“The Penta-S: A Scalable Crossbar Network for Distributed Shared Memory Multiprocessor Systems”

Abdulkarim Ayyad
Department of Computer Engineering, Al-Quds University, Jerusalem, P.O. Box 20002
Tel: 02-2797024, Fax: 02-2797023

Abstract: - It is well known that crossbar switch has the best performance among the multiprocessor interconnection networks. However expanding this switch so that the performance grows linearly with the number of processing elements (N) is costly and complicated. The size of the switch grows as a function of N^2. This paper presents an expansion scheme, which keeps the linearity between the performance, the number of processing elements and the number of switches.16x16 and 32x32 crossbar switch modules are used as building blocks of this scheme. It is a modular scheme as well and so nothing needs to be redesigned. The results of mathematical analysis show that the performance of this scheme is better than that of a Delta MIN built if 4x4crossbar switch modules (Delta-4) and comparable to that of MINs built with the same crossbar switch module size, i.e., Delta-16 and Delta-32.The cost efficiency of this scheme is better than that of any other network except Delta-4 MIN.

Key Words: Interconnection Networks, Multiprocessors, Scalability, Modularity, Crossbar, Performance, Cost Efficiency, Mathematical Model.

1. Introduction

In multiprocessor and multi-computer systems, the interconnection network is of prime importance. It is the medium through which the processing elements or the computers of the system exchange information. Over the last three decades, large number of network topologies were proposed, designed, implemented and tested. Many were mathematically analyzed or simulated [1, 2]. In all networks, the major concern was and is still, to achieve a better performance and a better scalability. Performance is measured in terms of bandwidth or throughput, probability of acceptance and latency. Scalability is a rather imprecise term used to indicate a design that allows the system to be increased in size and in doing so, obtain increased performance. This is similar to the definition of modularity given by Stenstrom [3], which states " for a multiprocessor to be modular, the bandwidth of the network must be proportional to the number of processors in the system". Increasing the system means expanding it by adding more functional units to obtain higher performance without redesigning these units (expandability).

The crossbar switch network was first proposed by Wulf and Bell in 1972, and implemented by Wulf and his team in 1981 as reported in [1]. From its early analysis and implementations [1,2,4], It became well known that the crossbar switch network has the best performance among the multiprocessor networks. Its bandwidth steadily increases after a certain size. However, for a full crossbar network (NXN), doubling the number of its inputs and outputs means increasing the number of switches four times in order to keep the scalability measures. This process necessitates the redesign of the system. Many attempts were made in order to design scalable and easily expandable crossbar-based systems. In all those attempts, the designers realized that they couldn’t achieve full scalability and expandability at the same time. So, they resorted to compromise the two measures. The most famous example is the large number and variety of multistage interconnection networks (MINs) that were designed using small crossbar modules as building blocks. Samples of these MINs are shown in [5-7]. The performance of the MIN network lies between the performance of the crossbar and that of the shared bus[1,2]. Delta network built with 4x4 crossbar units proved to be the most cost effective among the MIN networks if the cost is calculated in terms of the number of switches used in the network [5]. However, the advent in VLSI technology has made this measure of less importance.

In 1993, Barry Wilkinson proposed, analyzed and simulated an overlapped scalable topology for the crossbar switch. In this topology rhombic crossbar modules are connected to each other so that the processor can access two memory modules in two neighboring crossbar modules. As far as the processor is trying to access memories in neighboring modules, its performance matches that of a full crossbar, but if the range of requests reaches farther, the bandwidth will be degraded drastically [8].The main issue now is to achieve a better performance for a modular(scalable and easily expandable) design.

In this paper, the author presents a scalable crossbar and a cost effective scheme which is easy to expand, easy to control, and yet has a performance comparable, and sometimes better than those of MINs having the same number of inputs and outputs. The scheme will be analyzed for multiprocessor systems then, in a separate work, it will be shown how the scheme can be used for multi-Computer systems. This work was done eight months ago for two purposes. Firstly to model the proposed expansion of the traditional crossbar switch used in multiprocessor systems, and secondly, to use its results as an indicator for the expected results of the expansion scheme of a newly proposed STC104 like crossbar switch for multi-computer networks. The idea of this switch was first proposed, designed and simulated at The Jordanian University of Science and Technology (JUST) in Jordan [9]. The results obtained then bettered those of the STC104 switch shown in [10]. Currently the author is supervising an M.Sc project on an improved version of the JUST switch and its expansion according to this scheme (Penta-S) at Al-Quds University in Jerusalem. The preliminary simulation results have shown to be consistent with the results of the mathematical model presented in this paper and they will be submitted in separate papers in the near future.

