Hierarchical Adaptive Routing Under Hybrid Traffic Load

Ziqiang Liu*
Department of Teleinformatics, Royal Institute of Technology
Electrum 204, 164 40 Kista, Sweden

Abstract. In actual multicomputer networks, communications consist of *hybrid traffic*, a mix of short and long messages.Typically, the short messages are used to support synchronization, global combining, and multicasting where the latencies are critical to the execution time of whole parallel program. However, in normal wormhole routed networks without packetization, the presence of long messages degrades network performance of short messages dramatically, qualitatively changing network behaviour. In this paper, we extend one existing adaptive routing framework, Hierarchical Adaptive Routing (HAR) to a hybrid traffic model, to form a new simple and efficient adaptive routing framework called Hybrid-HAR. Hybrid-HAR has four important advantages. Firstly, without packetization, the impact of long messages on short messages is very small. Secondly, it supports fully adaptive routing to both short and long messages for high communication performance. Third, the implementation complexity of Hybrid-HAR is compatible to standard routing algorithms such as Dimension Order Routing. Fourth, Hybrid-HAR is applicable to a wide variety of network topologies. High level implementation and simulation studies of a Hybrid-HAR for 2D mesh networks are presented.

1 Introduction

In highly parallel machines, a collection of computing nodes works in concert to solve large application problems. The nodes communicate data and coordinate their efforts by sending and receiving messages through a routing network. Consequently, the achieved performance of such machines depends critically on the performance of their routing networks. Network performance depends on a variety of factors, not only network architecture (topology, routing and flow control), but also the characteristics of communication that is actually carried on the network such as message sizes and traffic patterns.

Many recent multicomputer networks use cut-through or *wormhole routing* [9, 3], a technique which reduces message latency by pipelining transmission over the channels along a message's route. In these networks, a message spans multiple channels which couples the channels tightly together, so blockage on one channel can have immediate impact on another. Tight coupling between channels means that one long message can block the progress of many other messages, and as a result, even a small fraction of long messages can affect overall network performance significantly. In these networks, channel coupling effects make performance quite sensitive to message size.

There are a number of factors which give rise to non-uniform message sizes. First, applications may generate a range of message sizes directly [2, 7]. Second, many

* Ziqiang Liu died in December 1994, leaving his work to the members of his research group.

synchronous message passing implementations implement a three-phase protocol to allocate buffer storage. This translates a large application message into two short control messages, followed by a long (application size) message. Third, typical message passing libraries implement some form of packetization, breaking long messages into some maximum physical transfer size (256, 512, or 1024 bytes, typically). In addition, message-passing library functions such as global synchronization, reduction and multicast operations give rise to short messages. Finally, system software needs to move large blocks of data to provide services such as parallel input/output and distributed virtual memory.

Typically, in multicomputer applications, the short messages are used to support synchronization, global combining, and multicasting where their latency is critical to the execution time of whole parallel program. However up to today, little work has been done to reduce the short message latency under a hybrid traffic model in wormhole routed networks. In this paper we propose *Hybrid-HAR* as a promising technique to reduce the impact of long messages on short messages. In the interconnection network, a certain amount of virtual channel resource is exclusively allocated to short messages, which almost eliminates the possibility that a short message will be blocked by a long message. In the network interface, separate injection and sink channels, and incoming and outgoing buffers are provided for short and long messages, which guarantee that the send and receive of short messages at the source and destination nodes will not be affected by long messages. Thus Hybrid-HAR dramatically decreases both the variance and the average of short messages without packetization. By providing fully adaptive routing, Hybrid-HAR also gives good overall network performance. The implementation complexity of Hybrid-HAR is compatible to standard routing algorithms such as DOR, provided the same number of virtual channels is used. Also Hybrid-HAR is applicable to a variety of network topologies.

The next section gives the relevant background and related work. After that, Section 3 describes the general Hybrid-HAR framework for a variety of network topologies with hybrid traffic model. Section 4 shows an application of Hybrid-HAR to a 2D mesh network. This produces a simple fully adaptive routing algorithm for hybrid traffic load yielding high performance. Section 5 examines the performance of Hybrid-HAR router in 2D mesh networks. Finally, Section 6 concludes the paper, summarizing the results.

