J. Indian Inst. Sci., Mar–Apr. 1991, 71, 105–116.

Multidimensional multilink multicomputer: A general-purpose parallel computer

RAJAT MOONA* AND V. RAJARAMAN
Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore 560012, India

Received on March 31, 1990; Revised on December 24, 1990.

Abstract

This paper describes the architecture of a parallel computer called multidimensional multilink system [MMS] which is designed and developed at the Indian Institute of Science. This system is a general-purpose multicomputer where each computing element comprises a processor and its local memory. The computing elements communicate using message passing. There is no shared memory in the system.

Key words: Parallel computer, computer architecture.

1. Introduction

Multidimensional multilink system [MMS] architecture is developed with the motivation of providing a testbed for developing and testing concurrent algorithms. It is a general-purpose message-passing multicomputer architecture. In this paper, we describe the architecture and topology of MMS.

First, we would like to make a distinction between a multicomputer and a multiprocessor. A multicomputer can broadly be characterized by two attributes. First, it is a network of multiple computing units, each with a local memory and processing power, and second, the computing units communicate and cooperate in solving a problem. Each of these computing elements (CEs) can also be used as a single computer. A multiprocessor system, on the other hand, has multiple processing units each of which is fed with data and an instruction stream through a controller. The controller may be either central or distributed among the processors. The systems where all communications are through a common memory are called shared memory multiprocessor systems. Multicomputer systems in which computing elements communicate by sending messages are called message-passing systems.

Examples of message-passing multicomputers are hypercube[1] where point-to-point

*Present address: Department of Computer Science and Engineering, Indian Institute of Technology, Kanpur 208016.

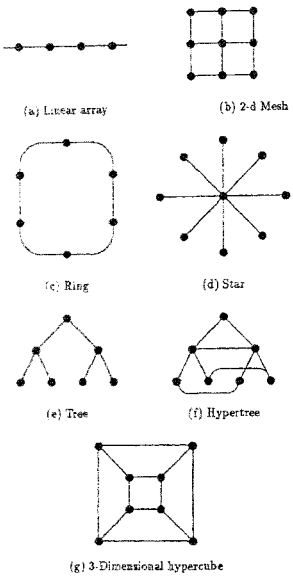


FIG. 1. Link-oriented static networks.

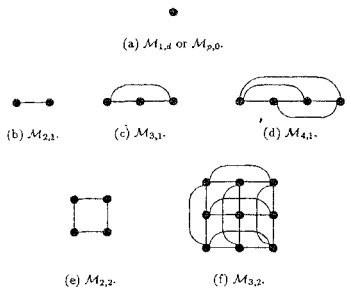


FIG. 2. MMS Configurations.

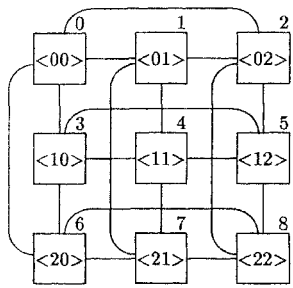


FIG. 3. Addressing of nodes in MMS. Addresses inside the box represent the addressing in radix 3 and the numbers outside in equivalent decimal.

communication links are used to interconnect processors ZMOB[2] and SMS[3,4] where processors are connected on a broadcast bus. Various other configurations of multicomputers use a variety of interconnection networks. Some of the popular networks are shown in fig. 1.

2. Architecture

MMS is an effort in the direction of developing a general-purpose multiprocessor system. We show some MMS configurations in fig. 2. It comprises a set of computers (referred as CEs throughout this paper), each with its private memory and a communication network

interconnecting them. All the CEs are alike and implement the same instruction set and interact by passing messages over the communication network. They are connected through a multidimensional network (described later). The architecture is scalable in terms of the hardware. An integrated software environment developed for the MMS architecture[5] makes it possible to write concurrent programs in any of the popular programming languages, namely, PASCAL, C, FORTRAN or PROLOG.

