



Optimal Workload Allocation in a Network of Computers with Single Class Job

Rahela Rahim

Faculty of Quantitative Sciences

University Utara Malaysia, 06010, Sintok, Kedah, Malaysia

Tel: 604-928405 E-mail: rahela@uum.edu.my

Ku Ruhana Ku-Mahamud

Faculty of Information Technology

University Utara Malaysia

Tel: 604-9283118 E-mail: ruhana@uum.edu.my

Abstract

Queueing models for multiple queue with multiple server are used to model workload allocation problems in a network of computers. The problem of determining optimal allocation of workload with single class jobs to a parallel of computers using optimization technique is presented. The generalized exponential (GE) distributional model has been used to represent general inter arrival and service time distributions as various jobs have various traffic characteristic. A close-loop expression is derived from a non-linear optimization problem based on a queueing theory objective function to obtain an optimal value for jobs arrival. The analysis of the recomputation has been done and has shown improvement.

Keywords: Workload allocation, Multi server queueing system, Optimization, Generalized Exponential distribution.

1. Introduction

In a distributed computer system, task generated by a user or a group of users can be allocated over a number of available computers. This situation is opposed to a system which a single computer provide its capacity for all users, or systems in which each user is provided with its own local processor, usually with very limited capacity. An operational aspect of such a distributed system is the availability of workload balancing policy. Such policy balances the workload over the available computers, aiming to optimize performance measures for the system. Most traffic allocation problems in the literature have been tackled by assuming that all jobs are identical or single class (Chow & Kohler, 1979; Ni & Hwang, 1981; Tantawi & Towsley, 1985; Ross & Yao, 1991; Chombe & Boxma, 1995, Tavana & Rappaport, 1997; Wolf & Yu, 2001). This assumption is normally done when the diversity of jobs is not of importance.

In this paper, we stress on the quantitative measures of workload allocation to a network of computers. Based on this, we show that by using quantitative modeling, arrival to a network of computers can be reallocated to get the optimal performance measure. We focus on the issue of job allocation in a network of computers where different computers have different job processing time. The optimization criterion studied here is to minimize the expected job response time in the systems to which jobs are allocated. Jobs arrive at a scheduler that allocates jobs to the computers according to a calculated arrival rate computed using Lagrange multiplier theorem. The paper is organized as follows: system model is described and the proposed GE optimization model of workload allocation is presented in section 2.0 and 2.1, followed by the computational results in section 2.2.

2. Optimal Workload Allocation in a Network of Computers with General Exponential (GE) Arrival and Service time Distribution

In this section, we consider workload allocation problems for static allocation protocol for the model of a single GE stream of jobs offered to a fixed number of computers. The allocation protocol which has been studied is static in the sense that only the total incoming traffic and information about basic characteristic, like arrival rate and service times are used. The objective of the network of computer systems studied here is to maximize user perceived performance, which is a function of the amount of time the user spends waiting for a file to download from a server. In this context, download refers to the actions from the time the user requests a file from a computer to the time the file or an error message is delivered to the user's terminal. The shorter the download time, the higher the user's perceived performance.

Most of the previous studies on workload allocation used exponential distribution for inter arrival and service time. The reason is that network traffic has long been assumed to have exponential behavior. However this situation is not always true since the number of network users is unpredictable. Furthermore mean queue length derived from exponential distribution does not factor in variation in the inter arrival and service time. Clearly, the more information is available for making decision, the better the workload allocation can be. So our proposed models include variation parameters in inter arrival and service time that we found lack in previous models. In this model, requests arrive at the system according to a GE process. They are numbered in the order that they arrive at the system. Once a request has entered the system, it does not leave until it completes service. The metric of interest is mean response time, the user spends on the system an amount relative to the download time upon the server's completion of the user's request. Here, the transit time required to send the result of the request back to user's interface has been ignored, and the response time is assumed to be achieved instantaneously upon the file's departure from the server or computer.

Criterion of Optimality

For a given total network traffic ϕ , find the optimal traffic workload λ_i , $i = 1, 2, \dots, N$ so that the expected response time incurred on any system is minimized.

