Optimal Server Allocation in General, Finite, Multi-Server Queueing Networks

J. MacGregor Smith
University of Massachusetts, Amherst, Massachusetts 01003, USA, jmsmith@ecs.umass.edu

F.R.B. Cruz
Federal University of Minas Gerais, Belo Horizonte, MG, Brazil, fcruz@est.ufmg.br

Tom van Woensel
Technische Universiteit Eindhoven, t.v.woensel@tm.tue.nl

Queueing networks with finite buffers, multiple servers, arbitrary topologies, and general service time distributions are considered in this paper. An approach to optimally allocate servers to series, merge, and split topologies and their combinations is demonstrated. The methodology builds upon two-moment approximations to the service time distribution embedded in the generalized expansion method for computing the performance measures in complex finite queueing networks and Powell's method for optimally allocating the servers within the network.

Key words: Multi-server; finite buffer; queueing networks; optimal server allocation

1. Introduction

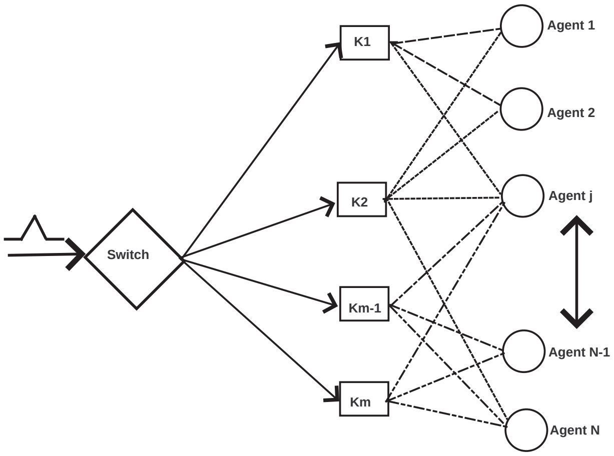

Figure 1: Call Center Topology

The design of finite buffer queueing networks with multiple servers is a difficult, challenging problem. Determining the optimal number of servers within the network is the central focus of this paper. Not only is the problem extremely complex, it is critically valuable to many industries and service sector activities such as: manufacturing, retail, transportation, and telecommunications. Suppose we have the task of designing a new network for a process which involves multiple nodes in a complex tree-topology much as in Figure 1. Recent examples of utmost importance include the design of call centers that normally employ hundreds of servers (agents), where finite buffer queues exist to hold the customers, and customer types can be handled by many different skilled servers. The dotted lines in the diagram indicate the necessary special skills available with the agents for handling the different customer types.

1.1. Motivation

Not only must we deal with the pattern of arrival and service rates as affected by the topology, let's argue that one key decision variable is to determine the number of servers at each node so as to effectuate the overall throughput performance of the network. We don't want to arbitrarily assign the number of servers, otherwise the desired throughput and queueing performance measures of the network will not be realized. How should we tackle this server allocation problem? This is the key design issue of this paper. We need a sensible methodology and a set of tools to carry out this task.

1.2. Outline of Paper

§2 presents a review of the problem and its difficulty, known results, along with the various references appropriate to its analysis. §3 presents the mathematical models appropriate to the optimization process. §4 delivers the algorithms and §5 the results for various topologies. One of the differences of our paper and those previous is our ability to model split, merge, and other complex topologies. §6 rounds out the paper with conclusions and open questions.

2. Problem Background

The determination of the number and the allocation of servers in an arbitrary topology queueing network is a complex problem. Many people have tackled this problem for single nodes, exponential service and infinite buffer queueing networks, yet not as much literature exists for the case when there are finite buffers, complex topologies, and general service time distributions in the network. The reason for this is basically due to the intractability of the problem for assessing the exact performance of a finite buffer queueing network of arbitrary topology, let alone optimizing them. While simulation could be a method of choice, employing simulation for large complex networks becomes prohibitive in terms of solution time, so we seek analytical approximations. We shall array some of the seminal analytical works in the area as well as outline the approaches to the problem utilized in the past.

