Requirement for High-Availability Communication Systems

Asst. Prof. Ismail M. Jaber
Electrical Eng. Dept.
Engineering College
Al- Mustansiriyah University
Baghdad-Iraq

Asst. Lect. Yasmin Abdul Ghani
Computer Eng. Dept.
Engineering College
Al- Mustansiriyah University
Baghdad-Iraq

Abstract

This study search in the availability requirement for the fault management server in high-availability communication systems. According to this study, one founds that the availability of the fault management server does not need to be 99.999% in order to guarantee a 99.999% system- availability ,as long as the fail-safe ratio (the probability that the failure of the fault management server will not bring down the system) and the fault coverage ratio (the probability that the failure in the system can be detected and recovered by the fault management server) are sufficiently high. Tradeoffs can be made among the availability of the fault management server, the fail-safe ratio and the fault coverage ratio to optimize system availability.

الخلاصة

هذه الدراسة تبحث في الاتاحية المطلوبة للمنظومة العاملة على إدارة الأخطاء في نظام اتصالات عالي الاتاحية. اعتمادا على هذه الدراسة, وجد انه ليس من الضروري أن تصل اتاحية المنظومة العاملة على إدارة الأخطاء إلى 99,999% لكي نحصل على اتاحية لنظام الاتصالات تساوي 99,999% (5 تسعات), طالما كانت نسبة السلامة في حالة الفشل(وهي ان تكون احتمالية فشل المنظومة لا تؤدي إلى فشل نظام الاتصالات بشكل كامل),وكذلك نسبة تغطية الخطأ(وهي احتمالية كشف الخطأ الحاصل في النظام وتغطيته من قبل المنظومة), احتماليات عالية نوعا ما.

سيتم اجراء مقارنة بين اتاحية المنظومة العاملة على إدارة الأخطاء مع نسبة تغطية الخطأ ونسبة السلامة في حالة الفشل لتحسين اتاحية نظام الاتصالات.

1. Introduction

Fault management plays an indispensable role in today's high-availability communication system. Fault management involves techniques for rapidly detecting, isolating and recovering system from faults, either automatically by the fault management software or manually by operators [1]. According to its function coverage, there are two levels of fault management in a communication system, i.e., equipment level and network level. At equipment level, fault management resides on the operational equipment, and detects, isolates and recovers failures in the equipment, e.g., brings up redundant power supply when the primary power supply fails. At the network level, fault management may adopt a server/client architecture with the server entity residing on specific equipment and the clients residing on the managed functional units.

The fault management server detects , isolates and recovers failures in the system, e.g., redirects traffic to redundant equipment when the primary one fails. Common intuition indicates that the server providing network-level fault management should be highly available in order to achieve higher system availability, for example, the fault management server should provide at least 5- nine (0.99999) availability in order to achieve 5-nine system availability. Based on this intuition, it was strongly recommended in [2] that the fault management software should run on fault-tolerant computers which can perform logic self-checking and have all of the main components (e.g., CPU, memory, I/O controller, bus, power supply and disk, etc.) physically duplicated. However, according to this study, this is not necessarily true. Besides the availability of the fault management server, there are another two parameters, i.e., the fault coverage ratio (the probability that the failures in the system can be detected and recovered) and fail- safe ratio (the probability that the failure of fault management server will not bring down the system), determining the availability of the system.

In order to dissipate this potential misconception, one needs to rephrase the system availability question as follows: what minimum level of availability needs to be achieved by the support systems (e.g., maintenance servers, network management servers, etc) in order to guarantee a 99.999% availability for the operational part of the system (that part in charge of delivering the main system functionality)?

The Markov models for a cluster of computers were constructed and system availability was studied as a function of fault coverage ratio and individual computer availability in [3][4]. It was found that higher system availability could be achieved over a cluster of computer with non- fault-tolerant architecture if higher coverage ratio could be well provided. However, the availability of the watchdog (i.e., the server) and its fail-safe characteristics were not considered in these two papers. Fail-safe behavior is the ability of a system to fail without producing a catastrophic result [5][6].

