Journal of Parallel and Distributed Computing
Vol. 67(1) pp. 13-32 (January 2007).
http://dx.doi.org/10.1016/j.jpdc.2006.08.012

DEADLOCK-FREE CONNECTION-BASED ADAPTIVE ROUTING WITH DYNAMIC VIRTUAL CIRCUITS

Yoshio Turner †
Enterprise Systems and Software Laboratory
Hewlett Packard Laboratories
Palo Alto, CA

Yuval Tamir
Concurrent Systems Laboratory
Computer Science Department
UCLA
Los Angeles, CA

Abstract

Virtual circuits can reduce routing overheads with irregular topologies and provide support for a mix of quality of service (QOS) requirements. Information about network loads and traffic patterns may be used during circuit establishment to utilize network resources more efficiently than is practical with packet routing. Most virtual circuit schemes are static — each established virtual circuit remains unchanged until the connection is no longer needed. In contrast, we propose the *Dynamic Virtual Circuit* (DVC) mechanism, which enables existing circuits to be quickly torn down in order to free up resources needed for other circuits or to re-establish circuits along routes that are better suited for current network conditions. We propose a deadlock avoidance technique, based on unconstrained routing of DVCs combined with a deadlock-free virtual network. We present a correctness proof for the scheme, describe key aspects of its implementation, and present performance evaluation results that explore its potential benefits.

Keywords: adaptive routing, deadlock, virtual circuits, interconnection networks.

† This work was done while the author was a graduate student at UCLA.

1. Introduction

The routing scheme used in multicomputer or high-end cluster interconnection networks should direct packets through the lowest latency paths from source to destination. It should take into account the topology of the network and *adapt* to the current workload and resource availability to route packets around congested or faulty areas [7, 12, 17, 22, 26, 33, 39]. Multicomputer and cluster networks use backpressure flow control in which packets that encounter congestion are blocked rather than discarded. A sequence of blocked packets can form a deadlock cycle unless sufficient constraints are imposed on packet routes and/or buffer usage [1, 6, 15, 16, 20, 31]. While the routing scheme should minimize routing constraints in order to maximize performance, it must also ensure deadlock-freedom.

The routing scheme should also minimize the overhead costs for routing and forwarding packets. In order to efficiently utilize link bandwidth, the addressing and control information that is sent with each packet should be minimized. Furthermore, in order to minimize communication latency, the processing to interpret and route each packet at each hop should be minimized. These goals can be achieved by using *connection-based* routing, in which resources are reserved in advance of communication at each switch along the path from source to destination. This approach reduces the overhead for forwarding each packet but introduces connection setup/teardown operations. Network efficiency is improved as long as the setup/teardown operations occur infrequently, i.e., for workloads in which each connection is used by several packets [24].

Virtual circuit switching [8, 13, 24, 38] is a form of connection-based routing which provides end-to-end connections called *virtual circuits*. The network's physical resources (packet buffers, link bandwidth, etc.) are multiplexed among active virtual circuits. Once a virtual circuit is established, packets can be sent without transmitting the addressing, sequencing, and QoS information needed with conventional packet switching. Instead, state information for the connection is maintained by switches along its route. This enables the use of smaller packet headers which provide more efficient link bandwidth utilization than with packet switching, particularly when packet data payloads are small. Even if the routes of the circuits are established using complex and relatively slow mechanisms, most packets are quickly forwarded along established virtual circuits by using a single lookup in a small virtual channel table at each hop. The benefits and applications of virtual circuit switching are discussed further in Section 3.

Most virtual circuit schemes have the limitation that each established circuit's route cannot be changed until the connection is no longer needed and the circuit is torn down. This prevents adaptation to changes in traffic patterns, and it prevents establishment of new virtual circuits once all the required

resources are assigned to existing circuits. Another limitation of traditional virtual circuit schemes is that the choice of routes for different flows is severely limited due to the need to ensure deadlock freedom. If all packets must follow fixed pre-determined routes, it must not be possible for such packets to form dependency cycles.

To solve the problems with conventional virtual circuit schemes, we have proposed the *Dynamic Virtual Circuits* (DVC) mechanism [34, 36] which combines adaptive routing and connection-based routing. With DVCs, a portion of a virtual circuit can be torn down from an intermediate node on the circuit's path and later be re-established, possibly along a different route, while maintaining the virtual circuit semantics of reliable in-order packet delivery. The DVC mechanism provides fast circuit establishment and re-routing by using only local operations at each node. Packets can proceed on new circuit paths without waiting for end-to-end handshakes. The DVC mechanism allows for routing of flows over low-contention routes, even if these paths do not ensure deadlock freedom. Nevertheless, deadlock freedom is ensured while preserving the benefits of virtual circuits.

The DVC mechanism presented in this paper is the first complete mechanism that demonstrates that virtual circuits and flexible adaptation are not mutually exclusive. We present the DVC algorithm and describe a practical hardware/firmware architecture for its implementation. The challenge in devising the algorithm is to simultaneously provide fully adaptive routing, deadlock-freedom, and the low per-hop overhead of static virtual circuit switching. Compared to pure packet switching networks, DVC networks pose a more complicated deadlock avoidance problem because DVCs introduce dependencies among circuit establishment and teardown operations and also dependencies between these operations and packet buffer resources.

The main contributions of the paper are as follows. First, we present the deadlock-free DVC algorithm. The algorithm allows virtual circuits to use any routes and imposes few constraints on packet buffer usage. A dependency cycle-free virtual network is used to avoid packet routing deadlocks, and a second virtual network decouples circuit manipulation and packet routing operations to avoid more complex deadlocks. Second, we present a description and evaluation of the hardware resources and mechanisms needed to implement the DVC algorithm and show that the scheme is simple to implement with modest hardware requirements. Third, a correctness proof of the DVC algorithm is presented. The proof analyzes the system's state space to show that each packet injected into the network using a DVC is delivered to the DVC destination in order.