2.1. Search for a Simple Formula

Within the infinite buffer queueing literature, there are certain simple formulas for determining the number of servers that are quite effective. Once such formula and its derivatives is often referred to as the square root rule (it actually goes back to Erlang) where ρ is the proportion of time each server is busy, c is the number of servers, and γ is a constant giving the rough grade of service (37, 3).

$$c = \rho + \gamma\sqrt{\rho} \tag{1}$$

We shall argue that the type of formula above becomes a useful bounding mechanism in the network topology search process and we will show that there is a reliable way to bound the optimal number of servers in finite queueing network topologies with the following expression. If the effective arrival rate to node i is $\tilde{\lambda}_i$, then the term on the left is a reliable lower bound and the term on the right provides an upper bound on $\mathbf{c}^*$.

$$\left\lceil \frac{\tilde{\lambda}_i}{\mu_i} \right\rceil \leq c_i^* \leq \left\lceil \frac{\tilde{\lambda}_i}{\mu_i} + \gamma\sqrt{\frac{\tilde{\lambda}_i}{\mu_i}} \right\rceil \tag{2}$$

Approaches for optimal server allocation are normally based on marginal allocation algorithms, convexity of the queueing performance parameters, and product-form properties of the queueing network system under study. In essence, as we shall argue, all these properties and concepts will be integrated in our solution methodology.

2.2. Literature Review

As argued earlier, there is a vast amount of literature for the optimal allocation of servers. Many studies have been done considering single nodes, open and closed networks, infinite and finite buffer waiting room, and exponential service systems. Of course, the latter topic of finite buffer, arbitrary topology, and general service networks is the most complex, and fewer studies have been accomplished.

Much of the earliest attempts to optimize the number of servers in infinite buffer exponential systems occured in the late 50's and early 60's. The 70's, 80's and 90's also experienced a number of efforts where general service was considered.

Interest in finite queueing systems did not really take off until the 80's when networks of finite queues began to be considered. Since then, there have been a number of publications in the 80's and 90's and beyond on finite queueing systems both for single nodes and serial/tandem networks. Figure 2 illustrates a sample view of the literature for this problem.