The concept of fail-safe systems has been well defined and widely used in LSI design [5]-[7] and railway applications [8], etc. The research in these areas focused on the conditions for achieving safety properties in a system, and its formal expression and proof. In this study, a new parameters are introduced, i.e., fail-safe ratio, to quantitatively investigate the impact of failsafe design to the system availability.

After constructing and solving a Markov chain for a communication system with N functional units (each functional unit configured as “1+1”) and one fault management server, the relationship of the server availability, fault coverage ratio and fail-safe ratio to the system availability will be studied. The study determines that the fault management server availability does not have to be 5-nine in order to achieve 5-nine operational system availability as long as the fail-safe ratio and coverage ratio are sufficiently high.

Availability along with the coverage ratio and fail-safe ratio are three important parameters of the management server, which can be traded off to achieve higher operational system availability. Because the fault-tolerant computers are very expensive, this observations lead to a more cost-effective design of the fault management server.

This study is organized as follows:

In section 2, it briefly introduce the fault management model used in this paper and the concepts of fail-safe, fail- safe ratio, and coverage ratio.

In section 3, we construct a Markov chain model and give a closed-form expression to the system availability.

2. Fault Management Model And Some Concepts

Server redirects traffic to the redundant equipment in case the primary equipment As shown in Figure 1, there are N functional units and one fault management server in the system. Each functional unit has “1+1” equipment, with one for primary and the other for hot standby. Each equipment element has a local (equipment-level) fault management. The fault management fails (the fault could not be recovered by the equipment-level fault management).

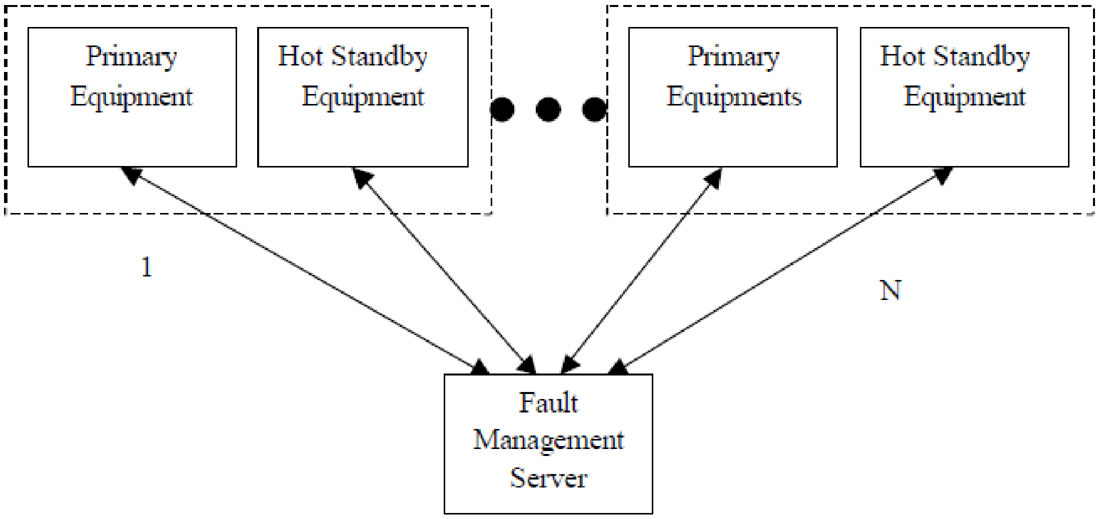


Figure 1. A Simple Model Assumed in the Paper

2.1 Fail-safe Ratio

Fail-safe is the ability of a system to fail but not produce a catastrophic result [5]. The concept on fail-safe system has been well defined and widely used in LSI design [5]-[7] and railway control [8]. This study revises the definition on fail-safe and extend it to the communication system.

A fault management server: is called fail-safe if the failure of the fault management server will not drag down the system[9].

In order to quantitatively evaluate the impact of the fail-safe property, must introduce a new parameter called fail-safe ratio, *pf* , which is defined as the probability that a failure of the fault management server will not bring the system down. Two extreme cases are pf =0 and pf =1[10].

pf=1 means that the failure of the fault management server will not drag down the system, which corresponds to the design that the fault management server is not involved in the normal operation of the system and only deals with the fault management functionality. pf =0 means that the failure of the fault management server will definitely drag down the system, which corresponds to the design that the fault management server is involved in not only the fault

management functionality but also the normal operation of the system, e.g., distributing traffic load and load balancing.

