BEHAVIORAL ANALYSIS OF FLEXIBLE MANUFACTURING SYSTEMS WITH MULTIPLE CLASSES AND FINITE CAPACITIES [1]

Mustafa Yuzukirmizi
Kirikkale University, Turkey

Corresponding author: M. Yuzukirmizi, Dep. of Industrial Engineering,Kirikkale University, Kirikkale/Turkey,
myuzukirmizi@kku.edu.tr

Abstract.

In this research, Flexible Manufacturing Systems(FMS) are studied. They are modeled as closed queueing networks with multiple chains and finite buffers with production blocking. The effects of limited buffer capacities on the performance of the customer chains as well as the system are investigated. It has been shown that some customer classes of finite networks can deliver higher throughput performance compared to identical networks with unlimited capacities. Several simulation studies are conducted to comprehend the causes of effects.

1 Introduction

Flexible Manufacturing Systems(FMS) are complex manufacturing systems where multifunctional computerized machines are united together with a material handling systems. Their costly investment requirement necessitate the study of system performance under operating conditions.

FMS systems are usually modeled as multi-chain closed queueing networks with a central server configuration[28][18][26]. Multi-chain queueing networks consist of different classes of customers which vary in their routes, service requirements, costs, waiting spaces etc.

With respect to single class models, multiple class models involve

- Performance measures of the individual customer classes
- Requirements of multiple sets of input parameters
- Difficult data collection due to the necessity of multiple sets of data
- More complicated solution techniques and the need for more computing resources

It is essential to study such models since they provide more accurate results for systems where jobs exhibit significantly different behavior. For instance, even without blocking, the multi-class networks may lead to inefficient systems and fail to deliver the maximal production rate.

In addition to other identified sources of instability, the blocking due to finite capacity of the workstations can be a destabilizing factor. In finite open multi-class networks, it has been shown through examples that FCFS, as well as other scheduling policies such as last come first serve (LCFS), earliest due date (EDD) and shortest remaining processing time (SRPT), can become unstable.

We consider closed queueing networks with multiple customers. Further, we assume that the stations may have limited waiting spaces and a customer can be blocked after service. With a FCFS queueing discipline, these networks may have the following surprising behaviors:

- Increasing the buffer space of a work station may not lead to better performance for the whole system
- The throughput of a class may decrease significantly even though its chain does not have a blocking station on its route.

We analyze these characteristics that arise due to blocking for several types of systems.

2 Background

2.1 Notation and Basic Definitions

First, we introduce the following notation that will be used in the description of multi-class queueing networks:

R := The number of job classes in the network.

M := The number of stations in the network.

[1] This research has been supported by the The Scientific and Technological Research Council of Turkey(TUBITAK) under the programme National Young Researchers Career Development (Career Programme) by grant number 106M367.

$\vec{N}$:= The population vector ($\vec{N} = (N_1, N_2, \ldots, N_R)$ where N_r is the number of jobs of the r^{th} class in the network, $r = 1, \ldots, R$.

$G(R, M, \vec{N})$:= The closed network with finite number of R classes, M stations and $\vec{N}$ population

$\vec{n}$:= The state of the network, ($\vec{n} = (n_1, n_2, \ldots, n_R)$).

μ_{ir} := The service rate of the r^{th} class of customer at center i .

$\mu_{ir}(j)$:= The service rate of the r^{th} class of customer at center i when there are a total of j customers at the center($j = \sum_{r=1}^{R} n_r$).

$\mu_{ir}(\vec{n})$:= The service rate of the r^{th} class of customer at center i when the center is in state $\vec{n}$.

$\lambda_r(\vec{n})$:= The throughput of class r customers when the network is in state $\vec{n}$.

$W_{ir}(\vec{n})$:= The mean response time of a class-r customer at service center i when the network is in state $\vec{n}$.

$Q_{ir}(\vec{n})$:= The mean number of class-r customers in center i when the network is in state $\vec{n}$.

$\pi_i(j|\vec{n})$:= The probability of j customers at center i when the network is in state $\vec{n}$.

$\pi_i(\vec{n}_k|\vec{n})$:= The probability of $\vec{n}_k$ customers at center i when the network is in state $\vec{n}$.

$P_i(\vec{n})$:= The blocking probability of center i when the network is in state $\vec{n}$.

V_{ir} := The r^{th} class customer visit ratio to the i^{th} node.

$\vec{1}_r$:= Unit vector in the r^{th} direction $(0_1, \ldots, 1_r, \ldots, 0_R)$.

K_i := The buffer space of node i.