2 Background

All networks examined in this paper are wormhole-routed, k-ary n-cube, direct networks. With *wormhole* routing, the packet is not completely buffered at each node. As soon as the header of a packet comes into a node, it is forwarded to the requested outgoing channel if free. If not, the packet is stopped in place. This achieves the pipelined packet transmission with only modest buffering requirements for each router. Typically in wormhole-routed network, each physical link is split into multiple virtual channels. Each virtual channel has its own buffer queue and control signals, sharing the bandwidth of the corresponding physical link [3]. Virtual channels can be used to avoid deadlock and support adaptive routing for high network performance.

In wormhole-routed networks, channel coupling can cause achieved network throughput to be much less than the network capacity. Several studies have shown that the channel coupling increases with message length, reducing the performance [1, 6]. There are three techniques to reduce the impact of long messages on short messages: *packetization*, *multipath routing* and *virtual channel allocation*.

Packetization reduces the impact of long messages on other messages because long messages are split into a number of small packets, reducing the maximum blocking time. However, packetization has two significant drawbacks. First, it requires a mechanism for conversion between messages and packets. Second, packetization increases the network load; each packet must contain routing and sequencing information in its header.

Multipath routing, virtual channels and adaptive routing can reduce the interference of long messages on short messages [11]. Virtual channels [3] virtualize the physical channels, multiplexing them among several messages. The multiplexing allows a short message to pass a blocked long message. Adaptive routing allows a message to use any one of several paths from source to destination [8, 5, 14, 10, 13]. This can allow a short message to circumvent a blocked long message. With multipath routing, the possibility that a short message will be blocked by a long message is reduced, but not totally eliminated, especially when the long messages dominate the traffic [11]. The combination of multipath routing and packetization can further eliminate the blocking and minimize the effect of the remaining blocking [11].

In the virtual channel allocation scheme, a certain amount of virtual channels are exclusively assigned to short messages. The remaining virtual channels are shared by both short and long messages [16]. This totally eliminates the possibility that a short message will be blocked by a long message. However, previous research works in this field have been limited to non-adaptive routing, and the design of router could be complicated when the number of virtual channels is large [15].

In this paper, Hybrid-HAR is introduced as a combination of multipath routing and virtual channel allocation. Hybrid-HAR has four important advantages. Firstly, one set of virtual channels is exclusively allocated to short messages, which almost eliminate the possibility that a short message will be blocked by a long message Secondly, fully adaptive routing is provided to both short and long messages to give good overall network performance; Third, the hierarchical design makes the implementation of Hybrid-HAR simple, which is compatible to the standard static routing such as DOR provided the same amount of virtual channels is used. Fourth, it is applicability to a wide variety of network topologies.

3 Hybrid Hierarchical Adaptive Routing (Hybrid-HAR)

In this section, Hierarchical Adaptive Routing (HAR) is extended with virtual channel allocation for hybrid traffic load. A new adaptive routing framework called Hybrid-HAR is introduced to increase the performance of short messages.

3.1 Hybrid-HAR Framework

Hybrid-HAR is a combination of hierarchical adaptive routing, as described in [13], and virtual channel allocation scheme, which provides a simple and efficient framework to networks with bimodal traffic load. As shown in Figure 1, Hybrid-HAR divides a physical network into several virtual networks. There are two *connection* channels between two adjacent virtual networks, one for short messages and another for long messages. The connection channel allows blocked messages in the higher level to move to the lower level. Two separate injection and sink channels are provided, both for short and long messages. Different routing algorithms can be used in different virtual networks for different purposes. Initially, when a message is generated, it is injected into the first level virtual network and routed towards the destination. However, if at some point there are no free output channels in the first level virtual network (due to congestion), the message can be moved to the second level virtual network through the connection channel. If this happens again at another node, message can then be moved into the third level virtual network and so on.