2.2 Fault Coverage Ratio

Because of the complexities and uncertainties of the network and its faults, and the design fault in the fault management software, a completely automated fault management is a very tough objective to reach. We assume the fault coverage ratio, i.e., the probability that the fault management server can detect the failure in the functional units and redirect traffic to the backup equipment automatically, to be *pc*., the coverage ratio is assumed to be the same for all the functional units. If the fault management server fails to detect and recover the failure in the functional units, it takes time Ts(i) for the operator to perceive the failure and manually switch the traffic for functional unit i. Improving the quality of fault management software and incorporating artificial intelligence [9][10]are potential solutions to enhance the fault coverage ratio.

3. Markov Model

To analysis Markov model for the system of this study, assumption below must taken in account:

1. The time to failure, repair and manual recovery are exponentially distributed.
2. If the failure can be detected and recovered by the fault management server, the recovery time is negligible.
3. All the functional units are configured as “1+1”.
4. Functional unit *i* is considered as working if there is one equipment at primary state.
5. The (operational) system is considered as working if all the functional units are working.

For functional unit *i* (with “1+1” equipment), consider the following two major cases:

- The fault management server is not working
- The fault management server is working

In the first case, functional unit *i* is working if and only if the primary equipment is working(no matter the standby one is working or not) and the failure of the fault management server does not bring down the primary equipment, with probability (1- Am)pf Ae(i), where :

$$Ae\,(i) = \mu\, f\,(i) / \,(\lambda f(i) + \mu\, f(i)) \quad \text{--------} \quad (1)$$

$$Am \;= \mu m \,/\, (\lambda m + \mu m) \quad \text{--------} \quad (2)$$

To evaluate the probability of the functional unit being working at the second case, we construct a Markov chain shown in Figure 2. We denote :

- (1,1) as the state where both equipments are working.
- (1,0) as the state where the primary equipment is working and the standby equipment is not functional.
- (0,1) as the state where the primary equipment fails and the standby has not taken it over.
- (0,0) as the state where neither equipment is working. The Markov chain in Figure 2 is obvious except the transition from (1,1) to (1,0) and (0,1), which is explained as follows. Starting from state (1,1), the system transits either
- to (0,1) with rate (1-pc) λf(i) if the primary equipment fails and the failure is not recovered automatically.
- to (1,0) if the backup equipment fails (with rate λf(i)) OR the primary equipment fails but the failure is recovered automatically (with rate pcλf(i)).

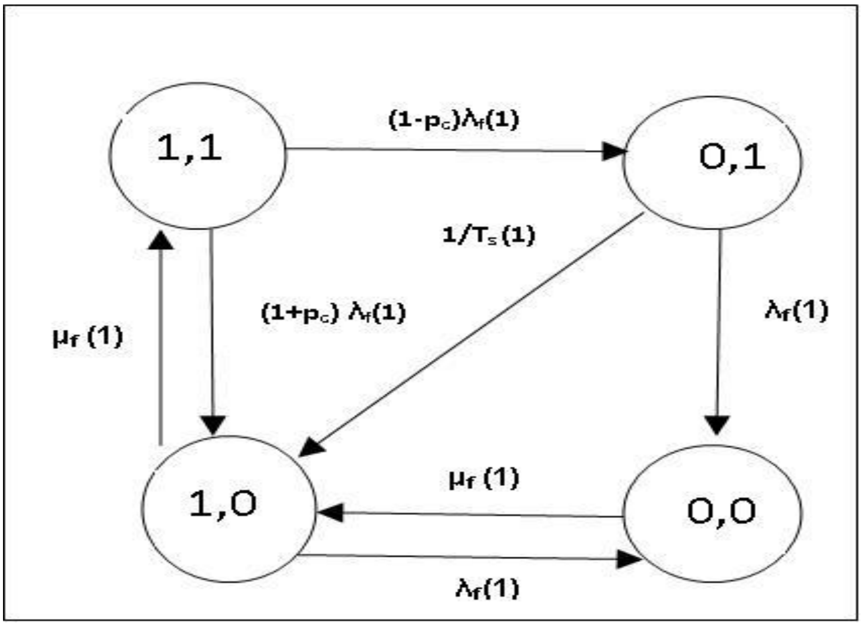


Figure (2) : Markov Chain For Functional Unit I With The Fault Management Server Being On