2.2 Model Description

We consider a multi-class queueing network with M work stations and R classes. The topology of the system is central-server. We, further assume:

- Fixed number of class-r customers, N_r
- Customers do not change classes
- Limited waiting spaces with BAS mechanism
- Single server nodes with FCFS queueing discipline
- Service rates of workstations are identical for each class
- Classes share waiting spaces
- The routing of a class is deterministic

2.3 Literature Review

Exact solution techniques exist for a class of separable networks [4], also referred as BCMP type networks. For these class of networks, the steady-state joint probability has a product form which can be expressed as in (1).

$$\pi(\vec{n}_1, \ldots, \vec{n}_M) = \frac{1}{G(\vec{N})} \prod_{i=1}^{M} f_i(\vec{n}_i) \tag{1}$$

where:

$\vec{N}$ is the number of jobs/customers in the various chains, known as the population vector ($\vec{N} = (N_1, \ldots, N_R)$),

R denotes the number of chains,

M is the number of stations,

$\vec{n}_i$ is the state of the i^{th} node ($\vec{n}_i = (n_{i1}, \ldots, n_{iR}), i = 1, \ldots, M$)

$G(\vec{N})$ is normalizing constant and defined as

$$G(\vec{N}) = \sum_{\sum_{i=1}^{M} \vec{n}_i = \vec{N}} \prod_{i=1}^{M} f_i(\vec{n}_i)$$

For BCMP type networks, the sum of class$-r$ jobs should be constant at any time. In closed queueing networks, if there is no class switching allowed, the number of chains are equal to the number of classes. Throughout the paper, we do not allow class switching, hence, we use the terms *class* and *chain* synonymously.

Some well-known product-form nodes are single-server and multiple-server first-come-first-served (FCFS), processor sharing (PS), infinite server (IS), and single server last-come-first-served (LCFS) queues. For nodes with FCFS queueing discipline, the service time distribution must be an exponential distribution that is identical for all classes.

There are efficient methods for analyzing the performance measures of product-form multiple-class closed queueing networks. To start with, the convolution algorithm, first proposed by Buzen [8] for single-chain networks and extended by Reiser [22] for multiple chains, is an efficient means of obtaining the normalization constant. Other performance measures such as mean queue lengths, server utilizations, mean waiting times and throughput are computed using the normalization constant. Another significant algorithm whose steps are described in latter sections is Mean Value Analysis (MVA) [23]. MVA can be applied to networks with multiple-chains and is based on the same fundamentals such as the Arrival theorem and Little's equation.

Another technique for closed product-form queueing networks is the Recursion by Chain Algorithm, or so called RECAL, developed by Conway and Georganas [9]. RECAL is well suited for networks with a large number of job classes but a small number of nodes. The idea of the RECAL is to break down each chain into constituent sub-chains so that each has a population of one. Then, the recursive expression relates the normalization constant of a network with r chains to those of a network with $(r-1)$ chain network.

The above mentioned algorithms calculate performance measures exactly. However, the memory requirements and computation time grow exponentially with the number of job classes in the system. For computationally difficult product-form closed networks, approximation methods such as Self-Correcting Approximation Technique [19], Summation Method [7], and Bottleneck Approximation [6] are also cited in literature.

Multiple class closed queueing networks with blocking have been shown to exhibit product-form solutions only in a few special cases: networks with reversible routing and self-dual networks. A survey and analysis of product-form queueing networks having finite capacities with various blocking mechanisms can be found in [2].

A finite multi-class closed queue with blocking-after-service (BAS) has been shown to have product forms in the following cases:

- The network consists of only two nodes [20].
- The total number of customers is one more than the minimum buffer space [20].
- The central server type network in which nodes have either FCFS or LCFS-PR with the condition that the neighbors must be priority or infinite servers.[1].

Unfortunately, other than these special cases closed queueing networks under the BAS mechanism can not be shown to have product-form distributions. Non-product-form types of networks are normally studied approximately. Two main approaches have been followed in the development of approximation methods for closed queueing networks. The first group of algorithms are heuristic extensions based on MVA. The generalization of the MVA algorithm to FCFS queues with class dependent service times with single servers was studied by Bard [3] and Reiser [21] and with multiple servers by Hahn [13] and Schmidt [25], to name a few.

The principle of the second group approaches is to approximate the performance of the original networks by that of an equivalent product-form network. An approximation for networks with general service times by Baynat and Dallery [5] can be categorized in this group.

Multi-chain closed queueing networks with finite buffers have not frequently occurred in the literature. Liebeherr and Akyildiz [17] investigated the deadlock properties of such systems assuming each node keeps separate buffers for jobs from different routing chains. They presented an optimization algorithm for finding deadlock-free capacity assignments with the least total capacity.

