How Much Adaptivity is Required for Bursty Traffic?

Ludmila Cherkasova* Al Davis† Vadim Kotov* Ian Robinson* Tomas Rokicki*
* Hewlett-Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94303
† Department of Computer Science, University of Utah, Salt Lake City, UT 84112

Abstract

Deterministic routing strategies are cheap and fast to implement but suffer from increased message latency due to contention for resources in a packet switching fabric. Adaptive routing strategies are inherently more complex which may result in slower routing. Our goal is to investigate the trade-offs involved in using different routing strategies. This paper presents the results of a simulation study designed to answer this question for realistic *bursty traffic* workloads.

1 Introduction

The work presented here is a natural extension of our previous work on a high perfomance router called the Post Office which was used to form the interconnect fabric for a scalable parallel multiprocessing system called Mayfly [4]. The Mayfly processing element (PE) architecture was designed to hide communication latency and hence the Post Office was designed primarily to provide a high capacity fabric. The Post Office was a fully adaptive *virtual cut-through* architecture [5]. Congested Post Office packets would not adapt immediately but would wait for a certain *stagnation delay* before choosing an alternate path.

We are now interested in creating an improved version of the Post Office which we call *PO2* that does not require a PE as complex as that provided in the Mayfly design. Since latency may be more difficult to hide in a more conventional PE design, low latency message traffic becomes the primary goal. Adaptivity is costly [2] both in terms of router complexity and in terms of latency when suboptimal paths are chosen. Several low latency deterministic routers have been developed [7, 3] but we are still interested in the potential use of limited adaptivity to bypass temporary congestion in the fabric rather than the added latency required to just wait for the resource.

A major concern we will address is how much routing adaptivity is enough for efficient transfer of different types of traffic. To this end we investigate a deterministic strategy and two forms of adaptive strategies.

Applications are primarily concerned with variable-length messages; the network interface must divide these into fixed-sized packets. This situation dramatically changes the traffic pattern because instead of uniformly distributed packets, there are variable-length bursts of packets going from some source node to some other destination node.

This shift in perspective—from packets to messages—has another effect for performance evaluation. Interconnect design usually focuses on minimizing the latency of individual packets through the interconnect, while the real goal is to minimize the latency of complete messages.

When compared with uniform random traffic, bursty traffic generates a different type of port contention. Instead of occasional packets competing briefly for the same port, two bursts compete for some port during a longer period of time. If this contention is not controlled, packets can build up in the preceding nodes, leading to more contention and eventually complete interconnect saturation [6]. To prevent this situation from happening a flow control mechanism based on "backpressure" is used.

However, the flow control provided by backpressure and the routing freedom provided by adaptivity are in conflict with each other. The remainder of the paper presents our results in more detail.

2 The PO2 Interconnect

The *PO2* interconnect topology is a continuous hexagonal mesh which permits each node in the fabric to communicate with its six immediate neighbors. Figure 1 illustrates a sample fabric containing nineteen nodes (only one axis is wrapped for clarity). The seventh port (the PE port, also not shown) connects each node to its corresponding processor. It is convenient to define the size of the interconnect by the number of nodes on each edge. For example, the interconnect shown in Figure 1 represents an $E3$ interconnect. The total number of nodes in an En interconnect is $3n(n-1)+1$. Thus an $E3$ interconnect has nineteen nodes, whereas an $E6$ consists of ninety-one nodes.

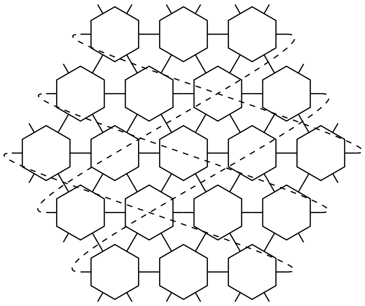

Figure 1: *PO2* Topology

Messages traveling through the interconnect are split into fixed-length *packets*. The first few words of a packet comprise the packet *header* which contains the source and destination addresses of the packet as well as a unique message and packet identifier.

The nodes in *PO2* are essentially buffered switches. The internal buffer pool receives packets from, and transmits them over, the seven ports. Each port is bidirectional, the link between them being half-duplex.

Routing logic decides which port or ports an arriving packet should be forwarded to. If the port is available, the packet transmission starts, even if it has not all been received. This virtual cut-through technique [5] leads to lower per-hop latencies than the alternative of store-and-forward. If the desired port or ports are not available then the packet waits in the buffer. Ports service waiting packets in a first-come, first-served manner.