Solving the Markov chain, we have the functional unit availability under condition that the fault management server is working,

$$\theta(i) = \frac{2 - A_e(i)}{A_e(i) + \frac{2}{A_e(i)} - 2 + (1 - P_c)\lambda_f(i)T_s(i)/(1 + \lambda_f(i)T_s(i))} \qquad -- \ (3)$$

Apparently, it is a decreasing function of switchover time Ts(i).

The functional unit availability is given as[12].

$$A_f = (1 - A_m)P_f A_e(i) + A_m \vartheta(i) \qquad ----- \qquad (4)$$

and the system availability is given as [12].

$$A_s = (1 - A_m)Pf \prod_{i-1}^{N} A_e(i) + A_m \prod_{i-1}^{N} \theta(i) \qquad -- ------ \qquad (5)$$

4. Numerical Results and Discussions

This study gives an example for a system with two functional units:

A_f (1)= A_f(2)=0.999429549, (1/ λ_f (1)= 1/ λ_f (2)= 1 year, 1/ λ_f (1)= 1/ λ_f (2)=5 hours) and Ts(1) = Ts(2) = 30 minutes. Table (1) gives As versus A_m , p_f and p_c. It studys five cases:

(p_f =0.9, p_c=0.9), (p_f=0.99, p_c=0.9), (p_f =0.9, p_c=0.99), (p_f =0.999, p_c=0.999), and (p_f=1, p_c=1).

Calculating the availability difference between the first two cases, it can easily confirm that as is an increasing functioning of p_f, and $\partial A_s / \partial p_f$ is a decreasing function of A_m, as shown in table (1). Using the programs of MATLP (version 7),it is easily to programming and solve Marko's equations, than find the results in table (1) and (2). In the example given in Table 1, suppose it have designed a system with pc =0.99, pf=0.9, Am =0.999, and As =0.99989653. . In order to achieve 5-nine system availability, that can be by increasing pc and pf to 0.999, or increase Am to 0.99999. Table 2 lists the system availability with lower manual switchover time, i.e., Ts(1) = Ts(2) = 10 minutes. All the other parameters are the same as the cases in Table 1. One can see that the system availability is a decreasing function of switchover time. One interesting observation is that the second case provides higher availability than the third case when Am is between 0.99 and 0.9999. The turning point Am in this case is between 0.9999 and 0.99999. Therefore, decreasing the manual switchover time may

Table (1) As Versus Am, Pf And Pc (Ts(1) = Ts(2) = 30 Minutes)

Am	0.99	0.999	0.9999	0.99999
A_s(pf=0.9,pc=0.9)	0.9989152	0.99988628	0.99997719	0.99998628
A_S(pf=0.99,pc=0.9)	0.999876125	0.99997617	0.99998618	0.99998718
A_s(pf=0.9,pc=0.99)	0.998987317	0.99989635	0.99998746	0.99999655
A_s(pf=0.999,pc=0.999	0.999977204	0.99999645	0.99999837	0.99999856
A_s(pf=1,pc=1)	0.999987306	0.99999756	0.99999848	0.99999869
2^{nd} case-1^{st} case	0.000898973	8.9897E-05	8.9897E-06	8.9897E-07
3^{rd} case-1^{st} case	1.01647E-05	1.0257E-05	1.0266E-05	1.0267E-05
3^{rd} case-2^{nd} case	-0.00088809	-7.964E-05	1.2767E-06	9.3683E-06

It can also see that when we improve the server availability from 0.99 to 0.999 (or from 0.999 to 0.9999), $\partial A_s / \partial p_f$ shrinks 90 percent. This is because that is $\partial A_s / \partial p_f$ proportional to 1-A_m. Calculating the availability difference between the first and third case, it can easily confirmthat as is an increasing function of pc, and $\partial A_s / \partial p_c$ is an increasing function of Am.