The paper is organized as follows. Section 2 reviews approaches for avoiding packet buffer deadlocks in traditional networks. Section 3 describes Dynamic Virtual Circuit (DVC) networks.

Section 4 presents the DVC algorithm including the proposed mechanisms for avoiding deadlocks in DVC networks. Section 5 presents a correctness argument for the scheme, and Section 6 describes practical implementation issues. Section 7 presents simulation results that explore the potential performance benefits of DVCs in one of the possible scenarios where DVCs can be beneficial. Specifically, this is done by considering limit cases, where the more sophisticated (complex) routing possible with DVCs leads to significantly higher performance than can be achieved with conventional packet switching networks, which typically must use simple (e.g., algorithmic) routing. Section 8 summarizes related work in connection-based and adaptive routing.

2. Background: Packet Buffer Deadlock Avoidance

To provide deadlock-freedom in DVC networks, our solution builds on previous deadlock avoidance techniques for traditional networks. Packet routing creates dependencies between packet buffer resources at adjacent switches. Cycles of these dependencies can cause deadlocks which prevent packets from making progress. A simple way to prevent deadlocks is to restrict packet routing such that dependency cycles do not exist, for example by using Dimension Order Routing (DOR) in a mesh network [11]. However, such restricted routing may result in poor performance because of insufficient flexibility to route packets around congested network links or buffers.

A network can provide higher performance by using a less restricted routing function that guarantees deadlocks are avoided despite having buffer dependency cycles. The key is to ensure that packets can escape from any dependency cycles they encounter. This approach was generalized by Duato, who determined necessary and sufficient conditions for deadlock-free routing in cut-through and wormhole networks [15, 16]. Stated informally, there must be a set of packet buffers that can be reached by packets in any buffer in the network, and this set of packet buffers acts as a deadlock-free escape path for the delivery of blocked packets.

An escape path from dependency cycles can be provided by embedding in the network a *virtual network*, consisting of a set of dedicated packet buffers, for which packet routing is free of dependency cycles [4, 12, 39]. For example, in the Disha scheme [4] the physical network is partitioned into two virtual networks, one which allows fully-adaptive routing with dependency cycles, and a second which provides a dependency cycle-free escape path. A packet that blocks in the fully-adaptive virtual network for a timeout period becomes eligible to be transferred to the dependency cycle-free network, which is guaranteed to deliver the packet without creating a deadlock. The timeout used by Disha can be viewed as a heuristic deadlock detection mechanism, and the transfer of timed-out packets to the dependency

cycle-free network can be viewed as a progressive deadlock resolution mechanism [32].

In the following sections we describe the DVC algorithm. The scheme avoids packet buffer deadlocks by using a dependency cycle-free virtual network. It also ensures that new types of dependencies introduced by virtual circuit setup and teardown procedures cannot cause deadlocks.

3. Dynamic Virtual Circuits (DVCs)

With both static virtual circuits and DVCs, a source host initiates communication with a destination host by establishing a new virtual circuit on some path to the destination. The source host then transmits one or more data packets over the new virtual circuit. The data packets carry only a few bits of overhead (packet header) information for addressing and sequencing, and are forwarded at each intermediate switch with very little processing. Finally, the source terminates (tears down) the virtual circuit. Each source and each destination may be end points of many established/active virtual circuits at the same time.

Virtual circuit switching can provide efficient support for traffic patterns that exhibit temporal locality such that each host communicates mostly with a slowly changing set of other hosts and where the size of this set is smaller than the number of nodes in the network. This characteristic enables the benefits of virtual circuits to be achieved because the overhead of frequently establishing and disestablishing virtual circuits is avoided. For example, some parallel applications of practical interest exhibit temporal locality appropriate for virtual circuits [24]. Such applications may execute in phases, where each phase has a stable traffic pattern which changes when the application moves to the next phase or when an application starts or ends. In a system where some traffic requires constant rapid adaptation inappropriate for virtual circuit switching, it would be straightforward to simultaneously support both DVCs and conventional packet routing.

Virtual circuit switching can lead to more efficient bandwidth utilization because most of the ‘‘overhead bits’’ of conventional packet headers need not be transmitted. These overhead bits include source identification (node ID and port ID), destination identification (node ID and port ID), sequencing information, and priority (QoS) control bits. The reduction in packet header size is significant since the payload of many packets is small (e.g., synchronization, acknowledgments, cache blocks).

For networks with irregular topologies [19], virtual circuit switching provides reduced latency for routing packets. Without virtual circuits, packets in irregular networks are routed by accessing large routing tables at each switch instead of using simple algorithmic routing. The routing tables are constructed at system initialization and may be changed over time to adapt to changing traffic conditions [1, 20, 33]. With a simple scheme, the number of entries in the routing table may equal the

number of nodes. However, with more sophisticated routing schemes, routes may also depend on the source node. Furthermore, based on different QoS requirements, it may be beneficial to route different flows from a particular source node to a particular destination node through different paths. Hence, the paths may depend on the source port number and destination port number. Thus, the number of entries may be up to $(\#nodes \times \#ports)^2$. Thus, the number of entries in the routing table may be very large and access to the table may require associative access or hashing. Hence, a lookup in the routing table is likely to be expensive. With virtual circuits, once a circuit is established, forwarding of each packet requires access to a *much* smaller routing table — the Input Mapping Table (IMT). The size of the IMT is related to the number of flows that may actually be maintained at a node at a particular point in time, not to the number of all possible flows that may ever pass through the node.