A *PO2* node can reject a packet if no buffers are available. An extension of this mechanism is used to provide some measure of backpressure-like flow control on message bursts: a node rejects a packet if it already contains a waiting packet from the same message. Our results indicate that the backpressure provided by this mechanism justifies the costs of implementation.

For a single hop, the local routing choices include the following options:

A *best path* direction sends a packet to a node which is closer to the packet's destination. There may be one or two *best path* directions.

A *no farther* direction sends a packet to nodes that are no farther from the destination than the current node, usually to bypass congested nodes.

We will investigate the following three global strategies:

The *Deterministic* strategy uses a single best path at each routing step, yielding a single minimal path through the interconnect to any destination.

The *Best Paths* strategy allows the choice of any best path at each hop. This is a minimal adaptive routing strategy.

The *Derouting* strategy allows a packet to use no farther directions as well. To prevent packets from continuously circulating without ever reaching their destination, packets are limited in the number of deroutes they can perform according to their original path length. Specifically, a packet that starts at a distance p from its destination can only be derouted on its first $p - 1$ hops. This full adaptivity permits each packet to potentially flow through a sizable fraction of the nodes in the interconnect.

The main parameters for the *PO2* model are:

- We assume that each port permits a byte of information to be transmitted in 1 time unit. Each standard packet is 160 bytes long and hence takes 160 time units to transmit.
- The PE port has an additional overhead of 80 time units to establish a connection into the interconnect, and 20 time units to establish a connection out of the interconnect. These overheads on PE ports occur before any real packet data is transferred; the actual packet data transfer occurs at the rated bandwidth of the port, and the additional delay is not propagated to the internal ports.
- To receive a packet header and to compute the next available direction takes 12 time units.
- There are 20 buffers in each node.

3 Uniform Random Traffic and Different Routing Strategies

Our first experiments considered uniform random traffic consisting of single-packet messages with a random source and destination node. Under such traffic, the three routing strategies are virtually indistinguishable with respect to message latencies.

The only significant observed difference between the three strategies was the internal port utilization. Under a traffic rate of 95%, the minimal routing strategies yielded an internal port utilization of 44%, while the Derouting strategy yielded an internal port utilization of 55%. The internal port utilization for the minimal routing strategies completely coincide for uniform random one-packet messages because each packet takes a minimal route to its destination. With the Derouting strategy, occasional derouting

		PE with overhead			PE without overhead		
	Nodes	Best Paths	Derouting	Penalty	Best Paths	Derouting	Penalty
E6	91	100.0%	100.0%		100.0%	100.0%	
E7	127	100.0%	100.0%		100.0%	97.7%	–2.3%
E8	169	100.0%	100.0%		100.0%	83.9%	–16.1%
E9	217	100.0%	96.5%	–3.5%	100.0%	73.5%	–26.5%
E10	271	100.0%	86.0%	–14.0%	94.7%	65.5%	–30.8%
E11	331	100.0%	77.4%	–22.6%	85.7%	59.0%	–31.2%
E12	397	100.0%	70.5%	–29.5%	78.3%	53.7%	–31.4%

Table 1: The theoretically maximum attainable PE port utilization for the minimal-adaptive (Best Paths) and non-minimal-adaptive (Derouting) strategies, and the throughput penalty for using the derouting strategy, for different network sizes. Where the values of the PE port utilization are less than 100%, the internal port utilization is 100%.

of packets leads to an increase in internal port utilization. This indicates that for larger interconnects and longer messages, the internal port utilization would become the bottleneck more quickly for the derouting strategy and perhaps make interconnect saturation more likely. For this reason, we investigated internal port utilization analytically.

For short path lengths, the PE ports dominate performance because there are fewer PE ports than internal ports. A message that requires $\bar{p}$ hops requires two units of PE port bandwidth for every $\bar{p}$ units of internal port bandwidth. The *PO2* interconnect has three times as many internal ports as PE ports (the internal ports are shared between two nodes), so if we assume the bandwidth of the PE ports and the internal ports are the same, as soon as the average path length exceeds six, a simple flow argument indicates that the internal ports become the performance bottleneck.

