AMVA Techniques for High Service Time Variability *

Derek L. Eager

Department of Computer Science
University of Saskatchewan
eager@cs.usask.ca

Daniel J. Sorin Mary K. Vernon

Computer Sciences Department
University of Wisconsin - Madison
sorin,vernon@cs.wisc.edu

Abstract

Motivated by experience gained during the validation of a recent Approximate Mean Value Analysis (AMVA) model of modern shared memory architectures, this paper re-examines the "standard" AMVA approximation for non-exponential FCFS queues. We find that this approximation is often inaccurate for FCFS queues with high service time variability. For such queues, we propose and evaluate: (1) AMVA estimates of the mean residual service time at an arrival instant that are much more accurate than the standard AMVA estimate, (2) a new AMVA technique that provides a much more accurate estimate of mean center residence time than the standard AMVA estimate, and (3) a new AMVA technique for computing the mean residence time at a "downstream" queue which has a more bursty arrival process than is assumed in the standard AMVA equations. Together, these new techniques increase the range of applications to which AMVA may be fruitfully applied, so that for example, the memory system architecture of shared memory systems with complex modern processors can be analyzed with these computationally efficient methods.

1. Introduction

Approximate Mean Value Analysis (AMVA) is a widely used approach to evaluating key computer system performance questions [1, 2, 5, 8, 11, 12, 16, 17, 18, 19, 20, 25, 26, 27, 28, 29]. The wide applicability of the AMVA technique is due to both its very low computational expense and its high degree of accuracy in producing performance estimates that agree with detailed system simulation or system measurement. These capabilities are achieved through the use of heuristic extensions to the Mean Value Analysis equations for product form queueing networks. The low computational expense is due to extensions, such as the Schweitzer approximation [22, 3, 4], that replace the exact equations which are recursive in the customer class populations with approximate equations that are solved iteratively, typically within a very small number of iterations. The high degree of accuracy is largely due to heuristic extensions for representing a number of important system features such as priority queueing disciplines, simultaneous resource possession, and FCFS queues with class-dependent mean service times [14].

*This research is supported in part by the Natural Sciences and Engineering Research Council of Canada under Grant OGP-0000264, DARPA/ITO under Contract N66001-97-C-8533, and the National Science Foundation under Grants CDA-9623632, MIP-9625558, and ACI 9619019.

The work in this paper is motivated by a recent highly efficient heuristic AMVA model for evaluating shared memory architectures that contain complex modern processors [24]. In that architecture model, each processor is modeled by a FCFS queue. Service times at the processor represent the time between memory requests that miss in the second level cache when the processor is active. The measured coefficient of variation (CV) of these times [24] was as high as 13 for the benchmarks and architecture that were modeled. For several of the benchmarks, Figure 1 shows the throughput (in units of instructions per cycle, IPC) obtained in [24] by (1) a detailed architecture simulator called RSIM, (2) the AMVA model with the standard AMVA approximation for FCFS centers with high service time CV [14], (3) the AMVA model with a new simple heuristic interpolation ("simple interp") for estimating the mean residual service time of the customer in service at an arrival instant at any of the processors. Note that the standard AMVA model provides system throughput estimates that have large error compared to the RSIM estimates.

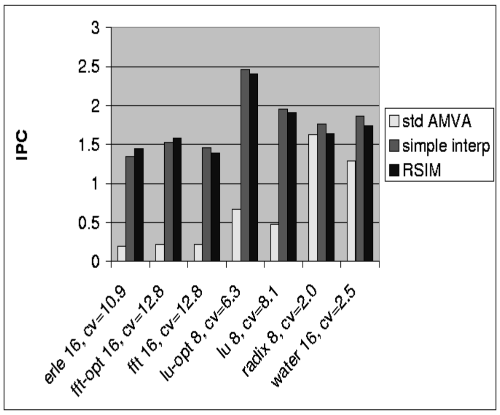


Figure 1: Architecture Throughput Estimates

The simple interpolation was found to be sufficiently accurate for evaluating the shared memory architecture performance over a fairly broad region of the design space [24], but the accuracy of the interpolation has not been investigated

Table 1: Notation

term	definition	introduced
N	number of customers in the (closed) network	Section 2
τ	mean service time at a queueing center	
CV_τ	coefficient of variation of service time	
L	mean residual service time at a random instant in time	
r	mean residual service time at an arrival instant	
Q	mean queue length at a queueing center	
R	mean residence time at a queueing center	
R_{other}	mean residence time in the rest of the queueing network	
R_{total}	mean total residence time in the queueing network	
U	server utilization at a queueing center	
τ_a	mean service time in path a of a hyperexponential distribution	
τ_b	mean service time in path b of a hyperexponential distribution	
p	probability of taking path a of a hyperexponential distribution	
S_d	mean service time at "other" delay center in the two-center network	Section 3
S_q	mean service time at "other" queueing center in the two-center network	
R_a	mean residence time at exponential queueing center with mean service time τ_a	
R_b	mean residence time at exponential queueing center with mean service time τ_b	
k	mean number of arrivals within a burst	Section 4
I	mean interarrival time within a burst	
B	mean time between bursts	
S_{down}	mean service time at downstream queueing center	
Q_b	mean queue length at downstream center during a burst	
Q_{nb}	mean queue length at downstream center during time intervals between bursts	

systematically. Furthermore, the architecture model used standard AMVA equations for the "downstream" processor bus queue, i.e., the center next visited by requests departing from the processor queue. Mean waiting time for the bus was significantly underestimated by the model because, due to the high service time CV at the processor, the arrival process to the bus is significantly more bursty than is assumed in the standard AMVA equations. We are not aware of any prior MVA approximations for estimating mean queueing time at such servers.

A previous paper by Bondi and Whitt [6] provides insight into the behavior of closed queueing networks that have FCFS centers with service time CV greater than one. Their paper evaluates the standard AMVA approximation for such networks, as well as several convolution-based approximate solution methods that have better accuracy. This paper considers improved AMVA techniques, which can be employed in the common case that there are other non-separable system features that are not easily represented in the convolution framework. Specifically, this paper makes the following contributions:

- The simple interpolation used in the architecture model is provided.
- A new heuristic AMVA interpolation is developed for estimating the mean residual service time at an arrival instant for a FCFS queue.
- A new AMVA technique, "AMVA-Decomp", is developed for estimating mean residence time at a FCFS center with service time CV greater than one.
- A new AMVA technique is developed for modeling the bursty arrivals and estimating the mean waiting time at a resource that is "downstream" from a FCFS center with high service time variability.
- The accuracy of each of the above techniques is systematically investigated by comparing the approximate estimates against the exact solution for queueing networks with two service centers, over a large parameter space for such networks.

A key result of the evaluations in this paper is that using the AMVA-Decomp technique, together with the technique for modeling bursty arrivals at the downstream queue, provides overall mean system residence time estimates that have error typically less than 5-10% and always less than 15% over the system parameter space examined.

The rest of this paper is organized as follows. Section 2 discusses previous AMVA and AMVA-compatible techniques for estimating mean residence time at FCFS queues with high service time variability. Section 3 proposes new AMVA techniques for estimating mean residence time at these queues, and evaluates their accuracy. Section 4 develops and evaluates the new AMVA technique for modeling bursty arrivals and estimating mean residence time at a downstream queue. Finally, Section 5 contains the conclusions of this work.