The reduction in packet header size discussed above also helps reduce latency. Specifically, packet routing can begin earlier after the arrival of the first byte of the packet.

Even with regular network, in order to minimize contention, it may be useful for some flows to be routed based on high-level global knowledge of the traffic patterns. As discussed in Section 7, such routes may not follow the routes obtained by simple algorithmic routing schemes. Hence, the routing of all flows or of some of the flows may be table-driven, leading to the same advantages to using virtual circuits as with irregular networks.

Virtual circuits can also be beneficial in large parallel systems where nodes can be dynamically grouped into partitions. Partitions are used to host multiple parallel applications or multiple parallel virtual machines. Over time the system can become fragmented, requiring each new partition to be assembled as a collection of non-contiguous regions in the topology. Virtual circuits can be created when a partition is set up and provide for efficient communication among the nodes that comprise the partition. Global knowledge of the communication requirements of the partitions in terms of bandwidth and QoS can be used to select network paths that minimize the interference among partitions.

Virtual circuit switching can be a valuable extension for current and emerging cluster and I/O interconnects. An important example is the industry standard InfiniBand architecture [1, 18], which is supported by major systems vendors (e.g., IBM, HP, and Dell) and operating systems (e.g., Linux, and announced support in Microsoft Windows), and is increasingly deployed in clusters (e.g., the NASA Columbia system, which as of November 2005 had the number four position on the TOP500 [2] list of supercomputer sites). InfiniBand matches our system model in its use of backpressure flow control and routing tables at each switch to support irregular topologies. In addition, one of the main communication mechanisms provided by InfiniBand requires establishing a connection between a ‘‘queue pair’’ at a

source host and a queue pair at a destination host. However, network switches do not reserve state for connections. Thus each packet must have routing and transport headers specifying source host and queue pair IDs, destination host and queue pair IDs, a service level indicating the scheduling priority for the connection, etc. If InfiniBand switches and protocols were extended to support establishment of virtual circuits for queue pair connections, much of this information could be eliminated, improving transmission efficiency.

As described in Section 1, the advantages of virtual circuits often cannot be fully realized because of the inability to adapt to changing traffic conditions. Unlike traditional static virtual circuits, our proposed DVC mechanism enables any intermediate switch on a circuit's path to independently tear down the circuit for the portion of the path from the switch to the destination (or from the switch to the first subsequent switch along the path that has also torn down the same circuit). The intermediate switch may later re-establish the portion of the torn down circuit from the intermediate switch to the destination in order to forward additional data packets that arrive. The new path chosen for the circuit may be different from the original path. For example, the new path may be chosen because it is less congested than the original path.

With DVCs, circuit teardown and re-establishment from an intermediate switch are fast, local operations that completely avoid time-consuming synchronization with the circuit's endpoints. A single DVC can even be torn down by multiple intermediate switches concurrently. However, without synchronization a race condition can develop in which data packets that traverse a new path after circuit re-establishment can arrive at the destination before older data packets that traverse one of the previously used paths. To preserve the FIFO delivery semantics of virtual circuits, the destination reorders the arriving packets that belong to the same circuit. As discussed below, to facilitate this packet reordering, each time a circuit is re-established one sequence number must be sent through the network. However, the overwhelming majority of the packets do not include sequence numbers [34]. Although the destination needs to reorder packets when circuits are torn down and later re-established, this is a rare event; packets almost always arrive in order. Hence, there is not need to devote resources to optimizing the performance of packet reordering at the receiver.

We next describe the basic steps of DVC establishment, disestablishment, and rerouting. Consider a virtual cut-through network composed of $n \times n$ switches interconnected with bidirectional point-to-point links. At each switch, one or more ports may connect the switch to one or more hosts. The switch is input buffered with an $n \times n$ crossbar connecting the n input ports to the n output ports.

The source host establishes a new DVC by injecting a Circuit Establishment Packet (CEP) into the

network. The CEP records the DVC's source and destination addresses, which are used to adaptively route the CEP. For example, CEP routing may be accomplished through the use of routing tables maintained at each switch, such as in the SGI Spider chip [20].

A CEP allocates for a new DVC one *Routing Virtual Channel* (RVC) on each link it traverses on its path from source to destination (including the source and destination host interface links). An RVC is an entry in a table at the switch input port that is connected to the link. Each physical link is logically subdivided into multiple RVCs. The number of RVCs that a switch provides for a link determines how many DVCs can be established at a time on the link. Each packet header has a field that identifies the RVC used by the packet. At each switch, the mapping from input RVC to output RVC is recorded in a small ''Input Mapping Table'' (IMT) at the input port. The IMT is indexed by the RVC value in the header of an arriving packet. An IMT entry records the following information about the DVC that has allocated the RVC: the output port, output RVC value, source and destination addresses, and sequence number.

Note that we use the term ''RVC'' instead of the more familiar ''virtual channel'' to distinguish it from the same term commonly used to refer to flow-controlled buffers that prevent deadlock and increase performance [11]. We call the latter *Buffering Virtual Channels*, or ''BVCs''. In contrast, ''RVCs'' simply identify DVCs, and they do not require separate flow-controlled buffers.

The source may transmit one or more data packets over a new virtual circuit. Each data packet is quickly routed at each switch, by accessing the IMT entry with the RVC value in the packet header. The RVC value in each packet's header is overwritten with the output RVC value recorded in the IMT entry, and the packet is enqueued for transmission to the next switch. The number of RVCs supported by the switch determines the IMT access time and therefore the latency of routing a packet. The switch designer must balance the goal of supporting a large number of established DVCs over each link with the goal of ensuring fast IMT access time for low latency packet routing. As described earlier, some systems can limit the number of entries needed in the IMT by using virtual circuit switching only for some of the flows while using conventional packet switching for the remaining traffic which does not exhibit temporal locality.