2. The Penta-S Scheme

The building block in this scheme is a full (nxn) crossbar module of the multiprocessor type. By multiprocessor type we mean that the data and the address are presented on separate parallel busses and each destination bus of the crossbar has its own arbiter. No input or output buffers are used. Each module can accommodate n processing elements (PEs) of a multiprocessor. An n-shuffle connection is used to connect up to (n+1) of these modules. For example, if 32X32 size crossbar modules are used as building blocks, then up to 1056 PEs can be connected using this scheme, see figure 1-A. Each PE in the network is provided with two ports; one to connect it to its crossbar module (hereinafter referred to as the crossbar port), and the other to connect it to a PE of another module in the scheme via the shuffle (hereinafter called the shuffle port). Each port has an input and an output see figure 1-B. The input of crossbar port has a bypass control unit to pass the information to the output of the shuffle port if the information belongs to a PE in another crossbar module. The input of the shuffle port is provided with a buffer in order to keep the incoming requests from other modules pending their service via the output of the crossbar port in the coming network cycles. This topology forms a Single Stage-Single Shuffle Scalable system, hence the name **Penta-S** scheme. The n-shuffle provides a direct link between each PE in the crossbar module and its correspondent (client) in another module. Each two correspondents are meant to provide a unique communication channel between two crossbar modules. For example, the shuffle port of PE5 of module 2 (PE25) is connected via the shuffle to the shuffle port of PE2 of module 5 (PE52). This makes PE25 and PE52 two correspondents, which serve the communication between the PEs of modules 2 and 5. Figure 1 shows a block diagram of the system and the two ports of the PE.

The shuffle link is designed to connect n+1 crossbar modules of nxn size. Note in figure 2 that the shuffle is designed so that the shuffle port output of PEij is connected to the shuffle port input of PEji and vice versa for all (i /= j). Note also that the ports of PEij are connected to PEkj for all (i = j),where k =n+1. In figure 2, PE11, PE22, and PE33 are connected to PE41, PE42 and PE43 respectively.

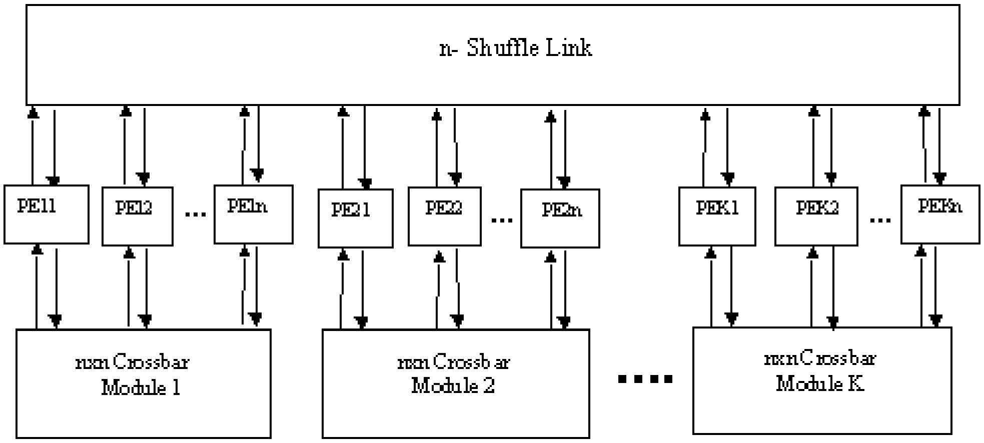


Figure 1-A: A block diagram of the system

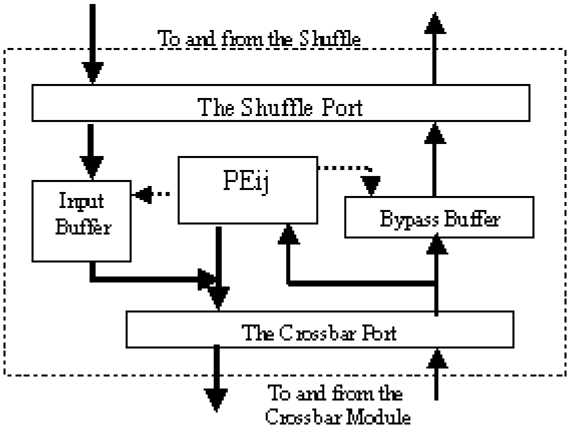


Figure 1-B: The Crossbar and the Shuffle ports of the processing element.

3. The Communication Process

In this system, all the requests are presented internally, i.e., within the crossbar module, regardless of whether the destination belongs to the same module as the source or belongs to another module. In case the destination belongs to another module, the bypass mechanism allows the data to go directly through the shuffle link to the client PE in that module. The data is stored in a temporary buffer in that PE in order to be sent to the final destination in the next cycle. For example, assume in figure 1 (see also figure 2 to follow the example) that the output of device PE12 shuffle port is connected via the shuffle to the input of device PE21 shuffle port. This means that device PE21 represents the client destination of module 1 in module 2. Assume now that device PE13 want to send data to device PE22.The communication process takes place as follows: In one network cycle PE13 sends the data to PE12. The Bypass mechanism of PE12 port recognizes that the data does not belong to PE12 and directs it to the output of the shuffle port. Thus the data goes directly to the input of PE21 shuffle port and stored in a local buffer. In the next cycle PE21 sends the data locally via module 2 to PE22.