2. Background: Previous Techniques

In this section, we review three previous techniques for estimating mean residence time at a FCFS center with high service time variability. These techniques are the "standard AMVA approximation" [14], a new simple interpolation used in the previous architecture model [24], and a decomposition technique proposed by Zahorjan et al. [31]. Other previous

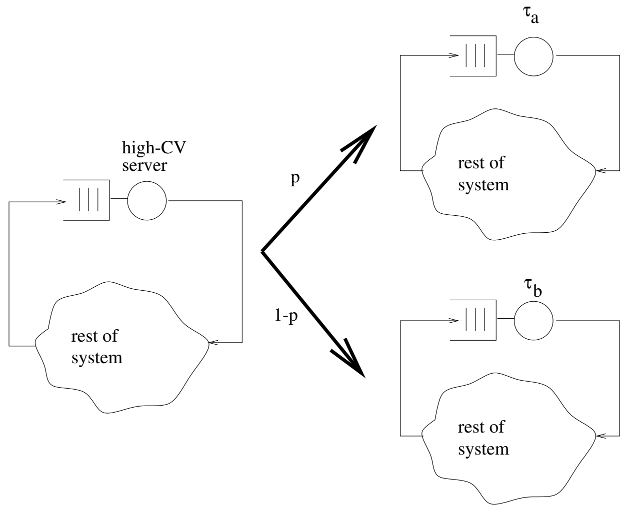


Figure 2: System Decomposition

approaches that are based on convolution or global balance are reviewed in [31, 6].

Throughout this section and the remainder of this paper, the AMVA techniques are defined and evaluated for single class queueing network models. There are extensions for some types of models with multiple customer classes, but evaluation of the accuracy of the approximations for multiple class models is beyond the scope of this paper. Without loss of generality, the techniques are defined assuming that the visit count for the given FCFS center with high service time variability is equal to one. Table 1 provides the notation that will be used throughout the paper.

2.1 The Standard AMVA Approximation

The Schweitzer AMVA equation [22, 3, 4] for the mean residence time R at a FCFS queueing center that has exponentially distributed service times with mean τ is:

$$R = \tau(1 + \frac{N-1}{N}Q) = \tau[1 + (N-1)\frac{R}{R_{total}}]. \quad (1)$$

In the above equation, Q denotes the mean queue length at the center, R_{total} denotes the mean total residence time in the queueing network, and N denotes the number of customers in the (closed) network.

When the service times at the FCFS center are not exponentially distributed and CV_τ represents the service time CV, the "standard" AMVA approximation for estimating mean residence time [14] is given by:

$$\begin{aligned} R &= \tau[1 + \frac{N-1}{N}(Q-U)] + \frac{N-1}{N}UL \\ &= \tau[1 + (N-1)\frac{R-\tau}{R_{total}}] + (N-1)\frac{\tau}{R_{total}}L, \quad (2) \end{aligned}$$

where U is the server utilization, and L is an estimate of the mean residual service time of the customer in service at an arrival instant assuming arrivals to the queue occur at random points in time, i.e., $L = \frac{\tau}{2}(1 + CV_\tau^2)$. To our knowledge, this heuristic approximation was first proposed by Reiser [21] in a paper that applied the approximation for solving a queueing network model containing FCFS centers with deterministic service times, in which case CV=0. The approximation has also been used in a number of other accurate models that contain FCFS centers with deterministic service times, such as those in [2, 8, 11, 17].

A problem with the accuracy of the above approximation arises for centers that have high service time CV. In this case, the estimated mean residual service time of the customer in service at a random point in time can be quite large (e.g., significantly larger than the mean service time at the queue). If the average residence time of a customer in the rest of the queueing network is smaller than this estimated mean residual service time, as was the case for memory requests in the architecture model discussed in Section 1, the customers do not arrive back at the high-CV center at a random point in time relative to the service times at the center. In this case, as will be shown in Section 3, the standard AMVA approximation (L) can greatly overestimate the mean residual service time at an arrival instant. This overestimation of the mean residual service time leads to a corresponding overestimation of the mean residence time at the center, which was the cause of the very pessimistic estimates of system throughput shown in Figure 1 for the model that used the standard AMVA approximation.

2.2 A Simple AMVA Interpolation

To develop a more accurate estimate of mean residual service time at an arrival instant for a FCFS queue with high service time CV, r, we define a simple interpolation between the two extremes, τ and L, where τ is the mean residual service time in the limiting case in which the time spent in the rest of the queueing network approaches zero, and L is the mean residual service time as would be seen by a random arrival. Letting R_{other} denote the mean residence time in the rest of the network, the simple interpolation is given by:

$$r \approx \frac{L}{L + R_{other}}\tau + \frac{R_{other}}{L + R_{other}}L. \quad (3)$$

This interpolation can be used in place of L in equation (2). The more accurate throughput estimates in Figure 1, produced by the "simple interp" model, were obtained using this simple interpolation at the processor queues.

2.3 The Decomposition Technique

The decomposition technique of Zahorjan et al. [31] estimates the performance measures for queueing networks with high-CV FCFS queues using weighted averages of those performance measures for simpler models. This technique is compatible with AMVA since each of these simpler models can be analyzed using standard AMVA techniques. In addition to providing accurate mean residence time estimates for high-CV FCFS queues, *this approach also approximately captures the impact of high service variability on other centers in the network.*

Consider a closed queueing network model in which there is one high-CV FCFS queue with a service time distribution that can be modeled with a two-stage hyperexponential distribution, with parameters p, τ_a, and τ_b. With probability p a given customer's service time is exponentially distributed with mean τ_a, and with probability $1-p$ it is exponentially distributed with mean τ_b. We assume, without loss of generality, that $\tau_a < \tau_b$, and we denote the average mean service time by $\tau = p\tau_a + (1-p)\tau_b$.

The technique of Zahorjan et al. decomposes this model into two simpler models, as shown in Figure 2. The simpler models are identical to the original model except that, in one of them, customers have mean service time τ_a at the FCFS center, while in the other they have mean service time τ_b. An estimate for the mean network residence time, or the mean residence time at a particular center, in the original model is given by the sum of p times the corresponding mean residence time in the first model and $1-p$ times the corresponding mean residence time in the second model.

The definitions of the simpler models, and the manner in which their performance measures are used to estimate those for the original model, are based on consideration of the transition rate matrices of such models and the use of the theory of near-complete decomposability [9]. This theory actually suggests a slightly different definition of the simpler models, but in the case where the theory directly applies (i.e., $p >> 1-p$ and $\tau_a << \tau_b$), the results are identical. Furthermore, Zahorjan et al. show that, with the simpler models defined as above, accurate results are obtained even when the theory does not strictly apply [31].

The decomposition technique has two key advantages. First, it has a firm theoretical foundation provided by the theory of near-complete decomposability [9], as explored in detail in [31]. Second, it was found to have high accuracy in [31].

The technique can be extended to networks with multiple high-CV FCFS queues and to high-CV service time distributions other than the two-stage hyperexponential distribution [31]. For example, general Coxian distributions [10, 13], in which there are a number of exponential stages of service connected by transition paths that have fixed probabilities, may be modeled by decomposing the original model into a number of simpler models equal to the number of paths. In some modeling applications, these more general distributions may be better able than the two-stage hyperexponential to capture important distributional characteristics of highly variable service times [15].

A principal disadvantage of the decomposition technique is the complexity of solving the model. For a model that contains H FCFS queues with service time distributions modeled by two-stage hyperexponential distributions, 2^H simpler models need be analyzed, one for each possible combination of service stages for each of the H centers. The complexity of the approach is increased when modeling more complex service time distributions. For some applications, particularly those that have many FCFS queues with high service time variability and also use decomposition for analyzing other non-product form system features (e.g., the model in [24]), the exponential cost in the number of high-CV FCFS queues may render the approach impractical.