The source host terminates the virtual circuit by injecting a Circuit Destruction Packet (CDP) into the network. The CDP traverses the circuit path, releasing at each hop the RVC that is allocated to the circuit after the data packets that use the circuit have departed the switch.

With DVCs, any intermediate switch may independently initiate a teardown at any time by inserting a CDP into the circuit path at the switch input port. An intermediate switch can tear down a circuit to

adapt to changing traffic conditions by shifting the circuit onto a lower latency path. Alternatively, the switch may tear down the circuit to free up an output RVC to allocate to a new circuit. Various policies could be used to select a victim output RVC for teardown. The implementation approach described later in Section 6.1 provides a selection mechanism that approximates Least Recently Used (LRU) replacement. A CDP traverses the circuit's path until it reaches either the destination or another switch that has also torn down the same circuit. The portion of the circuit from the source to the intermediate switch remains intact, unless the source or a switch along that portion of the circuit also initiates a teardown.

A data packet that arrives on the input RVC of a torn-down DVC triggers DVC re-establishment, in which the switch creates a new CEP from the information retained in the IMT. The CEP is routed to the destination, allocating RVCs on the new path. There are no restrictions on when to reroute DVCs or which new paths to take.

The adaptive rerouting of DVCs requires some packets to be stamped with sequence numbers for reordering at the destination. Specifically, each CEP is stamped with the sequence number for the next data packet of the circuit. Each switch along the path records the sequence number in the CEP as a circuit is established or re-established. The switch increments the sequence number for each data packet that subsequently arrives on the circuit.

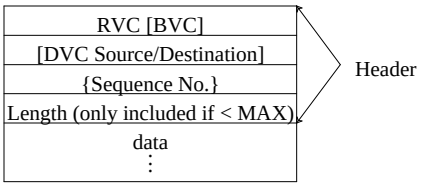


Figure 1: Packet Format. Fields in ''[]'' have the stated use only for diverted data packets (described in Section 4.1). The sequence number field is used only for CEPs, diverted data packets, and the next non-diverted data packet.

The general packet format is shown in Figure 1. A packet consists of a header followed by data phits. The first phit of the header records the RVC value. An additional four bits of the RVC field indicate packet type and whether the packet is of maximum length. If not, a length field is present in the header. For most packets, the header consists only of the RVC field and possibly the length field. For a minority of packets, the header includes additional fields. These additional fields are required for data packets that are diverted onto deadlock escape paths.

4. DVCs With Deadlock Avoidance

This section shows how the mechanism for tearing down and re-establishing Dynamic Virtual Circuits can be combined with a deadlock avoidance scheme that provides packets with an escape path from potential deadlock cycles. Data packets that take the escape path are routed individually to their destinations, independently of the virtual circuit paths.

4.1. Avoiding Deadlocks Involving Data Packets

The DVC mechanism supports adaptive routing by imposing no restrictions on the choice of path for any circuit, and by enabling circuits to be rerouted adaptively during their lifetimes onto new paths to minimize latency and maximize throughput. The flexibility of adaptive routing comes at the cost of possible deadlocks that involve the packet buffers. Deadlock cycles may also involve RVCs, since those resources are contended for by the circuit establishment and disestablishment operations at a switch.

To avoid deadlocks arising from the unrestricted paths of DVCs, we embed in the physical network two virtual networks: the *primary network* and the *diversion network* [16]. Each virtual network is composed of one *Buffering Virtual Channel* (BVC) per switch input port (i.e. per link). We name the two BVCs the *primary BVC* and the *diversion BVC*. Each is associated with a buffer (the ‘‘primary’’ and ‘‘diversion’’ buffers), which may be a FIFO buffer, or a more efficient Dynamically Allocated Multi-Queue (DAMQ) buffer [35], or a buffer with any other organization.

The primary network supports fully-adaptive routing. This allows data packets to follow the unconstrained paths that are assigned to virtual circuits, but it also creates dependency cycles among packet buffers. In contrast, routing in the diversion network is constrained such that dependency cycles do not exist (e.g., in a mesh topology, Dimension-Order Routing (DOR) could be used in the diversion network). In addition, the diversion network can accept blocked packets from the primary network to provide a deadlock-free escape path [16]. We say that a data packet is *diverted* if it takes a hop from the primary network into the diversion network.

Whereas virtual circuit forwarding is used to route data packets in the primary virtual network, traditional packet routing is used in the diversion network. Hence, while data packet headers in the primary network consist of only an RVC number, headers of packets in the diversion network must include source, destination, and sequencing information (Figure 1). When a data packet in the primary network becomes eligible to be diverted, the switch obtains this information from the IMT entry where this required information is recorded. To enable locating the correct IMT entry, the primary buffer retains each packet’s input RVC value until the packet is forwarded on the output RVC. As long as diversions

are rare, the slower forwarding of diverted packets at intermediate switches and the overhead of transmitting the larger headers of diverted packets will not significantly impact network performance. Furthermore, if diversions are rare, small diversion buffers are sufficient (e.g., capacity of one packet).

Data packets that become blocked in the primary network can time out and become eligible to enter the diversion network on the next hop. The routing function used for the diversion network determines which output links the blocked packet in the primary network can take to enter the diversion network. The packet eventually advances, either by taking one hop on its virtual circuit path within the primary network, or by taking one hop into the diversion network on an adjacent switch. To enable switches to identify which virtual network an arriving packet is using, one RVC is designated as special. A data packet arriving on this special RVC uses the diversion BVC, else it uses the primary BVC.