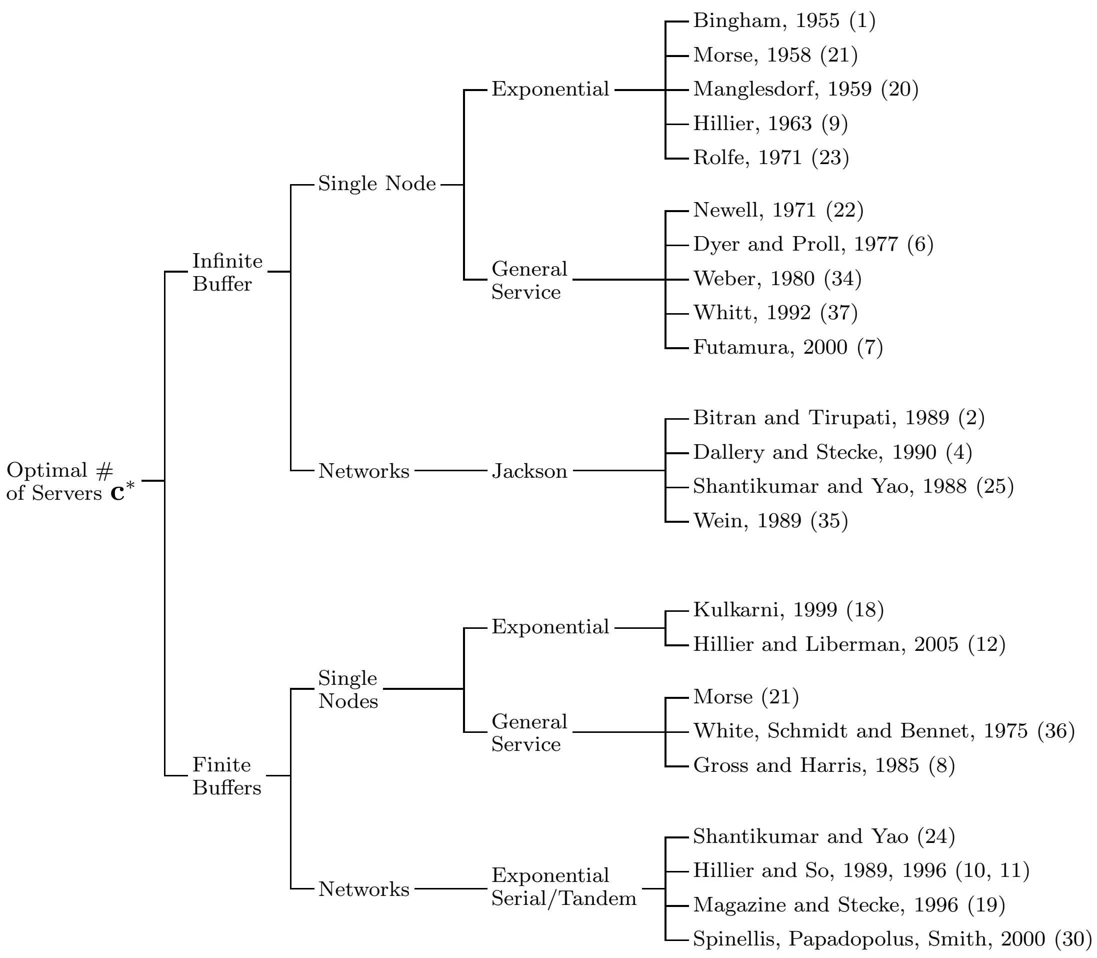

Figure 2 **Optimal Server Finite Queues and Networks References**

3. Mathematical Models

We assume Poisson arrivals and General Service Time distributions. We also assume blocking after service (BAS) (sometimes referred to as *production or transfer blocking*) which is a typical protocol in manufacturing and facility planning applications. Although communication networks often assume blocking before service (BBS) or *service blocking* and sometimes the protocol repetitive blocking (RPB) *rejection blocking*, the methodology used here assumes BAS.

3.1. Notation

This section presents most all of the notation we need for the paper:
Λ := External Poisson arrival rate to the network;
λ_j := Poisson arrival rate to node j;
μ_j := Exponential mean service rate at node j;
c := Number of servers;
$\epsilon \in (0,1)$:= Threshold for the blocking probability;
B_j := Buffer capacity at node j *excluding* those in service;
K_j := Buffer capacity at node j *including* those in service;
N := Number of stations in the network
p_K := Blocking probability of finite queue of size K;
p_0^j := Unconditional probability that there is no customer in the service channel at node j (either being served or being held after service);
$\rho = \lambda/(\mu c)$:= Proportion of time each server is busy;
$s^2 = \mathrm{V}ar(T_s)/\mathrm{E}(T_s)^2$:= Squared coefficient of variation of the service time, T_s;
$\mathbf{x}$:= Server allocation vector of decision variables in the optimization routine
Θ := Mean throughput rate.
Θ^τ := Threshold Mean throughput rate.

3.2. Optimization Problem

In this paper, we will consider the following type of optimization problem which also was the central objective used in previous buffer allocation papers (28, 29):

$$Z = \min\Big(f(\mathbf{x}) = \sum_i x_i\Big), \tag{3}$$

$$\text{s.t.} \quad \Theta(\mathbf{x}) \geq \Theta^\tau, \tag{4}$$
$$x_i \in \{1,2,3,\ldots\},\ \forall i, \tag{5}$$

that minimizes the total server allocation $\sum_i x_i$, constrained to provide the minimum throughput Θ^τ. In the above formulation Θ_j^τ is a threshold throughput value and $x_i \equiv c_i$ is the decision variable, which is the server allocation at the i-th queue.

4. Algorithms

This section illustrates briefly, the component algorithmic parts of the optimization methodology. First, we estimate the blocking probabilities with two-moment methods (5, 31, 33, 16). Then, we estimate the performance measures of the network topology with the general expansion method, and, finally, optimize the number of servers with Powell's algorithm.