3. FCFS Centers with High Service Time CV

In this section, we explore the accuracy of AMVA techniques for estimating mean residence time at a FCFS queue with high service time variability. The results in this section will show that the simple AMVA interpolation defined in Section 2.2 is considerably more accurate than the standard AMVA approximation, but that the error for the simple interpolation can be significant. Therefore, two new AMVA techniques will also be developed and evaluated.

The first new technique, defined in Section 3.1, is an improved interpolation for estimating the mean residual service time of the customer in service at an arrival instant. We evaluate the accuracy of the new interpolation, the previous simple interpolation, and the standard AMVA estimate of mean residual service time in Section 3.2.

The second proposed new technique, defined in Section 3.3, is a new heuristic method for estimating the mean center residence time. This technique, "AMVA-Decomp", is inspired by the decomposition method reviewed in Section 2.3. Section 3.4 compares the accuracy of the mean center residence time estimates obtained using this new "AMVA-Decomp" technique, the new interpolation for mean residual service time, and the previous techniques reviewed in Section 2.

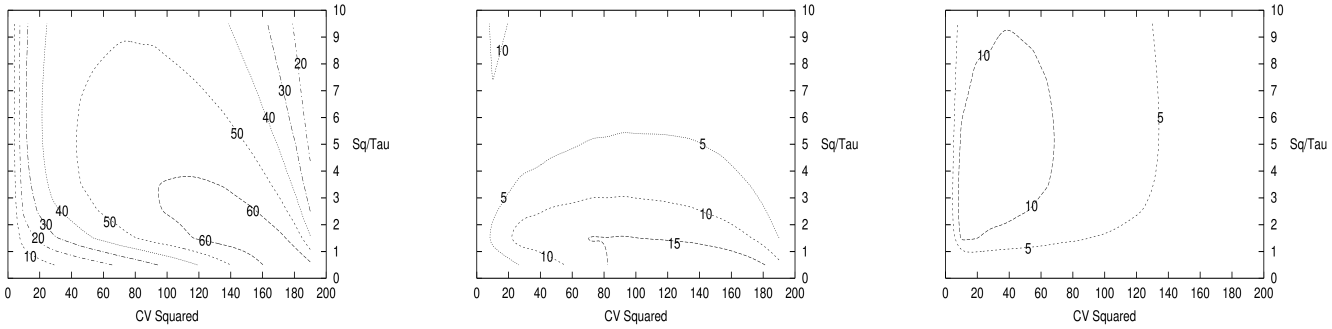
The techniques will be systematically evaluated using simple networks with two service centers. One center is a FCFS queue with service times modeled by a two-stage hyperexponential distribution that has mean τ and coefficient of variation CV_τ. The "other" service center, which has exponentially distributed service times, abstractly models customer sojourn times in the rest of the system. We consider two extreme cases of customer interference at this other service center. In one case, the other service center is a pure delay center where all customers receive service in parallel. In the second case, the other service center is a single-server queue with an arbitrary work-conserving scheduling discipline that is oblivious to actual customer service requirements. The mean service time at the other center is denoted by S_d or S_q, respectively. For these two-queue networks, we can use Markov chain techniques to compute the exact mean residual service time and mean queue residence time at the FCFS center with the high service time CV. More importantly, these networks are simple enough that we can explore the system parameter space fairly completely. This allows us to determine regions of the parameter space for which a given technique is least accurate.[1]

3.1 The New AMVA Interpolation

In this section, we define a new interpolation for estimating the mean residual service time at an arrival instant for a FCFS center that has a service time distribution that is reasonably well approximated by a two-stage hyperexponential distribution. As before, the parameters of the hyperexponential distribution are denoted by τ_a, τ_b, and p, such that $\tau = p\tau_a + (1-p)\tau_b$. The improved interpolation is obtained by replacing L in the weighting factors in equation (3) by $T = \frac{\tau_a \tau_b}{\tau}$, as follows:

$$r \approx \frac{T}{T + R_{other}}\tau + \frac{R_{other}}{T + R_{other}}\frac{\tau}{2}(1 + CV_\tau^2). \qquad (4)$$

[1]Evaluations for larger networks and for FCFS centers with other high-CV service time distributions can be found in [7]. Those evaluations show relative accuracies of each technique similar to the results in this paper.



N=2 N=5 N=20

Figure 3: % Relative Error of the New Interpolation for Mean Residual Service Time (two-center networks, "other" center is a queueing center, p=0.99)

This interpolation has the key property that it is exact when the mean delay in the rest of the network is exponentially distributed with mean R_{other}. To see this, consider a particular customer, named A, that is not at the high-CV FCFS center when another customer, named B, enters service at this center. If B has mean service time equal to τ_a, then with probability $\frac{\frac{1}{R_{other}}}{\frac{1}{\tau_a}+\frac{1}{R_{other}}}$, A will arrive at the high CV center before B completes service and, in this case, the mean residual life is τ_a. Similarly, if B has mean service time equal to τ_b, then with probability $\frac{\frac{1}{R_{other}}}{\frac{1}{\tau_b}+\frac{1}{R_{other}}}$, A will arrive at the high CV center before B completes service, and in this case the mean residual service time is τ_b. Thus,

$$\begin{aligned} r &= p\left[\frac{\frac{1}{R_{other}}}{\frac{1}{\tau_a}+\frac{1}{R_{other}}}\tau_a + \frac{\frac{1}{\tau_a}}{\frac{1}{\tau_a}+\frac{1}{R_{other}}}r\right] \\ &+ (1-p)\left[\frac{\frac{1}{R_{other}}}{\frac{1}{\tau_b}+\frac{1}{R_{other}}}\tau_b + \frac{\frac{1}{\tau_b}}{\frac{1}{\tau_b}+\frac{1}{R_{other}}}r\right], \end{aligned} \quad (5)$$

which reduces to the interpolation given in equation (4).

3.2 Mean Residual Service Time Accuracy

Table 2 provides results that illustrate the typical accuracy of each of three AMVA techniques for estimating the mean residual service time at a FCFS center with high service time CV: (1) the standard AMVA approximation, (2) the simple interpolation given in equation 3, and (3) the new interpolation given in equation 4 ("New Interp.").

The results in the table are for several parameter sets for the simple two-center networks with network population equal to 5 customers. Actual values of the mean residual service time are derived from numerical solutions of the corresponding Markov chains. For Table 2(a), in which the second center is a delay center, the new interpolation is exact as established in Section 3.1. In Table 2(b), the exact value of the mean residence time at the other queueing center (R_{other}), as obtained from the Markov chain analysis, is used in the interpolation formulas. Section 4 develops an accurate new AMVA technique for estimating this mean residence time.

In all cases shown in Table 2, the new interpolation is more accurate than the simple interpolation defined in Section 2.2, which is in turn more accurate than the "standard" AMVA estimate. Notably, the previous standard AMVA estimate can be extremely inaccurate (e.g., more than 1000% error). There are also cases where the simple interpolation overestimates the mean residual service time by more than 100%.

