AMCIS 1997 Proceedings

Americas Conference on Information Systems (AMCIS)

8-15-1997

The Impact of Breakdowns on the Decision to Consolidate or Cluster Computers

Hsing Kenneth Cheng
The College of William and Mary, cheng@cs.wm.edu

The Impact of Breakdowns on the Decision to Consolidate or Cluster Computers

Hsing Kenneth Cheng
Graduate School of Business Administration
The College of William and Mary
Williamsburg, VA 23187-8795
Phone: (757) 221-2879 Fax: (757) 221-2937
e-mail: cheng@cs.wm.edu

Abstract

More and more organizations are running a clustered twin-computer system to tackle the rapidly growing demand of computer capacity. A prevailing rationale for a clustered twin-computer system is that it is an effective way of coping with not only the capacity growth challenge but also the computer downtime problem.

This paper develops an analytical model for a clustered twin-computer system subject to breakdowns with an aim to identifying conditions under which the clustered twin-computer system is a better alternative to a consolidated system. This research finds that the clustered twin-computer system has a shorter expected time in system for most cases. The clustered twin-computer system also performs better when there is a heavy traffic intensity. For firms with a consolidated single computer system, this research finds that it pays off to reduce the mean time to repair than to increase the mean time between failures.

A complete version of this paper is cited as Cheng (1997) in the references and available from the author upon request.

1. Introduction

Many organizations face the challenge of keeping up with a vast growth of demand for computer capacity. A typical solution to this problem is to upgrade the firmís computer system. Another alternative is to cluster two identical computers together, a solution becoming more and more popular. One prevailing rationale for a clustered computer system is that it is an effective way of coping with not only the capacity growth challenge but also the computer downtime problem. In case one computer in the clustered system breaks down, the other computer takes over the processing until the failed computer becomes available, a feature generally described by the term ìautomatic failoverî in the industry.

While service interruptions and machine breakdowns is an active research area in manufacturing, e.g., Posner and Berg (1989), Federgruen and So (1990), Groenevelt, Pintelon, and Seidmann (1992), and Moinzadeh and Aggarwal (1996), little existing information systems literature considers the effect of computer breakdowns except perhaps the work by Cheng (1995) and Cheng (1996). The objective of this paper is to develop a queuing model for a clustered twin-computer system subject to breakdowns. The analytical results will be used to characterize conditions under which a clustered twin-computer system has a smaller expected time in the system than a consolidated system. This research provides useful results for decision makers especially when computer vendors have begun to deliver clustering capabilities for mid-range computers and local area network servers.

The rest of the paper is structured as follows. Section 2 develops analytical models for both a consolidated single computer and a clustered twin-computer system subject to the same breakdown parameters. The analytical results of the expected time in a twin-computer system subject to breakdowns are found too complicated for further analyses. Hence, a very accurate approximation is reported in this section as the basis of performance comparison. Section 3 summarizes findings from numerical experiments and Section 4 concludes the paper.

2. The Model

Consider a clustered twin-computer system where there is Poisson stream of arriving jobs requiring computer processing with parameter λ and arriving jobs have a homogeneous service requirement. Service times of arriving jobs are assumed to be independently and identically drawn from an exponential distribution with mean $1/\mu$ according to the First Come First Serve (FCFS) service discipline. Each of the twin computers has a capacity of processing μ jobs per unit of time. This corresponds to an M/M/2 queuing system with arrival rate λ and service rate μ for each server.

To model the computer breakdowns, assume that each computerís downtime and uptime follow exponential distributions with parameters γ and η respectively. Hence, the inverse of the downtime parameter, $1/\gamma$, can be interpreted as the mean time to repair (MTTR) and the inverse of the uptime parameter, $1/\eta$, amounts to the mean time between failures (MTBF). The MTTR and MTBF are common measures used in the industry to represent a computerís breakdown characteristics and availability. Each of the twin computers is assumed to break down and repair independently of each other. It is worth noting that computer system may break down due to hardware failures or software malfunctions. This model is general in nature and does not attempt to distinguish specific types of computer failures.