2.1 Mathematical Model Description

We will first present a mathematical description of the related workload allocation problems. Jobs arrive at a routing point according to a GE process with rate ϕ . At the instance of arrival, each job has to be assigned to one of N servers in parallel. The service rate μ_i that is assigned to server i has GE distribution as well. All service times are independent. Any job that is not fully processed, branches with certain probability and returns back to the scheduler for further processing. Otherwise the job is complete and exits the system. Let π denote an allocation policy and p_i , $i = 1, \dots, N$, be the fraction of the jobs that is routed to computer i under policy π . In our workload allocation problem, the aim is to minimize

$$\sum_{i=1}^N D_i W_i(\pi)$$

$W_i(\pi)$ denote the mean response time of a job assigned to computer i under allocation policy π . D_i is the cost associated with waiting one time unit at queue i . The objective function can have various interpretations, by varying D_i . Little's Law shows that the objective is to minimize the mean number of jobs in the queues. Instead of $W_i(\pi)$, we use $L_i(\pi)$, the mean queue length of queue i . The objective is to minimize a weighted sum of the mean queue length in the system. To obtain the assignment probabilities $p_i^* = \frac{\lambda_i^*}{\phi}$ which minimize this function, the following Mathematical Programming

problem has to be solved

$$\text{P1} \quad \text{Min} \quad \sum_{i=1}^N L_i = \sum_{i=1}^N D_i \left(\frac{\frac{\lambda_i}{\mu_i}}{1 - \frac{\lambda_i}{\mu_i}} \right) \left(\frac{\lambda_i}{\mu_i} \left(\frac{C_{si}^2 - 1}{2} \right) + \frac{C_{ai}^2 + 1}{2} \right) \quad (2.1)$$

$$\text{s.t} \quad \sum_{i=1}^N \lambda_i = \phi \quad (2.2)$$

$$0 \leq \lambda_i \leq \mu_i, \quad i = 1, K, N. \quad (2.3)$$

$$\lambda_i \geq 0 \quad (2.4)$$

$$\mu_i \geq 0 \quad (2.5)$$

The term L_i is strictly convex functions in λ_i and it can also be verified that the problem has a feasible solution provided that $\sum_{i=1}^N \mu_i > \phi$ that is the arrival rate does not exceed the total service capacity. Before analyzing the model, it is important to understand the meaning of the model parameters. Network Traffic (λ_i) is the average number of file request received by the computer each second. Service rate (μ_i) is the average rate the computer can serve. Obviously, this value will vary widely from one request to another.

Problem P1 allows an analytical solution. Using Lagrange multiplier theorem we obtain δ the Lagrange multiplier with the following first order Kuhn-Tucker constraints:

$$\frac{d}{d\lambda_i} \left\{ D_i \left(\frac{\frac{\lambda_i}{\mu_i}}{1 - \frac{\lambda_i}{\mu_i}} \right) \left(\frac{\lambda_i}{\mu_i} \left(\frac{C_{si}^2 - 1}{2} \right) + \frac{C_{ai}^2 + 1}{2} \right) \right\} = \delta \quad i = 1, K, N. \quad (2.6)$$

$$\sum_{i=1}^N \lambda_i - \phi = 0 \quad (2.7)$$

From (2.6) we find the unique optimal values λ_i^* as follows:

$$\lambda_i^* = \mu_i \left(1 - \left(\frac{C_{si}^2 + C_{ai}^2}{C_{si}^2 - 1 + 2\mu_i \delta} \right)^{1/2} \right), \quad (2.8)$$

Lagrange multiplier, δ is derived by solving the constraint equation below

$$\sum_{i=1}^N \mu_i \left(1 - \left(\frac{C_{si}^2 + C_{ai}^2}{C_{si}^2 - 1 + 2\mu_i \delta} \right)^{1/2} \right) = \phi \quad (2.9)$$

The computation has been developed in the MathCAD version 7 professional.

2.2 Computational result

In this section, numerical results are presented to assess the credibility of the GE distribution used. Two configurations will be shown. For the first configuration, service rate of the tasks are assumed to be $\mu_1=3, \mu_2=4, C_{a1}=0.5, C_{a2}=0.3, C_{s1}=0.2, C_{s2}=0.4$. The improvement of the performance measures is presented in Figure 2.1 and 2.3. To verify the results, we use simulation and the comparative results are presented in Figure 2.2 and 2.4. Further analysis of sample cases for $N = 2, 3, 4, 5$ and 6 computers are computed and the analysis shows that a larger range for the service rates results in greater percentage improvements of the aggregate objectives. For example, a two computer system with $\mu_1 = 2, \mu_2 = 1$ and $\rho = 0.9$ results in a 1.75 per cent improvement in mean queue length compared with 11.6 per cent improvement for a six computer system with $\mu_i, i = 1, 2, \dots, N$. The results of the analysis for such queueing systems are summarized in Figure 2.5.

