Moments of Response Time for Systems
With Hierarchical Memories

Kishori M. Konwar\(^*\)  Lester Lipsky  Marwan S. Sleiman
kishori@cse.uconn.edu  lester@engr.uconn.edu  marwan@engr.uconn.edu

Department of Computer Science and Engineering,
University of Connecticut, 371 Fairfield Rd., Unit 2155,
Storrs, CT 06269, USA.

Abstract

Storage systems in current day computing machines are in a variety of forms: registers, caches, main memory, disks, tapes, network storage systems, and so on. Each storage system provides the basic functions of storing data and holding the data until it is retrieved at a later time. The main differences between the various storage systems are their speed, cost, size, and volatility. The hierarchical nature of memory is prevalent in almost all computing systems ranging from the stand-alone desktop to grid computing environments. The rapid advancement of modern computer architecture and enterprise computing environments further extends this hierarchy. In this paper we abstract this hierarchical nature of memory in order to examine the variance, and in general, all moments of the distribution of time memory access. The existence of an unbounded variance for the response time can lead to extremely long wait queues and a particular class of distributions that has this property is the power tailed distribution. We present an analytical model to study the response time. We show that under reasonable assumptions for the values of the parameters involved, in the memory levels the response-time for such a hierarchical memory system is power tailed. \textbf{Mention about the } h_{\ell} > 1 \textbf{ as the ”rule of thumb” for the designers as a selling point}

1 Introduction

Computer systems have traditionally relied on a memory hierarchy in which large amounts of less expensive storage (disk, tape, SAN) have been used to retain the bulk of the stored information, while small amounts of fast storage (main memory, CPU cache memory) have

\(^*\)The authors’ names are in alphabetical order.
been employed to hold information while it is in active use. Fast memory and storage technology are vital to achieving superior system performance with the ever widening gap between the microprocessors speed and that of the memory technologies. Memory access time is increasingly the bottleneck in overall application performance. As a result, an application may spend a considerable amount of time waiting for data. This not only negatively impacts the overall performance, but the application cannot benefit much from a processor clock-speed upgrade either. One way to overcome this problem is to insert a small high-speed memory between the processor and main memory. The application can take advantage of this enhancement by fetching data from the cache instead of main memory. Unfortunately, little practical rules/metrics exists for systems designers and administrators to optimize their memory hierarchies. Exhaustive simulation takes far too long particularly with the memory hierarchies become more complex; trial and error on running systems is usually impossible; and prior mathematical analysis comes short of providing intuitive insight into memory sizing or assume the availability of memory technologies with arbitrary speeds and costs.

According to Greiner et al. a power-tail distribution is one for which the reliability function is of the form $x^{-\alpha}$, where $\alpha > 0$ for large $x$. These power tailed distributions have the singular property of unbounded $\mathbb{E}(X^\ell)$ for $\ell \geq \alpha$. This means these distributions may have second or higher order moments unbounded. Such distributions are abundant in the traffic arrivals in local and wide area networks, such as, video services, FTP data connections, NNTP, and WWW packet arrivals. We refer the inquisite reader to the papers for a comprehensive account on this. Experimental studies have revealed that in the presence of such power tailed arrival rates the number of arrival per unit time varies greatly. Moreover, the CPU time and file sizes are also shown to behave power tailed. Therefore the main question of concern is if the hierarchical nature of the memory produces such power tailed distributions. In this work we take a linear algebraic approach to make a general study of the current expanding memory hierarchy. The rest of this paper is organized as follows. In the next section we make a literature survey about previous work in this area and we present the power tailed formulation.

### Previous Work and Motivation

Many researchers studied memory hierarchy and some developed analytical models for designing memory hierarchies. Jacobs et al. proposed an analytical model to determine the optimal cache size to minimize the average memory reference time. For an excellent survey on CPU and disk cache performance the reader is referred to the articles generally focusing on a twolevel hierarchy. Tracedriven simulations are used extensively by researchers to investigate such aspects of cache performance as multiprocessor cache coherence and replacement strategies. Tracedriven studies are valuable for understanding cache behavior on specific workloads, but they are not easily applied to other workloads. Unlike traces, mathematical analysis lends itself well to understanding cache behavior on general workloads, though such generality usually leads to less accurate results. Du et al. showed the importance of the depth of memory hierarchy as a primary factor affecting the execution time on a cluster of workstations. But their results are dependent...
on the workload type which may vary depending on the application. Abraham et al. [?] show that using a cache simulator may take up to 466 days to simulate some cases which is clearly infeasible. Chow [?, ?] studied the scalability of the size of cache with the number of levels to conclude that the optimal number of cache levels scales with the logarithm of the capacity of the cache hierarchy. GarciaMolina and Rege [?, ?] demonstrated that for effective utilization of memory it is more suitable to use slower devices and less of faster device. Rege [?] showed that up to 3:1 advantage in performance can be achieved by using a three level rather than the two level hierarchy at the same total cost.

Welch [?] showed that the optimal speed of each level should be proportional to the amount of time spent servicing requests at that level. Jacob [?] pointed out that these studies have two major shortcomings:

(i) they assume the availability of memory technologies with arbitrary speeds and costs, and

(ii) they do not apply their analyses to a specific model of workload locality.

Being able to create and use technologies on a continuum of characteristics is convenient but too idealistic a viewpoint to be usable by system builders. Failing to apply a specific model of workload locality makes it impossible to provide an easily used, closedform solution for the optimal cache configuration [10], and so results from these papers have contained dependencies on the cache configuration—the number of levels, or the sizes and hit rates of the levels.

Figure ?? shows a typical hierarchy of the different memory components in computing systems.