Notation:
FAR: Fully Adaptive Routing (minimal); DOR: Dimension Order Routing
CCS: Connection Channel for Short Message; LVN: Lowest Level Virtual Network
CCL: Connection Channel for Long Message; OVC: Output Virtual Channel
SVN: Successive Virtual Network; DFR: Deadlock-Free Routing
C_1 : First Set of Virtual Channels in LVN; C_2 : Second Set of Virtual Channels in LVN

Hybrid-HAR Routing Algorithm

Upper Level Virtual Networks:

1 If one OVC is free by FAR, forward the message along that channel
2 else,
3.1 If the message is short and CCS is free, move it into SVN;
3.2 If the message is long and CCL is free, move it into SVN;
4 otherwise, block the message.

Lowest Level Virtual Network:

5.1 If the message is short and one OVC in C_2 is free by DFR, forward it;
5.2 If the message is short and one OVC in C_1 is free by DFR, forward it;
5.3 If message is long and one OVC in C_1 is free by DFR, forward it;
6 otherwise, block the message.

In any actual router design, there is only a finite number of virtual networks. When a packet enters into the lowest virtual network (LVN), it must be routed to the destination node entirely within that virtual network. The LVN must be deadlock-free for both short and long messages. However, the routing algorithms at the upper level virtual networks can be fully adaptive for higher performance. Also at LVN, one set of virtual channels is exclusively allocated to short messages to avoid that a short message will be directly blocked by a long message. As indicated in Figure 1, the virtual channels from the set C_2 are only used by short messages.

For any network topology, if there is a known deadlock-free routing algorithm, Hybrid-HAR can be used to produce a fully adaptive routing algorithm for hybrid traffic load. A virtual network with fully adaptive routing for both short and long messages needs to be added on the top of deadlock-free virtual network. One extra virtual channel needs to be added in the deadlock-free virtual network, which is used only by short messages. The routing algorithm on the extra virtual channel for short messages can be adaptive, as long as the whole routing algorithm is deadlock-free.

4 Hybrid-HAR in 2D Mesh

In this section, Hybrid-HAR is applied in a 2D mesh network to develop a simple and efficient fully adaptive routing algorithm for hybrid traffic load. A restricted Duato Adaptive Routing (DAR) algorithm is applied in the lowest virtual network to guarantee freedom from deadlock for both short and long messages, and to reduce the impact of long messages on short messages. We show that the possibility that a long message will be blocked by a long message is rare.

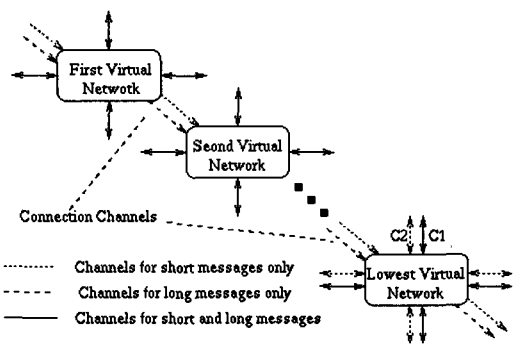

Fig. 1. Hybrid Hierarchical Adaptive Routing Framework

Hybrid-HAR in 2D mesh networks requires that each physical channel split into four virtual channels. The physical network is then divided into two equal virtual networks, which are connected through two internal connection channels, one for short and another for long messages. Separate injection and sink channels are used for short and long messages from and to the network interface. In the upper virtual network, fully adaptive minimum path routing is used on two sets of virtual channels (called FAR virtual network). At the lower level, the restricted Duato Adaptive Routing is used (called DAR virtual network) [5]. When a message, short or long, is injected into the network, it enters the FAR virtual network and uses the virtual channels there to move toward the destination. However, when there is congestion in the network, it is possible that all valid output virtual channels are busy. In such a case, if the corresponding connection channel to the DAR virtual network is free, the message is moved into the DAR virtual network.

In the lower level virtual network, the restricted DAR algorithm is applied. Both short and long messages can route over the first set of virtual channels $\langle C_1 \rangle$ with dimension

order routing; However, only short messages can route over the second set of virtual channels $\langle C_2 \rangle$ with fully adaptive minimum path routing. In the original DAR algorithm, there is no distinction between short and long messages. Dimension order routing and fully adaptive routing can be chosen by any message at any step according to the status of output channels.

