ANALYSIS AND PERFORMANCE EVALUATION OF A FAULT-TOLERANT MULTISTAGE INTERCONNECTION NETWORK

HUI SENG KHEONG

A thesis submitted in fulfilment of the requirements for the award of the degree of Doctor of Philosophy

Faculty of Electrical Engineering
Universiti Teknologi Malaysia

MAY, 2002

ABSTRACT

A single error occurs in the non fault-tolerant Multistage Interconnection Networks (MINs) render a catastrophe to the MINs. The new scheme is to design a fault-tolerant MIN. Multiple paths between an input port and output port in the proposed network are established by chaining switching elements which have the same partition in the same stage. To enhance the performance and reliability of the proposed switch, sub-switches are straddled across the stages in the proposed network. This thesis examines the performance and design issues of fault-tolerant MINs. It first presents a survey of the current state of the art in MINs. Then, it investigates one of the most important design issues: cost-effectiveness. Following this comprehensive study, the thesis proposed a fault-tolerant MIN model. Analytical models for evaluation of the proposed network survivability and performance are presented . Finally the thesis presents fault-diagnosis methods to locate possible single fault occurs in the proposed network. Because maximum alternatives paths in the network are exploited, the proposed fault-tolerant switch has long lifetime and high bandwidth. Compared to other switches, it performs better in term of cost-effectiveness and throughput. In conclusion, we have successfully developed a fault-tolerant MIN which is: high survivability, simple control algorithm, full connecting capability and high bandwidth.

TABLE OF CONTENTS

PART ONE
THESIS CONTEXT

CHAPTER I

INTRODUCTION

1.1 Introduction of Multistage Interconnection Networks (MINs)

The discipline of interconnection networks (including MINs) is based on a strong mathematical foundation. Today's multistage switching networks have their origins in the theory of circuit-switched telephone networks of the 1950's, pioneered by Clos and Benes [1-5]. Many problems in the topological design and analysis of the interconnection networks lend themselves to elegant formulations in terms of problems in graph theory, combinatorics, and group theory. In MINs, researchers use different mathematical models to analysis the properties and topological equivalence among MINs.

In spite of active research for more than four decades, research activities in interconnection networks (including MINs) remain strong today, with a large number of papers continued to be published in various journals and conference proceedings.

The interconnections designed for use in multiprocessor systems range from a common bus to a crossbar network. The network design involves trade-off between cost, performance, and control complexity. Ideally, we would like any processor in the system to connect to any other processor or memory unit so that many processors can communicate simultaneously without contention. Such an ideal network is the crossbar network [6].

With a crossbar network, N processors can simultaneously access data from N memory modules, and any degradation in performance occurs only due to memory conflicts. Such nonblocking crossbar networks [7], however, are difficult to design for large multiprocessors because of their hardware cost. Furthermore, a crossbar network is difficult to expand and its failure (single point) can seriously affect the reliability of the system. Therefore, to reduce the cost, many networks have been designed using large number of small crossbars organized in stages. The Clos [1] and Benes [2] networks are examples of such multistage networks.

Over the years, a variety of multistage interconnection networks were proposed for use in multiprocessor systems. These include the Omega network [8], indirect binary n-cube network [9], baseline network [10], delta networks [11], banyan networks [12], data manipulator type network [13], and a number of their variants and derivatives. These networks generally use fewer gates than the crossbar network, and are typically constructed for $N = 2^n$ inputs and outputs with multiple "stages" (columns) of switching elements; the networks differ in the interconnection pattern between stages, the type and operation of individual switching elements, and the control scheme for setting up the switching elements. A basic requirement in the design of these networks is to provide full access capability, which means that any input terminal of the network should be able to access any output terminal in one pass through the network.

Most of the multistage networks proposed in the literature can be constructed with a 2 x 2 crossbar switches as basic elements, and have $n = \log_2 N$ switching stages with each stage consisting of *N/2* elements. Thus, the total number of gates in these networks is $O(N \log N)$, as compared to $O(N^2)$ for a crossbar. These networks can also be constructed with larger switching elements, with corresponding fewer stages. The properties of these networks can be mathematically studied with respect to networks constructed from 2 x 2 switches, and the results are easily extended to network designed with larger switches.