In the *PO2* interconnect, the interface design imposes additional overhead on the PE ports, so the effective bandwidth is lower than for the internal ports. The flow reasoning remains the same, though the average path length at which the internal ports become the limiting factor increases in proportion to the factor by which the internal ports are faster than the PE ports. Assuming PE ports without overhead, for a given absolute traffic rate, the packet capacity will increase by the ratio that the overhead is decreased. The essential differences among the strategies remain the same, with or without PE port overhead.

With the Deterministic and Best Path routing strategies, the average path length is determined entirely by the distribution of message sources and destinations. If we assume these are random, then the average path length for an En interconnect is $(2n-1)/3$. Thus, with our PE ports approximately 1.31 times as slow as the internal ports, the internal ports should not become a bottleneck until the interconnect reaches a size of $E13$ with 469 nodes.

With the Derouting strategy, however, the average path length increases as more packets are routed along no-farther paths. Since such derouting is more likely as the traffic rate increases, the average path length depends on the traffic density. Since the average probability that a particular internal port is busy at a particular time is equal to the internal port utilization, we can calculate for a given utilization the likelihood of derouting at each step and thus the expected average path length. This effect tends to snowball; as the port utilization rises, so does the contention for ports, and thus the average path length, increasing port utilization further. Based on this observation, we have constructed an analytical model of internal port utilization based on the Derouting strategy, size of the internal interconnect, and traffic density. This analytical model can be solved iteratively, and we have used it to predict when the internal port utilization becomes the performance bottleneck. Table 1 summarizes the results.

For an E6-sized interconnect, the PE port is the bottleneck for all traffic densities. As the PE port utilization rises to 100%, the internal port utilization rises to 47% using one of the minimal routing strategies. For the Derouting strategy, however, the internal port utilization rises to 57%. If the PE ports were without overhead, the difference would be more striking; at PE port utilization attained 100%, the internal port utilization for the Best Paths and Deterministic strategies would reach 61%, while for Derouting it would reach 81%.

For an E8-sized interconnect and the Best Paths and Deterministic strategies, using a slow PE port, the internal port utilization rises to 63%. Under the Derouting strategy, the internal port utilization rises to 88%. With a PE port running at the same speed as the internal ports, the first two strategies yield an internal port utilization of 83%. For Derouting, in this case, the internal ports become the bandwidth-limiting factor, allowing the PE port to run at only 83.9% when the internal ports become fully saturated.

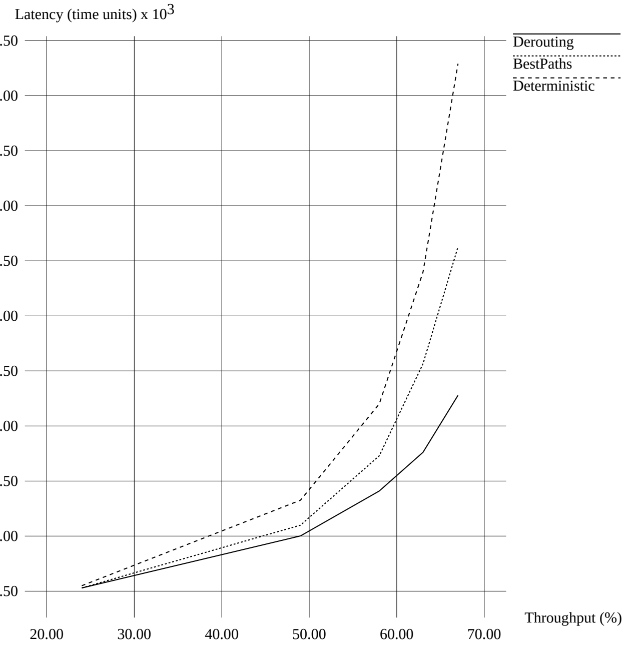

Figure 2: Normalised Average Message Latency for Different Routing Strategies and 10% Long Messages Workload

Table 1 illustrates that as the network size grows, using the Derouting strategy asymptotically causes an effective decrease of about 30% in overall network throughput.

4 Bursty Traffic and Different Routing Strategies

While performance evaluation of packet-switched interconnects has focused on the latency of packets, applications are more concerned with the overall latency of variable-sized messages. Thus, in order to obtain meaningful performance results, we need to define *bursty traffic* workloads. These workloads are defined primarily by a message length distribution. Rather than considering many different message length distributions, we consider only bimodal distributions consisting of short messages and long messages. We define short messages to be from one to five packets in length, and we give each length equal probability. We choose a size of twenty-five packets for long messages; this is about the size of a disk or memory page.