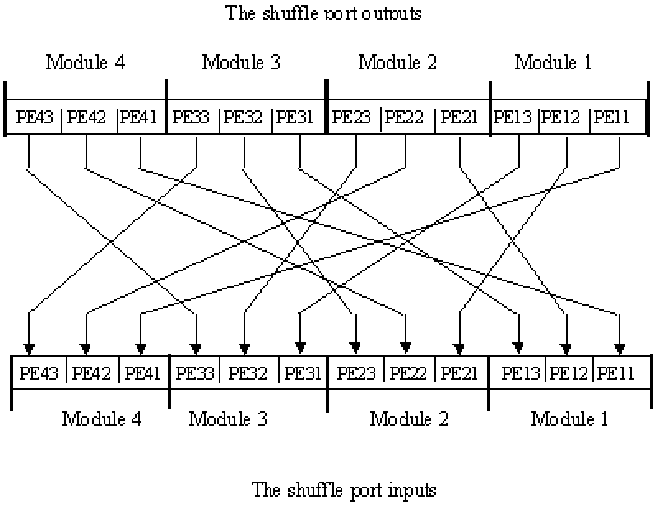


Figure 2: A 3-shuffle, which connects four 3X3 crossbar modules (a total of 12 PEs).

Regarding the priority, the devices give higher priority to their own requests in one cycle and a higher priority to the requests coming from other modules in the next. This pre-determined priority guarantees fairness between local and global requests. The priority for accessing the destination within the module depends on the priority algorithm used in the module design.

4. The System Mathematical Analysis

Mathematical analysis suits the tightly coupled multiprocessor networks more than multi-computer networks. This is due to the fact that in multiprocessor networks the request can be presented, arbitrated for and accepted or rejected in the same cycle. In multicomputer systems, other requests can arrive while previous requests are being served. This makes multi-computer networks difficult to be mathematically analyzed. However, for certain topology, analyzing the multiprocessor network gives a good indication of the corresponding multi-computer network behavior. So the author decided to mathematically analyze the multiprocessor version of the PENTA-S scheme. Simulation for the multi-computer version of PENTA-S will be presented in a near future work. In analyzing this scheme, the following assumptions are used:

1. nxn crossbar modules, which can connect n sources to n destinations, are used as a building block in this scheme.
2. k modules are used in building the system, where k can have a maximum value of n+1.
3. The devices present their own requests in one cycle and requests coming from other modules in the next.
4. The rejected requests will be discarded.
5. The cycle is defined as the time required for the request to be presented, arbitrated for and served.
6. The probability that a device is making a request during one cycle is r, and the probability that it is having a request coming from another module during a previous cycle is p.

4.1 Bandwidth and Probability of Acceptance Analysis

In the ith cycle the bandwidth of each module is given by the following equation:

$$BW(n,n) = n - n\left(1 - \frac{r}{n}\right)^n \qquad (1)$$

Assume k blocks are used then it is expected that $\frac{1}{k}BW(n,n)$ Requests are accepted for local PEs, and $\frac{k-1}{k}BW(n,n)$ requests are accepted and transferred to other blocks via the shuffle.

In the (i+1)th cycle, the probability that requests from other block have arrived is p, and the probability that a local processor has made a request is r. The priority is given for requests from other blocks. Then the probability that there is a request is given by:

$$R = r + p - rp \qquad (2)$$

So the expected bandwidth in the (i+1)th cycle is

$$BW_2(n,n) = n - n\left(1 - \frac{R}{n}\right)^n \quad (3)$$

So the total bandwidth for each block in both cycles

$$BW_{12}(n,n) = \frac{1}{k}\left[n - n\left(1 - \frac{r}{n}\right)^n\right] + \left[n - n\left(1 - \frac{R}{n}\right)^n\right] \quad \text{......... (4)}$$

where R is given by equation (2), and p is given by: $p = \frac{k-1}{kn}\left[n - n\left(1 - \frac{r}{n}\right)^n\right]$ i.e.,

$$p = \frac{k-1}{k}\left[1 - \left(1 - \frac{r}{n}\right)^n\right] \qquad (5)$$

Multiplying equation (4) by K gives the bandwidth of the system for two successive cycles

$$BW_{12}(k,n,n) = \left[n - n\left(1 - \frac{r}{n}\right)^n\right] + k\left[n - n\left(1 - \frac{R}{n}\right)^n\right] \quad \text{..........(6)}$$