Table (2) As versus Am, pf and pc (Ts(1) = Ts(2) = 10 minutes)

Am	0.99	0.999	0.9999	0.99999
A_s(pf=0.9,pc=0.9)	0.998984681	0.99989387	0.99998479	0.99999628
A_S(pf=0.99,pc=0.9)	0.999883655	0.99998377	0.99998822	0.99999718
A_s(pf=0.9,pc=0.99)	0.99898807	0.99989729	0.99998845	0.99999735
A_s(pf=0.999,pc=0.999	0.999977279	0.99999652	0.99999837	0.99999869
A_s(pf=1,pc=1)	0.999987306	0.99999756	0.99999848	0.99999869
2nd case-1st case	0.000898973	8.9897E-05	8.9897E-06	8.9897E-07
3rd case-1st case	3.3884E-06	3.4192E-06	3.4226E-06	3.4227E-06
3rd case-2nd case	-0.00089558	-8.647E-05	-5.567E-06	2.5236E-06

However, the change of w $\partial A_s / \partial p_c$ ith A_m is insignificant. The availability of the fault management server is lower than 0.999, and the third case provides higher availability when the availability of the fault management server is not less than 0.9999.

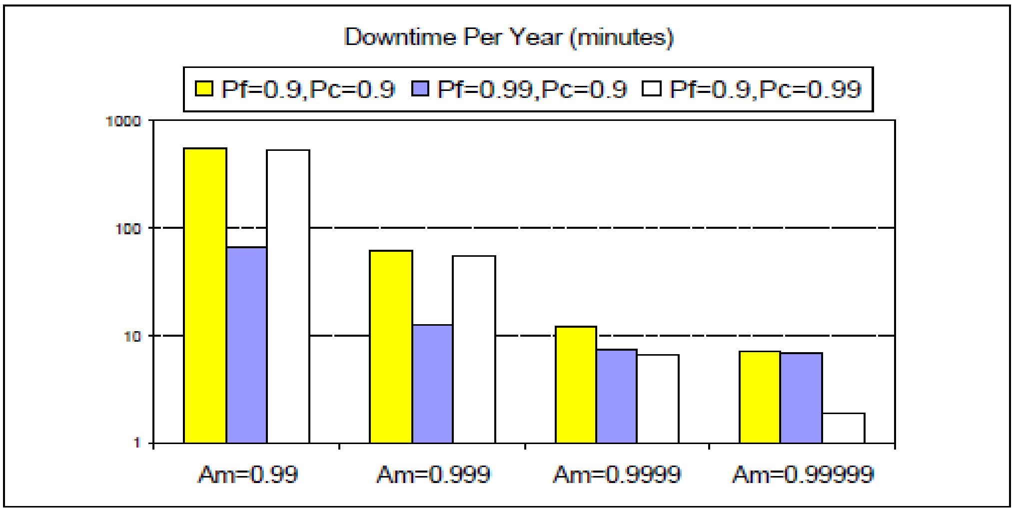


Figure (3) Comparison of Downtime for the First Three Cases (Ts(1) = Ts(2) = 30 minutes)

From figure (3),the aim of this study is clear, Am(Availability of the fault management server), is higher at 0.99 not at 0.99999(5 nine).So one thing should be noted, that improving system availability by improving the coverage ratio and fail-safe ratio does not mean the fault management server can be built on a very vulnerable platform.

System availability is an increasing function of the fault coverage ratio and fail-safe ratio. However, the improvement on system availability achieved by only increasing the fault coverage ratio and fail-safe ratio is limited. The best case is pf =pc.

Conclusion:

It can conjecture that when the availability of the fault management server is lower, fail safe ratio plays a more critical role than the coverage ratio to the system availability. On the contrary, when the availability of the fault management server is higher, coverage ratio plays a more critical role than the fail-safe ratio to the system availability.