We provide a relatively more general model to study the memory access time. First we motivate our formulation through an example. Consider the memory model depicted in Figure ?? in the form of a state diagram. $C_1$ represents the memory access by the CPU; this access has a hit ratio of $h$ and the average time of the access, $T$. In the case of a hit $C_2$ represents the read operation and in the case of a miss it accesses the next level of memory, denoted by $C_3$. At $C_3$ we assume that the average access time is $T$ and the memory update of $\tau / \lambda$. Here, $\lambda > 1$ is the speed factor. Now, we compare the average memory access time with and without the second level memory. The memory access time without the second level memory is $T + \gamma \tau$. Now, if we introduce the second level of memory to have the model as shown Fig. ?? then the access time is $T + h \tau + (1-h)T + (1-h)\gamma \tau + (1-h)\tau = (2-h)T + \tau [(1-h)\gamma + 1]$. The system with double memory levels has lesser average memory access time than the single-level system if

$$ (2 - h)T + \tau [(1 - h)\gamma + 1] < T + \gamma \tau $$

which reduces to $h > \frac{\theta + 1}{\theta + \gamma}$ where $\theta = \frac{T}{\tau}$. 


2 First Moment

We consider the model depicted in Fig. ??(a) Now we consider a permutation \( S \) to get the following (as in Fig. ??(b)) labeling of the same problem as above. Now we carry our computation based on the second model, i.e., after relabelling of the servers. Let \( S \) be the corresponding permutation matrix for the relabelling. Observe that \( S_{11} = 1 \) since we do not relabel the server \( C_1 \). Therefore,

\[
pS = p \quad \text{and} \quad S'\epsilon' = \epsilon'
\]

Now, we observe that

\[
\mathbb{E}[T^\alpha] = pV_n\epsilon' = pB_n^{-1}\epsilon' = p(M_n(I_n - P_n))^{-1}\epsilon' = p(I_n - P_n)^{-1}M_n^{-1}\epsilon = pS(I_n - P_n)^{-1}M_n^{-1}S'\epsilon' = pS(I_n - P_n)^{-1}S'SM_n^{-1}S'\epsilon' = p(S(I_n - P_n)S')^{-1}(SM_nS')^{-1}\epsilon' = p((SM_nS')(S(I_n - P_n)S'))^{-1}\epsilon' = \mathbb{E}[T^\alpha]
\]

Hence, the relabelling does not change our analysis regarding the moment of the service time

We denote by \( \tilde{h} = 1 - h \)

\[
P_n = \begin{pmatrix}
0 & \tilde{h} & 0 & 0 & 0 & \ldots & \ldots & \ldots & \ldots & \ldots & 0 & 0 & 0 & h & 0 \\
0 & 0 & \tilde{h} & 0 & 0 & \ldots & \ldots & \ldots & \ldots & \ldots & 0 & 0 & h & 0 & 0 \\
0 & 0 & 0 & \tilde{h} & 0 & \ldots & \ldots & \ldots & \ldots & \ldots & 0 & h & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & \ldots & \ldots & \ldots & 0 & \tilde{h} & 0 & h & 0 & \ldots & 0 & 0 & \ldots & 0 \\
0 & 0 & \ldots & \ldots & \ldots & 0 & 0 & 1 & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \\
0 & 0 & \ldots & \ldots & \ldots & 0 & 0 & 0 & 1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & \ldots & \ldots & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 1 & 0 & 0 \\
0 & 0 & \ldots & \ldots & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & \ldots & 1 \\
0 & 0 & \ldots & \ldots & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \\
0 & 0 & \ldots & \ldots & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & \ldots & 0
\end{pmatrix}
\]

\[
I_n - P_n = \begin{pmatrix}
1 & -\tilde{h} & 0 & 0 & 0 & \ldots & \ldots & \ldots & \ldots & \ldots & 0 & 0 & 0 & -h & 0 \\
0 & 1 & -\tilde{h} & 0 & 0 & \ldots & \ldots & \ldots & \ldots & \ldots & 0 & 0 & -h & 0 & 0 \\
0 & 0 & 1 & -\tilde{h} & 0 & \ldots & \ldots & \ldots & \ldots & \ldots & 0 & -h & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & \ldots & \ldots & \ldots & 1 & -\tilde{h} & 0 & -h & 0 & \ldots & 0 & 0 & \ldots & 0 \\
0 & 0 & \ldots & \ldots & \ldots & 0 & 1 & -1 & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \\
0 & 0 & \ldots & \ldots & \ldots & 0 & 0 & 1 & -1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & \ldots & \ldots & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 1 & -1 & 0 \\
0 & 0 & \ldots & \ldots & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 1 & -1 & 0 \\
0 & 0 & \ldots & \ldots & \ldots & 0 & 0 & 0 & 0 & 0 & \ldots & 0 & 0 & \ldots & 1
\end{pmatrix}
\]
where $\lambda$ is the service rate of the memories and $\gamma \geq 1$, which is the speed-up factor between two immediate levels of memory accesses during update and $\mu$ is the service rate of the slowest memory update.

Carrying out straightforward computation we get

$$B_n = M_n(I_n - P_n) = \begin{pmatrix}
\lambda & -\lambda h & 0 & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & -\lambda h & 0 \\
0 & \lambda & -\lambda h & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 & 0 & 0 & -\lambda h & 0 & 0 & 0 \\
0 & 0 & \lambda & -\lambda h & 0 & \ldots & 0 & \ldots & \ldots & \ldots & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \ldots & \ldots & 0 & \ldots & \ldots & \ldots & \ldots & 0 & 0 & 0 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu & -\mu \gamma & 0 & 0 & 0 \\
0 & 0 & 0 & \ldots & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\mu \gamma & 0 & \mu & 0 \\
0 & 0 & 0 & \ldots & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu & 0 & 0 & 0 \\
0 & 0 & 0 & \ldots & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu & -\mu \gamma & 0 & \mu & 0 \\
0 & 0 & 0 & \ldots & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \ldots & \ldots & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}$$