So the average bandwidth of the system per cycle is given by:

$$BW(k,n,n) = \frac{BW_{12}(k,n,n)}{2} \qquad (7)$$

This equation reduces to

$$BW(k,n,n) = \frac{N}{2}\left[1 + \left(\frac{p}{k-1}\right) - \left(1 - \frac{(r + p - rp)}{n}\right)^n\right] \quad \text{.........(8)}$$

Where $N = nk$

The probability of acceptance P_a is defined as the ratio between the bandwidth and the expected number of requests during a cycle.

$$P_a(k,n,n) = \frac{BW(k,n,n)}{A} \qquad (9)$$

Where $A = \frac{N(r+R)}{2}$

For the purpose of comparison, the bandwidth and the probability of acceptance of full crossbar and a delta-4 multistage interconnection network are stated in the following equations:

The bandwidth of NXN crossbar is given by

$$BW(N,N) = N - N\left(1 - \frac{r}{N}\right)^N \qquad (10)$$

where r is the probability that a PE is making a request during a cycle. The probability of acceptance of the crossbar is given by

$$P_a(N,N) = \frac{BW(N,N)}{rN} \qquad (11)$$

The bandwidth of a delta-b network which uses (bXb)crossbar modules as building blocks is given by:

$$BW_{delta}(b,b) = b^i r_i \qquad (12)$$

Where the number of inputs and outputs N is given by: $N = b^i$, where i is the number of stages in the network

The results obtained from equations 8 to 12 are plotted and discussed in the next section of this paper.

5. The Results of Mathematical Analysis

5.1 The Bandwidth

The bandwidth and the probability of acceptance of PENTA-S, NXN crossbar and delta-4 networks are obtained from the above equations over various values of N,two values of r (1 and 0.5), and two values of K (32 and 16). The results are shown in figures 3 to 17. Note that the term “Delta-n” means a Delta MIN built of nxn-crossbar modules, the term Penta-n means a Penta network topology built of nxn crossbar modules, BW means the bandwidth, BW(NXN) means the bandwidth of a full NXN crossbar and BW(NX0.5N) means the bandwidth of half full crossbar network.

Figure 3 shows the bandwidth of Penta-32 as compared to Delta and crossbar networks. The figure shows that Penta-32 has the same bandwidth as Delta-4 for N<=256 at r = 1 and a higher bandwidth than Delta-4 for N>256 at (r=1). It also has a better bandwidth at r = 1 than the half crossbar network (NX0.5N) at r =0.5.

The bandwidth of Penta-32 is also compareable to those of half crossbar, Delta-32 and Delta-16 for N<512 at r = 1. It must be noted that Penta network betters the Deltas and the crossbar networks in being modular and easily expandable. As most of MINs use 4X4 switches as building blocks, i.e., they are of Delta-4 type, we can say that Penta32 betters them in having higher bandwidth, being more modular and easier to expand. Most MINs are of Delta-4 type because the 4X4 switch prooved to be the most cost efficient switch size for building MINs [5].