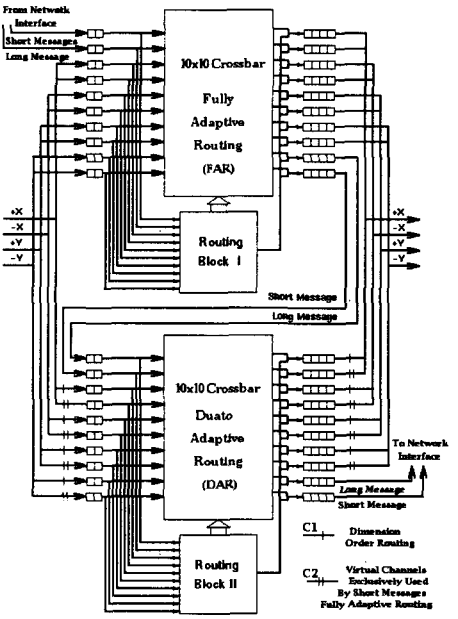

Fig. 2. Hybrid Hierarchical Adaptive Routing in 2D Mesh Network

In Hybrid-HAR, the probability that a short message will be blocked by a long message is very small. Firstly, a short message will never be blocked by a long message in the higher virtual network, since the connection channel will allow the blocked short message in the higher virtual network to move into the lower virtual network. Secondly, the chance that a short message will be blocked by a long message in the lower virtual channel is small. Two conditions must be satisfied. Firstly, a cycle dependency must exist in C_2 formed by several short messages due to the fully adaptive routing, which itself has no restriction on routing to avoid deadlock. Each message would occupy one virtual channel from C_2, requested by another short message. Secondly, all the corresponding virtual channels in C_1 must be occupied by long messages. In such a case, all those short messages in the dependency cycle are directly blocked by a long message. The simulation results in [10] have shown that the probability that a dependency cycle will occur with fully adaptive routing is actually very small. In Hybrid-HAR, it is even smaller with hierarchical adaptive routing and a larger number of virtual channels. Further more, the probability that the second condition will be true is small too. In one word, the probability that a short message will be blocked by a long message is very small in Hybrid-HAR. A proof that the restricted DAR algorithm is deadlock-free can be established using the theory described by Duato in [5].

5 Performance

In this section, we evaluate the performance of Hybrid Hierarchical Adaptive Routing compared to Dimension Order Routing (DOR) using a register-transfer level simulation. Firstly, the simulation modelling is presented. Then the simulation experiment results with uniform and non-uniform traffic patterns are given.

5.1 Simulation Modelling

The simulator is a 2000 line C program based on the HSMPL sequential discrete-event simulation environment [12]. It simulates 16x16 mesh networks at the flit-level. A flit transfer between two nodes and through the crossbar is assumed to take place in one clock cycle. The routing information fits in a single flit (the header flit), and a routing decision requires two clock cycles. In all cases, each physical channel has been split into four virtual channels. Routers are connected by dual unidirectional channels.

The models of DOR and Hybrid-HAR are similar to each other. The detailed modelling of Hybrid-HAR router has been presented in Section 3 and shown in Figure 2. Similar to Hybrid-HAR, the DOR router has two levels of virtual networks and each has almost identical 10x10 crossbars. Message has to finish routing in the X dimension before starts routing in the Y dimension. The Hybrid-HAR router has almost the same complexity as the DOR router, except routing decision is slightly complicated due to the fully adaptive routing and virtual channel allocation. Due to the same complexity, Hybrid-HAR and DOR routers will have same latency for data-through or cycle time. It is assumed that Hybrid-HAR router and DOR router has the same cycle in the simulation.

At each node, messages are generated by a Poisson process, whose mean value is determined by applied network load rate. Applied network loads are all normalized with respect to the network's maximum wire capacity, defined as all of the network channels transmitting simultaneously. Generated short or long messages which are not accepted by the network are source queued in a short or long messages queue which is allowed to grow without bound. Thus, the message latency numbers include the waiting time in the source queue. Three traffic patterns are considered: *uniform* and non-uniform *transpose* and *center-reflection.*

The traffic model is a bimodal distribution of message sizes, short and long. The length of short messages is fixed at 32 flits. If flits are eight bits, this short message size is comparable to a cache line or a procedure invocation record. The length of long messages is fixed to 256 flits. The traffic mixes contain 0%(short messages only), 25%, 50%, 75% and 100%(long messages only) long messages, respectively. The proportions are percentages by traffic volume.