Consider the matrix below

$$V_n = \begin{pmatrix}
1 & \frac{h}{\lambda} & \frac{h^2}{\lambda} & \ldots & \frac{h^{n-2}}{\lambda} & \frac{h^{n-1}}{\lambda} & \frac{\mu}{\gamma^{n-2}} & \frac{\mu}{\gamma^{n-1}} & \frac{\mu}{\gamma^n} & \frac{\mu}{\gamma^{n+1}} & \frac{\mu}{\gamma^{n+2}} & \frac{\mu}{\gamma^{n+3}} & \ldots & \frac{h}{\mu} & \frac{1}{\mu} \\
0 & \frac{h}{\lambda} & \frac{h^2}{\lambda} & \ldots & \frac{h^{n-2}}{\lambda} & \frac{h^{n-1}}{\lambda} & \frac{\mu}{\gamma^{n-2}} & \frac{\mu}{\gamma^{n-1}} & \frac{\mu}{\gamma^n} & \frac{\mu}{\gamma^{n+1}} & \frac{\mu}{\gamma^{n+2}} & \frac{\mu}{\gamma^{n+3}} & \ldots & \frac{1}{\mu} & \frac{1}{\mu} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \ldots & 0 & \frac{1}{\lambda} & \frac{1}{\mu} & \frac{\nu}{\gamma^{n-1}} & \frac{\nu}{\gamma^{n-2}} & \frac{\nu}{\gamma^n} & \frac{\nu}{\gamma^{n+1}} & \frac{\nu}{\gamma^{n+2}} & \frac{\nu}{\gamma^{n+3}} & \ldots & \frac{1}{\mu} & \frac{1}{\mu} \\
0 & 0 & \ldots & 0 & 0 & \frac{1}{\lambda} & \frac{\mu}{\gamma^{n-1}} & \frac{\mu}{\gamma^{n-2}} & \frac{\mu}{\gamma^n} & \frac{\mu}{\gamma^{n+1}} & \frac{\mu}{\gamma^{n+2}} & \frac{\mu}{\gamma^{n+3}} & \ldots & \frac{1}{\mu} & \frac{1}{\mu} \\
0 & 0 & \ldots & 0 & 0 & 0 & \frac{1}{\mu} & \frac{\nu}{\gamma^{n-2}} & \frac{\nu}{\gamma^n} & \frac{\nu}{\gamma^{n+1}} & \frac{\nu}{\gamma^{n+2}} & \frac{\nu}{\gamma^{n+3}} & \ldots & \frac{1}{\mu} & \frac{1}{\mu} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & \ldots & 0 & 0 & 0 & \ldots & \ldots & 0 & 0 & 0 & 0 & \frac{1}{\mu} & \frac{1}{\mu} \\
0 & 0 & \ldots & 0 & 0 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \frac{1}{\mu} & \frac{1}{\mu} \\
0 & 0 & \ldots & 0 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \frac{1}{\mu} & \frac{1}{\mu}
\end{pmatrix}$$
It is clear by straightforward computation that

\[ V_n B_n = B_n V_n = I_n \]

Hence, \( V_n = B_n^{-1} \). Now we define \( p = (1, 0, 0, \ldots, 0) \) and \( e' = (1, 1, 1, \ldots, 1) \).

\[
\mathbb{E}[T] = p V e' = \frac{1}{\mu} \frac{1 - (\tilde{h}\gamma)^n}{1 - \tilde{h}} + \frac{1}{\lambda} \frac{1 - \tilde{h}^n}{1 - \tilde{h}}
\]

where \( 0 \leq h = 1 - \tilde{h} \leq 1 \) and \( 0 < \mu < 1 \) and for \( \tilde{h} \gamma \geq 1 \) we have

\[
\mathbb{E}[(T - \bar{T})] = p V e' = \infty \text{ for } \tilde{h} \gamma \geq 1
\]

### 3 Second Moment

Now we want to see when the population variance \( \mathbb{E}[(T - \bar{T})^2] = p V^2 e' \) becomes unbounded. Observe that

\[
p V^2 e' = (p V V e')
\]

Observe that all the entries in the matrix \( V_n \) are non-negative and hence all the elements of the vectors \( p V \) and \( V e' \) are non-negative. Therefore,

\[
\mathbb{E}[T^2] \geq p V^2 e' \geq \sum_{i=n+1}^{2n} (p V)_{i} (V e')_{i}
\]

Observe that we enumerate \((p V)_{i}\) from \(2n\) to \(n\) as

\[
(p V)_{i} = \frac{\tilde{h}^{n-i}}{\gamma^{n-i} \mu} \quad \text{for} \quad i = 1, 2, \ldots, n
\]

and similarly \((V e')_{i}\)

\[
(V e')_{i} = \frac{1}{\gamma^{n-1} \mu} \frac{1 - \gamma^i}{1 - \gamma} \quad \text{for} \quad i = 1, 2, \ldots, n
\]

Hence,

\[
p V^2 e' \geq \sum_{i=1}^{n} \frac{\tilde{h}^{n-i}}{\gamma^{n-i} \mu} \times \frac{1}{\gamma^{n-1} \mu} \frac{1 - \gamma^i}{1 - \gamma}
\]

\[
= \frac{1}{\gamma^{2n-1} (1 - \gamma) \mu^2 \tilde{h}} \sum_{i=1}^{n} (\tilde{h}\gamma)^i (1 - \gamma^i)
\]

\[
= \frac{1}{\gamma^{2n-1} (1 - \gamma) \mu^2 \tilde{h}} \left[ \sum_{i=1}^{n} (\tilde{h}\gamma^2)^i - \sum_{i=1}^{n} (\tilde{h}\gamma)^i \right]
\]

\[
= \frac{1}{\gamma^{2n-1} (1 - \gamma) \mu^2 \tilde{h}} \left[ \tilde{h}\gamma^2 \frac{1 - (\tilde{h}\gamma^2)^n}{1 - \tilde{h}\gamma^2} - \tilde{h} \gamma \frac{1 - (\tilde{h}\gamma)^n}{1 - \tilde{h}\gamma} \right]
\]

6
4 Higher Moments

Definition 1 For any positive \( m \) and \( n \) let \( M_{m \times n} \) denote the set of matrices \( A_{m \times n} \), such that, \( M_{m \times n} = \{ A_{m \times n} : a_{ij} \in \mathbb{R}^0 \} \).