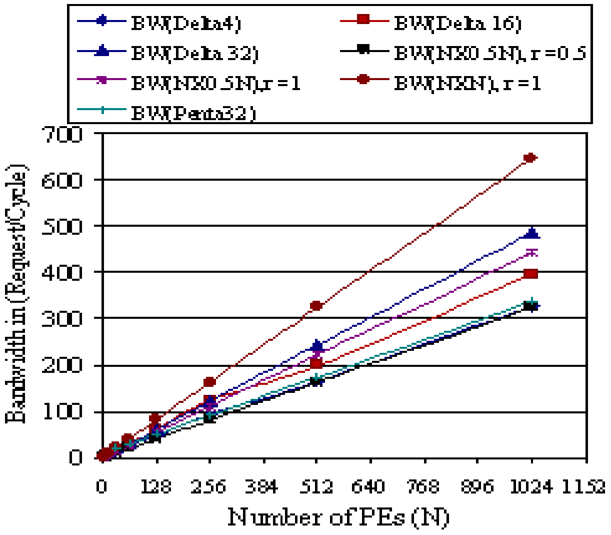


Figure 3: Bandwidth of Penta-32 as compared to other networks at r=1

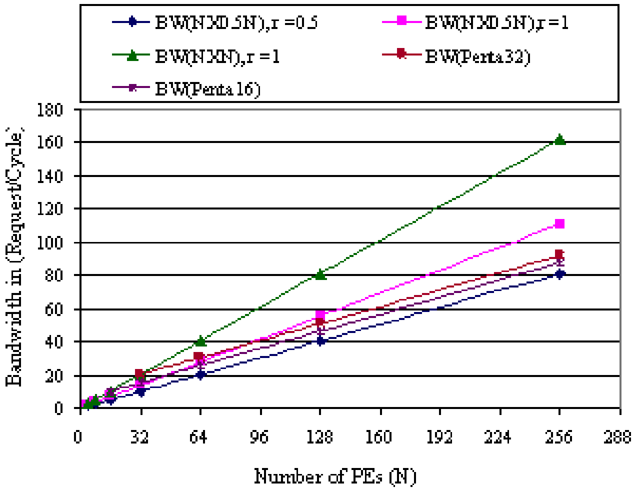


Figure 4:Bandwidth of Penta-16 as compared to Halfand Full-Crossbar networks

Figure 4 shows the bandwidth of Penta-16 as compared to other networks for 16=<N=<256 (The maximum size of Penta-16 is 272 PEs). Over this range, the bandwidth of Penta-16 at r = 1 is higher than bandwidth of the half crossbar at r = 0.5 and less than that of Penta-32 at r=1. Its bandwidth is also comparable to the bandwidth of the half crossbar at r = 1. Figure 5 shows the bandwidth of Penta-16 as compared to those of Delta networks at r=1. The figure shows clearly that the bandwidth of Penta-16 is nearly the same as that of Delta-4 and less than those of Deltas 16 and 32. Regarding the modularity and expandability, the above argument of Penta-32 applies to Penta-16 as well.

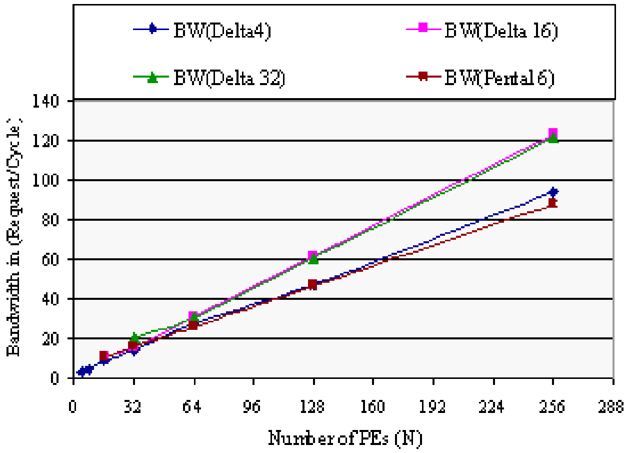


Figure 5: Bandwidth of Penta-16 as compared to Delta networks at r = 1.

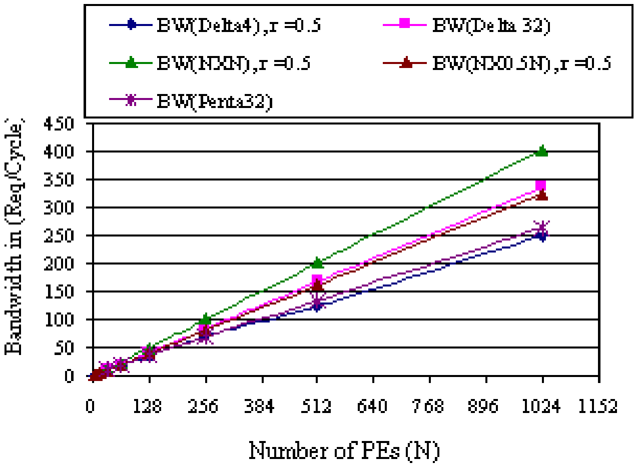
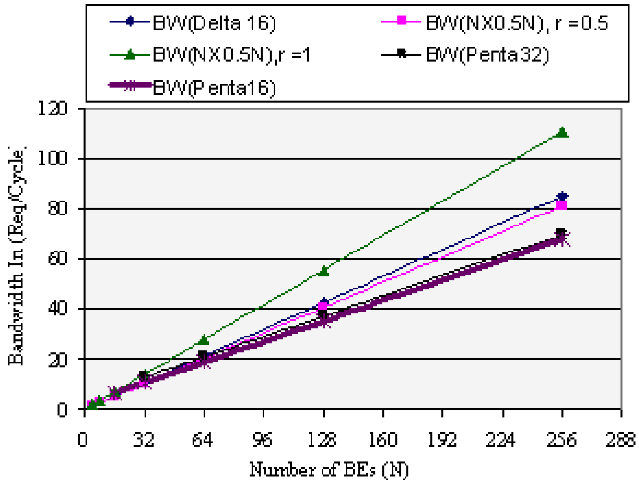


Figure 6 : The Bandwidth of Penta-32 as compared to other networks at r=0.5



Fifure 7 :Bandwidth of Penta-16 as compared to other networks at r=0.5

Figures 6 and 7compares the bandwidth of Pentas 32 and 16 to the bandwidth of other networks at r = 0.5. These figures show that reducing r does not change the position of Penta networks with respect to other networks regarding the bandwidth.

5.2 The Probability of Acceptance

The probability of acceptance is a measure which shows the probability of accepting a request during a cycle. The reciprocal of the probability of acceptance indicates the average latency, in cycles, of serving the request.