In the communication network topology of MMS, a number of CEs are organized in each dimension in a fully connected network using bidirectional communication links. A CE can perform broadcast operation by which data can be sent to all other CEs connected in this fully connected network. A CE can also do a selective broadcast (multicast) of data to a smaller set of CEs. This network is therefore referred as a multicast network. MMS architecture has a number of such fully connected multicast networks which are replicated in multiple dimensions, and are independent of each other. All the CEs in the system have access to equal number of multicast networks. Similarly, all multicast networks in the system connect equal number of CEs. These networks provide the same type of access mechanism for all CEs and are alike in terms of the method used for message passing between the CEs.

The multicast network has the following advantages over other structures.

1. It is simple to implement. It can be implemented using point-to-point communication links thus retaining the simplicity of a point-to-point communication network.
2. It allows the broadcast of data without any arbitration delay a message incurs when broadcast over a broadcast bus. The implementation of multicast network is less expensive than that of a broadcast bus, as the hardware costs in the latter are higher because of arbiters and priority-resolving logic required for each transmitting agent.
3. It is richer in connectivity than many other existing networks, and allows emulation of popular structures on MMS.
4. Communication bandwidth available for each of the transmitting agents is restricted by only the raw bandwidth of passive wires connecting various CEs. A CE can send data at any time over the multicast network and therefore does not wait for the availability of communication network.

The systems can be parameterized using two parameters.

Drop: Drop parameter of MMS architecture is the number of CEs that are connected by a single fully connected multicast network.

Dimension: In MMS, a CE is connected to multiple multicast networks. Dimension parameter of MMS architecture defines the number of fully connected multicast networks to which a CE in the system has access.

We denote the drop parameter of MMS architecture by p and the dimension parameter by d. An MMS structure with parameters p and d will be denoted by $\mathcal{M}_{p,d}$. In fig. 2, we show various MMS structures with different values of p and d.

MMS structure with p drops and d dimension, has p^d CEs. Each CE in an MMS structure with p and d parameters can be addressed uniquely by a set of labels containing p^d distinct labels. One such addressing scheme, which is followed throughout this paper, is to label the CEs with d digits in radix p. A CE therefore can be addressed using the following notations.

$$\text{Nodeaddress} = \overline{a_{d-1}a_{d-2}\cdots a_1 a_0} \quad p > a_i \geq 0 \forall i, d > i \geq 0.$$

We arbitrarily assign a set of multicast networks in one direction in the graphic representation of the MMS, as dimension 0 networks. Similarly, multicast networks in another direction are assigned dimension 1 networks. This way we assign all d dimension networks. CEs connected in one multicast network in dimension j are given the addresses such that only a_j differs in their node addresses. We use the following convention for addressing the CEs. A node address in angular braces denotes the node address in d radix p digits. Equivalent decimal address will be denoted by a number without any braces. In fig. 3, we show the node-addressing mechanism for a 3-drop, 2-dimension MMS structure. In this figure, multicast networks in dimensions 0 and 1 are shown by horizontal and vertical lines, respectively. For example, CEs $\langle 00\rangle$, $\langle 01\rangle$ and $\langle 02\rangle$ are connected fully here through bidirectional communication links in dimension 0. Similarly, the multicast network connecting $\langle 00\rangle$, $\langle 10\rangle$ and $\langle 20\rangle$ is a multicast network connecting these CEs in dimension 1.

Lemma 1. This addressing mechanism is a one to one and on to mapping of CEs on the address labels.

Proof: There are p^d CEs in an MMS structure with p drops and d dimension. To label the CEs, it is sufficient to have p^d labels. To label the CEs uniquely, all p^d labels in this set should be distinct. Addresses in radix p with d digits provide p^d labels. Hence, there exists a one to one correspondence between the set of all possible p^d addresses and p^d CEs. Since all these p^d labels are distinct, the mapping between CEs and the addresses is unique. Therefore, this addressing mechanism is a one to one and on to mapping of CEs on the address labels. □

Lemma 2. Two CEs, A and B, in the MMS structure $\mathcal{M}_{p,d}$ are connected iff their addresses differ in only one digit location.