We define our workloads by the percentage of long messages in the workload; this is the primary variable defining the workloads. For instance, a workload with 10% long messages has an average message length of 5.2 packets. Given a traffic density u between zero and one, we generate new messages using a negative exponential distribution with an average interarrival time of $5.2/u$ times the port bandwidth limit on the separation between packets.

A primary goal of the *PO2* design is to minimize the latency of short messages, possibly trading off long-message latency for short-message latency. Message latency is measured from the moment the message is sent by the application or operating system, as defined by the moment the message appears on the interconnect job list, to the moment all the packets of the message have been placed in the receiving (destination) PE's memory. Thus, this time includes queue wait time and time when some packets are in the interconnect. To compare different workload types and different message lengths, it is convenient to define a *normalized* average message latency which is not highly dependent on the message length. We define this normalized message latency as the total message latency divided by the message length.

We also measure and report the packet latency, as measured from the moment when PE port in the source node starts to inject the packet, until the moment when the packet is completely ejected from the interconnect at the destination node. This time is totally spent within the interconnect.

Figures 2 and 3 show the normalized average message latency and packet latency corresponding to a workload with 10% long messages using the three different routing strategies. The messages are injected into the interconnect in FIFO order. Figure 2 indicates that the Derouting strategy provides the best overall message latency, followed by the Best Paths and finally the Deterministic strategies. Interestingly, the packet latencies illustrated by Figure 3 are in precisely the opposite order, with Deterministic providing the best overall packet latency. This phenomena is partly explained by examining the port utilization under the different strategies, as shown in Figure 4. The solid black line represents ideal PE port utilization. The percentage of PE port utilization deviation from that line shows the frequency of packet rejection due to the flow control mechanism. For 67% traffic utilization, the PE port utilization for the Derouting strategy is 69%, while for the Best Paths strategy it is 73% and for the Deterministic strategy it reaches 77%. This shows that packets for the less adaptive strategies spend more of their time waiting in the message queue. The more adaptive strategies maintain fewer packets in the queue and more packets inside the interconnect for a given traffic load. With so many packets inside the interconnect, contention is higher, and the packets spend longer trying to reach the destination node and competing there for the destination PE port. Thus, the Derouting strategy leads to a higher overall utilization of interconnect fabric resources and provides better overall message latency. The backpressure mecha-

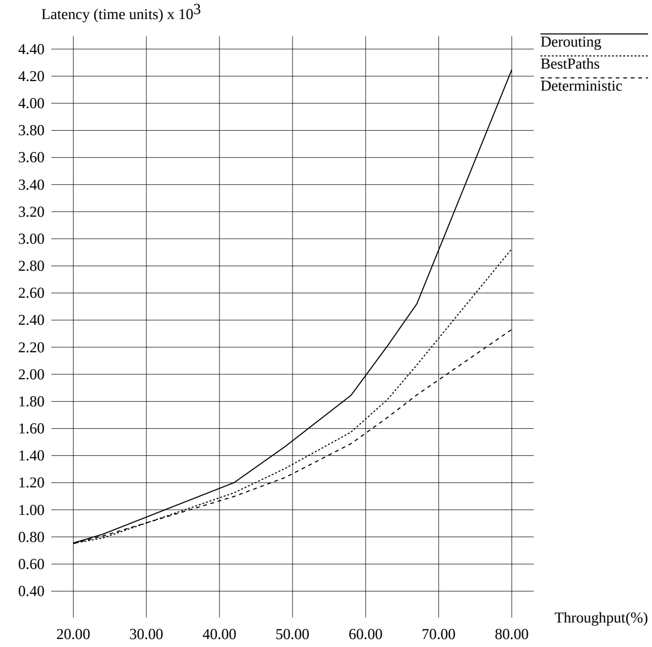

Figure 3: Packet Latency for Different Routing Strategies and 10% Long Messages Workload

nism under the Best Paths and Deterministic strategies has a significant impact, especially under heavier traffic, forcing packets and messages to wait outside the interconnect. This phenomena is even more pronounced with a higher percentage of long messages. Figure 5 and Figure 6 show the normalized average message latency and packet latency corresponding to a workload with 80% long messages. Figure 7 shows the PE and internal port utilization generated by this workload. However, the Derouting strategy has a few drawbacks. First, at high traffic loads and large interconnects, the internal port utilization rises above the PE port utilization, at which point the internal ports become the bottleneck of the interconnect. Figure 4 and Figure 7 illustrate this. In addition, the ability of a single message to distribute packets across a significant percentage of the interconnect fabric in the presence of contention raises the likelihood of interconnect saturation and deadlock.