Table 2: Mean Residual Service Time Estimates (two-center networks, N=5, p=0.99, τ=50)

(a) "Other" Center is a Delay Center

CV_τ^2	S_d/τ	Std. AMVA	Simple Interp.	New Interp./ and Actual
10	0.5	275	77.8	56.3
10	2	275	181.2	73.2
10	10	275	252.2	132.1
100	0.5	2525	87.0	108.1
100	2	2525	340.4	267.1
100	10	2525	1249.9	853.7

(b) "Other" Center is a Queueing Center

CV_τ^2	S_q/τ	Std. AMVA	Simple Interp.	New Interp.	Actual
10	0.5	275	86.6	63.0	62.0
10	2	275	187.4	124.6	117.8
10	10	275	252.2	215.9	231.0
100	0.5	2525	141.0	260.0	221.0
100	2	2525	440.6	824.0	727.4
100	10	2525	1269.0	1787.7	1777.3

To further investigate the reliability of the new interpolation, Figure 3 provides relative error contours over a large region of the parameter space for the two-center networks in which the second center is a queueing center. To obtain the contours, the percent error values were computed at a regular spacing equal to 9 on the x-axis (starting at $CV^2 = 1$ and ending at 190), and at a regular spacing of 1 on the y-axis (starting at 0.5 and ending at 9.5). The contours were computed using gnuplot [30]. Note that the range of CV^2 values on the x-axis covers the range of processor service time CV values observed in the architecture model benchmarks (see Figure 1). Each contour line corresponds to a particular absolute value of percent relative error for the mean residual service time estimated using the new interpolation. As in Table 2(b), the exact value of the mean residence time at the other queueing center is used in the interpolation formula.

Note that the new interpolation yields results within 15% of the exact values over large regions of the two-queue system parameter space. The interpolation is inaccurate only for quite small N and maximal interference in the rest of the network (i.e., in unlikely contexts). Should this case be of interest, however, the accuracy can be greatly improved through a modification of equation (4) in which R_{other} is computed for the network with one fewer customer, rather than for the full population. (That is, since the mean residual service time is conditioned on at least one customer being at the high-CV center, at most $N-1$ customers can be in the remainder of the network.) With this modified version of equation (4), accuracy is significantly improved for small N. In particular, this method gives exact results for the $N = 2$ case considered in Figure 3. Moreover, a simple approximation for R_{other} with population $N-1$ is sufficient to achieve these accuracy improvements. For example, it is sufficient to compute an estimate of $R_{other}(N-1)$ by letting the arrival instant mean queue length with population $N-1$ be approximated by $\frac{N-2}{N-1}\frac{N-1}{N}Q(N) = \frac{N-2}{N}Q(N)$ for each queueing center in the rest of the network. (In the case of the two-queue model, there is only one such center.)

Although the new interpolation estimate of mean residual service time is quite accurate, using this value in equation (2) can give inaccurate estimates of the mean center residence time, because the standard AMVA estimates of the mean queue length at an arrival instant (i.e., $(N-1)\frac{R-\tau}{R_{total}}$) and the probability that an arrival finds the server busy (i.e., $(N-1)\frac{\tau}{R_{total}}$) can be inaccurate for a FCFS center with high service time CV. This motivates the new AMVA technique for estimating mean queue residence time developed next.

3.3 The New AMVA-Decomp Technique

A key hypothesis leads us to adapt the decomposition technique reviewed in Section 2.3, to the AMVA context. That is, it may be sufficiently accurate to apply the decomposition only at the level of the individual center at which there is a high service time CV. Thus, for a FCFS center with a service time distribution modeled by the two-stage hyperexponential distribution defined previously, we estimate the mean residence time at that center, R, using:

$$R = pR_a + (1-p)R_b, \tag{6}$$

where

$$R_a = \tau_a(1 + \frac{N-1}{N}Q_a), \tag{7}$$

$$R_b = \tau_b(1 + \frac{N-1}{N}Q_b), \tag{8}$$

$$Q_a = N\frac{R_a}{R_a + R_{other}}, \tag{9}$$

$$Q_b = N\frac{R_b}{R_b + R_{other}}. \tag{10}$$

In the above equations, R_{other} is the mean total residence time spent at the other centers in the network, which is computed iteratively together with R within the usual AMVA iterative solution framework.

Note that the above approach yields identical results to the Zahorjan et al. decomposition technique, given that the two simpler models of the latter technique are solved using AMVA, if the mean total residence time at the other centers in the network is identical in each of the two simpler models. In Section 3.4, we examine the accuracy of this approximation for two-center networks, including systems for which this property does not hold.

A principal advantage of this new AMVA technique is that there is no need to solve 2^H separate models to obtain the solution for a system that includes H FCFS centers with high service time variability. Only one model is solved, with the above modified mean residence time equations at each of the H centers.

As with the decomposition technique, the above technique is easily extended to FCFS servers with general Coxian service time distributions [10, 13]. In this case, the mean residence time at the high-CV FCFS center is expressed as a weighted sum of conditional mean residence times, with one term for each path through the stages of service defining the Coxian distribution, and weight equal to the probability of following the path. For a given path consisting of multiple (exponential) stages of service, the mean residence time can be estimated using the standard (and quite accurate) AMVA approximation for service times with low variability. As in the case of the two-stage hyperexponential service time distribution, the average residence time in the rest of the system is assumed to be the same regardless of which path is active. Investigating the accuracy that is achieved for such service time distributions is beyond the scope of this paper, but results in [7] indicate that the accuracy is very similar to the results reported in this paper for the two-stage hyperexponential distribution.

3.4 Mean Queue Residence Time Accuracy

Table 3 provides typical results for the accuracy of five techniques for estimating the mean residence time at a FCFS center with high service time variability. Those techniques are: (1) the standard AMVA technique, (2) use of the simple interpolation to estimate mean residual service time in equation 2, (3) use of the new interpolation to estimate mean residual service time in equation 2, (4) the new technique proposed in Section 3.3 ("AMVA-decomp"), and (5) the decomposition approach ("Decomp.") [31].

The techniques are compared for the same two-queue network parameter sets that were used in Table 2. The exact values for the mean residence time at the FCFS center with high CV service times are derived from numerical solutions of the corresponding Markov chains. For Table 3(b), the exact value of R_{other} is used in the calculations for all of the techniques except for the decomposition approach, in which this quantity is not used. R_{other} could instead be computed using the accurate approximation developed in Section 4.

For the models in which the "other" center is a delay center, the AMVA-Decomp approximation yields the same results as the decomposition approach on which it is based, since R_{other} is identical in the decomposed submodels. For the networks in which the second center is a queueing center,

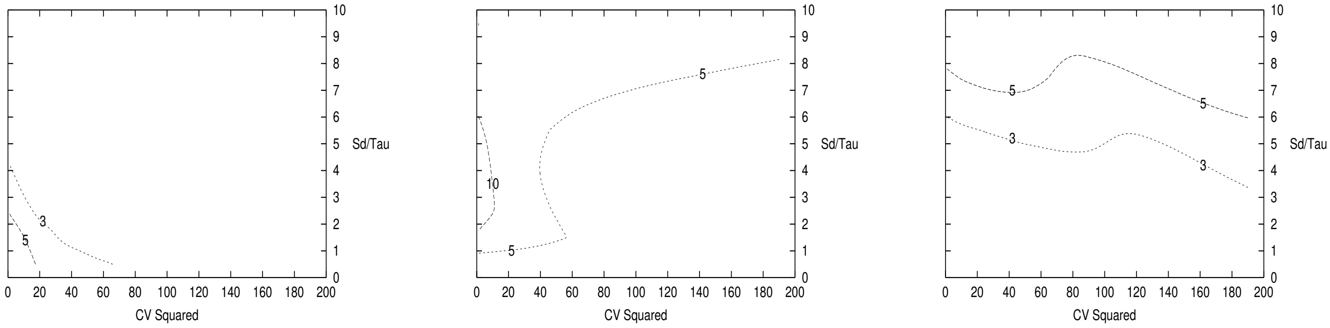
Table 3: Mean Residence Time Estimates for FCFS Queue with High Service Time CV (two-center networks, $N = 5$, p=0.99, τ=50)