Proof: This follows from the addressing scheme used to label the CEs in the MMS structure. Any two CEs are connected iff they share a single multicast network. Node addresses of all CEs connected through the same multicast network differ only in one digit location. Two CEs cannot be connected by more than one multicast network. This establishes the fact that two CEs $A(=a_{d-1}a_{d-2}\cdots a_1 a_0)$ and $B(=b_{d-1}b_{d-2}\cdots b_1 b_0)$ are connected if and only if there exists a unique j such that

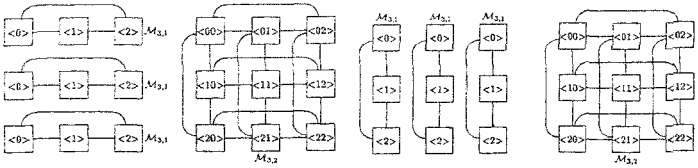
$$a_j \neq b_j, \quad d > j \geq 0$$

and

$$a_i = b_i \quad \forall i,\ i \neq j.$$

□

There are many other addressing mechanisms for the CEs in the MMS structure. For



(a) Prefix dimension expansion. (b) Postfix dimension expansion.

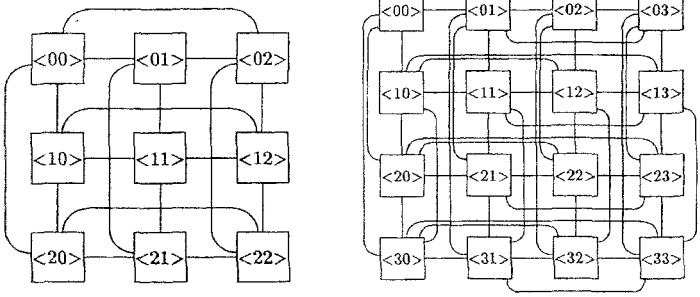
FIG. 4. Dimension expansion of MMS.

example, a new set of addresses may be obtained by interchanging the addresses of CEs in row i and row j and simultaneously the addresses of CEs in column i and column j.

An MMS structure $\mathcal{M}_{p,d}$ of dimension d and drops p can be obtained from p MMS structures $M^0_{p,d-1}, \mathcal{M}^1_{p,d-1}, \cdots, \mathcal{M}^{p-1}_{p,d-1}$, each of dimension $d-1$ and drops p. This is done by appending the address of CEs in the MMS structure by one more radix p digit. This digit can be prefixed or postfixed to the original addresses. Figure 4 shows such expansions.

Similarly, an MMS structure $\mathcal{M}_{p,d}$ can be obtained from an MMS structure $\mathcal{M}_{p-1,d}$ by adding $p^d-(p-1)^d$ CEs. These additional CEs can be addressed by extending the addressing scheme from radix $p-1$ to radix p. We call this a drop expansion. In fig. 5, we show such an expansion where an MMS structure $\mathcal{M}_{4,2}$ is obtained from another MMS structure $\mathcal{M}_{3,2}$ by adding seven CEs.

Each CE in MMS has $p-1$ neighbors in each dimension. In a d dimensional structure, a CE is connected to $d.(p-1)$ CEs. The switching distance between two CEs is defined as the number of routing steps a message undergoes. Switching distance between any pair of



(a) $\mathcal{M}_{3,2}$ MMS Structure. (b) $\mathcal{M}_{4,2}$ MMS Structure.

FIG. 5. Drop expansion of MMS.

CEs which are connected is 0. Maximum switching distance between any arbitrary pair of CEs is $d-1$. The topological distance between two CEs is one greater than the switching distance between them. Maximum topological distance between any two CEs in the MMS structure is d. An MMS with 2 drops is a hypercube structure where each CE has d neighbors. Thus hypercube structures fall in a class of MMS configurations $\mathcal{M}_{2,d}$.