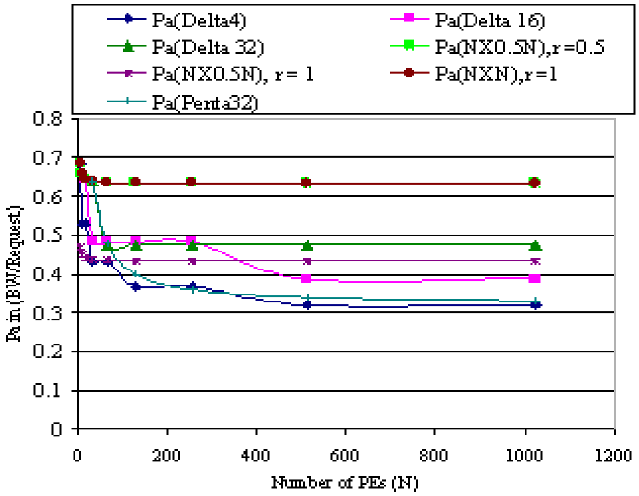


Figure 8:Probability of Acceptance of Penta-32 as compared to other networks at r=1

Figure 8 shows the probability of acceptance of Penta-32 as compared to other networks at r =1. It is clear from the figure that Penta-32 has a higher probability of acceptance than Delta-4 network for all values of N exept N =256 where they have equal probability of acceptance. We can note from the figure that Delta-16, Delta-32, and Full–crossbar networks have higher probability of acceptance than Penta-32. The probability of acceptance of Half –Crossbar network falls between the probability of acceptance of Delta-32 and that of delta-16 for N>384 and r =1.

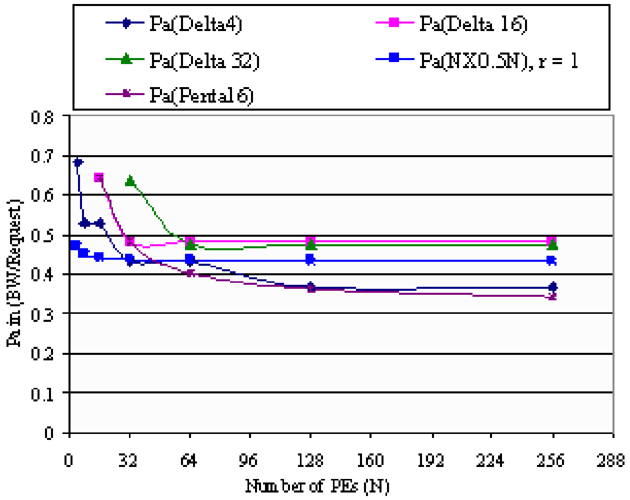


Figure 9:Probability of Acceptance of Penta-16 as compared to other networks at r=1

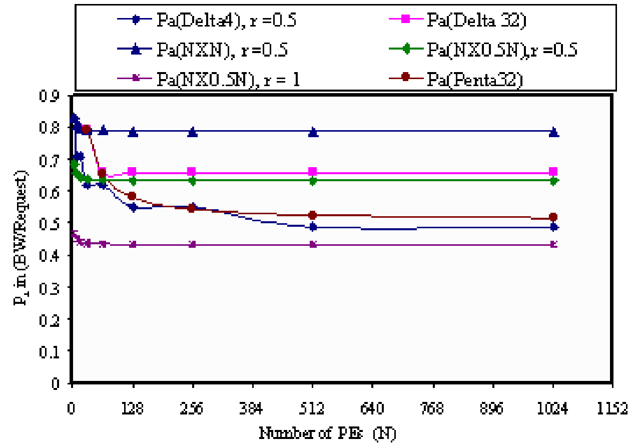


Figure 10:Probability of Acceptance of Penta-32 as compared to other networks at r=0.5

For N<384 Half-crossbar has less probability of acceptance than both Delta-32 and Delta-16. Also, we can note that in Delta networks, the larger the size of the switch, the higher the probabilty of acceptance. Figure 9 shows that Penta-16 and Delta-4 have , more or less, the same probability of acceptace for N > 48. Figures 10 and 11 show that changing the value of r to 0.5 does not change the position of Penta networks with respect to other networks regarding the probability of acceptance.