(a) "Other" Center is a Delay Center

CV_τ^2	S_d/τ	Std. AMVA	Simple Interp.	New Interp.	AMVA-Decomp. and Decomp.	Actual
10	0.5	355.2	242.2	235.7	230.5	225.1
10	2	310.9	156.3	198.6	180.7	164.8
10	10	167.6	71.8	113.8	105.7	96.1
100	0.5	825.7	249.1	272.2	231.7	226.2
100	2	786.3	234.2	307.7	203.6	199.3
100	10	606.8	157.1	323.8	189.9	177.3

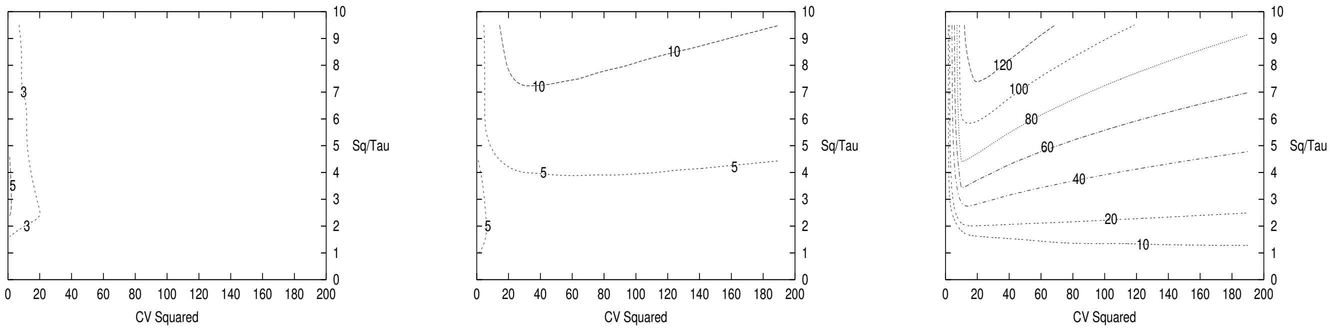
(b) "Other" Center is a Queueing Center

CV_τ^2	S_q / τ	Std. AMVA	Simple Interp.	New Interp.	AMVA-Decomp.	Decomp.	Actual
10	0.5	337.8	238.5	220.5	210.0	213.9	204.5
10	2	182.8	148.6	120.8	109.3	110.6	112.2
10	10	73.8	71.8	68.7	79.6	90.8	73.2
100	0.5	787.9	246.9	306.4	204.2	203.5	206.0
100	2	617.2	226.4	325.4	190.2	193.1	186.2
100	10	251.8	155.5	196.1	173.3	187.9	154.8



$N = 2$ N=5 N=20

Figure 4: % Relative Error of New AMVA-Decomp Mean Residence Time Estimate (two-center networks, "other" center is a delay center, p=0.99)



N=2 N=5 N=20

Figure 5: % Relative Error of New AMVA-Decomp Mean Residence Time Estimate (two-center networks, "other" center is a queueing center, p=0.99)

we might expect the AMVA-Decomp technique to be less accurate, as we would not expect the average residence time in the rest of the system (i.e., the mean residence time at the second queueing center) to be the same in each of the decomposed submodels. However, for the cases considered in Table 3, the AMVA-Decomp technique appears to be quite accurate in spite of this possibility for error.

The accuracy of the AMVA-Decomp approach is evaluated over a wide range of the parameter space of the two-center networks in the contour plots of Figures 4 and 5. In Figure 5, as in Table 3(b), the exact value of the mean residence time at the other queueing center is used where needed. The key conclusions from Table 3 and Figures 4 and 5 are:

- The standard AMVA estimate of mean queue residence time is not very robust.
- The interpolation techniques can also yield inaccurate estimates of the mean queue residence time, since the standard AMVA estimates of the mean arrival queue length and the probability that the server is busy at an arrival instant are inaccurate.
- For most regions of the system parameter space, the new AMVA-Decomp technique yields estimates of mean residence time that have under 10% error, which is similar to the accuracy provided by the significantly more costly decomposition approach. Note, however, that we have not yet established that AMVA can estimate the mean residence times at downstream centers as accurately as can the decomposition approach; this question is addressed in Section 4.

The one context in which the accuracy of AMVA-Decomp is substantially poorer than the decomposition technique is the case, illustrated in Figure 5, of large N and $\frac{S_q}{\tau} > 2$. However, the accuracy of the estimated mean residence time at the high-CV center may not be very important in this case, as the average residence time in the system is dominated by the queueing delay at the "other" center. Section 4 (in particular, Figure 8) will show that the overall mean system residence time is accurately predicted for these cases where the AMVA-Decomp estimate of mean residence time at the FCFS queue with high CV is inaccurate.

In some cases, the simple interpolation technique for estimating mean residual service time yields a more accurate estimate of mean residence time in the queue than the new interpolation technique. This is consistent with the observations in [24] that the simple interpolation was useful for the architecture model. However, since the simple interpolation is less accurate than the new interpolation in estimating the mean residual service time, the cases where it leads to higher accuracy in predicting mean queue residence time are due to fortuitously compensating errors in the standard approximation of mean arrival queue length and/or the probability that the server is busy at an arrival instant for the center. Furthermore, the results for the simple interpolation in Figure 1 are perhaps more accurate than one would expect from the results in Table 3(b). This is due to the fact that, in cases where the simple interpolation overpredicts mean residence time, the error is partially compensated in the throughput

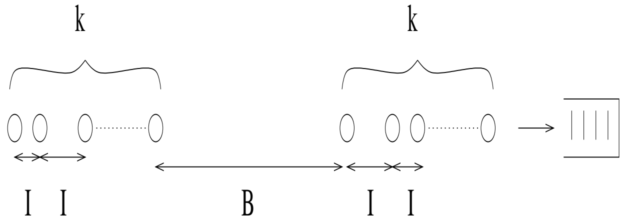


Figure 6: Model of a Bursty Arrival Process

estimate because, as noted in Section 1, the bursty arrivals at the downstream queues were not modeled.

The next section considers how to estimate mean residence time at a downstream queue that is visited by customers departing from the FCFS center with high service time CV.

4. Downstream Center Residence Time

The AMVA approximations for modelling non-exponential service times at FCFS queues consider only the impact of the service time distribution on the local mean queue residence time. When service times are highly variable, however, there may be a substantial impact on the "downstream" centers. In particular, as noted in [6], if the server utilization is at least moderate, a FCFS center with high service time variability generates bursty departures, leading to bursty arrivals and increased queueing at downstream centers.

In this section, we develop a new AMVA technique that captures the impact of bursty arrivals on the mean residence time at a center downstream from a high-CV FCFS center. Section 4.1 develops a model of the bursty arrivals, and Section 4.2 develops the new AMVA estimate of mean residence time assuming the proposed model of the bursty arrivals.

4.1 A Model of the Bursty Arrivals

The arrival process at a center downstream from a high-CV FCFS service center is modeled as consisting of bursts of customer arrivals, with relatively short interarrival times during a burst and relatively long gaps between bursts. The inter-burst gaps, and the interarrival times within a burst, are modeled with exponential distributions (with different means). The number of customer arrivals within a burst is modeled with a geometric distribution.[2]

Figure 6 illustrates the following three parameters that are used to characterize the arrival process in this model:
k: the average number of customer arrivals within a burst,
I: the mean interarrival time within a burst, and
B: the mean time between bursts.