3. Conclusions

The proposed solution mechanism focuses on the workload allocation of single class jobs through the use of optimal GE arrival rate in the workload allocation scheme. The key idea of optimizing the workload allocation scheme is to send a disproportionately fraction of workload to the computers with known capacities. GE component processes are expected to be more regular than a Poisson process, in the sense of having variation parameters.

We have described models that consider the workload allocation decision for the single class job case. The problem of workload allocation in an open network of queues is formulated as a non-linear optimization problem. The problems of maximizing system's mean queue length and mean response time for a given total arrival rate, and a specified arrival and service variation, are found to have the optimality condition. These optimality conditions are used to prove that, for queueing networks with unbalanced configuration of computer capacity, the optimal allocation of workload is unbalanced. A larger per computer share of workload goes to a larger capacity of computers. The unbalanced allocation result is related to efficiencies gained from computer pooling systems, we show that, holding utilization per computer constant, increasing the number of computers in the network reduces the average number in queue.

References

- Allen, A. O. (1990). *Probability, Statistics and Queueing Theory with Computer Science Applications*, 2nd ed. San Diego: Academic Press.
- Bell, S. L. & Williams, R. J. (1999). Dynamic scheduling of a system with two parallel servers: Asymptotic optimality of a continuous review threshold policy in heavy Traffic. *In Proceedings of the 38th Conference on Decision and Control*, Phoenix, Arizona, 1743-1748.
- Boxma, O. J. (1995). Static Optimization of Queueing Systems, *CWI Report*. BS-R 9302.
- Chombe, M. B. & Boxma, O. J. (1995). Optimization of Static Traffic Allocation Policies, *Theoretical Computer Science*, 125, 17-43.
- Chandy, K. M., Herzog, U. & Woo, L. (1975). Parametric Analysis of Queueing Networks, *IBM J. Research and Development*, 19(1), 36-42.

- De Jongh, J. F. C. M. (2002). *Share Scheduling in Distributed System*, PhD Thesis University of Technische, Netherland.
- Gelenbe, E & Mitrani, I. (1980). *Analysis and Synthesis of Computer Systems*, London: Academic Press.
- Gunther, N. J. (2000). *The Practical Performance Analyst*, McGraw Hill.
- Harrison, P. G. & Patel, N. M. (1992) *Performance Modeling of Communication Networks and Computer Architecture*, Addison Wesley.
- Hsiao, M. T. & Lazar A. A. (1990). Optimal Flow Control of Multiclass Queueing Networks with Partial Information, *IEEE Transaction on Automatic Control*, 35(7) 855-860.
- Hsiao, M. T. & Lazar, A. A. (1991). Optimal Decentralized Flow Control of Markovian Queueing Networks with multiple Controllers, *Performance Evaluation*, 13(3), 181-204.
- Ross, K. W. & Yao, D. D. (1991). Optimal Load Balancing and Scheduling in a Distributed Computer System, *Journal of the ACM*, 38(3), 679-690.
- Klienrock, L. (1975). *Queueing Systems Volume 1: Theory*, John Wiley Inc.
- Kobayashi, H. (1974). Application of the Diffusion Approximation to Queueing Networks I: Equilibrium Queue Distribution, *J. of the Association for Computing Machinery*, 21(2), 316-328.
- Koole, G. (1999). On the Static Assignment to parallel Servers, *IEEE Transaction on Automatic Control*, 44, 1588-1592.
- Kouvatsos, D. D. (1986). A Maximum Entrophy Queue Length Distribution for A G/G/1 Finite Capacity Queue, *Journal of ACM*, 224-236.
- Kouvatsos, D. D. & Othman, A. T. (1989). Optimal Flow Control of a G/G/1 Queue, *International J. of Systems Science*, 20(2), 251-265.
- Kouvatsos, D. D. & Othman, A. T. (1986). Optimal Flow Control of a G/G/C Finite Capacity Queue, *J. Operational Research Society*, 40(7), 659-670.
- Kouvatsos, D. D. & Othman, A. T. (1989). Optimal Flow Control of end to end Packet Switched Network with Random Routing, *IEEE Proceeding*, 136(2), 90-100.
- Ku Mahamud, K. R. (1993). *Analysis and Decentralized Optimal Flow Control of Heterogeneous Computer Communication Network Models*, PhD thesis, Universiti Pertanian Malaysia.
- Lazar, A. A. (1981). Optimal Control of an M/M/1 Queue, *In Proc. 19th Allerton Conf. On Communication, Control and Computing*, 279-289.
- Lazar, A. A. (1982). Centralized Optimal Control of a Jacksonian Network, *In Proceedings of the Sixteenth Conference on Information Science and Systems*, 316-324.
- Lazar, A. A. (1983). The Throughput Time Delay Function of an M/M/1 Queue, *IEEE Transaction on Information Theory*, 6, 1001-1007.
- Lazar, A. A. (1984). Optimal Control of an M/M/m Queue, *Journal of the Association for Computing Machinery*, 86-98.
- Lin, W. & Kumar, A. (1984). Optimal Control of a Queueing System with Two Heterogeneous Servers, *IEEE Trans Automatic Control*, 29(8), 696-703.
- Liu, J. B. (1999). A Multilevel Load Balancing Algorithm in a Distributed System, *Proceedings of the 19th annual conference on Computer Science*, 35-142.
- Menasce, D. A. and Almeida, V. A. F. (2000). *Scaling for E-Business*, Prentice Hall.
- Ni, L. M., & Hwang, K. (1985). Optimal Load Balancing in a multiple Processor System with Many Job Classes. *IEEE Trans. Software Engineering*, 491-496.
- Smith, C. U. & Williams, L. G. (2001). *Performance Solutions, A Practical Guide to Creating Responsive, Scalable Software*, Pearson Education.