1.2 Major Characteristics of MINs

In the past, many parallel/distributed computing systems processing and ATM switches used Multistage interconnection networks (MINs) due to their simplicity and distributed processing capability [14]. Recently, optical MINs draw much attention from research communities [15-18]. In this thesis, we only focus on fault-tolerant switches which are based on MINs. The interconnection of several basic switching building blocks (also called switching elements) in a network formed MIN. The size of MIN ranges from a few inputs and outputs to some hundred or tens of thousands thereof. Basically MIN is composed of a large number of identical basic switching building blocks. Figure 1.1 shows a particular type of MIN called baseline network [10] with 8 input ports and 8 output ports (i.e. 8 x 8 switch).

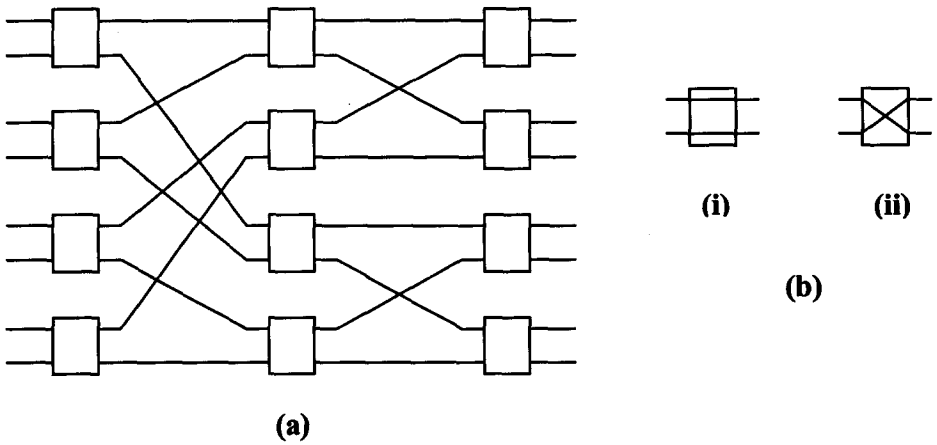

Fig. 1.1. : (a) Baseline network, (b) Operation of basic switching in (a): (i) Direct connection, (ii) Crossed connection.

Basically all MINs have the same major characteristics, i.e. to build an N x N switching, and they are as follows:-

(1) Only a unique path is provided between each input-output pair of the network.

(2) They are constructed of identical b x b switching elements (SEs).

(3) Their regularity and interconnection pattern makes them very suitable for very large scale chip integration.

(4) They have the self-routing property, requiring $\log_b N$ (where N is the number of inputs) digits to route a cell from input to output.

(5) They consist of $\log_b N$ stages, each stage having N/b basic switching elements.

1.3 Problem Formulation

From chapter 2, we conclude that all the networks under consideration are not perfect if the network is examined based on (i) high reliability; (ii) simple control algorithm; (iii) full connecting capability; (iv) low hardware complexity; (v) high bandwidth; and (vi) short average delay. Some networks achieve high reliability by sacrificing its simplicity and vice versa. Some networks even double the hardware cost to achieve higher reliability. Moreover, also a network fortified its reliability within SEs only and not the overall network.

By observing the weaknesses of these networks, We derive our proposed network. Our proposed network is based on solving some questions exist in the weaknesses of these networks. They are: (i) It is possible to mitigate the looping problems in Tzeng's and Cyclic networks? (ii) It is possible to increase the reliability in Tzeng's and Cyclic networks? (iii) How to maintain the simplicity in routing in our proposed network? (iv) How to keep the hardware cost as low as possible without double the hardware cost? (v) How to keep the out of sequence problem as minimum as possible in Itoh's network?

1.4 Contributions of this Thesis

The purpose is to modify and improve the weaknesses in these networks at the same time maintain its advantages. Analytical methods will be used in our analysis and comparison is made between these networks architectures based on the same parameters. Hence, this thesis will make the following 4 major contributions

Contribution 1: Creation of a new fault-tolerant network architecture based on Tzeng's network architecture. However, the concept of our proposed network can also be extended to other networks like Banyan-class networks and cyclic network.

Contribution 2: Development of the analytical model that is, our proposed network for evaluating the parameters like expected faulty element, and cost-effectiveness.

Contribution 3: Development of the analytical model for evaluating its performance in term of throughput under error-free and single-error environment.

Contribution 4: Development of a diagnostic method to locate the corresponding error in a specific location.

As shown in the following chapters of this thesis, contributions 1,2, and 3 will surpass Tzeng's network architecture.

1.5 Objective and Scope

The main objective is to solve the weaknesses of networks in chapter 2 (especially Tzeng's network architecture). The weaknesses will be improved to build a robust new MIN fault-tolerant switch. Since there are many MIN switches in existing literature, it is impossible to cover all of them, so our scope is narrowed down to fault-tolerant MIN switches by maintaining their advantages and improve their disadvantages.