The value of I is (heuristically) determined from the service time distribution(s) at the center(s) that generates the arrivals. For example, consider the simple case in which arrivals are generated by departures from a single FCFS queueing center with service times modeled by a two-stage

[2]Fairly straightforward generalizations of the analysis are possible for other distributions of interarrival times within a burst (such as deterministic) and other distributions of the number of arrivals within a burst.

hyperexponential distribution. In this case, I is equal to the smaller of the mean service times of the two stages.

Let X denote the overall arrival rate, equal to the center throughput which is iteratively computed during the iterative AMVA solution. If CV_a^2 denotes the squared coefficient of variation of interarrival times (to be derived below), k and B can be computed from the following two equations:

$$\frac{k}{(k-1)I+B} = X, \tag{11}$$

$$\frac{\frac{k-1}{k}2I^2 + \frac{1}{k}2B^2}{(\frac{(k-1)I+B}{k})^2} - 1 = CV_a^2. \tag{12}$$

To estimate CV_a^2, we employ the method proposed by Sevcik et al. [23] for approximating the arrival processes within a general network of queues. In this method, CV_a^2 is approximated by a simple function of the routing probabilities, utilizations, and the coefficients of variation of the interarrival and service times at the "upstream" centers whose departing customers may next visit the center of interest. For example, consider again the case that a single FCFS queueing center with high-CV service times generates arrivals to a single "downstream" center. Assuming that the arrivals to the upstream center are not substantially more (or less) bursty than Poisson, an approximation for CV_a^2 at the downstream center, expressed in terms of the squared coefficient of variation of service times, CV_τ^2, at the upstream center and the utilization, U, of the upstream center, is as follows [23]:

$$CV_a^2 = 1 + U^2(CV_\tau^2 - 1). \tag{13}$$

In a closed network, the length of the queue is bounded by the size of the customer population. This is the well-known "limited damage" argument (perhaps first articulated by Buzen, as referenced in [15]) concerning the impact of highly variable service times in closed networks. Applying this limited damage argument to the bursty arrival model, we constrain k so that, for an arrival that occurs during a burst, the average number of customers that arrived previously in the burst plus the average number found at the downstream center by the first arrival in the burst is at most $N-1$. Since the burst size in the arrival model is geometrically distributed with mean k, on average there are $k-1$ prior customer arrivals within the burst. Letting Q_{nb} denote the mean queue length during the time intervals between bursts (to be derived in Section 4.2), and using the standard AMVA approximation for the arrival instant mean queue length seen by the first arrival in a burst, we obtain:

$$k \leq N - \frac{N-1}{N}Q_{nb}. \tag{14}$$

If the bound on the value of k is lower than the value of k determined by equations 11 and 12, k is set equal to the bound and the value of B is computed from equation (11), so as to ensure that the basic "arrival rate = throughput" constraint is satisfied.

4.2 New Mean Residence Time Estimate

The proposed new technique for estimating mean residence time at a queueing center downstream from a high-CV FCFS queue employs the model of bursty arrivals described in Section 4.1. The analysis below uses the arrival model that has exponential inter-burst gaps, exponential interarrival times within a burst, and geometric number of customer arrivals within a burst, which has parameters k, I, and B.

For simplicity, the mean residence time approximation is developed for the case of exponentially distributed service times at the downstream queue, although it can be modified for other service time distributions. For clarity and without loss of generality, we also assume that the visit count at the center of interest (i.e., the downstream queueing center in this case) is equal to one.

To develop the mean residence time approximation, we make the assumption that the downstream center never idles during a burst of arrivals. We expect that the assumption will be fairly reasonable in many if not most cases where burstiness in the arrival process has significant impact, due to two key observations. First, in the cases where burstiness is most pronounced, $I << B$. Second, if $I << B$ and k is reasonably small (e.g., due to k being constrained by the size of the network customer population) then if there is substantial queueing at the downstream center, I can be expected to be significantly smaller than the mean service time at the downstream center, S_{down}. Conversely, if $I > S_{down}$, we can expect that either there is not much queueing at the downstream center or the arrivals are not very bursty. In that case, R can simply be estimated using equation (1).

Under the assumption that the downstream queueing center never idles during a burst of arrivals, the residence time of a customer is equal to the sum of (1) the customer's own service time, (2) the service times of those customers found at the center by the first arrival in the burst (less any service time already acquired by the customer in service at the lead arrival instant), and (3) the service times of the prior customers within the same burst, minus the time from the start of the burst until the customer's arrival.[3] Since on average there are $k-1$ prior arrivals within the burst, the average time from the start of the burst until the customer's arrival is $(k-1)I$. Thus,

$$R = S_{down}\left(1 + \frac{N-1}{N}Q_{nb} + (k-1)\right) - (k-1)I, \tag{15}$$

where Q_{nb} is the mean queue length during time intervals between bursts. Note that this equation only makes sense if $I < S_{down}$. As noted above, if $I > S_{down}$, R can be estimated using equation (1).

To obtain an expression for Q_{nb}, we first note that there is a simple relationship between the overall mean queue length ($Q = RX$, as computed during the iterative AMVA solution), Q_{nb}, and the mean queue length during a burst, Q_b. Since the average duration of a burst is $(k-1)I$, and the average time between bursts is B, we have:

$$Q = Q_{nb}\frac{B}{B+(k-1)I} + Q_b\frac{(k-1)I}{B+(k-1)I}. \tag{16}$$

Furthermore, the assumption that the center never idles in the midst of a burst of arrivals allows Q_b to be written in

[3] Our description assumes FCFS service, although if service times are exponentially distributed, any work conserving scheduling discipline that is oblivious to actual customer service requirements will give the same mean residence time.

Table 4: Estimates of Mean Residence Time at Queueing Center with Bursty Arrivals (two-center models with queueing center, N=5, p=0.99, τ=50)

CV_τ^2	S_q/τ	Std. AMVA	AMVA-Decomp.	AMVA-Decomp.-Bursty	Decomp.	Actual
10	0.5	33.8	40.7	40.7	47.8	53.4
10	2	353.8	416.3	464.6	452.2	431.5
10	10	2441.4	2436.8	2495.0	2449.6	2439.4
100	0.5	28.3	40.4	78.6	104.1	96.4
100	2	176.4	361.3	482.1	483.0	473.2
100	10	2293.4	2362.8	2496.2	2468.8	2450.7

Table 5: Estimates of Mean System Residence Time (two-center models with queueing center, N=5, p=0.99, τ=50)

CV_τ^2	S_q/τ	Std. AMVA	AMVA-Decomp.	AMVA-Decomp.-Bursty	Decomp.	Actual
10	0.5	383.5	259.6	259.6	261.7	257.9
10	2	558.1	526.5	572.0	562.8	543.7
10	10	2515.9	2516.3	2574.1	2540.4	2512.7
100	0.5	852.6	262.9	286.5	307.6	302.4
100	2	923.7	553.1	672.3	676.1	659.4
100	10	2558.3	2536.8	2669.2	2656.7	2605.5

terms of Q_{nb} and the parameters of the arrival model, as follows. Consider a burst that consists of j customer arrivals (and thus $j-1$ interarrival periods each of average length I). The mean queue length during this burst, $Q_{b|j}$, is given by the mean number in the queue at the beginning of the burst (Q_{nb}), plus the time average of the number of customers that arrive to the queue during the burst, minus the time average of the number of customers that depart the queue during the burst. Since the queue is draining at a rate of one unit of work per unit of time during the entire burst period, the time average of the number of customers that depart during the burst is equal to $\frac{1}{S_{down}}$ times the expected age, or residual life, of the burst at a random instant. Since the duration of a burst has a $j-1$ stage Erlang distribution, the second moment of the burst duration is $(j-1)jI^2$. Thus,

$$\begin{aligned} Q_{b|j} &= Q_{nb} + \frac{\sum_{i=1}^{j-1} iI}{(j-1)I} - \frac{1}{S_{down}} \frac{(j-1)jI^2}{2(j-1)I} \\ &= Q_{nb} + \frac{j}{2} - \frac{jI}{2S_{down}}. \end{aligned} \tag{17}$$