Table 2.1 Results for the classical and proposed approaches of a dual GE/GE/1 with $Ca_1^2 = 0.5$, $Ca_2^2 = 0.3$, $Cs_1^2 = 0.2$, $Cs_2^2 = 0.4$ and $\mu_1 = 3$, $\mu_2 = 4$.

Classical		Proposed		Classical		Proposed	
λ_1	λ_2	λ_3	λ_4	L	W	L	W
1.6	2.1	1.578	2.122	1.158	0.313	1.153	0.312
1.8	2.4	1.776	2.424	1.47	0.35	1.465	0.349
2.0	2.7	1.981	2.719	1.896	0.403	1.891	0.402
2.2	2.9	2.15	2.95	2.396	0.47	2.377	0.466
2.4	3.2	2.367	3.233	3.36	0.6	3.342	0.597
2.6	3.4	2.544	3.456	4.86	0.81	4.775	0.796

Table 2.1 provides numerical results after recomputation of arrival rate and the improvement in performance measures of mean queue length and waiting time for the stated parameters.

Table 2.2. Results for the classical and proposed approaches of a dual GE/GE/1 with $Ca_1^2 = 0.1$, $Ca_2^2 = 0.2$, $Cs_1^2 = 0.4$, $Cs_2^2 = 0.3$ and $\mu_1 = 3$, $\mu_2 = 4$.

Classical		Proposed		Classical		Proposed	
λ_1	λ_2	λ_3	λ_4	L	W	L	W
1.6	2.1	1.465	2.235	0.906	0.245	0.897	0.212
1.8	2.4	1.698	2.502	1.14	0.27	1.132	0.253
2.0	2.7	1.931	2.769	1.455	0.31	1.449	0.298
2.2	2.9	2.117	2.983	1.82	0.357	1.805	0.342
2.4	3.2	2.35	3.25	2.52	0.45	2.507	0.446
2.6	3.4	2.536	3.464	3.599	0.6	3.539	0.598

Table 2.2 provides numerical results after recomputation of arrival rate and the improvement in performance measures of mean queue length and waiting time for the stated parameters.