Now, we define a partial order \( \succeq \) in \( M_{m \times n} \) so that for any \( A, B \in M_{m \times n} \) we say \( A \succeq B \) iff \( a_{ij} \leq b_{ij} \) for all \( i = 1, \ldots, m \) and \( j = 1, \ldots, n \).

Lemma 1 If for any \( A, B \in M_{m \times n} \) and \( A \succeq B, C, D \in M_{\ell \times m} \) and \( C \succeq D \) and \( E, F \in M_{n \times p} \) and \( E \preceq F \), then \( CA \preceq DB \)

Proof: The proof is trivial once we observe that all the entries of the matrices are non-negative real numbers. \( \square \)

Let us denote for any \( n \in \mathbb{N} \) the \( n \times n \) matrices

\[
A_n = \frac{1}{\lambda} \begin{pmatrix}
1 & \tilde{h} & \tilde{h}^2 & \ldots & \tilde{h}^{n-2} & \tilde{h}^{n-1} \\
0 & 1 & \tilde{h} & \ldots & \tilde{h}^{n-3} & \tilde{h}^{n-2} \\
0 & 0 & 1 & \ldots & \tilde{h}^{n-4} & \tilde{h}^{n-3} \\
0 & 0 & 0 & \ldots & 1 & \tilde{h} \\
0 & 0 & 0 & \ldots & 0 & 1
\end{pmatrix}
\]

\[
B_n(\gamma) = \frac{\gamma^{n-1}}{\mu} \begin{pmatrix}
\tilde{h}^{n-1} & \tilde{h}^{n-2} & \tilde{h}^{n-3} & \ldots & \tilde{h} \gamma^{-n} & 1 \\
\tilde{h}^{n-2} & \tilde{h}^{n-3} & \tilde{h}^{n-4} & \ldots & 1 & \gamma^{-n} \\
\tilde{h}^{n-3} & \tilde{h}^{n-4} & \tilde{h}^{n-5} & \ldots & 1 & \gamma^{-n} \\
\vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\
\tilde{h} & 1 & \frac{1}{\gamma} & 1 & \ldots & \frac{1}{\gamma} \\
1 & \frac{1}{\gamma} & \frac{1}{\gamma^2} & \ldots & \frac{1}{\gamma^{n-2}} & \frac{1}{\gamma^{n-1}}
\end{pmatrix}
\]

and

\[
C_n(\gamma) = \frac{\gamma^{n-1}}{\mu} \begin{pmatrix}
1 & \frac{1}{\gamma} & \frac{1}{\gamma^2} & \ldots & \frac{1}{\gamma^{n-2}} & \frac{1}{\gamma^{n-1}} \\
0 & \frac{1}{\gamma} & \frac{1}{\gamma^2} & \ldots & \frac{1}{\gamma^{n-2}} & \frac{1}{\gamma^{n-1}} \\
0 & 0 & \frac{1}{\gamma} & \ldots & \frac{1}{\gamma^{n-2}} & \frac{1}{\gamma^{n-1}} \\
0 & 0 & 0 & \ldots & \frac{1}{\gamma^{n-2}} & \frac{1}{\gamma^{n-1}} \\
0 & 0 & 0 & \ldots & 0 & \frac{1}{\gamma^{n-1}}
\end{pmatrix}
\]

Now, we can write the \( V_n \) above as

\[
V_n = \begin{pmatrix}
A_n & B_n(\gamma) \\
O_n & C_n(\gamma)
\end{pmatrix}
\]

where \( O_n \) is the \( O_{n \times n} \) matrix will all the entries 0. Now, consider the matrices \( S_n \) and \( T_n \) defined as

\[
S_n = \begin{pmatrix}
O_n & B_n(\gamma) \\
O_n & O_n
\end{pmatrix}
\]

\[
T_n = \begin{pmatrix}
O_n & O_n \\
O_n & O_n
\end{pmatrix}
\]
and
\[ T_n = \begin{pmatrix} O_n & O_n \\ O_n & C_n(\gamma) \end{pmatrix} \]
and since all the elements of \( V_n \) are non-negative we clearly have \( S_n \preceq V_n \) and \( T_n \preceq V_n \). But now by Lemma ?? we have \( V_n^2 \succeq S_n T_n \) and hence
\[ S_n T_n = \begin{pmatrix} O_n & B_n(\gamma) \\ O_n & O_n \end{pmatrix} \begin{pmatrix} O_n & O_n \\ O_n & C_n(\gamma) \end{pmatrix} = \begin{pmatrix} O_n & B_n(\gamma) C_n(\gamma) \\ O_n & O_n \end{pmatrix} \]
clearly by induction we can show that for any \( \alpha \in \mathbb{N} \) we have
\[ V_n^\alpha \succeq S_n T_n^\alpha \succeq \begin{pmatrix} O_n & B_n(\gamma) C_n^\alpha(\gamma) \\ O_n & O_n \end{pmatrix} \]
Now, by Lemma ?? we have
\[ pV^\alpha e' = p \begin{pmatrix} O_n & B_n(\gamma) C_n^\alpha(\gamma) \\ O_n & O_n \end{pmatrix} (0, \ldots, 0, 1, \ldots, 1)' \]

**Lemma 2** For \( \ell = 1, \ldots \) we have
\[ B_n(\gamma) C_n^\ell(\gamma) \succeq \left( \frac{1}{\mu} \right) \ell B_n(\gamma^{\ell+1}) \]