Since diverted data packets may arrive out of order at the destination, each diverted data packet is stamped with a packet sequence number for use by the packet reordering at the destination. After a packet is diverted, the next data packet on the same circuit is also stamped with a sequence number. At each subsequent switch the non-diverted data packet visits along the circuit path, the switch reads the sequence number and locally updates its record to account for the diverted data packets. As long as only a minority of packets require the escape path, most of the advantages of static virtual circuits are maintained with DVCs.

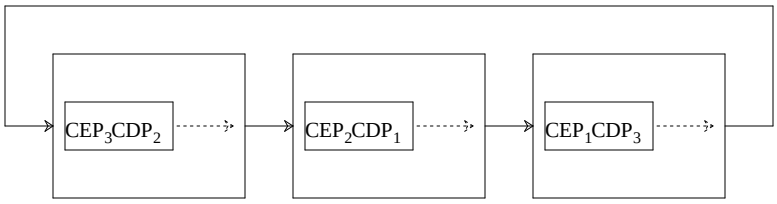


Figure 2: Deadlock involving only control packets in three switches. Packet buffer capacity is two packets. Each CEP_i establishes a unique virtual circuit. The matching CDP_i will disestablish the circuit set up by CEP_i.

4.2. Avoiding Deadlocks Involving Control Packets

Whereas data packets are diverted from their virtual circuit paths to avoid deadlock, control packets (CEPs and CDPs) cannot deviate from virtual circuit paths. Instead, each CEP must traverse the path that is selected for it by the network's fully adaptive routing function, and each CDP must traverse the path used by the virtual circuit it is disestablishing. A deadlock forms when a cycle of packet buffers fills with CEPs and CDPs, as shown by example in Figure 2.

To avoid such deadlocks we develop a mechanism which prevents any buffer from filling with

control packets and blocking. The mechanism is derived by analyzing all possible sequences of arrivals of control packets on a single RVC to identify the storage required for each arrival sequence. Deadlocks are prevented by providing switches with packet buffers that are large enough to store the control packets of the worst case sequence without filling and blocking.

We initially examine arrival sequences consisting only of control packets on a single RVC to find the per-RVC storage requirement. Control packets arrive in alternating order (CDP, CEP, CDP, etc.) on an RVC. If a CDP arrives and the CEP that matches it is present at the switch without an intervening data packet, then the CEP and CDP are deleted from the network (freeing buffer slots). If they were not deleted, the circuit establishment and disestablishment would be wasted operations since the circuit is not used by any data packet. Thus if the first packet in an arrival sequence is a CEP, then the CDP that follows it causes both packets to be deleted. One other case of packet deletion is possible; a CDP is deleted if it arrives at a switch where its circuit is already torn down as a victim. If the first packet of a sequence is a CDP that tears down an existing circuit, then the second packet is a CEP that establishes a new circuit. The third packet is a CDP, which matches the CEP and causes both packets to be deleted. Therefore, in the worst case storage is required for one CDP and one CEP for each RVC. One additional buffer slot is needed that is common to all the RVCs. This slot accommodates the arrival on any RVC of a CDP that will be deleted along with its matching CEP. Hence an input port that supports R RVCs requires storage for R CEPs plus $(R+1)$ CDPs to handle arrival sequences without data packets.

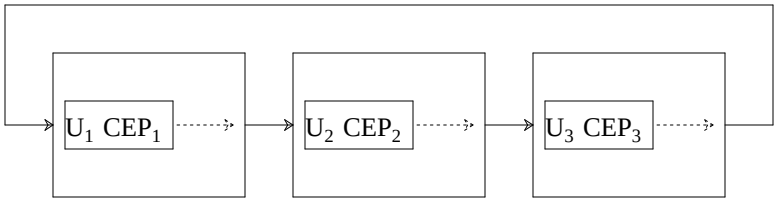


Figure 3: Example deadlock involving three switches. Each U_i is an unmapped data packet that is waiting for CEP_i to establish an RVC mapping. When a CEP transmits to the next switch, the unmapped data packet is converted into a mapped data packet. So long as the unmapped data packets are present at the input ports, subsequent arrivals on the primary Buffering Virtual Channel (BVC) are blocked. Therefore, the CEPs cannot make progress, and a deadlock results.

We next consider arrival sequences that include data packets. A data packet that uses a circuit that is currently allocated an output RVC at the switch is called ‘‘mapped’’. A mapped data packet eventually frees its buffer slot via normal forwarding or via diversion. Therefore, including a mapped data packet in an arrival sequence does not increase the storage required to prevent deadlock. A data packet that uses a circuit that is not currently allocated an output RVC is called ‘‘unmapped’’. It is complex to divert an

unmapped data packet because its circuit identification information is not recorded in the IMT entry, as with mapped data packets, but rather in a CEP that is somewhere in the packet buffer [37].

To avoid this complexity, we prohibit diverting unmapped data packets. However, this causes the storage requirement to become unbounded because an infinite repeating pattern of packet arrivals of the following form may arrive to a single RVC: CEP, followed by unmapped data packets, followed by CDP. Matching CEPs and CDPs cannot be deleted because unmapped data packets intervene.

To restore a bound on the storage requirement, we restrict each input to store at most one unmapped data packet at a time (having a larger limit on the number of unmapped packets would also work but would increase the storage requirements). If a data packet arrives and the Input Mapping Table lookup reveals that the RVC is free or allocated to a different circuit, then the primary buffer asserts flow control to block subsequent packet arrivals.

Blocking flow when an unmapped data packet arrives prevents control packets from arriving in addition to data packets. Blocking control packets can cause deadlocks, as shown in Figure 3.