For comparison purpose, also consider a consolidated single computer system having twice the processing capacity of each of the twin computers. That is, this single computer system has a capacity 2μ. This single computer service system has an identical stream of arriving jobs and breakdown characteristics. The expected time in the consolidated single computer system subject to breakdowns is readily available from Cheng (1996) as follows.

$$T_{SS}(\lambda, \mu, \gamma, \eta) = \frac{1 + \frac{2\eta\mu}{(\gamma + \eta)^2}}{2\mu\frac{\gamma}{(\gamma + \eta)} - \lambda} \quad (2.1)$$

where the subscript *ss* in Equation (2.1) stands for Single Server.

One of the earliest theoretical work on a multiple-server system subject to breakdowns is due to Mitrany and Avi-Itzhak (1968). A similar but simpler method, probability generating function method, is used to derive the expected time in the clustered twin-computer system subject to breakdowns. The derivation of the expected time in a clustered twin-computer system is a rather lengthy process. Interested readers are referred to Cheng (1997), the complete version of this paper.

The closed-form results for the clustered twin-computer system are found too complicated for subsequent analyses. Hence, an accurate approximation for the expected time in a clustered twin-computer system is derived in Cheng (1997) and reported as follows:

$$T_{tc}(\gamma, \eta, \lambda, \mu) = \frac{1 + \frac{\eta\mu}{(\gamma + \eta)^2}}{\frac{\gamma}{(\gamma + \eta)}2\mu - \lambda} + \frac{\frac{\gamma}{(\gamma + \eta)}}{(2\mu + \gamma)} \quad (2.2)$$

Apparently, the clustered twin-computer system has a shorter expected time in the system than the consolidated single computer system *if and only if*

$$\frac{\frac{\gamma}{(\gamma + \eta)}}{(2\mu + \gamma)} < \frac{\frac{\eta\mu}{(\gamma + \eta)^2}}{\frac{\gamma}{(\gamma + \eta)}2\mu - \lambda} \quad (2.3)$$

In order to gain further insights, numerical experiments were conducted and reported in the following section.

3. Numerical Explorations

Four parameters under consideration in the numerical experiments include (1) the arrival rate of jobs for computing processing, λ, (2) the computer capacity, μ, (3) MTBF, and (4) MTTR. Only one parameter was changed each time while the other three remained the same. The numerical experiments aimed at uncovering the conditions under which the clustered twin-computer system generates a lower expected time in system, or higher for this matter, than the consolidated single computer system. Numerical experiments were conducted according to empirical data of realistic business computer systems.

Figures 3.1 plots the expected time of the two different configurations where the broken line represents the consolidated single computer and the solid line is the clustered twin-computer result. The expected time in the system is plotted as a function of the mean time between failures. The arrival rate is held constant at 480 jobs per minute and each configuration has a total capacity of processing 900 jobs per minute. Figure 3.1 shows the comparison when the mean time to repair, MTTR, equals 60 minutes. It shows that the clustered twin-computer has a lower expected time in the system, even when the mean time between failures equals *one year*. The difference of the expected time in system becomes smaller as the mean time between failures gets longer as predicted by Equation (2.3). In general, the more reliable the computer system is, the better a consolidated single computer will be. However, a computer is considered extremely reliable if it only breaks down about once a year. Figure 3.1 clearly demonstrates the value of having a clustered twin-computer system even when the computer system is rather reliable.

For firms with a consolidated single computer system, numerical experiments show that it pays off to reduce the mean time to repair rather than to increase the mean time between failures. For example, when the mean time to repair is 60 minutes, it does not help the firm with a single computer system to extend the mean time between failures, even to a year. However, reducing MTTR to 10 minutes has an apparent benefit in terms of a shorter expected time in system than the clustered twin-computer system.

