. Mitrani and R. Chakka. Spectral expansion solution for a class of Markov models: Application and comparison with the matrix-geometric method. *Perf. Eval.*, 23(3):241–260, 1995.

[19] N. Miyoshi and T. Rolski. Ross-type conjectures on monotonicity of queues. *Australian & New Zealand J. of Stat.*

[20] M. Neuts. The single server queue with Poisson input and semi-Markov service times. *J. Appl. Prob.*, 3:202–230, 1966.

[21] M. Neuts. The m/m/1 queue with randomly varying arrival and service rates. *OPSEARCH*, 15(4):139–168, 1978.

[22] G. F. Newell. Queues with time-dependent arrival rates I – the transition through saturation. *J. Appl. Prob*, 5:436–451, 1968.

[23] G. F. Newell. Queues with time-dependent arrival rates II – the maximum queue and the return to equilibrium. *J. Appl. Prob*, 5:579–590, 1968.

[24] G. F. Newell. Queues with time-dependent arrival rates III – a mild rush hour. *J. Appl. Prob*, 5:591–606, 1968.

[25] V. Ramaswami. The N/G/1 queue and its detailed analysis. *Adv. Appl. Prob.*, 12:222–261, 1980.

[26] K. L. Rider. A simple approximation to the average queue size in the time-dependent M/M/1 queue. *JACM*, 23(2):361–367, 1976.

[27] T. Rolski. Queues with non-stationary input stream: Ross's conjecture. *Adv. Appl. Prob.*, 13:603–618, 1981.

[28] S. Ross. Average delay in queues with non-stationary Poisson arrivals. *J. Appl. Prob.*, 15:602–609, 1978.

[29] B. Sengupta. A queue with service interruptions in an alternating random environment. *OR*, 38(2):308–318, 1990.

[30] B. Sengupta. The semi-Markov queue: Theory and applications. *Commun. Statist.-Stoch. Mod.*, 6(3):383–413, 1990.

[31] T. Takine, Y. Matsumoto, T. Suda, and T. Hasegawa. Mean waiting times in nonpreemptive priority queues with Markovian arrival and i.i.d. service processes. *Perf. Eval.*, 20:131–149, 1994.

[32] T. Takine and B. Sengupta. A single server queue with service interruptions. *QUESTA*, 26:285–300, 1997.

[33] Y. Yang and C. Knessl. Asymptotic analysis of the M/G/1 queue with time-dependent arrival rates. *QUESTA*, 26:23–68, 1997.

[34] U. Yechiali and P. Naor. Queueing problems with heterogeneous arrivals and service. *OR*, 19(3):722–734, 1971.

A Proofs

A.1 Proof of Lemma 6.1

By combining (4.41) and (4.43) using the notation introduced in Definition 5.1, we get the following equation relating Δ and r:

$$\Delta^2 r^2 v \left[\rho^A + \frac{rv(\rho^A - \rho^L)(u - \rho^A)}{c_1}\right] - \Delta r \left[rvc_2 + c_1(v + \rho^A)\right] + c_1^2(1 - r) = 0 \qquad \text{(A.67)}$$

The above is a quadratic equation in Δ with coefficients involving r. We can solve for Δ in terms of r in closed form as

$$\begin{aligned}\Delta &= \frac{rvc_2 + c_1(v + \rho^A) - \sqrt{\left[rvc_2 + c_1(v + \rho^A)\right]^2 - 4c_1^2(1 - r)v\left[\rho^A + \frac{rv(\rho^A - \rho^L)(u - \rho^A)}{c_1}\right]}}{2rv\left[\rho^A + \frac{rv(\rho^A - \rho^L)(u - \rho^A)}{c_1}\right]} \\ &= \frac{\frac{2c_1^2(1-r)}{r}}{rvc_2 + c_1(v + \rho^A) + \sqrt{\left[rvc_2 + c_1(v + \rho^A)\right]^2 - 4c_1^2(1 - r)v\left[\rho^A + \frac{rv(\rho^A - \rho^L)(u - \rho^A)}{c_1}\right]}}\end{aligned} \qquad \text{(A.68)}$$