We prevent such deadlocks by introducing a new Buffering Virtual Channel (BVC) for control packets. We call the new BVC the *Control BVC*. The Control BVC is the third and final BVC, in addition to the Primary BVC and the Diversion BVC. The buffering associated with the Control BVC is dedicated to storing control packets. Introducing the Control BVC allows control packets to arrive even when the Primary buffer is full with data packets. The deadlock of Figure 3 does not occur since the CEPs in the figure can all advance using the Control BVC.

Prohibiting unmapped data packets from being diverted and restricting each input port to store at most one unmapped data packet cause the storage requirement for control packets to increase slightly over the case with no data packets. The additional storage is common to all the RVCs at a switch input port and can store one CDP and one CEP. This increases the storage requirement to $(R+1)$ CEPs plus $(R+2)$ CDPs for an input port that supports R RVCs. The two additional control packets arrive after the unmapped data packet and on its RVC. We present elsewhere (Section 5.1.2 in reference [37]) an exhaustive examination of the storage requirements for all possible packet arrival sequences. The worst case sequence of packet arrivals on the RVC that is used by the unmapped data packet is as follows (in order of arrival): mapped data packets, CDP_1 (will release the RVC), CEP_2 (will allocate the RVC for the next data packet), unmapped data packet, CDP_2 (will release the RVC), and CEP_3 (will allocate the RVC for yet another circuit). Since the unmapped data packet cannot be diverted, all four control packets in the sequence are necessary and cannot be deleted. After this sequence of arrivals, CDP_3 may arrive which matches CEP_3. In this case, both CDP_3 and CEP_3 are unnecessary and are deleted.

It is possible for the initial CEP and final CDP that are injected by a source host for a circuit to be deleted in the network. This occurs if all the data packets that use the circuit are diverted, and the CDP catches up to the CEP at the switch, with no intervening data packet. Thus, the destination host must not depend on the arrival of these control packets to trigger the allocation and deallocation of host memory resources needed for the circuit to deliver packets. To resolve this potential problem, the source host interface ensures that the first data packet that uses a circuit replicates the information in the initial CEP, and the final data packet of a circuit consists only of a header (no data payload). Since data packets are never deleted, these packets are guaranteed to reach the destination. The destination host allocates resources for the circuit when either the initial data packet or the initial CEP arrives. Since the final CDP of the circuit can be deleted in the network, the destination host releases host resources for a circuit only after all the data packets of the circuit, including the special final data packet, have arrived. The destination host uses the sequence numbers stamped on all out-of-order data packets to detect whether any data packets are in transit.

4.3. Algorithm for Handling Control Packets with DVCs

As described in the previous subsections, deadlock-freedom is maintained using data packet diversion through a deadlock-free virtual network. The DVC mechanism imposes two restrictions that simplify the algorithms and their implementations: each input port accommodates at most one unmapped data packet at a time, and unmapped data packets cannot be diverted. Dedicated control packet buffers are used to prevent deadlocks that include control packets. These control packet buffers are large enough to ensure that they can never fill up and cause blocking.

The control packet buffer is organized as follows. For each RVC i, a logical queue is maintained of the control packets in the order they will be transmitted to the next switch. The head of the logical queue for RVC i is comprised of a queue H_i with space for one CEP and one CDP. Whenever an unmapped data packet is present on RVC i, the tail of the logical queue is extended by a queue T_* with space for one CEP and one CDP. Queue T_* is shared by all RVCs at an input port but used by only one RVC at a time.

The algorithm for managing DVCs and control packets is listed in Figure 4. Details of data packet forwarding are not shown because the focus is on circuit manipulation. The algorithm listing shows the key DVC control events that occur at a switch and the actions the switch takes in response. For example, the first event shown is ‘‘Output port Z becomes free’’. The following lines 1 through 15 show how the switch decides which next packet to transmit on output Z and the actions that are taken if the transmitted packet is a CDP or a CEP. The algorithm listing uses the convention that i and j refer to input RVCs, k

is an output RVC, X is an input port, and Z is an output port. Also, N_X is a primary input buffer, D_X is a diversion buffer, and H_i and T_* are control packet buffers.

Conditions/Events and Actions (atomic):

1. Output port Z becomes free

```
1       Arbitrate access (assume RVC i of input
        port X maps to RVC k of output port Z),
        but also obey BVC flow control.
        Arbitration priority:

2          A. a packet in some DX buffer
              waiting for port Z or a mapped
              packet on RVC k from buffer NX
3          B. a CEP/CDP on RVC k from head of Hi

4       when a CDP from RVC i is transmitted
        on RVC k:
5          delete mapping from i to k
6          if a CEP is waiting at the head of
           some Hj for RVC j {
7            set up new mapping from RVC j
             to RVC k
8          }
9          if required, transfer head of
           T* to Hi

10      when a CEP from input RVC i is
        transmitted on output RVC k:
11         if an unmapped packet on RVC i
           exists {
12            convert unmapped packet to a
              mapped packet
13            unblock flow to primary input
              buffer NX
14         }
15         if required, transfer head of
           T* to Hi
```

2. CEP arrives on RVC i of input port X

```
16      if Hi is not full {
17         enqueue CEP at tail of Hi
18      } else {
19         place CEP at tail of T*
20      }
```

3. CEP at head of H_i and RVC i does not map to an output RVC

```
21      record SRC, DST in Input Mapping Table
        entry for RVC i

22      output port Z <-
        route(CEP's destination field)

23      if free RVC k exists on output port Z {
24         set up mapping from i to k
25      } else if the number of RVCs on Z for
        which there are CDPs at the switch is
        less than the number of CEPs that are on
        unmapped RVCs and are routed to output
        port Z {

26         // it is necessary to select and
           // teardown a victim

27         select RVC k (that reverse maps to
           input RVC j != i) such that no CDP
           resides at Hj

28         create CDP, place at tail of Hj

29         (note: the next data packet to
           arrive on RVC j must be considered
           unmapped even though RVC j stays
           mapped to RVC k until the
           transmission of the CDP in Hj)
30      }
```