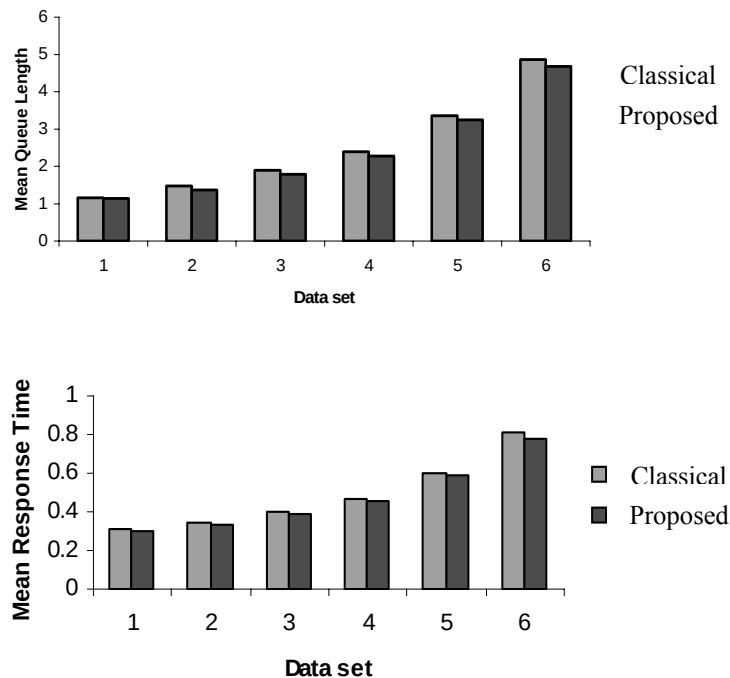


Figure 2.1. Performance improvement of a dual GE/GE/1 with $Ca_1^2 = 0.5$, $Ca_2^2 = 0.3$, $Cs_1^2 = 0.2$, $Cs_2^2 = 0.4$ and $\mu_1 = 3$, $\mu_2 = 4$.

Figure 2.1 shows the improvement of the two computers system's mean response time using dataset in Table 2.1.

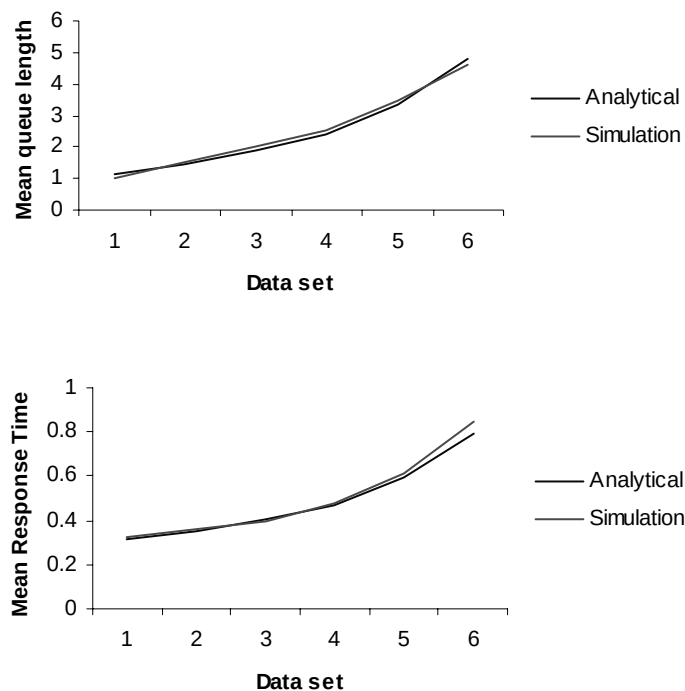


Figure 2.2. Analytical versus simulation result for a dual GE/GE/1 with $Ca_1^2 = 0.5$, $Ca_2^2 = 0.3$, $Cs_1^2 = 0.2$, $Cs_2^2 = 0.4$ and $\mu_1 = 3$, $\mu_2 = 4$

Figure 2.2 shows verification of the results with simulation using the stated parameters.