1.6 Thesis Outline

This thesis is divided into seven chapters. This chapter is an introduction to the thesis, and it describes briefly about MINs. Following that, we outlined the major characteristics of MINs. Then, we outlined the development of our proposed network and its 4 major contributions. Finally in this chapter, we defined the objective and scope of this thesis.

Chapter 2 is a literature review of various fault-tolerant MINs, and there we investigate existing fault-tolerant MINs, understand their characteristics; and give their advantages and disadvantages. From there, we consider Tzeng's network architecture design and study its nitty-gritty; and the weaknesses in its design.

Chapter 3, sets out our proposed fault-tolerant MIN network that is based on Tzeng's architecture. Its configuration and its types of SEs are discussed in detail. Some calculations of the numbers of various SEs to be used in our proposed network is presented.

Chapter 4 is focused on devising an analytical model for later use in calculating parameters like expected faulty element $\bar{k}$, and cost-effectiveness in our proposed network architecture. We will devise it by using probabilistic approach. We will formulate the formulas for calculating corresponding parameters. Since the calculations are colossal and complicated, we will program them using *Mathematica (version 4)*. The code is given in Appendix of this thesis.

Chapter 5 is focused on devising another analytical model for evaluating the performance of our proposed network architecture. We also focus on our performance analysis in term of throughput based on some stated assumptions. The performance analysis is divided into two sub-analysis, one in an error-free environment and the other in a single-error environment.

Chapter 6 is a fault-diagnosis. We devise procedures to detect single error in our proposed network architecture and outline the corresponding fault model. Finally, methods for detecting the error are presented.

Chapter 7 sets out conclusions and further work of this thesis. We will also give some suggestions for further research on our proposed network architecture.

REFERENCES

1. Clos, C. (March 1953). "A Study of Non-Blocking Switching Networks." The Bell System Technical Journal, volume xxx11. 406-424.

2. Benes, V. E. (1965). "Mathematical Theory of Connecting Networks and Telephone Traffic." Academic Press, New York.

3. Benes, V. E. (1975). "Proving the rearrangeability of connecting networks by group calculations." The Bell System Technical Journal, 54: 421-434.

4. Benes, V. E. (1975). "Towards a group-theoretic proof of the rearrangeability theorem for Clos network." The Bell System Technical Journal, 55:797-805.

5. Benes, V. E. (1975). "Applications of group theory to connecting networks." The Bell Technical Journal, 54:407-420.

6. Chao, H. J., Lam, C.H. and Oki, E. (2001). "Broadband Packet Switching Technologies: A Practical Guide to ATM Switches and IP Routers." John Wiley & Sons, Inc. ISDN."

7. http://people.deas.harvard.edu/~jones/cscie129/lectures/lecture10/images/clos_network.html.

8. Lawrie, D. H. (February 1973). "Memory-Processor Connection Networks." University of IIIinois, Urbana-Champaign, Dep. Comput.Sci.Rep. UIUCDCS-R-73-557.

9. Pease, M. C. (May 1977). "The Indirect Binary *n*-Cube Microprocessor

Array." IEEE Transactions on Computers, vol. c-26, no.5. 458-473.

10. Wu, C. L. and Feng, T. Y. (August 1980). "On a Class of Multistage Interconnection Networks." IEEE Transactions on Computers, vol c-29, no. 8. 694-702.

11. Patel, J. H. (April 1979). "Processor-Memory Interconnection for Multiprocessors." Proc of the 6th Annual Symposium on Computer Architecture. 168-177.

12. Goke, L. R. and Lipovski, G. J. (December 1973). "Banyan Network for Partitioning Multiprocessing Systems." Proc. 1 st Annu. Comput. Architecture Conf. 21-28.

13. Siegel, H. J. and McMillen, R. J. (February 1981). "Using the Augmented Data Manipulator Network in PASM." IEEE Computer, vol. 14. 25-33.

14. Adams, G. B. Agrawal, D. P. and Siegel, H. J. (June 1987). "A Survey and Comparison of Fault-Tolerant Multistage Interconnection Networks." IEEE Computer, vol. 20, no. 6. 14-27.

15. Yi, P., Chunming, Q. and Yuanyuan, Y. (February 1999). "Optical Multistage Interconnection Networks: New Challenges and Approaches." IEEE Communications Magazine. 50-56.

16. Yuanyuan, Y., Jianchao, W. and Yi, P. (2000). "Permutation Capability of Optical Multistage Interconnection Networks." Journal of Parallel and Distributed Computing. Academic Press. 72-91.