5 Conclusion

With uniform random traffic, non-minimal routing does not provide a significant performance advantage over minimal adaptive or deterministic routing strategies. With bursty traffic, however, the use of non-minimal routing can yield a significant decrease in message latency.

The Derouting strategy reveals some potentially dan-

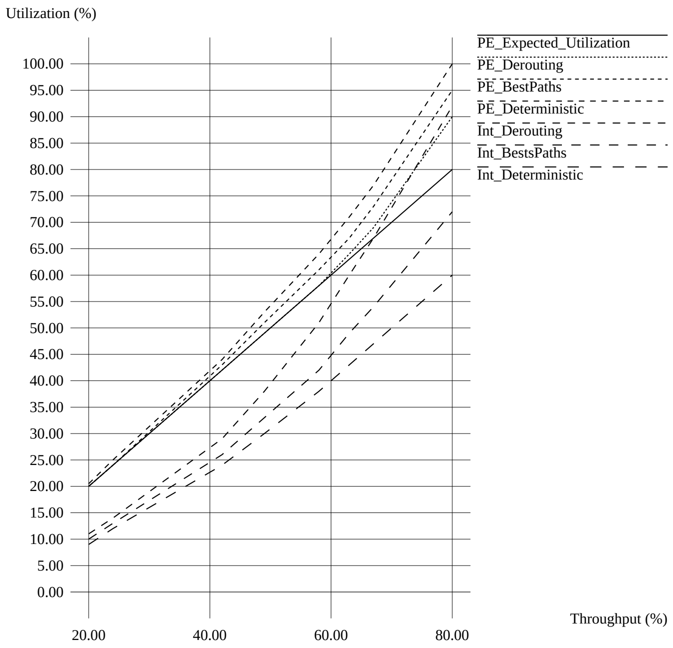

Figure 4: Port Utilization for Different Routing Strategies and 10% Long Messages Workload

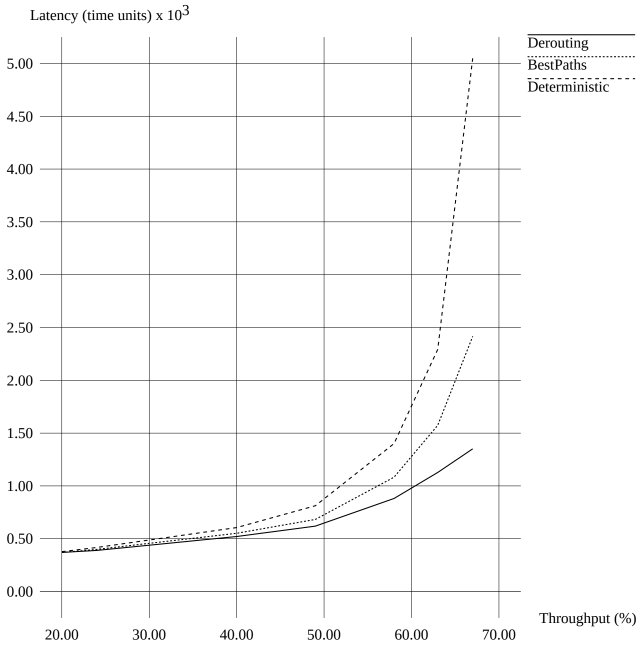

Figure 5: Normalised Average Message Latency for Different Routing Strategies and 80% Long Messages Workload