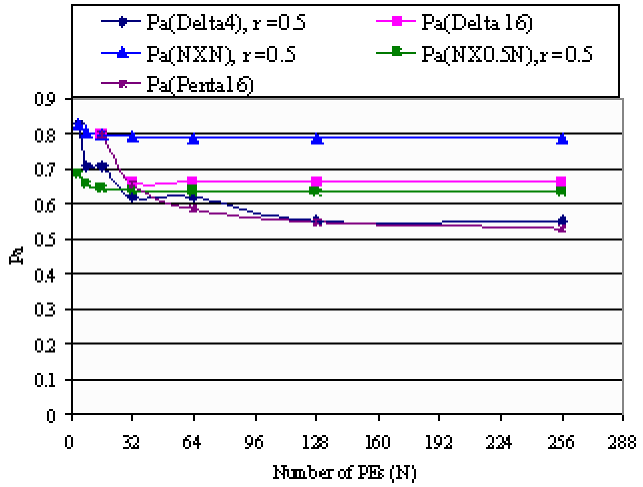


Figure 11 :Probability of Acceptance of Penta 16 as compared with other networks at r = 0.5

5.3 The Cost Efficiency

In this study the author calculates the cost efficiency in terms of the bandwidth per simple switch. This is simply because the networks used in this study utilizes various switch sizes. However, regardless of the switch size, all the switching elements are built of simple switches.

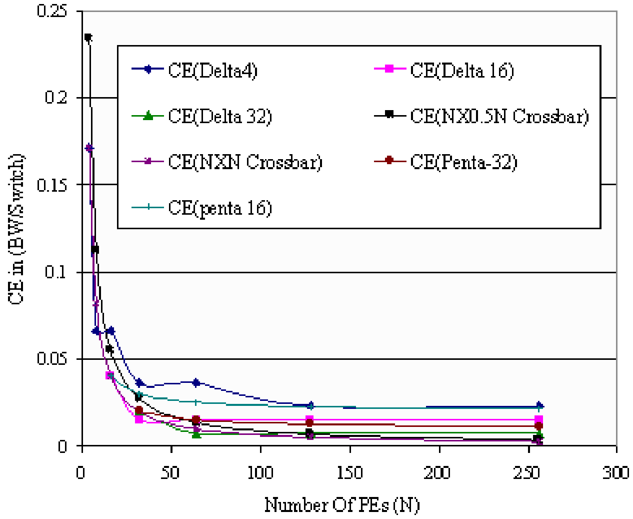


Figure 12:Cost Efficiency of Pentas 16 and 32 as compared to other networks

For example the 4X4 switch used in delta-4 is a 4X4 crossbar switch that is made of 16 simple switches and the 16X16 switch used in Delta-16 and Penta-16 is a 16X16 crossbar switch made of 256 simple switches.

Figure12 shows the cost efficiency of Penta-32 and Penta-16 as compared to those of other networks. It is clear that the most cost efficient Network is Delta-4, the thing with agrees with the calculations presented in [5]. However as now we have 16X16 and 32X32 crossbar switches integrated on one chip it is more tempting to build Delta MINS of these chips, i.e., Delta-16 & Delta-32, to have a better performance. Figure 13 compares the cost efficiency of Penta-32 to those of Delta-16 and Delta-32. It is clear that the cost efficiency of Penta-32 is higher than that of Delta-32. Delta-16 has a higher cost efficiency that that of Penta-32 only for number of PEs<100.

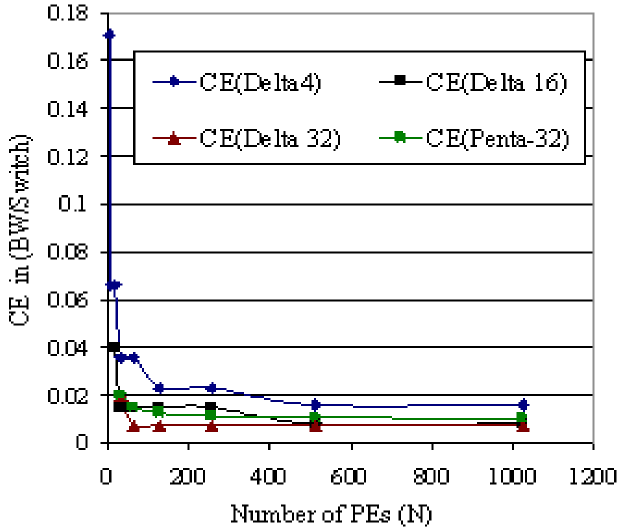


Figure 13:Cost Effeciency of Penta-32 as compared to Delta Networks

5.4 The Cost Per Processing Element

Figures 14 and 15 show two things, Firstly, Penta networks have a fixed cost per processing element regardless of the size of the networks, 16 simple switches per PE for Penta-16 and 32 simple switches per PE for Penta-32. For number of processing elements N>=128, Delta-4 matches Penta-16, in this measure. For number of PEs N>=64 Delta-16 matches Penta-32. So Penta network allows the usage of larger switch size without the penalty of the cost. Note that the cost per processing element for the full crossbar C/PE(FCB) and the half crossbar C/PE(HCB) grows with the same rate as the number of PEs, i.e., the size of the switch grows as a function of N^2, where N is the number of the PEs the switch designed to serve.