**Proof:**
We show the proof by induction on \( \ell \).

**Base case:** \( \ell = 1 \)
\[ B_n(\gamma) C_n(\gamma) \]
\[ = \frac{\gamma^{2(n-1)}}{\mu^2} \begin{pmatrix} \hat{h}^{n-1} & \frac{1}{\gamma} (\hat{h}^{n-1} + \frac{\hat{h}^{n-2}}{\gamma}) & \frac{1}{\gamma^2} (\hat{h}^{n-1} + \frac{\hat{h}^{n-2}}{\gamma} + \frac{\hat{h}^{n-3}}{\gamma^2}) & \cdots & \frac{1}{\gamma^{n-1}} (\hat{h}^{n-1} + \frac{\hat{h}^{n-2}}{\gamma} + \cdots + \frac{\hat{h}}{\gamma^{n-2}} + \frac{1}{\gamma^{n-1}}) \\ \hat{h}^{n-2} & \frac{1}{\gamma} (\hat{h}^{n-2} + \frac{\hat{h}^{n-3}}{\gamma}) & \frac{1}{\gamma^2} (\hat{h}^{n-2} + \frac{\hat{h}^{n-3}}{\gamma} + \frac{\hat{h}^{n-4}}{\gamma^2}) & \cdots & \frac{1}{\gamma^{n-1}} (\hat{h}^{n-2} + \frac{\hat{h}^{n-3}}{\gamma} + \cdots + \frac{\hat{h}}{\gamma^{n-2}} + \frac{1}{\gamma^{n-1}}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{h} & \frac{1}{\gamma} (1 + \frac{1}{\gamma}) & \frac{1}{\gamma^2} (1 + \frac{1}{\gamma} + \frac{1}{\gamma^2}) & \cdots & \frac{1}{\gamma^{n-1}} (1 + \frac{1}{\gamma} + \cdots + \frac{1}{\gamma^{n-2}} + \frac{1}{\gamma^{n-1}}) \\ 1 & \frac{1}{\gamma} (1 + \frac{1}{\gamma}) & \frac{1}{\gamma^2} (1 + \frac{1}{\gamma} + \frac{1}{\gamma^2}) & \cdots & \frac{1}{\gamma^{n-1}} (1 + \frac{1}{\gamma} + \cdots + \frac{1}{\gamma^{n-2}} + \frac{1}{\gamma^{n-1}}) \end{pmatrix} \]
\[ \succeq \frac{\gamma^{2(n-1)}}{\mu^2} \begin{pmatrix} h^{n-1} & h^{n-2} & h^{n-3} & \cdots & h \\ h^{n-2} & h^{n-3} & h^{n-4} & \cdots & h \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h & 1 & \frac{1}{\gamma} & \frac{1}{\gamma^2} \cdots \frac{1}{\gamma^{n-2}} \\ 1 & \frac{1}{\gamma} & \frac{1}{\gamma^2} \cdots \frac{1}{\gamma^{n-2}} \end{pmatrix} \]
\[ = \left( \frac{1}{\mu} \right) B_n(\gamma^2) \]
which we get by retaining only the last term in each of the expansion of the bracketed terms and considering the fact that all the terms are non-negative. Suppose for some \( \ell \) we have \( B_n(\gamma) C_n^\ell(\gamma) \succeq
\[(\frac{\gamma^{n-1}}{\mu})^\ell B_n(\gamma^{\ell+1})\] Next consider,

\[B_n(\gamma)C_{n+1}^\ell(\gamma) = B_n(\gamma)C_{n+1}^\ell(\gamma)C_n(\gamma) \geq \frac{1}{\mu^\ell} B_n(\gamma^{\ell+1})C_n(\gamma)\]

by applying induction hypothesis.

\[
\begin{pmatrix}
\tilde{h}_n^{-1} & \frac{1}{\gamma} \left(\tilde{h}_n^{-1} + \frac{\tilde{h}_n^{-2}}{\gamma} + \frac{\tilde{h}_n^{-3}}{\gamma^2} + \cdots + \frac{1}{\gamma^{\ell+1}}\right) \\
\tilde{h}_n^{-2} & \frac{1}{\gamma^2} \left(\tilde{h}_n^{-2} + \frac{\tilde{h}_n^{-3}}{\gamma} + \frac{\tilde{h}_n^{-4}}{\gamma^2} + \cdots + \frac{1}{\gamma^{\ell+1}}\right) \\
\vdots & \vdots \\
\tilde{h} & \frac{1}{\gamma^\ell+1} \left(\tilde{h} + \frac{1}{\gamma} + \frac{1}{\gamma^2} + \cdots + \frac{1}{\gamma^{\ell+1}}\right) \\
1 & \frac{1}{\gamma^{\ell+1}} \left(1 + \frac{1}{\gamma} + \cdots + \frac{1}{\gamma^{\ell+1}}\right)
\end{pmatrix}
\geq \frac{\gamma(\ell+1)(n-1)}{\mu^{\ell+1}}
\begin{pmatrix}
\tilde{h}_n^{-1} & \frac{\tilde{h}_n^{-3}}{\gamma(\ell+1)} & \frac{1}{\gamma(n-1)(\ell+1)} \\
\tilde{h}_n^{-2} & \frac{\tilde{h}_n^{-4}}{\gamma(\ell+1)} & \frac{1}{\gamma(n-1)(\ell+1)} \\
\vdots & \vdots & \vdots \\
\tilde{h} & \frac{1}{\gamma^{\ell+1}} & \frac{1}{\gamma(n-1)(\ell+1)} \\
1 & \frac{1}{\gamma^{\ell+1}} & \frac{1}{\gamma(n-1)(\ell+1)}
\end{pmatrix}
\]

\[= \left(\frac{1}{\mu^{\ell+1}}\right) B_n(\gamma^{\ell+2})\]

\[\mathbb{E}[T^\ell] = \ell! p V^\ell e''
\geq \ell! p \left(\frac{1}{\mu^{\ell+1}} B_n(\gamma^{\ell})\right)(0, \ldots, 0, 1, \ldots, 1)
\geq \frac{\ell!}{\mu^{\ell+1}} \times \frac{n(\ell)}{\mu} \left(\tilde{h}_n^{-1} + \frac{\tilde{h}_n^{-2}}{\gamma} + \cdots + \frac{\tilde{h}}{\gamma^{n-1}} + \frac{1}{\gamma^{n-1}}\right)
\geq \frac{\ell!}{\mu^{\ell+1}} \left((\tilde{h}\gamma)^n - 1 + (\tilde{h}\gamma)^{n-2} + \cdots + (\tilde{h}\gamma) + 1\right)
\]

In the above we used the relations \(\geq\) and \(\geq\) interchangeably since the terms are scalars.