4.1. P_K Calculations

If one starts with the blocking probability of the $M/M/1/K$ system and treats K continuously, one can generate an expression for the continuous optimal buffer size as a function of p_K and s^2. If one fixes the number of servers, one can solve for the blocking probability of the system. In the case of $c = 1$, the following expression is obtained for the blocking probability:

$$p_K = \frac{\rho^{\frac{-\sqrt{\rho\, e^{-s^2}}+2\,B+\sqrt{\rho\, e^{-s^2}}\,s^2}{2+\sqrt{\rho\, e^{-s^2}}\,s^2-\sqrt{\rho\, e^{-s^2}}}}\,(\rho-1)}{\left(\rho^{2\,\frac{-\sqrt{\rho\, e^{-s^2}}+B+\sqrt{\rho\, e^{-s^2}}\,s^2+1}{2+\sqrt{\rho\, e^{-s^2}}\,s^2-\sqrt{\rho\, e^{-s^2}}}}-1\right)}$$

This expression of the blocking probability is especially useful when $0 \leq s^2 \leq 1$. This process can be extended for $c > 1$, in fact, expressions for p_K of up to $c = 500$ have been found. Please see some of the other references for further details, (27, 28).

4.2. General Expansion Method

The Expansion Method is a robust and effective approximation technique developed by Kerbache and Smith (14). As described in previous papers, this method is characterized as a combination of repeated trials and node-by-node decomposition solution procedures. The Expansion Method uses blocking after service (BAS) type blocking, which is prevalent in most production and manufacturing, transportation and other similar systems.

4.3. Optimization Algorithm

In order to couple the optimization problem with the performance algorithm, the Expansion Method described in §4.2 previously, Powell's algorithm will be used to search for the optimal server vector(s) while the Expansion Method computes the performance measure of throughput. Powell's method, as presented in Himmelblau (13), locates the minimum of $f(\mathbf{x})$ of a non-linear function by successive unidimensional searches from an initial starting point $\mathbf{x}^{(0)}$ along a set of conjugate directions. These conjugate directions are generated within the procedure itself. Powell's method is based on the idea that if a minimum of a non-linear function $f(\mathbf{x})$ is found along p conjugate directions in a stage of the search, and an appropriate step is made in each direction, the overall step from the beginning to the p^{th} step is conjugate to all of the p subdirections of the search. We have had remarkable success in the past with coupling Powell's algorithm and the Expansion Method (28, 29).

5. Experimental Design

The completed results will be presented in our final paper for series, merge, splitting and other complex topologies.

6. Summary and Conclusions

Briefly, we have presented a viable approach to the design of allocating an optimal number of servers to complex topologies of finite queueing networks with general service time distributions.

References

[1]Bingham, G., 1955. "On a Congestion Problem in an Aircraft Factory", *Oper. Research 3*, 412-428.

[2]Bitran, G.R. and D. Tirupati, 1989. "Tradeoff Curves, Targeting and Balancing in Manufacturing Queueing Networks." *Oper.Res.***37**(4), .

[3]Borst, S., A. Mandelbaum, and M. Reiman, 2004. "Dimensioning Large Call Centers," *Operations Research, 52*(1), 17-34.

[4]Dallery, Y. and K.E. Stecke, 1990. "On the Optimal Allocation of Servers and Workloads in Closed Queueing Networks." *Oper. Res.***38**(4), 694-703.

[5]De Kok, A.G. and H. Tijms, 1985. "A Two-moment approximation for a Buffer Design Problem Requiring a Small Rejection Probability," *Performance Evaluation* **5**, 77-84.

[6]Dyer, M.E. and Proll, L.G., 1977. "On the validity of Marginal Analysis for Allocating Servers in $M/M/m$ queues. *Mgmt. Sci.***23**(9), 1019-1022.

[7]Futamura, Kenichi, 2000. "The multiple server effect: Optimal allocation of servers to stations with different service time distributions in tandem queueing networks." *Annals of OR*,**93,** 71-90.

[8]Gross, D. and C. Harris, 1985. **Fundamentals of Queueing Theory.** Wiley.

[9]Hillier, F.S., 1963. "Economic Anlaysis for Industrial Waiting Line Models," *Mgmt. Sci. 10*, 119-130.

[10]Hillier, F.S. and K.C. So, 1989. "The Assignment of Extra Servers to Stations in Tandem Queueing Systems with Small or No Buffers." *Perform. Eval.***10**, 219-231.