In this study, we analyze closed queueing networks with multiple chains. We consider systems with finite capacities and present an approximation method based on the Mean Value Analysis(MVA) algorithm. The basic idea is to utilize Bard's [3] approximation for networks with different service times at FCFS nodes along with our estimation of effective service rates and the blocking probability.

3 Behaviors of Multi-class networks with finite buffers

In this section, we give some insights into the behavior of multi-class networks and how these systems are affected by problem parameters.

3.1 Benefiting from Finite Capacity

An interesting, observation in closed finite queueing networks with multiple customer classes is that increasing the buffer size of a workstation may not necessarily lead to better performance of the system. So far, as conjectured and generally believed, the throughput of a finite capacity network is less than or equal to the same network without

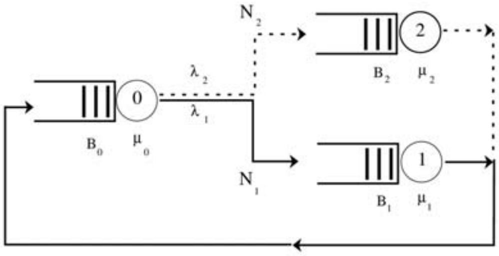

Figure 1: A closed queueing network with three service centers and two classes

blocking. That may be the case in a single class network, however, the customers in multiple class networks can benefit from finite capacities and may have higher throughput than their infinite node capacity counterparts.

In this context, we propose the following result for a multiple class closed queueing network with FCFS queueing discipline:

Theorem 1. *A multi-class closed queueing network with infinite node capacities does not provide an upper bound for the throughput of the customer classes as the same network with finite node capacities.*

Proof: It is sufficient to show this property through an example. Consider the multi-class network in Figure 2 with $M = 3$ nodes and $R = 2$ classes and N_1 and N_2 customers in each class. After receiving service at the first node class-1 customers proceed node-2, while second class customers proceed to node-3. Further, assume that service rates are independent of classes ($\mu_{ir} = \mu_i$) with $\mu_3 \gg \mu_2 \gg \mu_1$.

In this network, node-1 arises as a bottleneck station. With a FCFS queueing discipline, entering customers from their respective chains will move towards having service and will be sequentially queued at this station. Arriving from a slower stream, class-1 customers will only find waiting spaces towards the back of the queue. For μ_3 large enough, $\mu_3 \gg \mu_2$, class-1 customers will receive service after all class-2 customers have finished. The residence time of class-1 customers at the first station, W_{11}, would be:

$$W_{11} > N_2 \frac{1}{\mu_1}$$

Now, assume we impose a finite capacity on node-1 with K_1. Finite space will limit the number of class-2 customers waiting in the queue. Since blocking will also occur on a FCFS basis, class-1 customers will be able to find waiting spaces in between class-2 customers. The residence of class-1 customers at node-1, W_{11}, would be in the interval of :

$$N_2 \frac{1}{\mu_1} > W_{11} > (K_1 + 1) \frac{1}{\mu_1}$$

Specifically, for $N_1 = 1$, the residence time at node-1 for class-1 customers would be in the non-blocking case:

$$W_{11} = N_2 \frac{1}{\mu_1} + \frac{1}{\mu_1}$$

and in the blocking case:

$$W_{11} = (K_1 + 1) \frac{1}{\mu_1} + \frac{1}{\mu_1}$$

For $K_1 + 1 < N_2$ the residence time of class-1 will be lower than the infinite network which will result in higher throughput for this class.

3.2 Detriments of Finite Capacity

Another behavior which contrasts with the previous property is the negative effect of finite buffers on chains. Due to the interactions of customers arising from use of finite capacities, the performance of a class can be affected significantly, although, there is no blocking station on its route. This is intuitively understandable, especially in a network with single servers, FCFS scheduling and BAS. Once a customer is blocked by its downstream station, it continues to occupy the upstream station. During this time, the server cannot provide service to other customers although it is idle. The blocked customer functions as a barrier forcing the customers in the queue to wait longer.

4 Design Experiments

In previous sections, we have presented interesting effects of finite buffers on class outputs in multiclass closed queues. Nevertheless, the question of how these effects occur remains to be answered.

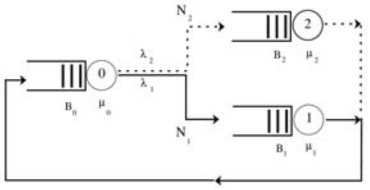

Figure 2: A closed queueing network with three service centers and two classes

Number of Customers		Service Rates			Buffer Sizes		
N_1	N_2	μ_0	μ_1	μ_2	B_0	B_1	B_2
5	8	1	3	5	∞	∞	∞

Table 1: Default control values of simulation experiments

We focus on a small queueing model with two-classes and three stations (Figure 2), and investigate the effects of the system design parameters on the throughput of the classes.