17. Maier, G. and Pattavina, A. (July 2001). "Design of Photonic Rearrangeable Network With Zero First-Order Switching-Element-Crosstalk." IEEE Transactions on Communications, vol. 49, no. 7. 1268-1279.

18. Vaez, M. M. (September 1997). "Nonblocking Banyan-Type Optical Switching Networks Under Crosstalk." Georgia Institute of Technology, School of Electrical and Computer Engineering: Ph.D Thesis.

19. Tzeng, N.F. (1986). "Fault-tolerant Multiprocessor Interconnection Networks And Their Fault-Diagnosis." University of IIIinois at Urbana-Champaign: Ph.D. Thesis.

20. Rooholamini, R, and Garver, M. (Apr. 1994). "Finding the right ATM switch for the market." IEEE Computer. Vol. 27. 16-28.

21. Gaughan, P.T., and Yalamanchili. (May 1993). "Adaptive routing protocols for hypercube interconnection networks." IEEE Computer. Vol. 26. 12-24.

22. Ayanoglu, E. (1989). "Signal flow graphs for path enumeration and deflection routing analysis." IEEE Globecom. 1022-1029.

23. Lea, C.-T, A. (Dec 1986). "Load-sharing Banyan network." IEEE Transactions on Computer. vol. c-35. 1025-1034.

24. Lee, K.Y., and Yoon, H. (April 1990). "The B-network: A multistage interconnection network with backward links." IEEE Transactions on Computer. vol. 39. 966-969.

25. Tzeng, N.F., Yew, P.C. and Zhu,C,Q. (April 1988). "Realizing Fault-tolerant Interconnection Networks via Chaining." IEEE Transactions on Computers,vol. 37, number 4. 458-462.

26. Park, J-H., Yoon, H., and Lee, H-K. (August 1999). "The deflection self-routing Banyan network: A large-scale ATM switch using the fully adaptive self-routing and its performance analyses." IEEE/ACM Transactions on Networking, vol.7. no. 4.

27. Anan, M. and Guizani, M. (2000). "A fault-tolerant ATM switching architecture." Conference Proceeding of the IEEE International on Performance, Computing, and Communications. IEEE IPCC. 295-301.

28. Kevin, B.T. (1995). "The W-network: A low-cost fault-tolerant Multistage Interconnection Network for fine-grain multiprocessing." Proceedings of High Performance Computing Symposium. 81-92.

29. Tagle, P.U., and Sharma, N.K. (October 1996). "Performance of fault-tolerant ATM switches." IEE Proc.-Communication. vol. 143. no. 5. 317-324.

30. Li, J.J., and C.M. Weng (February 1994). "B-tree: a high-performance fault-tolerant ATM switch." IEE Proc.-Communication. vol. 141. no. 1. 20-28.

31. Somani, A.K., and Zhang, T. (March 1998). "DIRSMIN: A fault-tolerant switch for B-ISDN application using dilated reduced-stage MIN." IEEE Transactions on Reliability. vol. 47. no. 1.

32. Kamiura, N., Kodera, T. and Matsui, N. (2000). "Design of a Fault Tolerant Multistage Interconnection Network with Parallel Duplicated Switches." IEEE International Symposium on Defect and Fault Tolerance in VLSI system." 143-151.

33. Wang, K., and Lin, F.M. (1999). "Design and Implementation of a Fault-tolerant ATM switch." Journal of Information Science and Engineering 15. 521-541.

34. http://www.intel.com/research/silicon/mooreslaw.htm.

35. Johnson, B, W. (1989) "Design and Analysis of Fault-Tolerant Digital Systems." Addison-Wesley Publishing Company. 47-155.

36. Huei, L, S., and Jye, C.J. (1992). "Fault tolerance of Banyan using multiple-pass." IEEE INFOCOM. 867-875.

37. Cusani, R., and Sestini, F. (1991). "A recursive multistage structure for multicast ATM switching." IEEE INFOCOM. 1289-1295.

38. Chen, X., and Kumar, V. (1994). "Multicast routing in self-routing multistage networks." IEEE INFOCOM. 306-314.

39. Uematsu, H., and Watanabe, R. (1988). "Architecture of a packet switch based on Banyan switching network with feedback loops." IEEE J. Selected Areas in Communications. vol. 6. no. 9. 1521-1527.

40. Mohsen, G. and Ammar, R. (1999). "Designing ATM switching networks." McGraw-Hill.

41. Kumar, V.P., and Wang, S.J. (October 1991). "Reliability Enhancement by Time and Space Redundancy in Multistage Interconnection Networks." IEEE Transactions on Reliablity. vol. 40. no. 4. 461-473.