The equation for Q_b can be obtained by forming a weighted average with the above expression, where the weight for the j^{th} term is the fraction of the time occupied by arrival bursts that consists of bursts of size j, and is given by the probability of a burst of size j multiplied by $\frac{(j-1)I}{(k-1)I}$. For geometrically distributed burst sizes, this yields, after simplification:

$$Q_b = Q_{nb} + k - \frac{kI}{S_{down}}. \tag{18}$$

Equations (11), (16), and (18) yield:

$$Q_{nb} = Q - X\frac{I(k-1)(S_{down} - I)}{S_{down}}. \tag{19}$$

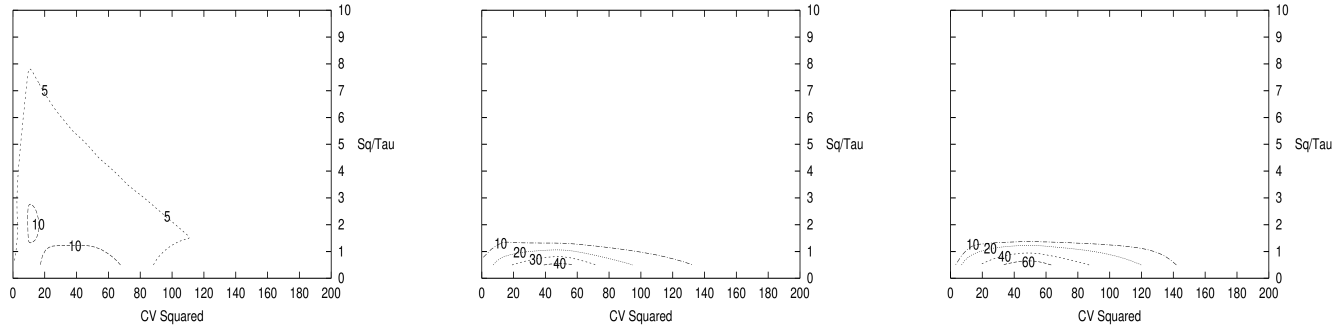
4.3 Validation Results

Tables 4 and 5 provide results that illustrate the typical accuracy of four approaches for capturing the performance impact of highly variable service times at a FCFS center. These approaches are: (1) the standard AMVA approximation at the high-CV FCFS center, with no attempt to model the bursty arrivals generated downstream, (2) the approach proposed in Section 3.3 for estimating mean residence time at the high-CV FCFS center with no attempt to model the bursty arrivals ("AMVA-Decomp"), (3) the approach proposed in Section 3.3 for mean residence time at the high-CV FCFS center together with the AMVA technique proposed in this section for estimating mean residence time at the downstream queueing center ("AMVA-Decomp-Bursty"), and (4) the decomposition approach of Zahorjan et al. ("Decomp.") [31].

Table 4 provides, for each technique, the estimates of the mean residence time at the downstream queueing center with bursty arrivals, while Table 5 provides the estimates of the mean total system residence time. These tables are for the same two-center model configurations that were used in Tables 2(b) and 3(b). As before, exact values of the performance metrics are derived from numerical solution of the corresponding Markov chains.

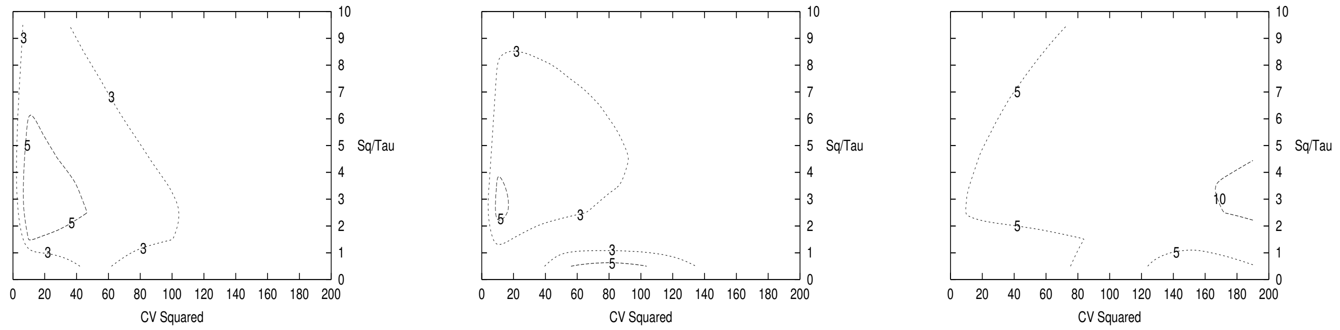
The reliability of the AMVA-Decomp-Bursty approach, which uses the decomposition-based AMVA approximation for estimating mean residence time at the high-CV center and the proposed AMVA technique for modelling bursty arrivals at the downstream queue, is explored more fully for the two-center systems in Figures 7 and 8. These figures show contours for the absolute value of the percent relative error in the mean residence time at the center with bursty arrivals, and the mean system residence time, respectively.

As shown in Table 4 and Figure 7, the combined use of the new decomposition-based AMVA technique and the proposed technique for modelling bursty arrivals at the downstream center generally yields an accurate estimate of mean residence time at a queueing center downstream from a high-



N=2 N=5 N=20

Figure 7: % Relative Error of AMVA-Decomp-Bursty: Bursty Arrival Center Mean Residence Time (two-center models with queueing center, p=0.99)



N=2 N=5 N=20

Figure 8: % Relative Error of AMVA-Decomp-Bursty: Mean System Residence Time (two-center models with queueing center, p=0.99)

CV FCFS center. The only cases in which relative errors are large are when S_q/τ is small (and thus I is close to or larger than S_q). In such cases, the mean residence time at the downstream center is only a small contributor to the overall mean system residence time.

The results in Table 5 and Figure 8 show that, for the cases considered, the combined use of the new AMVA-Decomp technique and the new technique for modeling bursty arrivals yields mean system residence time estimates that are within 10% of the exact values over much of the two-queue model parameter space. Thus, these methods together have accuracy similar to the more costly decomposition approach.

5. Conclusions

This paper has examined AMVA approximations for FCFS queues with non-exponential service times. Prior case studies that we are aware of in which the previous standard AMVA technique has been applied successfully to such queues, have been only for the case of low service time variability. As also observed in [6], the results in this paper have shown that the standard AMVA approximation cannot be recommended for high service time CV.

To model FCFS centers with high service time variability, this paper proposes new AMVA approximations that capture both the impact of high service time variability on local queueing delays and the impact of such variability on queueing delays at downstream queueing centers. The new approximations are simpler to apply than the previously proposed AMVA-compatible decomposition technique in [31], particularly if the model includes multiple FCFS queues with high service time variability, or if decomposition is needed for other non-separable system features.