Our main purpose is; to determine the causes of effects by thoroughly analyzing a small system. Afterwards, using these outcomes, we intend to derive categorical results for a larger system. For this purpose we conduct several simulation experiments. Although, every instance of the experiments can be solved by a Markovian Chain approach, there is no-known product form formulation for this system and the population vector grows exponentially with the number of customers. Hence, determining the steady-state probabilities may not be feasible. Moreover, since we are only interested in calculating throughput rates and conduct a high number of experiments, we prefer the simulation approach.

Simulation experiments are conducted with ARENA version 8.0 with 10.000 time units after a 1000 warm-up period and 20 replications. System throughput of classes, λ_1 and λ_2, are the primary measures. The standard deviation of simulations are very small, on the order of [.0013,.0091] for the throughput values, and omitted for presentation purposes.

We have carried out several simulation experiments with varying buffer sizes, customer quantities and service rates as control parameters of classes and stations. For a reference the default values are displayed in Table 1. Just to give insight, we include a comparison study of buffer sizes of Station-1 and Station-2.

4.1 Simulation Results for Buffer Sizes

In these first set of experiments (Figure 3-5), the throughput values are examined against the change of buffer sizes. Other model parameters are as in Table 1.

In Figure 3, B_0 and B_1 are the control variables. While, B_0 is varied from 1 to 12, B_1 is increased from 1 to 7. Consequently, 84 simulation experiments are run and resultant throughput space have been shown. As seen in the figure, in some cases the throughput values are not available due to deadlock of the system. For feasible values, the throughput of class 2 increases with the increase in B_0. At the same time, class 1 benefits from limited buffer size of B_0 and rises for lower values of 6. This gain of class 1 can be as much as %23.8 compared to the value in no blocking case. We also would like to note that B_1 has limited or no effect,-other than causing deadlock- in this