42. Neeraj K.S. (1996). "Space and Time Redundant Broadband Packet Switch." Conference Proceedings of the Fifteenth Annual International Phoenix Conference on Computers and Communications.

43. Adams, G.B. and Siegel, H.J. (May 1982). "The Extra Stage Cube: A Fault-Tolerant Interconnection Network for Supersystems." IEEE Transactions on Computers. 443-454.

44. Reddy,S.M. and Kumar,V.P. (August 1984). "On Fault-Tolerant Multistage Interconnection Networks." Proc. Int. Conf. Parallel Process.

45. Kumar, V.P. and Reddy,S.M. (June 1987). "Augmented Shuffle-Exchange Multistage Interconnection Networks." Computer. 30-40.

46. Parker, D.S. and Raghavendra, C.S. (April 1984). "The gamma network." IEEE Transactions on Computers, vol. c-33, no. 4. 367-373.

47. Itoh, A. (Oct 1991). "A Fault-Tolerant Switching Network for B-ISDN." IEEE Journal on Selected Areas in Communication, vol. 9, no. 8. 1218-1226.

48. Park, J-H., Yoon, H., and Lee, H-K. (1995). "The Cyclic Banyan Network: A Fault tolerant Multistage Interconnection Network with the Fully-Adaptive self-routing." Seventh IEEE Symposium on Parallel and Distributed Processing.

49. Blake, J.T., and Trivedi, K.T. (November 1989). "Multistage Interconnection Network Reliability." IEEE Transactions on Computers. vol. c-38. 1600-1604.

50. Shyy, D. J., and Lea, C. T. (1991). "Log_2(N, m. p) strictly nonblocking networks." IEEE Transactions on Communications." vol. 39. no. 10. 1502-1510.

51. Choi, S. B., and Somani, A. K. (May 1996). "Design and Performance Analysis of Load-Distributing Fault-Tolerant Network." IEEE Transactions on Computers. vol. 45. no. 5.

52. Kruskal, C. P., and Snir, M. (December 1983). "Performance of Multistage Interconnection networks for multiprocessors." IEEE Transactions on Computers. vol. c-32. no. 12. 1091-1098.

53. Chin, C. Y., and Hwang, K. (November 1984). "Packet Switching Networks for Multiprocessors and Data Flow Computers." IEEE Transactions on Computers. vol. 33. no. 11. 991-1003.

54. Kamiura, N., Kodera, T., and Matsu, N. (November 2001). "Design of Fault Tolerant Multistage Interconnection Networks with Dilated Links." IEICE Tans. Inf & Syst. vol. E-84-D. no. 11.

55. Lawrie, D. K. (December 1975). "Access and alignment of data in an array processor." IEEE Transactions on Computers, c-24(12). 1145-1155.

56. Kim, H. S. and Leon, G.A. (May 1990). "Performance of Buffered Banyan Network Under Nonuniform Traffic Patterns." IEEE Transactions on Communications. vol. 38. no.5. 648-658.

57. Hui, S, K., Kamaruzzaman, S, and Jasmy, Y. (July 3-5,2002) "An Augmented Chained Fault-tolerant ATM Switch." IEEE 5th International Conference on High Speed and Multimedia Communications (HSNMC 2002), Jeju Island, Korea. 397-400.

58. Dias, D. M. and Jump. J.R. (April 1981). "Analysis and Simulation of Buffered Delta Networks," IEEE Transactions on Computers, vol. c-30, 273-282.

59. Weiss, C. W. (Mar 1972). "Bounds on the Length of Terminal Stuck-fault Tests." IEEE Transactions on Computer, vol. c-21, No.3. 305-309.

60. Padmanabhan, K. (May 1984). "Fault Tolerance and Performance Improvement in Multiprocessor Interconnection networks." University of IIIinois at Urbana-Champaign: Ph.D. Thesis.

61. Wu, C.L. and Feng, T.Y. (October 1981). "Fault-Diagnosis for a Class of Multistage Interconnection Networks." IEEE Transactions on Computers, vol c-30, no. 10. 351-366.

62. Davis, IV N.J., Hsu, W.T.Y, and Siegel, H.J. (October 1985). "Fault Location Techniques for Distributed Control Interconnection Networks." IEEE Transaction on Computers, vol. c-34, 902-910.

63. Lin. W, Wu, C.L. (June 1982). "Design of a 2x2 Fault-Tolerant Switching Element." The Proceedings of the 9th Annual International Symposium on Computer Architecture, 181-189.