Lemma 3. Topological distance between any pair of CEs 'A' and 'B' in an MMS structure with p drops and d dimension is the number of digits in the node addresses of 'A' and 'B' which do not match.

Proof: We prove this lemma with the addressing scheme described earlier. Two CEs are connected only if their node addresses in radix p differ exactly in one digit location. Topological distance between two connected CEs is one.

Now let's consider two CEs, 'A' and 'B', with node addresses $a_{d-1}a_{d-2}\ldots a_0$ and $b_{d-1}b_{d-2}\ldots b_0$. If these addresses differ in exactly two digit locations, say $i1$ and $i2$, then it is possible to find a third CE 'C' such that node addresses of CEs 'A' and 'C' differ only in digit location $i1$ and addresses of CEs 'B' and 'C' differ only in digit location $i2$. As CEs 'A' and 'B' are not connected, topological distance between 'A' and 'B' is greater than 1. Topological distance between CEs 'A' and 'C' and CEs 'C' and 'B' is 1. Therefore, topological distance between CEs 'A' and 'B' is 2.

We now prove this lemma with induction. Let's assume that the distance between CEs 'A' and 'B' is i if their node addresses differ in exactly i digit locations and this assumption is true for any $j < i$. Now, we consider the case in which node addresses differ in exactly $i+1$ digit locations. Clearly the distance between these two CEs cannot be less than or equal to i. Therefore, the distance between two CEs is more than or equal to $i+1$. Now it is possible to find a CE 'C' such that node address of 'C' differs in exactly one digit location from the node address of CE 'A' and i digit locations from the node address of CE 'B'. The distance between CEs 'A' and 'C' is 1 and between CEs 'B' and 'C' is i. Therefore, we conclude that distance between CEs 'A' and 'B' is $i+1$. As this lemma is true for $i=1$ and 2, it is true for any value of i.

This also proves the following lemma.

Lemma 4. Maximum topological distance between two CEs 'A' and 'B' in the MMS structure $\mathcal{M}_{p,d}$ is d.

Average topological distance is the distance a message travels, on an average, in the communication network of any system. The average distance[6], in terms of communication links, is defined as,

$$\text{Avg Dist} = \frac{\sum_{r=1}^{\text{MaxDist}} r \cdot N_r}{N-1}$$

where N_r is the number of CEs at a distance r, MaxDist, the diameter (maximum of the minimum distance between any two pairs of CEs) of the system, N, the total number of CEs in the system. Factor at the denominator does not take into account the CE from

which the AvgDist is being computed. By noting that this CE is at distance 0 from itself, the above equation can be modified as follows.

$$\text{Avg Dist} = \frac{\sum_{r=0}^{\text{MaxDist}} r \cdot N_r}{N}.$$

In MMS, all CEs have the same AvgDist as the topology of the system is symmetric with respect to the CEs. In an MMS structure $\mathcal{M}_{p,d}$, the number of CEs at a topological distance 1 from a CE is $\binom{d}{1} * (p-1)$. Similarly, the number of CEs at a topological distance 2 from any CE is $\binom{d}{2} * (p-1)^2$. In general, the number of CEs at a topological distance i is $\binom{d}{i} * (p-1)^i$. Therefore, the AvgDist for MMS can be given by

$$\begin{aligned}\text{Avg Dist} &= \frac{\sum_{i=0}^{d} i \cdot \binom{d}{i} \cdot (p-1)^i}{p^d} \\ &= \frac{d \cdot (p-1) \cdot (1 + (p-1))^{d-1}}{p^d} \\ &= \frac{d \cdot (p-1)}{p} \\ &= d \cdot (1 - 1/p).\end{aligned}$$

Communication network in MMS is a regular one and AvgDist from each CE is the same irrespective of its location in the network. A CE in the MMS structure has $d \cdot (p-1)$ links for accessing the communication network.