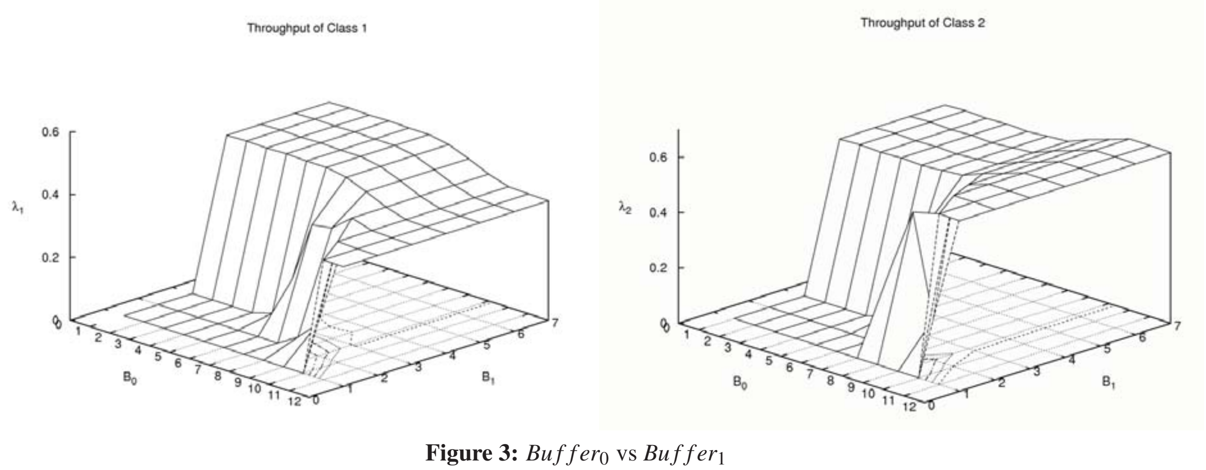

Figure 3: $Buffer_0$ vs $Buffer_1$

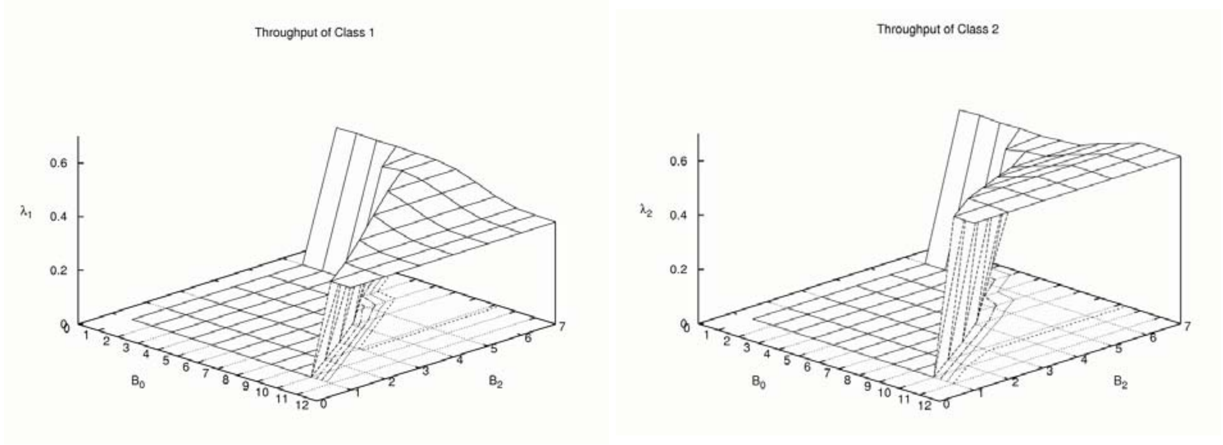

Figure 4: Buffer_0 vs Buffer_2

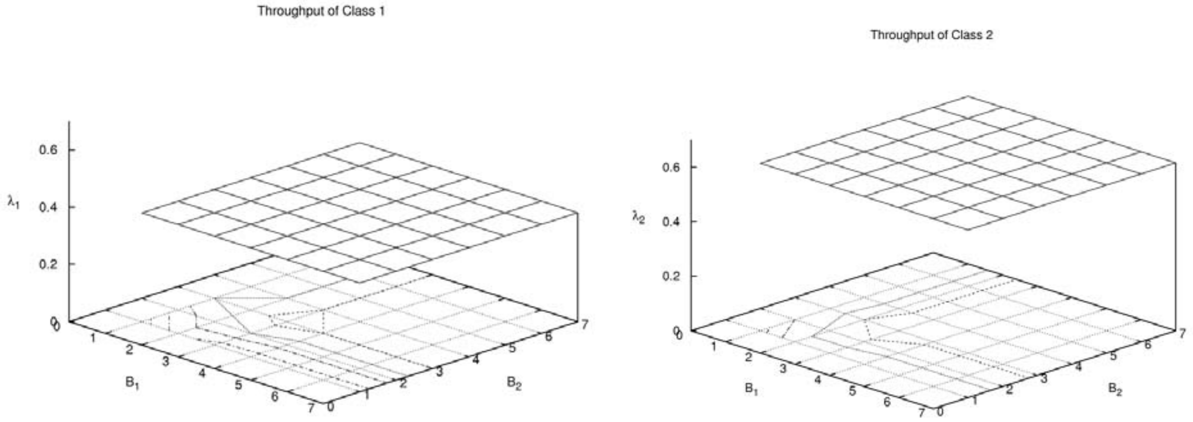

Figure 5: Buffer_1 vs Buffer_2

system setting.

Similar observations can be deducted from Figure 4. Class-1 customers benefit from limited buffer sizes of B_0. The achievable throughput is same as Figure 3.

In Figure 5,the influence of B_1 and B_2 are plotted keeping other parameters at their default values. However, the throughput space of class-1 and class-2 are planar, and we can presume that these buffers do not have significant effects.

Overall, the shared buffer of B_0 arises as the most influential buffer amongst all.

4.2 Simulation Results for the Number of Customers

In this section, we examine the behavior of the blocking network with different number of customers(Figure 6). The number of customers of class-1 and class-2 are varied in 64 simulation experiments while B_0 is the blocking buffer with a value of 1. For a comprehensive comparison we plot the non-blocking network, $B_0 = \infty$, with the same parameters.

Interestingly, the throughput space of classes are planar in the system with limited buffers. Moreover, both classes may benefit from blocking effect. If their throughput is low, i.e. they have less customers, the throughput gain is greater. The throughput gain for class-1 is the highest in combination of $N_1 = 1$,$N_2 = 8$ with %148 relative to the non-blocking case. Similarly, class-2 customers can have a %176 higher throughput in combination of $N_1 = 8$,$N_2 = 1$ if limited buffers are employed.