[11]Hillier, F.S. and K.C. So, 1996. "On the Simultaneous Optimization of Server and Work Allocations in Production Line Systems with Variable Processing Times." *Oper. Res.***44**(3), 435-443.

[12]Hillier, F.S. and G. Lieberman, 2005. **Introduction to Operations Research.** 8th ed. McGraw Hill: New York.

[13]Himmelblau, D. M., 1972, *Applied Nonlinear Programming*, McGraw-Hill Book Company, New York.

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

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

[16]Kimura, T., 1996. "Optimal Buffer Design of an M/G/s Queue with Finite Capacity," *Commun. Statist.-Stochastic Models,* **12**(1), 165-180.

[17]Kimura, T., 2000. "Equivalence Relations in the Approximations for the $M/G/s/s+r$ Queue," *Mathematical and Computer Modelling.* **31** 215-224.

[18]Kulkarni, V.G., 1999. **Modeling, Analysis, Deign and Control of Stochastic Systems.** Springer-Verlag: New York.

[19]Magazine, M.J. and K.E. Stecke, 1996. "Throughput for Production Lines with Serial Work Stations and Parallel Service Facilities."*Perform. Eval.25*, 211-232.

[20]Manglesdorf, T.M., 1959. "Waiting Line Theory Applied to Manufacturing Problems," in **Analysis of Industrial Operations.** eds. Bowmand and Fetter, Richard D. Irwin: Homewood, Illinois.

[21]Morse, P., 1958. **Queues, Inventories and Maintenance.** Wiley: New York.

[22]Newell, G.F., 1971. **Applications of Queueing Theory.** London: Chapman and Hall.

[23]Rolfe, A.J., 1971. "A Note on Marginal Allocation in Multiple Server Systems," *Mgmt. Sci.,***9**, 656-658.

[24]Shantikumar, G. and D. Yao, 1987. "Optimal Server Allocation in a System of Multi-server Stations." *Mgmt. Sci.* **33** (9), 1173-1180.

[25]Shantikumar, G. and D. Yao, 1988. "On Server Allocation in Multiple Center Manufacturing Systems." *Oper. Res.***36**(2), 333-342.

[26]Smith, J. MacGregor, 2003. $M/G/c/K$ "Blocking Probability Models and System Performance", *Performance Evaluation***52,** 237-267.

[27]Smith, J. MacGregor, 2004. "Optimal Design and Performance Modelling of $M/G/1/K$ Queueing Systems," *Mathematical and Computer Modelling***39**, 1049-1081.

[28]Smith, J. MacGregor and F.R.B. Cruz, 2005. "The Buffer Allocation Problem for General Finite Buffer Queueing Networks," *IIE Transactions: Design and Manufacturing***37**(4), 343-365.

[29]Smith, J. MacGregor, F.R.B. Cruz, and T. van Woensel, 2006. "M/G/c/K Performance Models in Manufacturing and Service Systems." Under review.

[30]Spinellis, D., Papadapoulos, and J. MacGregor Smith, 2000. "Large Production Line Optimization Using Simulated Annealing." *Int.J.Prod.Res.***38**(3), 509-541.

[31]Tijms, Henk, 1986. **Stochastic Modeling and Analysis.** New York:Wiley.

[32]Tijms, Henk, 1992. "Heuristics for Finite-Buffer Queues," *Probability in the Engineering and Informational Sciences* **6**, 277-285.

[33]Tijms, Henk, 1994. **Stochastic Models: An Algorithmic Approach.** New York:Wiley

[34]Weber, R., 1980. "On the Marginal Benefit of Adding Servers to $G/GI/m$ Queues." *Mgmt. Sci.***26**(9), 946-951.

[35]Wein, L.M., 1989. "Capacity Allocation in Generalized Jackson Networks," *Oper. Res. Lett.***8**, 142-146.

[36]White, J.A. J.W. Schmidt, and G.K.Bennet, 1975. **Analysis of Queueing Systems.** Academic Press, Inc: London.

[37]Whitt, W., 1992. "Understanding the Efficiency of Multi-server Service Systems." *Mgmt. Sci.*38(5), 708-723.