The sign of the discriminant in the first expression has to be negative because $\Delta \to 0$ as $r \to 1$. This gives:

$$\begin{aligned}\frac{d\log\Delta}{d\log r} &= \frac{d\log\left(\frac{1-r}{r}\right)}{d\log r} - \frac{d\log p(r)}{d\log r} \\ &= -\frac{1}{1-r} - \frac{r}{p(r)}\frac{dp(r)}{dr}\end{aligned}$$

where $p(r)$ is the expression in the denominator of (A.68). Since $0 \leq r \leq 1$, $\frac{-1}{1-r} \leq -1$. Therefore we only need to prove that $\frac{dp(r)}{dr} > 0$.

Case 1: $c_2 \geq 0$

The term outside the radical in $p(r)$ is increasing in r. The polynomial inside the radical in $p(r)$ is a quadratic where the coefficient of r^2 is positive. It will suffice to show that the coefficient of r in this quadratic polynomial is also positive. The coefficient of r in the quadratic polynomial inside the radical of $p(r)$ is:

$$
\begin{aligned}
&2vc_1\left[c_2(v+\rho^A)+2c_1\rho^A-2v(\rho^A-\rho^L)(u-\rho^A)\right]\\
=&2vc_1\left[\left\{(u-\rho^A)(v-\rho^L)-(v-u)(\rho^A-\rho^L)\right\}(v+\rho^A)+2c_1\rho^A-2v(\rho^A-\rho^L)(u-\rho^A)\right]\\
=&2vc_1\left[(u-\rho^A)\left\{(v-\rho^L)(v+\rho^A)-2v(\rho^A-\rho^L)\right\}+(v-u)\left\{2\rho^A(v-\rho^L)-(\rho^A-\rho^L)(v+\rho^A)\right\}\right]\\
=&2vc_1\left[(u-\rho^A)(v-\rho^A)(v+\rho^L)+(v-u)(\rho^A+\rho^L)(v-\rho^A)\right]\\
>&0
\end{aligned}
$$

Since the coefficients of all terms involving r in $p(r)$ are positive, $\frac{dp(r)}{r}>0$.

Case 2: $c_2<0$

In this case the coefficient of r in the term outside the radical is negative and a more tedious analysis is required. Let,

$$
\begin{aligned}
f(r)&=rvc_2+c_1(v+\rho^A)\\
g(r)&=4c_1^2(1-r)v\left[\rho^A+\frac{rv(\rho^A-\rho^L)(u-\rho^A)}{c_1}\right]
\end{aligned}
$$

Now,

$$
\begin{aligned}
\frac{d}{dr}p(r)\geq 0 &\iff \frac{d}{dr}\left(f(r)+\sqrt{f^2(r)-g(r)}\right)\geq 0\\
&\iff f'(r)+\frac{2f(r)f'(r)-g'(r)}{2\sqrt{f^2(r)-g(r)}}\geq 0\\
&\iff (g'(r))^2+4\left(f'(r)\right)^2g(r)-4f(r)f'(r)g'(r)\geq 0
\end{aligned}
$$

We first show that $g'(r)\leq 0$:

$$g'(r)=4c_1v\left[v(\rho^A-\rho^L)(u-\rho^A)-\rho^Ac_1-2rv(\rho^A-\rho^L)(u-\rho^A)\right]$$

Since $g''(r)=-2v(\rho^A-\rho^L)(u-\rho^A)<0$, to prove $g'(r)<0$ it suffices to show that $g'(0)\leq 0$.