4.3 Simulation Results for the Service Rates

In this section, we study the effect of service rates (μ_0, μ_1, μ_2) on the blocking network (Figure 7-9). We have set the buffer size of station-0 to 1 and examined the class throughputs. To achieve a significant divergence, 7 level of control parameters on service rates are selected as .33, .50, 1.00, 1.50, 3.00, 5.00, 10.00. The stations are compared in dual in 3 cases where each consists of 49 simulation experiments. The remainder of the parameters are at their default level. The corresponding outcomes of throughput rates are presented whereas the same network

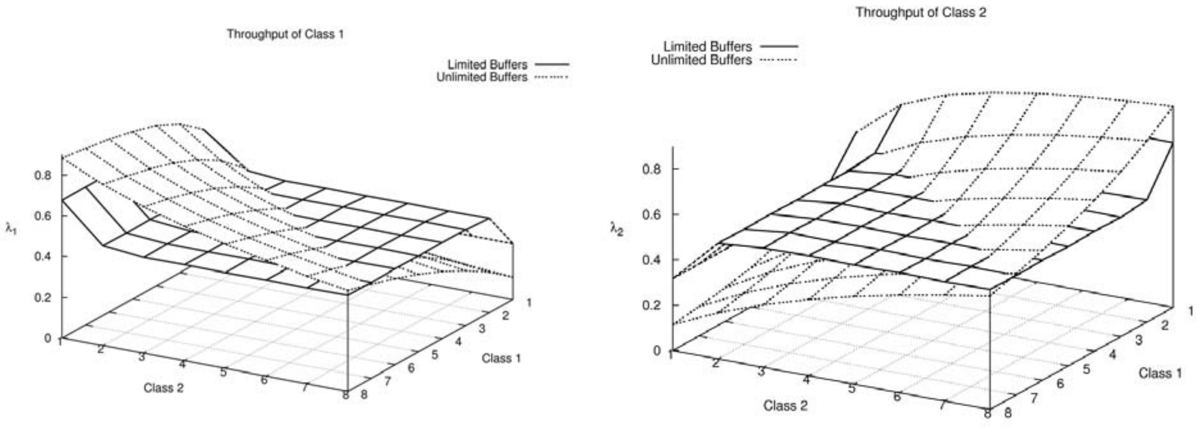

Figure 6: $Class_1$ vs $Class_2$

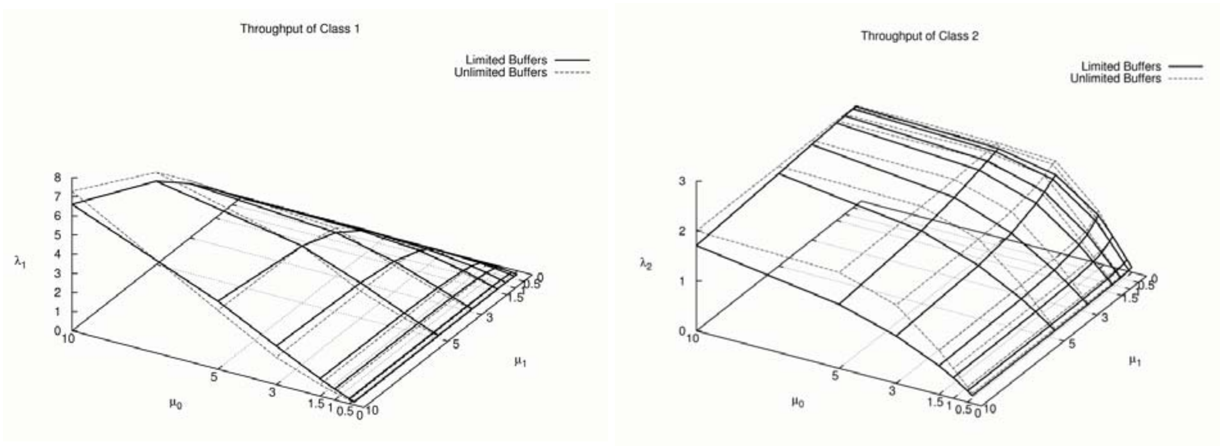

Figure 7: μ_0 vs μ_1

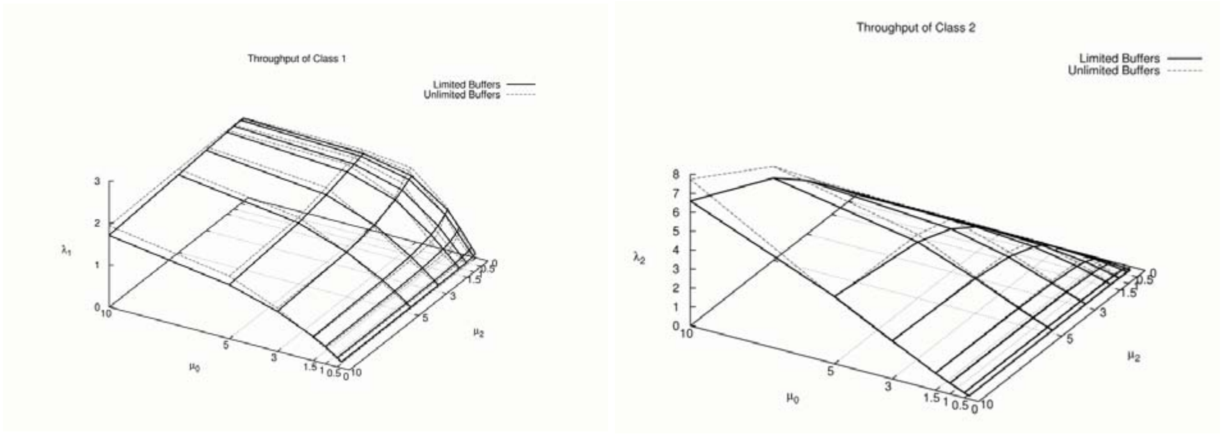

Figure 8: μ_0 vs μ_2

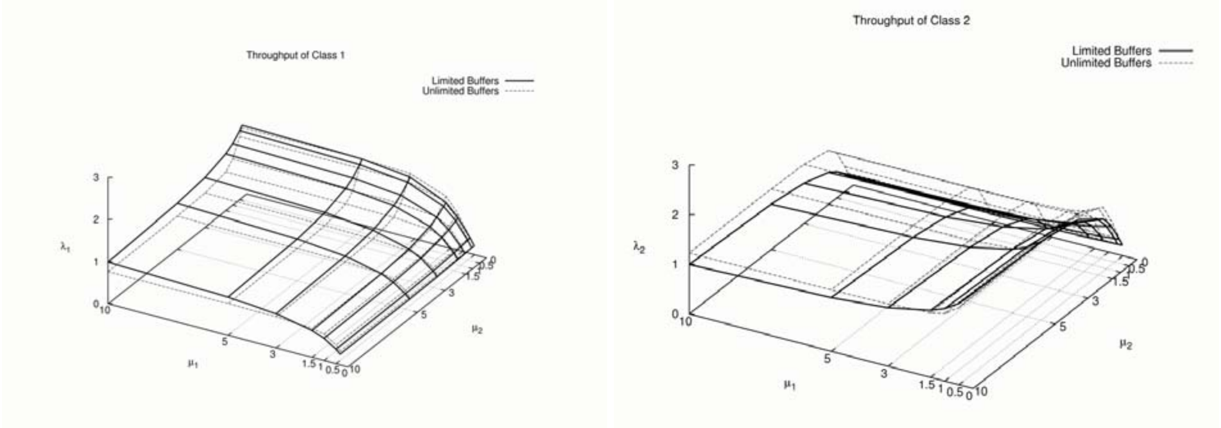

Figure 9: μ_1 vs μ_2

with unlimited buffers are also plotted.

The phenomenon of finite system performing better than unlimited buffer case can also be observed in these 3-graphs.

5 Conclusion and Future Research

In this research, we have investigated the behavior of multi-class finite closed queueing networks and showed that they can benefit from existence of finite buffers. On the other hand, we also have illustrated that finite capacities may significantly reduce the performance of the system if not carefully positioned.

As demonstrated in numerical experiments, the relative percentage difference of finite networks compared to non-blocking ones can be very high. Estimating the performance of the finite buffer queueing from non-blocking network is not suitable, and hence, finite networks require a particular treatment.