For an architecture, cross-section bandwidth is defined as the number of wires crossing a section which divides the number of nodes in the architecture into two equal size partitions. This is a very important measure for an architecture as it gives a rough estimate of the layout complexity of the interconnection network.

Lemma 5. For p drop and d dimension MMS, the cross-section bandwidth is

$$\frac{p^{d+1}}{4}, \quad \text{if } p \text{ is even}$$

$$p^{d-1}\left(\frac{p^2-1}{4}\right), \quad \text{otherwise.}$$

Proof:

Case 1: p is even.

Take a one-dimensional MMS structure. Divide it into two halves. Each half contains $p/2$ nodes. As one-dimensional structure is a fully connected network of p nodes, each of

the p 2 nodes in a partition connects each of $p/2$ nodes in the other partition. Therefore, p^2 4 links connect these partitions. It implies that the cross-section bandwidth of one dimensional MMS is p^2 4. Take p of these one-dimensional MMS structures and connect them in a two-dimensional structure. Divide this structure into two halves such that only links in dimension 0 are disconnected. As there are exactly p networks in dimension 0, the total number of links from one partition to another is $p*p^2/4$, or $p^3/4$. Now, take an arbitrary MMS structure with p drop and d dimension. Partition this structure into two halves by dividing each network in dimension 0 into two. There are p^{d-1} networks and the cross-section bandwidth of each network is $p^2/4$. Therefore, the cross-section bandwidth of this structure is $p^{d-1}*p^2$ 4, or $p^{d+1}/4$.

Case 2: p is odd.

Consider a one-dimensional MMS structure. Partition this structure into two such that one partition contains $(p-1)$ 2 nodes and the other $(p+1)/2$ nodes. The number of links from one partition to another is $(p-1)(p+1)/4$, or $(p^2-1)/4$. Following the proof for case 1 for a p drop and d dimension MMS structure, it is possible to prove that the cross-section bandwidth is $p^{d-1}(p^2-1)$ 4 by dividing each network in dimension 0 to partition MMS structure into two. □

2.1. *System components*

We have already introduced the CE which is the most important system component of the MMS. The other system components are host processor, and communication links.

We now describe the structure of a CE.

2.1.1. *Computing element (CE)*: In an MMS structure, all CEs are homogeneous and incorporate the same instruction set. Code for one CE can be executed by any other CE at the same speed. Further, all CEs in MMS have the same structure.

A CE in MMS is made of the two major modules. One of them is responsible for computation and we call it the arithmetic processing unit (APU). The other module is responsible for communication among CEs and operates under the control of APU. We call it a communication module (CM). The structure of a CE is shown in fig. 6.

2.1.2. *APU*: The arithmetic processing unit in a CE is used for all computations and implements a general-purpose instruction set. All APUs in the MMS structure are built around a general-purpose processor. An APU executes the tasks loaded by the host processor. It communicates with other APUs using the communication network and the communication module.

For a communication instruction, APU decides the multicast network on which the messages are to be sent or received. An appropriate packet is sent by the APU to the communication unit in its CM, responsible for data transfer on that multicast network. Communication modules work in parallel with the APU under its control.

An APU contains a general-purpose processor with a local memory. It also contains an

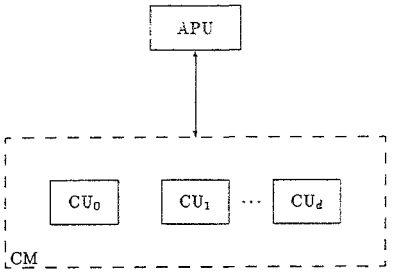


FIG. 6. Structure of a computing element (CE) in MMS

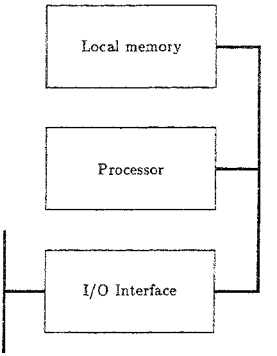


FIG. 7 Structure of an APU.