The new techniques have been evaluated using two-center queueing networks that represent a wide range of contexts in which the techniques might be applied. The results in this paper as well as in [7] show that the new AMVA-Decomp technique, together with the proposed model and analysis of bursty arrivals at the downstream queueing centers, can be expected to be quite accurate in practice for these contexts. In particular, the estimates of mean system residence time have less than 15% error over the entire parameter space considered in this paper, and less than 10% error over most of that parameter space. The results in this paper show that the new techniques greatly increase the range of applicability of AMVA as compared with the previous standard AMVA approximation for FCFS centers with high service time CV.

Future research includes: (1) developing and evaluating the accuracy of the new AMVA techniques for multi-class models, (2) further evaluating the AMVA-Decomp technique for

FCFS centers with other Coxian service time distributions that have high coefficients of variation, (3) extending the model of the bursty arrivals at the downstream center for more complex distributions at the upstream high-CV FCFS center and for the case that arrivals occur from multiple upstream queues or to multiple downstream queues.

Acknowledgments

We would like to thank John Zahorjan for his helpful comments on an earlier draft of this paper, and we would like to thank Jianjun Chen and Jin Zhang for their further validations of the AMVA-Decomp technique in [7].

6. References

[1] S. V. Adve, V. S. Adve, M. D. Hill, and M. K. Vernon. Comparison of Hardware and Software Cache Coherence Schemes. In *Proc. 18th Annual Int'l. Symp. on Computer Architecture*, pages 298–308, May 1991.

[2] V. Adve and M. K. Vernon. Performance Analysis of Multiprocessor Mesh Interconnection Networks with Deterministic Routing. *IEEE Trans. on Parallel and Distributed Systems*, 5(3):225–246, Mar. 1994.

[3] Y. Bard. A Model of Shared DASD and Multipathing. *Comm. ACM*, 23(10):564–572, Oct. 1980.

[4] Y. Bard. A Simple Approach to System Modeling. *Performance Evaluation*, 1(3):225–248, Aug. 1981.

[5] G. E. Bier and M. K. Vernon. Measurement and Prediction of Contention in Multiprocessor Operating Systems with Scientific Application Workloads. In *Proc. 1988 Int'l. Conf. on Supercomputing*, pages 9–15, July 1988.

[6] A. B. Bondi and W. Whitt. The Influence of Service-Time Variability in a Closed Network of Queues. *Performance Evaluation*, 6(3):219–234, Sept. 1986.

[7] J. Chen and J. Zhang. Further Evaluation of the Recent AMVA-Decomp Technique. Tech. Report 1413, Computer Sciences Dept., Univ. of Wisconsin - Madison, Mar. 2000.

[8] M. Chiang and G. Sohi. Evaluating Design Choices for Shared Bus Multiprocessors. *IEEE Trans. on Computers*, 41(3):297–317, Mar. 1992.

[9] P. J. Courtois. *Decomposability: Queueing and Computer System Applications*. Academic Press, New York, 1977.

[10] D. R. Cox. The Use of Complex Probabilities in the Theory of Stochastic Processes. In *Proc. Cambridge Phil. Soc. 51*, 1955.

[11] A. J. Field and P. G. Harrison. An Analytical Model of the Standard Coherent Interface 'SCI'. In *Proc. 1995 International Conference on Parallel Processing*, pages I:173–177, Aug. 1995.

[12] M. I. Frank, A. Agarwal, and M. K. Vernon. LoPC: Modeling Contention in Parallel Algorithms. In *Proc. 6th ACM SIGPLAN Symp. on Principles and Practices of Parallel Programming*, pages 276–287, June 1997.

[13] L. Kleinrock. *Queueing Systems Volume 1: Theory*. John Wiley and Sons, New York, 1975.

[14] E. Lazowska, J. Zahorjan, G. Graham, and K. Sevcik. *Quantitative System Performance, Computer System Analysis Using Queueing Network Models*. Prentice-Hall, Englewood Cliffs, NJ, 1984.

[15] E. D. Lazowska. The Use of Percentiles in Modeling CPU Service Time Distributions. In K.M. Chandy and M. Reiser (eds.), *Computer Performance,* North-Holland, pages 53–66, 1977.

[16] S. T. Leutenegger and M. K. Vernon. A Mean Value Performance Analysis of a New Multiprocessor Architecture. In *Proc. 1988 ACM SIGMETRICS Conf. on Measurement and Modeling of Computer Systems*, pages 167–176, May 1988.

[17] D. A. Menasce, O. I. Pentakalos, and Y. Yesha. An Analytic Model of Hierarchical Mass Storage Systems with Network-Attached Storage Devices. In *Proc. 1996 ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 180–189, May 1996.

[18] S. S. Owicki and A. Agarwal. Evaluating the Performance of Software Cache Coherence. In *Proc. 3rd Int'l. Conf. on Architectural Support for Programming Languages and Operating Systems*, pages 230–242, April 1989.

[19] J. Patel, M. J. Carey, and M. K. Vernon. Accurate Modeling of the Hybrid Hash Join Algorithm. In *Proc. 1994 ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 56–66, June 1994.

[20] X. Qin and J.-L. Baer. A Performance Evaluation of Cluster-Based Architectures. In *Proc. 1997 ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 237–247, June 1997.

[21] M. Reiser. A Queueing Network Analysis of Computer Communication Networks with Window Flow Control. *IEEE Trans. on Commun.*, 27(8):1199–1209, Aug. 1979.

[22] P. Schweitzer. Approximate Analysis of Multiclass Closed Networks of Queues. In *International Conference on Stochastic Control and Optimization*, 1979.

[23] K. Sevcik, A. Levy, S. Tripathi, and J. Zahorjan. Improving Approximations of Aggregated Queuing Network Subsystems. In K.M. Chandy and M. Reiser (eds.), *Computer Performance,* North-Holland, pages 1–22, 1977.

[24] D. Sorin, V. Pai, S. Adve, M. Vernon, and D. Wood. Analytic Evaluation of Shared-Memory Systems with ILP Processors. In *Proc. 25th Annual Int'l Symp. on Computer Architecture*, pages 180–191, July 1998.

[25] J. Torrellas, J. L. Hennessy, and T. Weil. Analysis of Critical Architectural and Program Parameters in a Hierarchical Shared Memory Multiprocessor. In *Proc. 1990 ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 163–172, May 1990.

[26] T. Tsuei and M. K. Vernon. A Model of Multiprocessor Memory and Bus Interference Validated by System Measurement. *IEEE Trans. on Parallel and Distributed Systems, Special Issue on Measurement and Evaluation of Parallel and Distributed Systems*, 3(6):712–727, Nov. 1992.

[27] M. K. Vernon, R. Jog, and G. Sohi. Performance Analysis of Hierarchical Cache-Coherent Multiprocessors. *Performance Evaluation*, 9(4):287–302, Aug. 1989.

[28] M. K. Vernon, E. D. Lazowska, and J. Zahorjan. An Accurate and Efficient Performance Analysis Technique for Multiprocessor Snooping Cache-Consistency Protocols. In *Proc. 15th Annual Int'l. Symp. on Computer Architecture*, pages 308–315, May 1988.

[29] G. M. Voelker, H. A. Jamrozik, M. K. Vernon, H. M. Levy, and E. D. Lazowska. Managing Server Load in Global Memory Systems. In *Proc. 1997 ACM Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 127–138, June 1997.

[30] T. Williams and C. Kelley. gnuplot: An Interactive Plotting Program. Manual, version 3.7, Dec. 1998.

[31] J. Zahorjan, E. Lazowska, and R. Garner. A Decomposition Approach to Modelling High Service Time Variability. *Performance Evaluation*, 3(1):35–54, Feb. 1983.