There are several important applications of such networks: packet-switching communication networks with different types of packets and priorities, job-shop manufacturing systems and multi programmed computers systems, to name a few.

The findings here open new areas of study in closed queueing networks. One possible area is that of studying the achievable performance bounds of customer chains. Since, the classes may exceed their non-blocking network counterparts in performance they do not provide an upper threshold anymore. Specific calculations and distinct algorithms should be derived for finite buffer systems in order to predict their performance.

Another area is optimization. While optimizing the system performance or/and determining the number of customers in each chain, the influence of the finite buffers may contribute to the objective function. Hence, the finite buffer factor should be included in decision process.

6 References

[1] I. F. Akyildiz and C. C. Huang, *Exact analysis of queueing networks with multiple job classes and blocking after service*, Queueing systems,13,(1992), 427-440.

[2] S. Balsamo, *Properties and Analysis of queueing network models with finite capacities*,Lecture Notes in Comp. Sci.,729,(1993), 21-52.

[3] Y. Bard, *Some extensions to Multiclass Queueing Network Analysis*, 4th Int. Symp. on Modelling and Performance Evaluation of Computer Systems, in Queueing Networks and Markov Chains by G. Bolch, S. Greiner, H. de Meer and S. Trivedi, John Wiley and Sons, 1998,pages 469-470, 1,(1979), 51-62.

[4] F. Baskett, K. M. Chandy, R. R. Muntz and F. Palacios-Gomez, *Open, closed and mixed networks for queues with different classes of customers*, Journal of A.C.M.,22-2,(1975), 248-260.