I/O interface which is used for handling communication instructions. APU sends and receives messages through its I/O interface to the communication units. Figure 7 shows the structure of an APU.

CM: A communication module in a CE is responsible for handling communication between CEs. It provides the necessary buffers for incoming and outgoing messages and relieves the APU from supervising the execution of communication instruction.

A CM contains d communication units (CU) each comprising $(p-1)$ communication links and buffers for incoming messages. All CUs work asynchronously and independent of each other. All CUs are controlled by the APU. In fig. 8, the structure of a link in CU is shown. All the outgoing messages from the APU are sent directly to the CU which connects to the InBuf of another CU at the other end of the communication link. Similarly, all incoming messages are stored in the buffer InBuf till APU takes it out of this buffer. A handshaking scheme between CEs ensures the integrity and reliability of data transmission.

2.1.3. *Host processor*: The host processor is a very important component in the MMS architecture. It is responsible for all input and output of a program that is to be executed on MMS. A programmer interacts with the system through the host processor. It is also responsible for loading the tasks on various CEs in the system.

In MMS, the host processor has the secondary storage for the system. A secondary storage in MMS may be distributed among all CEs but its presence with the host processor is absolutely essential for the system's operation. At power on time, all CEs in the system are booted by the host processor which boots from its secondary storage. The presence of a distributed secondary storage helps only in increasing the data storage space at each CE

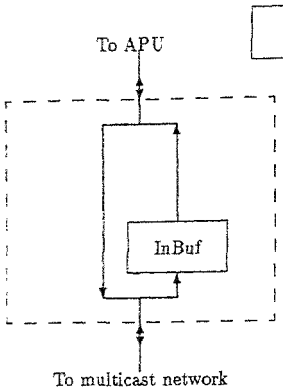


FIG. 8. Structure of a link in communication unit.

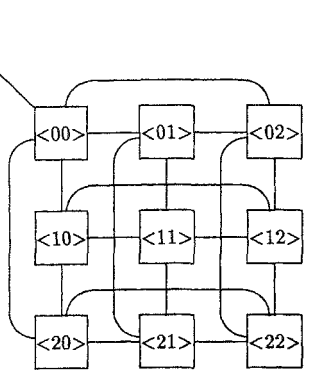


FIG 9. $\mathcal{M}_{3,2}$ Structure with separate host processor.

but cannot be used for booting these CEs. Host processor is also equipped with all input output peripherals like keyboard and display. As a programmer interacts with the system only through the host processor, the presence of I/O peripherals with host processor is essential.

Programs for MMS are written in a high-level language and are compiled on the host processor. In this process, a task mapping is also generated by a scheduler running on the host processor. Host processor then loads various compiled tasks into the memory of CEs in the system prior to program execution. In MMS, a program is statically scheduled and at runtime no scheduling is attempted.

During the execution of a program, the host processor also sends the input to the program and collects the output of the CEs. Any interaction with the programmer is through the host processor. Therefore, all interactive input and output of the program are done by the host processor only.

In MMS, usually the CE with nodeaddress 0 does all functions of a host processor. However, a separate host processor may be attached to the CE with nodeaddress 0 which may be used for interaction of a programmer and rest of the system. Such a structure is shown in fig. 9 for a 3-drop, 2-dimensional system.

2.2. *Architectural features*

To summarize, the architectural properties of the MMS structure are (assuming an MMS structure with p drops and d dimension):

1. The MMS structure comprises p^d CEs. All CEs in the MMS structure are identical and have local memory. There is no global memory in the MMS architecture.