5.2 Uniform Traffic

In the uniform traffic pattern, each node sends messages to all other nodes with equal probability and there is no congestion in the network. DOR has slightly better performance than adaptive routing. The reason is that with uniform traffic, non-adaptive routing preserves the traffic's uniformity, while adaptive routing disturbs it.

Figure 3 shows average short message latency (S-latency). DOR has slightly lower S-latency than Hybrid-HAR when the percentage of long messages (P) is less than 75%. As P is increased from 0% to 50%, S-latency increases in both DOR and Hybrid-HAR due to the increasing impact of long messages on short messages. However, further increasing P to 75%, S-latency is decreased in both DOR and Hybrid-HAR. Especially in Hybrid-HAR, S-latency is very low when P is 75% due to less congestion on those virtual channels exclusively allocated to short messages.

Figure 4 shows the average long message latency (L-latency). DOR has slightly lower L-latency than Hybrid-HAR. L-latency increases with the increased P in both DOR and Hybrid-HAR. Figure 5(a) shows that DOR has higher throughput than Hybrid-HAR with all traffic mixes. Throughput decreases with increased P, since channel coupling increases with message length, reducing the performance.

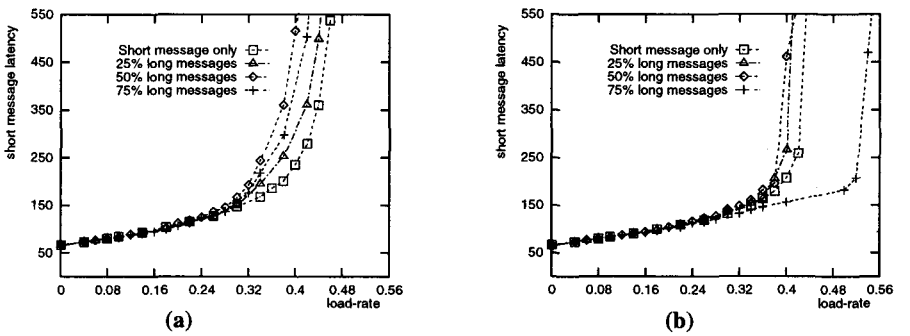

Fig. 3. The average short message latency under uniform traffic pattern when the routing algorithm is (a) DOR and (b) Hybrid-HAR.

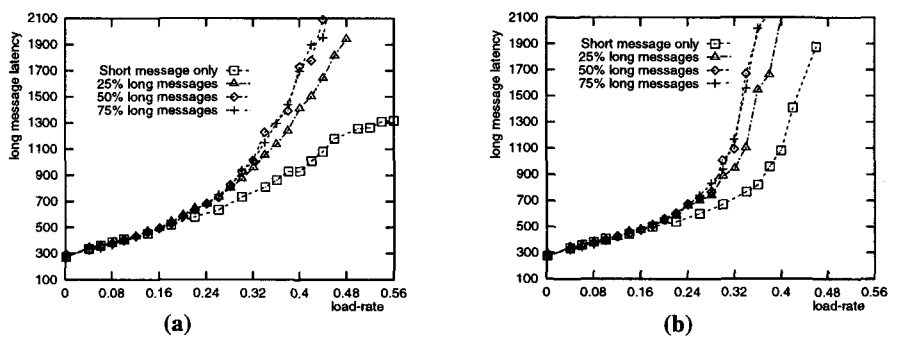

Fig. 4. The average long message latency under uniform traffic pattern when the routing algorithm is (a) DOR and (b) Hybrid-HAR.

The short message latency distribution is shown in Figure 6. All short messages in Hybrid-HAR are delivered with lower latency than in DOR. When there is no long messages in the network, all short messages are delivered in less than 900 cycles in Hybrid-HAR, versus 1250 in DOR. Fully adaptive routing in Hybrid-HAR reduces the possibility that short messages will block each other. With long messages present, a substantial fraction of short messages are delivered with extremely high latencies in

DOR. For example when P is 50%, the largest delay in DOR reaches 3400 clock cycles, where only 1300 in Hybrid-HAR.