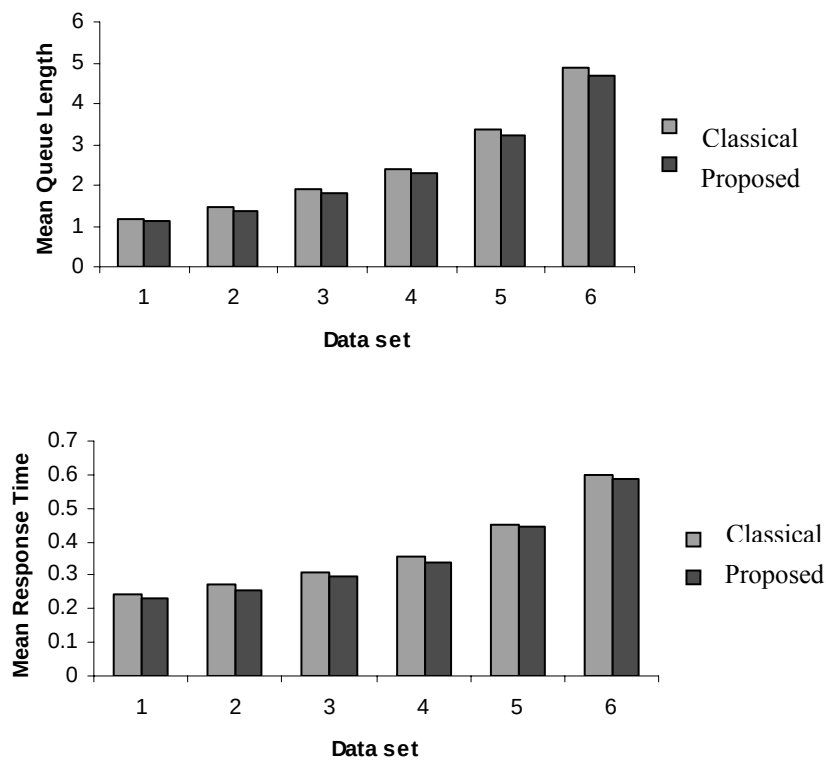


Figure 2.3. Performance improvement of a dual GE/GE/1 with $Ca_1^2 = 0.1$, $Ca_2^2 = 0.2$, $Cs_1^2 = 0.4$, $Cs_2^2 = 0.3$ and $\mu_1 = 3$, $\mu_2 = 4$.

Figure 2.3 shows the improvement of the two computers system's mean response time using dataset in Table 2.2.

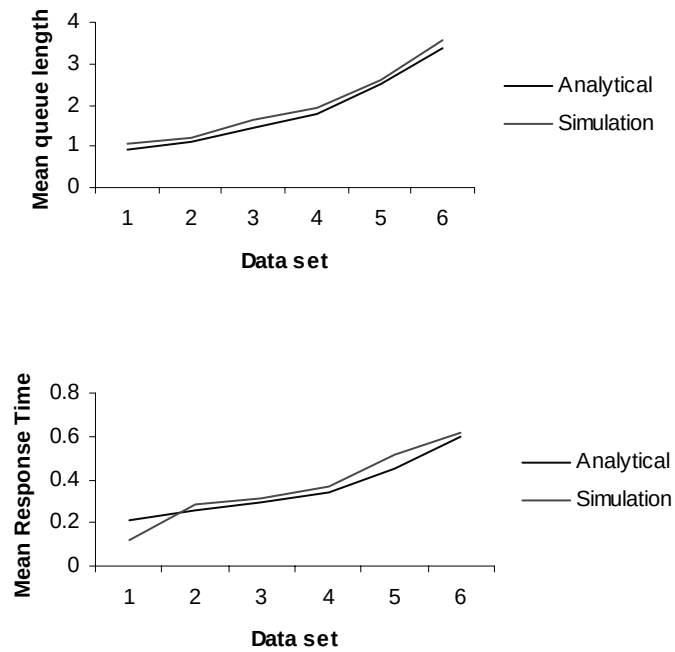


Figure 2.4. Performance improvement of a dual GE/GE/1 with $Ca_1^2 = 0.1$, $Ca_2^2 = 0.2$, $Cs_1^2 = 0.4$, $Cs_2^2 = 0.3$ and $\mu_1 = 3$, $\mu_2 = 4$.

Figure 2.4 shows the verification of the results with simulation using the stated parameters.

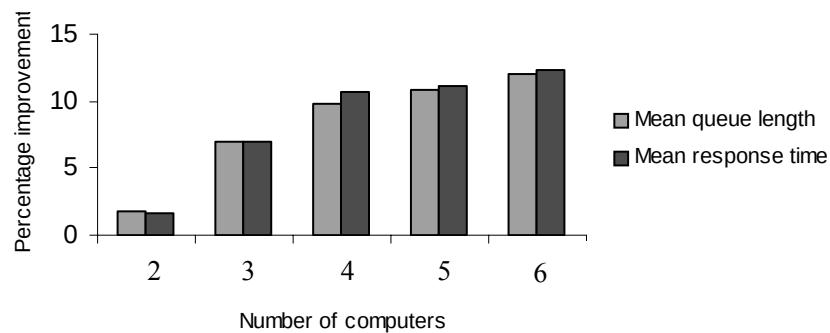


Figure 2.5. Performance improvement for a sample number of computers where $\rho = 0.9$.

Figure 2.5 shows the results of the analysis for sample cases for $N = 2, 3, 4, 5$ and 6 computers. The analysis shows that a larger range for the service rates results in greater percentage improvements of the aggregate objectives.