[5] B. Baynat and Y. Dallery, *A product-form approximation method for general closed queueing networks with several classes of customers*, Performance Evaluation, 24,(1996), 165-188.

[6] G. Bolch and M. Fischer, *Botapproz: Eine engpassanalyse fur geschlossene warterschlangennetze auf der basis der summations- methode. in H. Dyckhoff, U. Derigs, M. Salomon and H. Tijms, editors, Operations Research Proc. 1993, Amsterdam, Berlin, August 1993*, Informatik-Fachberichte (1993), 511-517.

[7] G. Bolch, G. Fleischmann, and R. Schreppel, *Ein funktionales konzept zur anlayse von warteschlangennetzen und optimierung von leistungsgrossen*, Informatik-Fachberichte 154 (1987), 327-342.

[8] J. Buzen, *Computational algorithms for closed queueing networks with exponential servers*, Communications of A.C.M. 16-9 (1973), 527-531.

[9] E. Conway and D. Georganas, *Recal-a new efficient algorithm for the exact analysis of multiple chain closed queueing networks*, Journal of A.C.M. 33-4 (1986), 768-791.

[10] Y. G. Dai,*On positive harris recurrence of multiclass queueing networks:a unified approach via fluid limit models*, Annals of Applied Probability 5 (1995), 49-77.

[11] E. A. Gonzales, *Optimal Resource Allocation in Closed Finite Queueing Networks with Blocking After Service*, Ph.D. thesis, University of Massachusetts- Amherst, Department of Mechnical and Industrial Engineering, 1997.

[12] W. G. Gilland,*Analysis of optimal and nearly optimal sequencing policies for a closed queueing network*, Operations Research Letters 33 (2005),9-16.

[13] T. Hahn, *Implementierung und Validierung der Mittelwertanalyse fur hohere Momente und Veteilungen. Studienarbeit, Universitat Erlangen- Nurnberg*, IMMD IV, 1988.

[14] L. Kerbache and J. M. Smith, *The generalized expansion method for open finite queueing networks*, European Journal of Operational Research 32 (1987), 448-461.

[15] L. Kerbache and J. M. Smith,*Asymptotic behavior of the expansion method for open finite queueing networks*, Computers and Operations Research 15-2 (1988), 157-169.

[16] P. R. Kumar, *A tutorial on some new methods for performance evaluation of queueing networks*, IEEE Journal of Selected Areas in Communications 13-6 (1995), no. 6, 970-980.

[17] J. Liebeherr and I. F. Akyildiz, *Deadlock properties of queuing-networks with finite capacities and multiple routing chains*, Queueing Systems 20 (1995), 409-431.

[18] A. Matta, Q. Semeraro, T. Tolio,*Configuration of Advanced Manufacturing Systems*, Chapter 4 in Design of Advanced Manufacturing Systems(Eds. A. Matta and Q Semeraro), Springer 2005

[19] D. Neuse and K. Chandy, *Scat: A heuristic algorithm for queueing network models of computing systems*, ACM Sigmetrics Performance Evaluation Review 10-3 (1981), 59-79.

[20] R. Onvural, *A note on the product form solutions of multiclass closed queueing networks with blocking*, Performance Evaluation 10 (1989), 247-253.

[21] M. Reiser, *A queueing network analysis of computer communication networks with window flow control*, IEEE Trans. Commun 27 (1979), 1199-1209.

[22] M. Reiser, *Mean value analysis and convolution method for queue-dependent servers in closed queueing networks*, Perfomance Evaluation 1 (1981), 7-18.

[23] M. Reiser and S. Lavenberg, *Mean value analysis for closed multi-chain queueing networks*, Journal of A.C.M. 27-2 (1980), 313-322.

[24] S. P. Reveliotis, *The destabilizing effect of blocking due to finite buffering capacity in multi-class queueing networks*, IEEE Trans. on Automatic Control 45-3 (2000), 585-588.

[25] R. Schmidt, *An approximate mva algorithm for exponential, class dependent multiple servers*, Performance Evaluation 29 (1997), 245-254.

[26] C. S. Sung and S.T. Kwon,*Performance Modeling of an FMS with finite input and output buffers*, Int. J. Production Economics 37 (1994),161-175

[27] M. Yuzukirmizi, *Finite Closed Queueing Networks with Multiple Servers and Multiple Chains*, Ph.D. thesis, University of Massachusetts-Amherst, Department of Mechnical and Industrial Engineering, 2005.

[28] L. Zhuang, K. S. Hindi *Approximate MVA for Closed Queueing Network Models of FMS with a Block-and-Wait Mechanism*,Computers Ind. Eng. 20 1(1991) 35-44