$$
\begin{aligned}
g'(0)\leq 0 &\iff v(\rho^A-\rho^L)(u-\rho^A)-\rho^A(v-u)(v-\rho^L)\leq 0\\
&\iff (u-\rho^A)\left[(v-\rho^L)-\left\{v(1-\rho^A)+\rho^L(v-1)\right\}\right]\\
&\qquad -(v-u)\left[(\rho^A-\rho^L)+\left\{\rho^A(v-1)+\rho^L(1-\rho^A)\right\}\right]\leq 0\\
&\iff c_2-\left[(u-\rho^A)\left\{v(1-\rho^A)+\rho^L(v-1)\right\}+(v-u)\left\{\rho^A(v-1)+\rho^L(1-\rho^A)\right\}\right]\leq 0\\
&\Longleftarrow c_2\leq 0
\end{aligned}
$$

Let

$$q(r)=(g'(r))^2+4\left(f'(r)\right)^2g(r)-4f(r)f'(r)g'(r)$$

Proving $\frac{d}{dr}p(r) \geq 0$ is equivalent to proving $q(r) \geq 0$. For this, it suffices to show that $q'(r) \geq 0$ and $q(0) \geq 0$.

Claim: $q'(r) \geq 0$
Proof:

$$q'(r) = 2g''(r)[g'(r) - 2f(r)f'(r)]$$

since $f''(r) = 0$. We know $g''(r) \leq 0$. Further consider

$$s(r) = g'(r) - 2f(r)f'(r)$$

Now,

$$\begin{aligned} s'(r) &= g''(r) - 2\left(f'(r)\right)^2 \\ &\leq 0 \end{aligned}$$

and,

$$\begin{aligned} s(0) &= g'(0) - 2f(0)f'(0) \\ &= 2c_1 v\left[2v(\rho^A - \rho^L)(u - \rho^A) - 2\rho^A c_1 - c_2(v + \rho^A)\right] \\ &= 2c_1 v\left[2v(\rho^A - \rho^L)(u - \rho^A) - 2\rho^A(v - u)(v - \rho^L)\right. \\ &\quad \left. -(u - \rho^A)(v - \rho^L)(v + \rho^A) + (v - u)(\rho^A - \rho^L)(v + \rho^A)\right] \\ &= -2c_1 v(v - \rho^A)\left[(u - \rho^A)(v + \rho^L) + (v - u)(\rho^A + \rho^L)\right] \\ &\leq 0 \end{aligned}$$

Therefore, $s(r) \leq 0$ and combining $q'(r) = 2g''(r)s(r) \geq 0$.

Claim: $q(0) \geq 0$
Proof:

$$\frac{q(0)}{16c_1^2 v^2} = \left(v(\rho^A - \rho^L)(u - \rho^A) - \rho^A c_1\right)^2 - c_2(v + \rho^A)\left(v(\rho^A - \rho^L)(u - \rho^A) - \rho^A c_1\right) + c_2^2 v\rho^A$$

Now the polynomial $x^2 - c_2(v + \rho^A)x + c_2^2 v\rho^A$ is negative only in the interval $\left(c_2 v, c_2\rho^A\right)$. But,

$$\begin{aligned} & v(\rho^A - \rho^L)(u - \rho^A) - \rho^A c_1 \\ =& v(\rho^A - \rho^L)(u - \rho^A) - \rho^A(v - u)(v - \rho^L) \\ =& v(u - \rho^A)\left\{(v - \rho^L) - (v - \rho^A)\right\} - (v - u)\left\{v(\rho^A - \rho^L) + \rho^L(v - \rho^A)\right\} \\ =& c_2 v - (v - \rho^A)\left[v(u - \rho^A) + \rho^L(v - u)\right] \\ \leq& c_2 v \end{aligned}$$

Therefore, $q(0) \geq 0$.

Combining $q(0) \geq 0$ and $q'(r) \geq 0$, $q(r) \geq 0$ and hence $\frac{d}{dr}p(r) \geq 0$.

Hence,

$$\begin{aligned}
\frac{d \log \Delta}{d \log r} &= -\frac{1}{1-r} - \frac{r}{p(r)} \frac{dp(r)}{dr} \\
&\leq -\frac{1}{1-r} \\
&\leq -1
\end{aligned}$$