4. Unmapped packet arrives on RVC i of input port X

```
31      block flow to primary input buffer NX
32      if Hi has no CEP {
33         create CEP using circuit
           info in RVC i's Input Mapping
           Table (IMT) entry

34         place CEP at tail of Hi
35      }
```

5. CDP arrives on RVC i of input port X

```
36      if CDP is redundant {
37         delete arriving CDP
38      } else if CDP matches a CEP not
        associated with an unmapped packet {
39         delete both the CEP and the
           arriving CDP
40      } else if Hi not full {
41         place arriving CDP at tail of Hi
42      } else {
43         place arriving CDP at tail of T*
44      }
```

Figure 4: Algorithm for Handling Control Packets with DVCs

5. Algorithm Correctness

We consider the algorithm to be *correct* if the following statement is true: *All data packets injected into the network on a DVC are delivered eventually to the DVC destination in the order injected*. Eventual delivery to the DVC destination is guaranteed if packets make progress (the network is deadlock-free) and data packets are never delivered to incorrect destinations. In-order delivery is guaranteed by attaching a sequence number to each packet that may arrive at the destination out of order.

Section 5.1 below proves that network deadlocks are impossible. Section 5.2 proves that data packets are associated with the same DVC from injection to delivery. Together with the FIFO delivery mechanism described in Section 4, these results show the algorithm satisfies the statement of correctness.

5.1. Proof of Deadlock-Freedom

To prove the network never enters a deadlocked configuration, we examine data packets and control packets separately. Theorem 1 below shows that diverted and mapped data packets cannot be part of a deadlock. Theorems 2 and 3 prove that control packets make progress which in turn guarantees that mappings are eventually provided for unmapped data packets. Together, the theorems show that every packet, regardless of its type, is guaranteed to make progress.

Theorem 1: Every mapped data packet in a primary buffer N_X and every data packet in a diversion buffer D_X is eventually forwarded.

Proof: This follows from Duato's sufficient condition for deadlock freedom in a virtual cut-through packet-switching network [16]. Deadlock is avoided if there exists a set C of BVCs such that all packets can reach set C in one hop, routing in set C reaches all destinations from all switches, and the set C is free of buffer dependency cycles. The set of diversion buffers D_X meets the definition of set C. A packet in any primary buffer N_X can enter the diversion network by taking one hop and is routed to its destination with a routing policy that is free of dependency cycles (Section 4.1). □

Theorem 2: Control packets do not block in deadlock cycles that involve multiple switches.

Proof: Assume a deadlock involves a control packet and an entity at another switch. By examining all possible chains of dependencies from a control packet to entities at neighboring switches, we show that each chain includes an entity that cannot be in a deadlock cycle, contradicting the assumption.

A control packet may wait directly for one of the following five entities: a buffer slot at the next switch, the output link, a mapped data packet at the same switch, a control packet at the same switch, or an unmapped data packet at the same switch. We examine each entity in turn. First, dedicated control packet buffers at the neighbor cannot be part of a multiple switch deadlock cycle because they have sufficient capacity to avoid filling and blocking. Second, eventual access to the output link is guaranteed for bounded length packets with virtual cut-through forwarding and starvation-free switch crossbar scheduling. Third, mapped data packets cannot participate in deadlock cycles (Theorem 1). Fourth, all control packets have the same set of possible dependencies to other entities, hence a dependence of one control packet on another control packet at the same switch does not introduce the possibility for a deadlock cycle that spans multiple switches. Fifth, a CDP may directly wait for an unmapped data packet

at the same switch and on the same RVC. In turn, the unmapped data packet waits for a CEP at the same switch to transmit, which will cause the unmapped data packet to become mapped. The chain of dependencies (from the CDP to the unmapped data packet to the CEP at the same switch) is equivalent to case four above (direct dependence from one control packet to another at the same switch). □

Theorem 3: Control packets do not enter intra-switch deadlocks.

Proof Sketch (full proof in Section 5.2.1 of reference [37]): Such deadlocks would arise from cycles of dependencies among entities within a single switch. We construct a graph of all the dependencies at a switch that involve control packet buffers. The resulting graph is acyclic, hence intra-switch deadlock is impossible.

The dependency graph is constructed from the control packet handling algorithm (Figure 4) along with an enumeration of all possible buffer states. The buffer space enumeration is presented in Sections 5.1.2.1 and 5.1.2.2 of reference [37]. The dependency graph characterizes a mapped RVC i at some input port X and an unmapped RVC j at some input port Y. RVCs i and j are arbitrary representatives of their classes: mapped and unmapped RVCs, respectively. The set of all dependencies associated with these representative RVCs captures all the dependencies associated with all mapped RVCs and all unmapped RVCs because there are no dependencies between different RVCs in the same class. □

Theorem 4: The network is deadlock-free.

Proof: By Theorem 1, mapped data packets and diverted data packets always make progress. By Theorems 2 and 3, control packets always make progress, hence if a switch creates a CEP to establish a mapping for an unmapped data packet, the CEP is guaranteed to establish the mapping and transmit to the next switch. Once that occurs, the unmapped data packet becomes a mapped data packet. Therefore, all types of packets make progress, and the network is deadlock free. □

5.2. Correct Delivery

To complete the correctness discussion, we show that the deadlock-free network delivers each data packet to the destination of its virtual circuit (instead of some other destination). We also show that the destination can associate each data packet with its virtual circuit. This association is needed to deliver the contents of a data packet to the host application to which the packet's circuit is assigned.