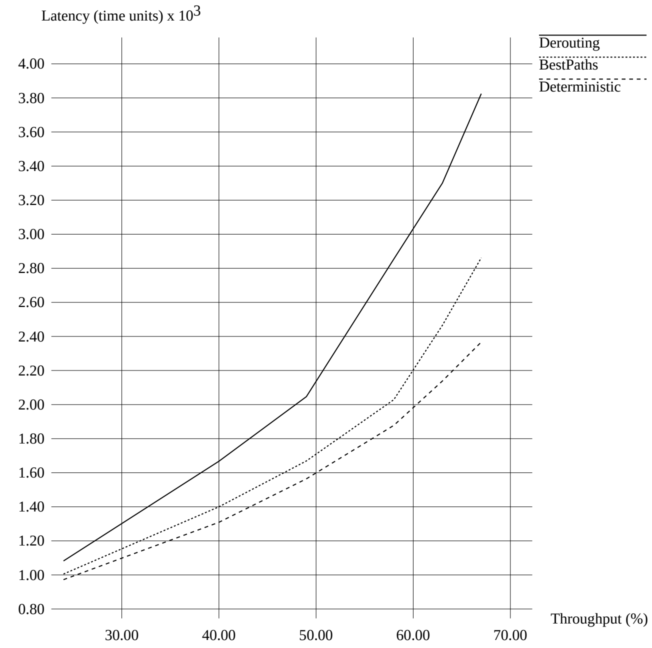

Figure 6: Packet Latency for Different Routing Strategies and 80% Long Messages Workload

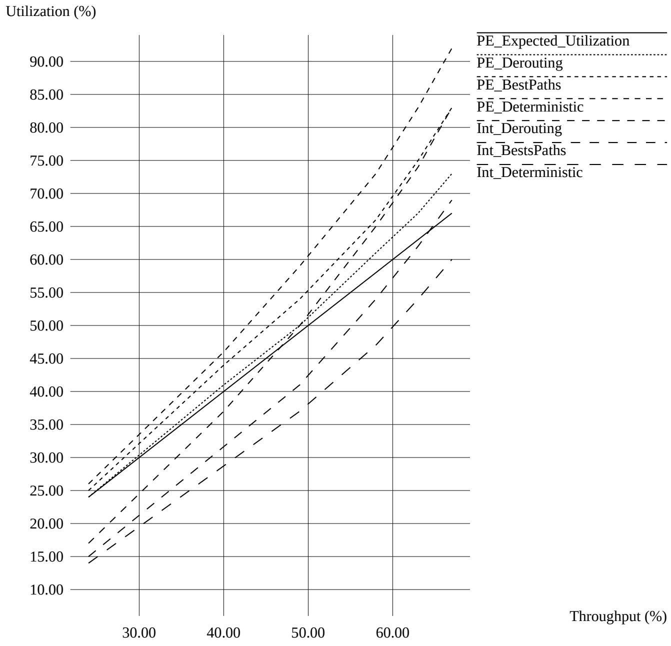

Figure 7: Port Utilization for Different Routing Strategies and 80% Long Messages Workload

gerous drawbacks. Adaptivity decreases the efficiency of backpressure because of the larger number of nodes that can be populated with packets from a particular message. In addition, the internal port utilization rises as derouting frequency increases, lowering the effective maximum throughput of the interconnect for large networks. In addition, this higher port utilization threatens a higher likelihood of network saturation and deadlock.

These conclusions are based on a simulation results as well as an analytical model, using the wrapped hexagonal mesh topology as a case study. However they reveal a general phenomena valid for many different types of interconnect, namely, the conflict between the flow control provided by backpressure and the routing freedom provided by adaptivity.

This paper does not take into account the effects of intelligently scheduling the packets from various messages for injection into the interconnect. Early results [1] show that via the scheduling strategy the workload can be made to appear closer to a random traffic model. This has the effect–as shown earlier–of reducing the performance advantage of the Derouting strategy.

References

[1] Cherkasova, L. and Rokicki, T.: Alpha Message Scheduling for Packet-Switched Interconnects. To be published.

[2] Chien, Andrew A.: A Cost and Speed Model for k-ary n-cube Wormwhole Routers. In *Proceedings of Hot Interconnects'93, A Symposium on High Performance Interconnects*, 1993.

[3] Dally, W. J. et al.: The J-Machine: A Fine-Grain Concurrent Computer. In *Proceedings of the IFIP Conference*, North-Holland, pp. 1147–1153, 1989.

[4] Davis A., Mayfly: A General-Purpose, Scalable, Parallel Processing Architecture. J. *LISP and Symbolic Computation*, vol.5, No.1/2, pp. 7–48, 1992.

[5] Fujimoto R. M. VLSI Communication Components for Multicomputer Networks.Ph.D. Thesis,University of California at Berkeley,1983.

[6] Jain, R.: Myths About Congestion Management in High-speed Networks. Internetworking: Research and Experience, Vol.3, pp. 101–113, 1992.

[7] Seitz, C.: The Cosmic Cube. J.*Communications of the ACM*, Vol.28, No.1, pp. 22-33, 1984.