5.3 Non-Uniform Traffic

Two non-uniform traffic patterns, transpose and center-reflection, have been examined to compare the performance of Hybrid-HAR with DOR. In the transpose traffic pattern, node (x,y) sends messages to node (y,x). For DOR, there are congestions in the lower left and upper right corners. In the center-reflection traffic pattern, node (x,y) sends messages to node (16-x,16-y). There is traffic jam in DOR at the left and right edges of the network. Due to the fully adaptive routing, Hybrid-HAR reduces the traffic jams significantly and has much better performance than DOR. The performance under transpose traffic pattern is similar to the one under center-reflection. All simulation results with center-reflection except the throughput have been omitted in the paper.

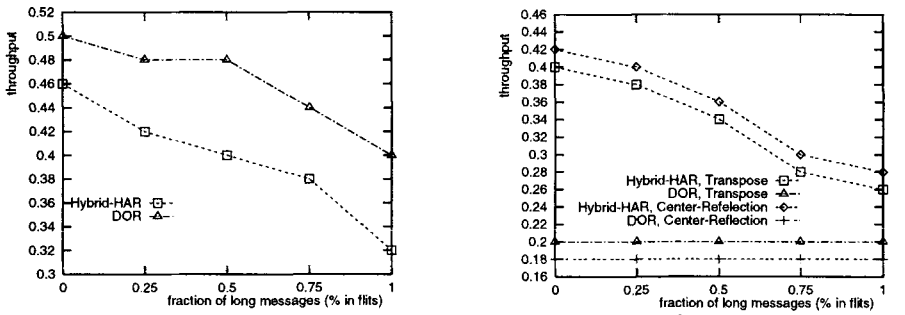

Fig. 5. The overall throughput under traffic with a range of traffic mixes. The traffic pattern are (a) Uniform and (b) Transpose and center-reflection

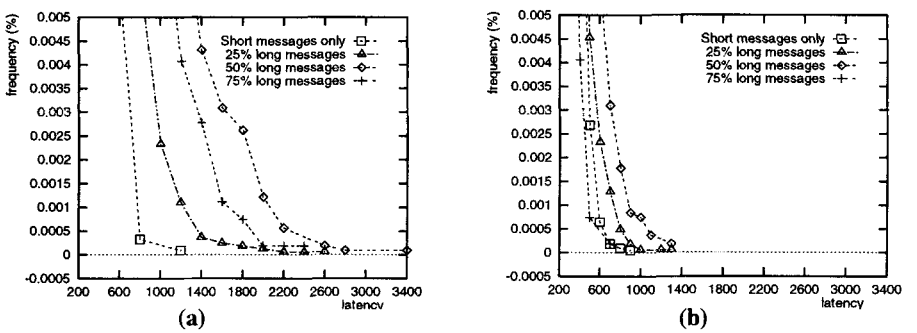

Fig. 6. The short message distribution under uniform traffic pattern when the routing algorithm is (a) DOR and (b) Hybrid-HAR. The applied load-rate is 0.36.

Figure 5(b) describes the throughput of DOR and Hybrid-HAR under two non-uniform traffic patterns. Hybrid-HAR out-performs DOR significantly, especially when P is low. The throughput of DOR in non-uniform traffic patterns is very low, only 0.20 and 0.18 respectively. It tends to be constant with all traffic mixes. The reason is that the congestion remains almost the same in the DOR network, regardless of the traffic mix. In Hybrid-HAR, throughput decreases with increased P, since channel coupling increases with message length, reducing the performance.

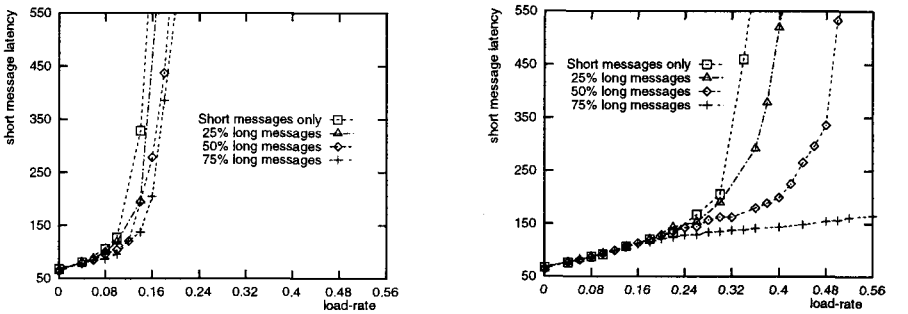

Fig. 7. The average short message latency under transpose traffic pattern when the routing algorithm is (a) DOR and (b) Hybrid-HAR.