## 5 Non-exponential Servers

In the previous sections we assumed that each component in the cache hierarchy had exponentially distributed service times. Here we derive an expression which is valid for any network where all of the servers have ME distribution times. We now must deal with three levels of transition and other matrices, so we first introduce our generalized notation.
In Fig. ?? we show a system with $m$ non-exponential service time subsystems, or servers. We denote the $m$ subsystems as $S_1, S_2, \ldots, S_m$ shown as squares inside the system. We denote by $T$ the random variable for the total time a customer spends in the entire system $S$, and by $T_i$ the time spent in subsystem $S_i$, each time the customers it. The distribution of $T_i$ (level one) is represented by $m_i$ exponential phases, where the $j^{th}$ phase in $S_i$ has service rate $[M_i]_{jj}$. That is, $M_i$ is a diagonal $m_i \times m_i$ matrix. Similarly, the transition probabilities of travelling within $S_i$ are denoted by the elements of matrix $P_i$. Thus, $[P_i]_{jk}$ is the probability that a customer, upon finishing at phase $j$ in $S_i$ will go to $k$, also in $S_i$. The probability that the customer will leave $S_i$ is given by the column vector, $q_i$, where $[q_i]_{jj}$ is the probability that a customer, on finishing at phase $j$ in $S_i$ will leave $S_i$. Obviously, we have the property that

$$P_i \epsilon_i + q_i = \epsilon_i, \text{ or } q_i = [I - P_i] \epsilon_i.$$ 

The $j^{th}$ element of the (row) entrance vector, $p_i$, for entering $S_i$ is defined as $[p_i]_j$, the probability that a customer, upon entering $S_i$ will go to phase $j$. Therefore, $p_i \epsilon_i = 1$. It is well known that these matrices allow the distribution function of the time spent in $S_i$ to be computed (see, e.g., ??). First define

$$B_i := M_i(I - P_i), \text{ and } V_i := B_i^{-1}.$$ 

Note that since $q_i = [I - P_i] \epsilon_i$, we have

$$M_i q_i' = B_i \epsilon_i.'$$ 

Then from [???],

$$R_{T_i}(t) := \Pr(T_i > t) = p_i[\exp(-B_i t)] \epsilon_i,'$$

and

$$\mathbb{E}[T_i] = \ell_i p_i [V_i] \epsilon_i.'$$

In particular, the mean time spent in $S_i$ is $T_i = p_i[V_i] \epsilon_i.$

The second level of matrices governs the traffic between servers. Thus the element $[P]_{ij}$, in the $m \times m$ transition matrix $P$, denotes the probability that the customer will go to subsystem $S_j$ when the service is completed in subsystem $S_i$. The element $[p]_i$ of the entry vector $p$ denotes the probability that the customer goes to subsystem $S_i$ upon entering the system. While the element $[q]_j$ of the exit vector $q$ denotes the probability that the customer leaves the system when service is completed at subsystem $S_j$. Clearly, the total number of phases in $S$ is

$$M = \sum_{i=1}^{m} m_i.$$ 

It then follows that

$$p \epsilon' = 1 \text{ and } P \epsilon' + q' = \epsilon'$$

We next define the diagonal matrix, $T$, with elements $[T]_{ii} = T_i$. This is the second level equivalent of $M_i^{-1}$. If all the servers are exponential ($m_i = 1 \ \forall \ i$), then $M = m$ and we have the system as described in the previous sections. We denote by $T_e$ the time spent in such a system, and can write:

$$\mathbb{E}[T_e] = p[(I - P)^{-1} T]\epsilon' =: p[V] \epsilon'.$$
In what follows we will prove that $\mathbb{E}[T_e] = \mathbb{E}[T]$ irrespective of the distributions of the $T_i$’s.

We now discuss level three, the composite of levels one and two, i.e., the full description of $S$. For the sake of simplicity first we consider a system with only two subsystems $S_1$ and $S_2$ but generalization to $m > 2$ subsystems is straightforward. The state-transition matrix for the entire system is

$$ P = \begin{pmatrix} P_{11} + q_{11}' P_1 & P_{12} q_{12}' P_2 \\ P_{21} q_{21}' P_1 & P_2 + P_{22} q_{22}' P_2 \end{pmatrix} $$

where $P_{ij} := [P]_{ij}$. Note that $P$ is an $M \times M$ matrix even though it is written as a $2 \times 2$ matrix, where the $ij$th element is an $m_i \times m_j$ matrix. The components are interpreted in the following way. For instance, consider the $k\ell$th element of block $P_{11}$, corresponding to the probability that a customer, having left a phase in $S_1$ will either remain in or return to $S_1$. That is, a customer, upon leaving phase $k$ in $S_1$ either goes to phase $\ell$ in $S_1 \{[P_1]_{k\ell}\}$, or leaves $\{[q_1]\}$ and immediately returns to $S_1 \{P_{11}\}$ and goes to phase $\ell \{[p_1]\}$. Also, we have the service-time matrix for the system $S$ as

$$ \mathcal{M} = \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix} $$

Before going on we review our somewhat complex notation. Objects that relate to transitions within subsystem $S_i$ ($m_i$-dimensional) are denoted by bold-faced Roman letters with bold-faced subscripts denoting the subsystem, namely: $p_i$, $P_i$, $B_i$, $q_i'$, $\epsilon_i'$, etc. Objects that relate to transitions to or between subsystems are denoted by bold-faced Italic letters, ($m$-dimensional) such as: $p$, $P$, $B$, $V$, $T$, $\epsilon'$, etc. Finally, objects relating to the overall system (with dimension $M$) are denoted by bold-faced characters of the form: $p$, $P$, $B$, $V$, $\epsilon'$.