2. Total number of neighbors of a CE is $d.(p-1)$.
3. There are three types of communication through each multicast network in the system. In point-to-point communication, a CE communicates data to another CE connected by a multicast network. In the second type, a CE multicasts data to other CEs connected to it through a multicast network. A message can be passed to selective CEs connected through multicast network in a single instruction. For example, in a one-dimensional structure with four CEs, a message from processing node 0 can be sent to processing nodes 1 and 3 but not to processing node 2. However, multicast communication can be achieved in a single instruction. The third type of communication allows *broadcast* of data to all CEs connected through a multicast network.
4. A CE in MMS is connected to d multicast networks.
5. A multicast network in MMS connects p CEs together in a fully connected network. This network allows a multicast or a broadcast communication among CEs in a single instruction.
6. Total number of multicast networks in the system is $d.p^{d-1}$.
7. AvgDist parameter for the MMS is $d.(1-1/p)$.
8. Interconnection network in the MMS is a passive network connecting CEs and is very reliable.
9. An existing configuration of MMS can be extended either by increasing the number of drops in a multicast network or by increasing the dimension of MMS.
10. Maximum switching delay in MMS is $d-1$.

Lemma 6. In MMS, there are $d(p-1)p^d/2$ bidirectional communication links.

Proof: In MMS, each CE has $d(p-1)$ neighbors connected through d multicast networks. As each of these neighbors is connected by a bidirectional communication link, there are $d(p-1)p^d/2$ bidirectional communication links. □

Lemma 7. The MMS communication network can carry $d(p-1)p^d$ messages simultaneously.

Proof: This follows from the previous lemma. Each bidirectional communication link can have two messages simultaneously and, therefore, there can be $d(p-1)p^d$ messages in the communication network. □

3. Conclusion

In this paper, we described the MMS architecture for general-purpose parallel computing. It has a good connectivity and can be used to simulate a variety of architectures and to try out the algorithms designed for these architectures. The architecture has a good expansion capability. There are more than one path connecting two CEs making the system fault tolerant. A memory element in the MMS architecture can be addressed by a tuple

containing CEs address and memory address within that CE. This way a global addressing scheme for memory can be used.

Three versions of the MMS architecture have been implemented using 'off-the-shelf' components. The first version is a fully connected network of 4 CEs ($\mathcal{M}_{4,1}$). The second system is a two-dimensional network of 9 CEs ($\mathcal{M}_{3,2}$). Finally, the third system is an 8-processor hypercube ($\mathcal{M}_{2,3}$). Each link in these implementations provides an 8-bit-wide parallel communication between nodes and operates at processor-memory speed. The system provides a facility to selectively broadcast (multicast) data to a few nodes. This improves the usable bandwidth of the link. In a hypercube, broadcast of a datum to 2 nodes requires twice the time of communication to one node, whereas in MMS, this can be done in a single step. Thus an effective bandwidth becomes p times the processor memory bandwidth. An integrated software environment allows a programmer to program the machine independent of the structure in any of the programming languages like Pascal, C, Fortran and Prolog.

Acknowledgements

Authors would like to acknowledge the Department of Electronics, Government of India, and UNDP who have jointly sponsored this project.

References

1. SEITZ, C. L. The cosmic cube, *Commun. ACM*, 1985, **28**, 22–33.

2. RUGER, C. ZMOB: Hardware from a user's viewpoint, *Proc. IEEE Comput. Soc. Conf. on Pattern Recognition and Image Processing*, 1981.

3. KOBER, R. The multiprocessor system SMS 201 – Combining 128 microprocessors to a powerful computer, *Proc. COMPCON*, Fall 1987.

4. KOBER, R. AND KUZNIA, CH. SMS–A multiprocessor architecture for high speed numerical calculations, *Proc. Int. Conf. on Parallel Processing*, 1978.

5. RAJAT MOONA AND RAJARAMAN, V. A software environment for general purpose MIMD multiprocessors, *Proc. IEEE TENCON 89*, Bombay, Nov. 22–24, 1989.

6. BHUYAN, L. N AND AGARWAL D. P. A general class of processor interconnection strategies, *Proc. 9th A. Symp. Computer Archit.*, Austin, Texas, pp 90–98, April 26–29, 1982.