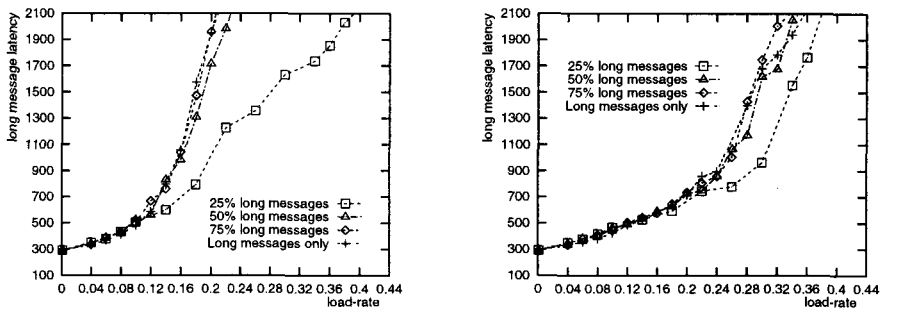

Fig. 8. The average long message latency under transpose traffic pattern when the routing algorithm is (a) DOR and (b) Hybrid-HAR.

Figure 7 shows that Hybrid-HAR has much lower S-latency than DOR with all traffic mixes. In DOR, S-latency decreases slightly with the increased P due to lower short messages traffic load. In Hybrid-HAR, however, S-latency decreases much quickly with the increased P. As P is increased, the short message traffic load decreases dramatically. There is much less congestion on those virtual channels exclusively allocated to short messages. Figure 8 shows that Hybrid-HAR has slightly lower L-latency than DOR with all traffic mixes.

Under transpose traffic pattern, all short messages in Hybrid-HAR are delivered with much lower latency than in DOR, as shown in Figure 9. All short messages in Hybrid-HAR are delivered in less than 600 clock cycles with all range of traffic mixes, versus at least 1600 in DOR. The distributions of short message latency differ dramatically as traffic mix changes in DOR. The largest delay in DOR is 9000, 1600, 7600 and 1600 clock cycles when P is 0, 0.25, 0.50 and 0.75 respectively. However, the longest delay in Hybrid-HAR is less than 600 clock cycles with all traffic mixes. Fully adaptive routing and virtual channels allocation scheme in Hybrid-HAR dramatically reduces the impact of long messages on short messages under non-uniform traffic pattern.

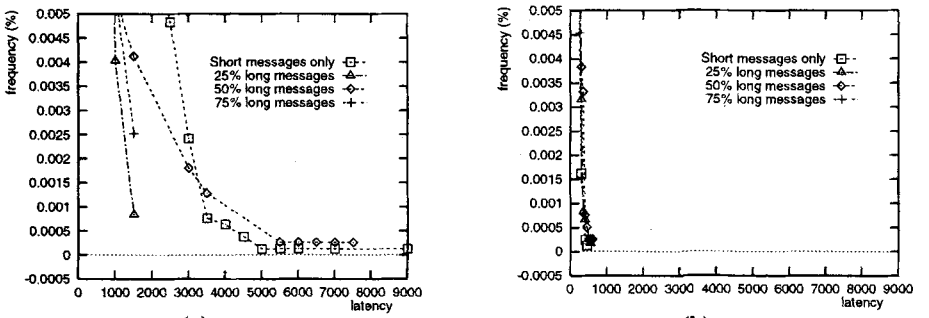

Fig. 9. The short message distribution under transpose traffic pattern when the routing algorithm is (a) DOR and (b) Hybrid-HAR. The applied load-rate is 0.20.