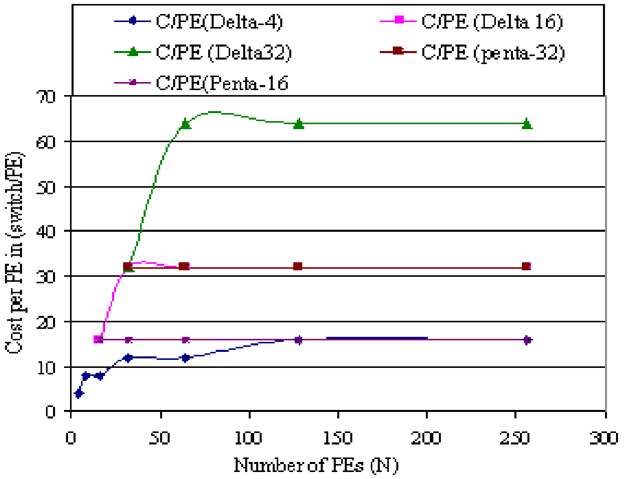


Figure 14:Cost per PE of Pentas 16 and 32 as compared to Delta networks

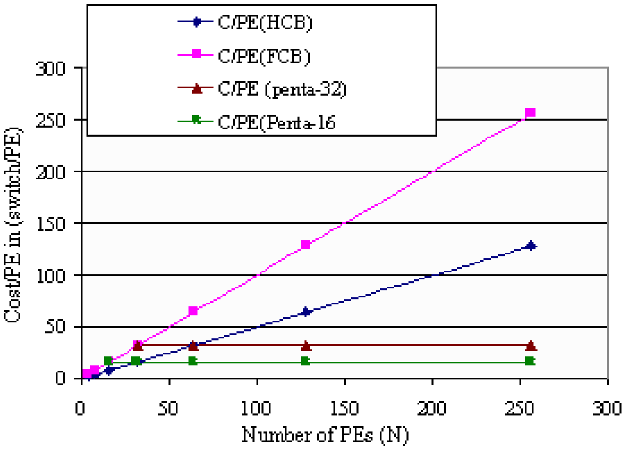


Figure 15:Cost per PE of Pentas 32 and 16 as compared to Half and Full Crossbar networks

6. Conclusion

In this paper we have mathematically analyzed the behavior of a new scheme of connecting clusters of multiprocessor systems. The term "scheme" means the topology plus its usage (especially, the local and global addressing plus assigning a client for each module in all other modules of the network).

This proposed scheme is a mixture of circuit switching and wormhole routing schemes, in the sense that inside the crossbar module, the Network behaves in a circuit switching fashion, where as in the case the request is made for other modules, the request bypasses th the local client PE of the destination module to the buffer associated with the client PE in the destination module. The client in the destination module allows the buffer to present its request internally in order to address its final destination. This is very much like the wormhole routing but with a single middle stage in the way to its destination. The performance of such a scheme proved to be better than that of the traditional MINs built of (4X4) crossbar switch modules, and comparable to that of modern MINs built of 16X16 and 32X32 crossbar modules. It is cost effective, easy to expand, and has a reasonable aggregate scalability

References

[1] K. Hwang and F. Briggs, " Computer Architecture and Parallel Processing", McGraw-Hill, USA, 1984.

[2] B. Wilkinson, "Computer Architecture; Design and Performance", 2^nd edition, Prentice Hall Europe, Hertford shire, U.K,1996.

[3] P. Stenstrom," Reducing the contention in Shared memory Multiprocessors", IEEE Computer, V. 21, No. 11,Nov, 1988, PP 26-37.

[4] B. Wilkinson and H. Abachi, " Crossbar Switch Multiprocessor System", Microprocessors and Microsystems, Vol. 7, No. 2, March 1983, PP 75-79

[5] L. Bhuyan and D. Agrawal, " Design and Performance of Generalized Shuffle Networks", IEEE Transaction on Computers, Vol. C-32, No. 12, Dec., 1983, PP 1081-1089.

[6] C-L. Wu and T-Y. Feng, "On a Class of Multistage Interconnection Networks", IEEE Transaction on Computers, Vol. C-29, No. 8, Aug. 1980, PP 694-702.

[7] J. Patel, " Performance of Processor-Memory Interconnection for Multiprocessors" IEEE Transaction on Computers, Oct., 1981, PP 771-780.

[8] A.B. Wilkinson, " On Crossbar Switch and Multiple Bus Interconnection Networks With Overlapping Connectivity", IEEE Trans. Compu., V. 41, No. 6, PP. 738-46.

[9] Nabeel Hasasneh, "Router Architecture With Collision Avoidance Control Using Crossbar Switch in Multicomputer Systems", M.Sc Thesis, Supervised by T. El-Dos and A. Ayyad, Computer Engineering Department, JUST University, Jordan, July 2000.

[10] P.W Thompson and J.D Lewis, "The STC104 Packet Routing Chip", SGS-Thomson Microelectronics Limited, Bristol,England, 1997, PP. 1347-1356.