4. Concluding Remarks

This paper presents an analytical model to examine the impact of breakdowns on the decision whether to consolidate or cluster computing capacity. When breakdowns are not considered, standard queueing result favors consolidating server capacity. However, computers do break down in reality. This research finds that in most cases a clustered twin-computer system has a smaller overall time in the system, including both queueing and service time, than a consolidated single computer system having identical breakdown characteristics. Another key finding is that for firms with a consolidated single computer system, numerical experiments show that it pays off to reduce the mean time to repair rather than to increase the mean time between failures in terms of shorter time jobs spend in the system. For example, adequate backup of databases may not extend the mean time between failures of computer operations but will contribute to a speedy recovery. Of course, the benefit derived from reducing the mean time to repair should be weighted against the cost of doing so.[1]

Lastly, some interesting observations regarding computer breakdowns are in order. In todayís personal computer environment, computer downtime results more from software related problems. Sometimes, software malfunction may cause hardware failures. For example, Windows 95 operating system sometimes caused RAM burnouts.[2] As mentioned before, the model used to examine the impact of breakdowns is rather general and is not restricted to specific sources of computer failures due to hardware or software malfunctions. Moreover, the numerical experiments in this paper were conducted according to empirical data of mostly mainframes. Future research addressing personal computer systemsí breakdown characteristics and their impact should be of great interests.

Endnotes:

1,2 The author is indebted to the insightful comments of a referee.

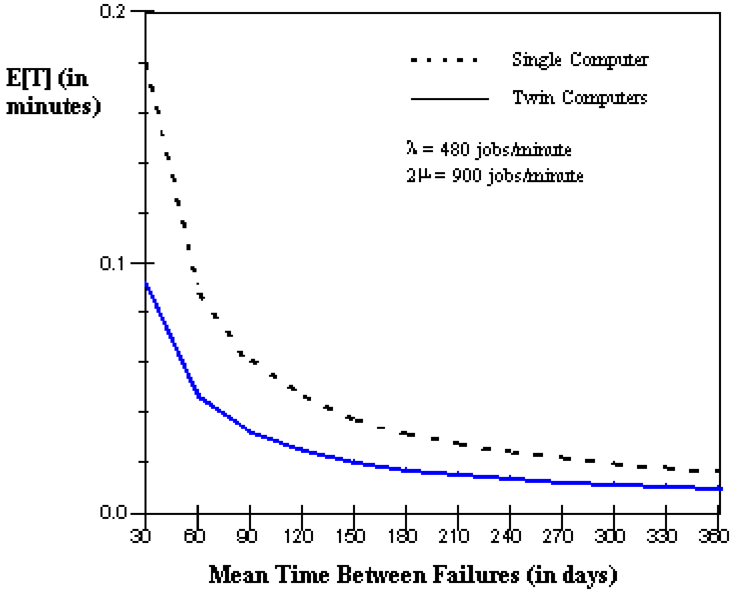

Figure 3.1 Expected time in the system as a function of MTBF (MTTF = 60 minutes)

References

Cheng, Hsing Kenneth, ìOptimal Capacity of a Firmís Computer Backup Center,î *Computers and Operations Research*, Vol. 22, No. 10, pp. 1015-1029, 1995.

Cheng, Hsing Kenneth, ìOptimal Internal Pricing and Backup Capacity of Computer Systems Subject to Breakdowns,î forthcoming in The Special Issue on Economics of Information Systems, Journal of Decision Support Systems, 1996.

Cheng, Hsing Kenneth, ìPricing and Capacity Decisions of a Twin-Computer System Subject to Breakdowns,î Working Paper, The College of William and Mary, 1997.

Federgruen, Awi and K. C. So, ìOptimal Maintenance Policies for Single-Server Queuing Systems Subject to Breakdowns,î Operations Research, 38 (2), 1990, 330-343.

Groenevelt, Harry, Liliane Pintelon and Abraham Seidmann, ìProduction Lot Sizing with Machine Breakdowns,î Management Science, 38 (1), 1992, 104-123.

Mitrany, I. L. and B. Avi-Itzhak, ìA Many-Server Queue with Service Interruptions,î Operations Research, 16, 1968, 628-638.

Moinzadeh, Kamaron, and Prabhu Aggarwal, ìAnalysis of a Production-Inventory System Subject to Random Interruptions,î Working Paper Series, School of Business, University of Washington, 1996, forthcoming in Management Science.

Posner, M. J. M. and M. Berg, ìAnalysis of a Production-Inventory System with Unreliable Production Facility,î OR Letters, 8, 1989, 339-345.