6 Conclusion

In this paper, we have extended the hierarchical adaptive routing to a hybrid-traffic model and produced a framework for fully adaptive deadlock-free wormhole routing (Hybrid-HAR). Hybrid-HAR divides the physical network into two levels of virtual networks. Fully adaptive routing is used in the higher level virtual network for higher performance. Another fully adaptive routing (by Duato) is used in the lower level virtual network to avoid deadlock. One set of virtual channels with fully adaptive routing are exclusively allocated to short messages in the lower level virtual network, which gives very good performance for short messages with all traffic mixes.

A draft implementation outline of Hybrid-HAR router for 2D mesh network has shown that it has almost the same complexity as the DOR router, when the same amount of virtual channels is used. All simulation results have confirmed the high performance of Hybrid-HAR. Under non-uniform traffic pattern, with fully adaptive routing, Hybrid-HAR out-performs DOR substantially. Under uniform traffic pattern, Hybrid-HAR gives compatible performance. Under both uniform and non-uniform traffic patterns, Hybrid-HAR can guarantee that short messages are delivered with much lower latency than DOR, with all traffic mixes. The conclusion is that Hybrid-HAR is a new fully adaptive routing scheme worth to be introduced for actual multicomputer application, where the traffic model is hybrid.

References

1. A. Agarwal. Limits on interconnection network performance. IEEE Transactions on Parallel and Distributed Systems 2(4):398--412, 1991.

2. R. Cypher, A. Ho, S. Konstantinidou, and P. Messina. Architectural requirements of parallel scientific applications with explicit communication. In Proceedings of the International Symposium on Computer Architecture, pages 2--13, 1993.

3. W. Dally and C. Seitz. Deadlock-free message routing in multiprocessor interconnection networks. IEEE Transactions on Computers, C-36(5):547--53, 1987.

4. W. J. Dally, et. al. Design and implementation of the message-driven processor. In Proceedings of the 1992 Brown/MIT Conference on Advanced Research in VLSI and Parallel Systems, T. Knight and J. Savage, eds., pages 5--25. MIT Press, 1992.

5. J. Duato. On the design of deadlock-free adaptive routing algorithms for multicomputers: design methodologies. In Proceedings of Parallel Architectures and Languages Europe, pages 390--405, 1991.

6. F. Hady. A Performance Study of Wormhole Routed Networks Through Analytical Modelling and Experimentation. PhD thesis, University of Maryland, 1993.

7. J. M. Hsu and P. Banerjee. Performance measurement and trace driven simulation of parallel cad and numeric applications on a hypercube multicomputer. IEEE Transactions on Parallel and Distributed Systems, 3(4):451--464, 1992.

8. C. R. Jesshope, P. R. Miller, and J. T. Yantchev. High performance communications in processor networks. In Proceedings of the International Symposium on Computer Architecture, pages 150--7, 1989.

9. P. Kermani and L. Kleinrock. Virtual cut-through: A new computer communications switching technique. Computer Networks, 3(4):267--86, 1979.

10. Jae Kim, Ziqiang Liu, and Andrew A. Chien. Compressionless routing: A framework for adaptive and fault-tolerant routing. In Proceedings of the International Symposium on Computer Architecture, pages 289--300, 1994.

11. Jae H. Kim and Andrew A. Chien. Network performance under bimodal traffic loads.To appear in the Journal of Parallel and Distributed Computing.

12. Z. Liu, L-E Thorelli, and H. Wu. Hsim: A hybrid sequential and parallel simulation. In Proceedings of the Information Processing, pages 372--378, 1992.

13. Ziqiang Liu and Andrew A. Chien. Hierarchical adaptive routing: A framework for fully adaptive and deadlock-free wormhole routing. In Sixth IEEE Symposium on Parallel and Distributed Processing, 1994.

14. L. Ni and C. Glass. The turn model for adaptive routing. In Proceedings of the International Symposium on Computer Architecture, pages 278--87, 1992.

15. Abdel-Halim Smai and Lars-Erik Thorelli. Dynamic allocation of communication bandwidth in multicomputers. In Parallel Architectures and Languages Europe, pages 677--687, 1994.

16. Abdel-Halim Smai and Handong Wu. Evaluation of a priority flow control scheme for multicomputer networks. In Euromicro Workshop on Parallel and Distributed Processing, pages 111-116, 1994.