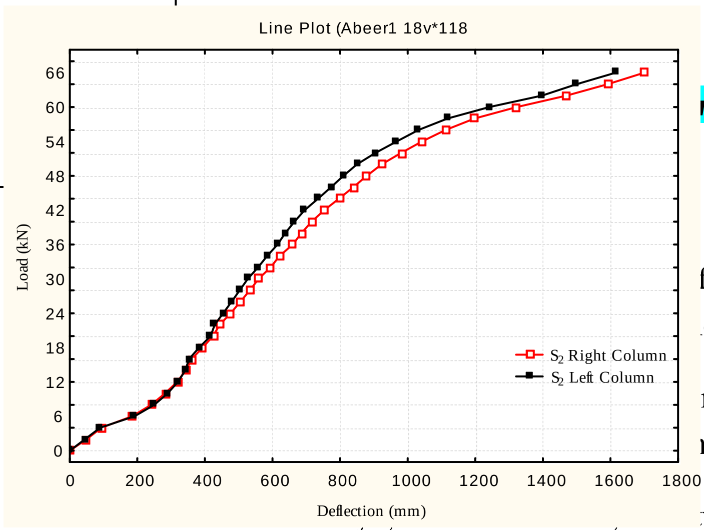
From table (1), one can see that 5-nine (0.99999645) system availability is achieved when pf=0.999, pc=0.999 and Am =0.999, which means that the fault management server is not necessarily to be 5-nine (3-nine is sufficient in this example) in order to achieve 5- nine system availability if the fail-safe ratio and coverage ratio are high enough.

Therefore, availability along with the coverage ratio and fail-safe ratio are three important parameters of the fault management server, which can be traded off to achieve higher system availability.

The findings are intuitively explainable. The system level availability eventually comes from the functional units. The presence of fault management server has both positive and negative impacts on the availability of the functional units. The server can help in that it increases the chance of recovering a failed unit (modeled as the coverage ratio), which otherwise would have to go through a longer manual repair.

The server presents negative impact as well if its own failure affects the rest of the system (the effect captured by fail-safe ratio). This explains why coverage and fail- safe ratios are so vital to system availability. At the same time, also keep in mind that the chance that both the equipment and the fault management server fail at the same time is very slim. It implies that failure on the server shall not present the major downtime for any functional units. This seems to explain why the server’s own availability does not appear to be a dominant negative factor.

push the turning point Am higher, which implies that decreasing the manual switchover time ma y weaken the impact of the coverage ratio.



for Supporting Fault Isolation and Recovery”, d Management Symposium, 2004, pp729-742

ment for Management System”, Stratus Computer /whitep/openvie.htm.

ware Fault Tolerance in a Clustered Architecture: Techniques and Reliability Modelling”, Aerospace Conference, 1999, pp141-149.

4. V.B.Mendiratta, “Reliability Analysis of Clustered Computing Systems”, The Ninth International Symposium on Software Reliability Engineering, 1998, pp268-272.

5. N. Kamiura, Y. Hata, K. Yamato, “On Design of Fail-safe Cellular Arrays”, Proc. of IEEE ATS’96, 2001, pp107-112.

6. M. Nicolaidis, S.Noraz, B.Courtois, “A Generalized Theory of Fail-safe Systems”, Proc. Of FTCS’25, 2001, pp273-281.

7. M. Lubaszewski, B. Courtois, “A Reliable Fail-safe System”, IEEE Trans. Computer, Vol.47, No. 2, Feb. 1999, pp236-241.

8. D.Essame, J.Arlat, D.Powell, “Available Fail-safe Systems”, Proceedings of the Sixth IEEE Computer Society Workshop on Future Trends of Distributed Computing Systems, 1999,pp176-182.

9. G. Pemido, J.Nogueira, C.Machado, “An Automatic Fault Diagnosis and Correction System for Telecommunications Management”, Proceedings of the Sixth IFIP/IEEE International Symposium on Integrated Network Management, 1999, pp777 – 791.

10. L.F.Pau, “Artificial Intelligence in Communications, Network Monitoring, Diagnosis and Operations”, In Knowledge-based System Diagnosis, Supervision, and Control, Plenum Press, New York, 1989.

11. J.S.Baras, M.Ball, S.Gupta, P.Viswanathan, P.Shah, “Automated Network Fault Management”, Proc. of MILCOM, 1997, pp1244-1250.

12. R.A.Sahner, K.S.Trivedi, A.Puliafito, “Performance and Reliability Analysis of Compute Systems: An Example-based Approach Using the SHARPE Software Package”, Kluwer Academic Publishers, Boston, MA, 1996.