We are now ready to build the matrices that describe $T$, the time a customer spends in $S$.

$$ B := \mathcal{M}(I - P) = \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix} \begin{pmatrix} I_1 - P_1 - P_{11} q_{11}' P_1 & -P_{12} q_{12}' P_2 \\ -P_{21} q_{21}' P_1 & I_2 - P_2 - P_{22} q_{22}' P_2 \end{pmatrix} $$

$$ = \begin{pmatrix} M_1(I_1 - P_1) - P_{11} M_1 q_{11}' P_1 & -P_{12} M_1 q_{12}' P_2 \\ -P_{21} M_2 q_{21}' P_1 & M_1(I_2 - P_2) - P_{22} M_2 q_{22}' P_2 \end{pmatrix} $$

$$ = \begin{pmatrix} B_1 - P_{11} B_1 q_{11}' P_1 & -P_{12} B_1 \epsilon_{12}' P_2 \\ -P_{21} B_2 \epsilon_{21}' P_1 & B_2 - P_{22} B_2 \epsilon_{22}' P_2 \end{pmatrix} $$

$$ = \mathcal{B}_0 \begin{pmatrix} I - \begin{pmatrix} P_{11} \epsilon_{11}' P_1 & P_{12} \epsilon_{12}' P_2 \\ P_{21} \epsilon_{21}' P_1 & P_{22} \epsilon_{22}' P_2 \end{pmatrix} \end{pmatrix} $$

where we have used $M_i q_i' = B_i \epsilon_i'$, and

$$ \mathcal{B}_0 := \begin{pmatrix} B_1 & 0 \\ 0 & B_2 \end{pmatrix} $$

Note that the vector-matrix pair, $p$, $B$ generates the distribution function, $R_T(t)$ in the same way that the pair, $p_i$, $B_i$, generates the time spent in $S_i$. We pause for a moment to introduce some new notation and accompanying algebraic lemmas we will use to find explicit expressions for $V$, as well as some other properties of $S$. 

11
Definition 2  For the any $2 \times 2$ matrix $A$ we define the following:

$$
\langle A \rangle := \begin{pmatrix}
A_{11}e_1^1p_1 & A_{12}e_1^1p_2 \\
A_{21}e_2^1p_1 & A_{22}e_2^1p_2
\end{pmatrix}
$$

Keep in mind that $\langle \text{sym}BA \rangle$ is an $(m_1 + m_2) \times (m_1 + m_2)$ matrix.

Lemma 3  For every pair of $2 \times 2$ matrices $A$ and $B$ we have $\langle A \rangle \langle B \rangle = \langle AB \rangle$

Proof:  It is straight-forward to verify that (using $p_1e_1^1 = 1$)

$$
\langle A \rangle \langle B \rangle = \begin{pmatrix}
A_{11}e_1^1p_1 & A_{12}e_1^1p_2 \\
A_{21}e_2^1p_1 & A_{22}e_2^1p_2
\end{pmatrix} \begin{pmatrix}
B_{11}e_1^1p_1 & B_{12}e_1^1p_2 \\
B_{21}e_2^1p_1 & B_{22}e_2^1p_2
\end{pmatrix}
= \begin{pmatrix}
(A_{11}B_{11} + A_{12}B_{21})e_1^1p_1 & (A_{11}B_{12} + A_{12}B_{22})e_1^1p_1 + A_{11}e_1^1p_1B_{12}e_1^1p_2 + A_{12}e_1^1p_2B_{22}e_2^1p_2 \\
(A_{21}B_{11} + A_{22}B_{21})e_2^1p_1 & (A_{21}B_{12} + A_{22}B_{22})e_2^1p_2 + A_{21}e_1^1p_1B_{12}e_1^1p_2 + A_{22}e_2^1p_2B_{22}e_2^1p_2
\end{pmatrix}
= \begin{pmatrix}
(A_{11}B_{11} + A_{12}B_{21}) & (A_{11}B_{12} + A_{12}B_{22}) \\
(A_{21}B_{11} + A_{22}B_{21}) & (A_{21}B_{12} + A_{22}B_{22})
\end{pmatrix}
= \langle AB \rangle
$$

\[\square\]

Lemma 4  For every pair of square matrices $A$ and $B$ we have $\langle A \rangle + \langle B \rangle = \langle A + B \rangle$

Proof:  This is also straight-forward to verify.

Now, we want to evaluate $V$, first noting that

$$
B = M(I - P) = B_0 - B_0(P) = B_0(I - \langle P \rangle)
$$

Since $V = B^{-1}$, we can get it by using the formal Taylor expansion of $(I - \langle P \rangle)^{-1}$. That is,

$$
V = (I - \langle P \rangle)^{-1}V_0 = (I + \langle P \rangle + \langle P \rangle^2 + \ldots) V_0 = (I + \langle P \rangle + \langle P^2 \rangle + \ldots) V_0 = (I + \langle P(I + P + P^2 + \ldots) \rangle) V_0 = (I + \langle P(I - P)^{-1} \rangle) V_0
$$

where $V_0 = B_0^{-1}$. To prove that this is truly the inverse of $B$ we merely show that $VB = I$:

$$
VB = (I + \langle P(I - P)^{-1} \rangle) V_0 B_0(l - \langle P \rangle) = (I + \langle P(I - P)^{-1} \rangle)(I - \langle P \rangle) = I - \langle P \rangle + \langle P(I - P)^{-1} \rangle(l - \langle P \rangle) = I - \langle P \rangle + \langle P(I - P)^{-1} \rangle(I - \langle P \rangle) \text{ by Lemma } ?? = I - \langle P \rangle + \langle P(I - P)^{-1}(I - P) \rangle \text{ by Lemma } ?? = I
$$
and similarly

$$BV = I$$

Hence $$V = B^{-1}$$.

Now we have

$$E[T] = pV\epsilon' = pV_0\epsilon' + p(P(I - P)^{-1})V_0\epsilon'$$

where $$p = [a_1p_1, a_2p_2]$$. First let us consider the leading term, $$pV_0\epsilon'$$

$$pV_0\epsilon' = (p_1p_1, p_2p_2) \begin{pmatrix} V_1 & 0 \\ 0 & V_2 \end{pmatrix} \begin{pmatrix} \epsilon'_1 \\ \epsilon'_2 \end{pmatrix} = (p_1[p_1V_1, p_2[p_2V_2]]) \begin{pmatrix} \epsilon'_1 \\ \epsilon'_2 \end{pmatrix} = p_1[p_1V_1\epsilon'_1] + p_2[p_2V_2\epsilon'_2] = p_1T_1 + p_2T_2 = pTe'$$

Before moving on we need yet another definition.

**Definition 3** Any $$M \times M$$ matrix, $$X$$ can be partitioned into $$m^2$$ sub-matrices of dimension $$m_i \times m_j$$. Each sub-matrix (call it $$X_{ij}$$) can be reduced to a scalar by the operation: $$p_iX_{ij}\epsilon'_j$$. Then define:

$$[X]_{ij} := p_iX_{ij}\epsilon'_j$$

**Lemma 5** For any $$m \times m$$ matrices $$A$$ and $$C$$, and any $$M \times M$$ matrix $$X$$, we have

$$\langle A \rangle X\langle C \rangle = \langle AXC \rangle$$

**Proof:** This can be done by direct substitution. \(\square\)

**Lemma 6** For any matrices $$A$$ and $$X$$ (with the appropriate dimensions implied by their symbols) we have

$$\langle A \rangle X\epsilon' = \langle AX \rangle \epsilon'.$$

Similarly,

$$pX\langle A \rangle = p\langle AX \rangle.$$

**Proof:** Again by direct substitution. \(\square\)

**Lemma 7** For any $$m \times m$$ matrix, $$X$$,

$$pX\epsilon' = pX\epsilon'$$

**Proof:**

Going back to the expression for $$E[T]$$, we can now write

$$p\langle P(I - P)^{-1}\rangle V_0\epsilon' = p\langle P(I - P)^{-1}\rangle T\epsilon' = pP(I - P)^{-1}Te' = pPV\epsilon'$$

Hence, we have the mean time of service for the system $$S$$ as

$$E[T] = pV\epsilon' = pV_0\epsilon' + pP(I - P)^{-1}Te' = pTe' + pP(I - P)^{-1}Te' = p[I + P(I - P)^{-1}]Te' = p[(I - P)^{-1}Te'] = pV\epsilon' = E[T_e].$$

13
Thus, the mean time spent in the system depends only on the mean times spent in each of the $S_i$'s, and not on their distributions.

To compute the second moment we proceed as follows:

$$pV = pV_0 + p(P(I - P)^{-1})V_0$$

$$V_{e'} = V_{0e'} + (P(I - P)^{-1})V_0 e' = V_{0e'} + (P(I - P)^{-1}T)e'$$

and since $pVVe' = pV^2 e'$, we can write:

$$pV^2 e' = \left(pV_0^2 e' + pV_0(P(I - P)^{-1})V_0 e' + p(P(I - P)^{-1})V_0^2 e' + p(P(I - P)^{-1})V_0(P(I - P)^{-1})V_0 e' \right)$$

Therefore,

$$(I) = pT^{(2)}e'$$

where $T^{(2)}$ is a diagonal matrix with $[T^{(2)}]_{ii} = p_i V_i^2 e'_i$.

$$(II) = pV_0(P(I - P)^{-1})V_0 e'$$

$$= pV_0(P(I - P)^{-1}T)e' \text{ by Lemma ??}$$

$$= p(TP(I - P)^{-1}T)e' \text{ by Lemma ??}$$

$$(III) = p(P(I - P)^{-1})V_0^2 e'$$

$$= p(P(I - P)^{-1}T^{(2)})e'$$

$$= pP(I - P)^{-1}T e'$$

$$(IV) = p(P(I - P)^{-1})V_0(P(I - P)^{-1}T)e'$$

$$= p(P(I - P)^{-1}P(I - P)^{-1}T)e' \text{ by Lemma ??}$$

$$= pP(I - P)^{-1}TP(I - P)^{-1}T e' \text{ by Lemma ??}$$

$$(I) + (III) = pT^{(2)}e' + pP(I - P)^{-1}T^{(2)}e'$$

$$= p(I + P(I - P)^{-1}T^{(2)})e'$$

$$= p(I - P)^{-1}T^{(2)}e'$$

$$(II) + (IV) = pTP(I - P)^{-1}T e' + pP(I - P)^{-1}TP(I - P)^{-1}T e'$$

$$= p(T(I + P(I - P)^{-1}TP)T) e'$$

$$= p\bar{V}PVe'$$

$$E[T^2] = 2pV^2 e' = p(I - P)^{-1}T^{(2)}e' + 2pVPVe'$$

$$= p(I - P)^{-1}T^2(C^2 + I)e' + 2pVPVe'$$

$$= pVT(C^2 + I)e' + 2pVPVe'$$

where $C^2$ is a diagonal matrix with $[C^2]_{ii} = C_i^2 := a_i^2 / T_i^2$, and $2T^{(2)} = T^2 C^2 + T^2 = T^2(C^2 + I)$

Now,
\[ \sigma^2 = \mathbb{E}[T^2] - (\mathbb{E}[T])^2 \]
\[ = pVT(C^2 + I)e' + 2pVVP\epsilon' - (pT\epsilon')^2 \]

and for the pure exponential service time case

\[ \sigma_e^2 = pV^2\epsilon' - (pV\epsilon')^2 \]

Finally, we have

\[ \sigma^2 = \sigma_e^2 + pVT(C^2 + I)e' + 2pVVP\epsilon - 2pV^2\epsilon' \]
\[ = \sigma_e^2 + pVT(C^2 + I)e' - 2pV(I - P)V\epsilon' \]
\[ = \sigma_e^2 + pVT(C^2 - I)e' \]
\[ = \sigma_e^2 + pVT\Gamma\epsilon' \]

where

\[ \Gamma = \begin{pmatrix} C_1^2 - 1 & 0 \\ 0 & C_2^2 - 1 \end{pmatrix} \]

6 Conclusion

The high variance in the time of memory access is not recommended. Our approach is independent of the load and other constraints and the moments of response time are higher.