The proof is based on a state transition table derived from an elaboration of the possible states of a switch. From the perspective of some RVC i of input port X of a switch, the switch state is the combination of the states of the buffers H_i, T_*, the primary buffer N_X, and the record in the IMT entry of a

mapping to an output RVC. For our purpose and from the perspective of RVC i, the state of buffers at other input ports does not matter: the state of those buffers does not determine the destination to which a data packet on RVC i is forwarded. The state of the diversion buffer D_X is also irrelevant because diverted data packets do not participate in circuit manipulation operations.

In Section 5.2.2 of reference [37], we present in detail the elaboration of the state space as a list of sets of states, where each set of states is classified as either ‘‘reachable’’ or ‘‘unreachable’’. Reachable states arise through error-free operation of the switch, and the switch never enters unreachable states. A set of states is labeled unreachable if it contradicts steps of the control packet handling algorithm (Figure 4) or basic properties of DVCs such as the alternating arrival of CEPs and CDPs on an RVC.

To verify that the manual elaboration of the state space is complete and consistent, we wrote a program that exhaustively generates all combinations of all values each buffer may hold. The program verifies that each generated state matches at least one set of states in the list and that all matching sets are labeled identically (all reachable or all unreachable).

Using the set of reachable states and the control packet handling algorithm, we next derive the possible state transitions. Each state is represented by the contents of H_i, T_*, and two more fields that capture the relevant state of N_X and the IMT entry for RVC i. H_i can hold one CEP and one CDP on RVC i. T_* can hold one CEP and one CDP on any RVC. The state of N_X and the IMT entry for RVC i are too numerous to list exhaustively, but the relevant states are determined by the mapping status of RVC i, and whether an unmapped packet is present at input port X. That information is represented by using two symbols: map_i and U_X. Table 1 shows the state transitions for all reachable states. The state notation used in Table 1 is described in Table 2 which lists the possible states of a slot for H_i and T_* and defines the symbols used in columns U_X and map_i. In the state transition table, the actions that trigger a transition are as follows: k (RVC i acquires mapping to output RVC k), D (CDP arrives on RVC i), E (CEP arrives on RVC i), U (unmapped packet arrives on RVC i), T (transmission of packet at head of H_i), D' (CDP arrives on RVC $i' \neq i$), E' (CEP arrives on RVC $i' \neq i$), $U_{i'}$ (arrival of unmapped packet on some RVC $i' \neq i$), $TE_{i'}$ (transmission of CEP at head of $H_{i'}$), and $TD_{i'}$ (transmission of CDP at head of $H_{i'}$). Each entry of the table is derived by tracing through the control packet handling algorithm to determine what the next state would be given each possible action. Note that each empty entry in Table 1 means the column’s action cannot occur in the present state for the row. For example, for rows 0–9, there is no packet in buffer H_i. Therefore, action T, transmission of the packet at the head of H_i, cannot happen.

In some cases, two next states are indicated for some transitions, indicating that either of the two next states may become the new present state. This is because the next state depends on the present state

Present State						Action/Next State									
ID	H_i		T_*		$U_X map_i$	k	D	E	U	T	D′	E′	$U_{i'}$	$TE_{i'}$	$TD_{i'}$
0	–	–	–	–	– F		0	10	12		0	0	2	0	0
1	–	–	–	–	– T		26				1	1	3	1	1
2	–	–	–	–	i′ F		2	14			2,6	4		0	2
3	–	–	–	–	i′ T		27				3,7	5		1	3
4	–	–	CEP′	–	i′ F		4	16			2			0	
5	–	–	CEP′	–	i′ T		28				3			1	
6	–	–	CDP′	–	i′ F		6	18			6	8			2
7	–	–	CDP′	–	i′ T		29				7	9			3
8	–	–	CDP′	CEP′	i′ F		8	20			6				4
9	–	–	CDP′	CEP′	i′ T		30				7				5
10	CEP	–	–	–	– F	11	0		12		10	10	14	10	10
11	CEP	–	–	–	– T		0		13	1	11	11	15	11	11
12	CEP	–	–	–	i F	13	22				12	12		12	12
13	CEP	–	–	–	i T		23			1	13	13		13	13
14	CEP	–	–	–	i′ F	15	2				14	16		10	14
15	CEP	–	–	–	i′ T		3			3	15	17		11	15
16	CEP	–	CEP′	–	i′ F	17	4				14			10	
17	CEP	–	CEP′	–	i′ T		5			5	15			11	
18	CEP	–	CDP′	–	i′ F	19	6				18	20			14
19	CEP	–	CDP′	–	i′ T		7			7	19	21			15
20	CEP	–	CDP′	CEP′	i′ F	21	8				18				16
21	CEP	–	CDP′	CEP′	i′ T		9			9	19				17
22	CEP	CDP	–	–	i F	23	22	24			22	22		22	22
23	CEP	CDP	–	–	i T		23	25		26	23	23		23	23
24	CEP	CDP	CEP	–	i F	25	22				24	22		24	24
25	CEP	CDP	CEP	–	i T		23			31	25	23		25	25
26	CDP	–	–	–	– T		26	31	32	0	26	26	27	26	26
27	CDP	–	–	–	i′ T		27	33		2	27,29	28		26	27
28	CDP	–	CEP′	–	i′ T		28	34		4	27			26	
29	CDP	–	CDP′	–	i′ T		29	35		6	29	30			27
30	CDP	–	CDP′	CEP′	i′ T		30	36		8	29				28
31	CDP	CEP	–	–	– T		26		32	10	31	31	33	31	31
32	CDP	CEP	–	–	i T		37			12	32	32		32	32
33	CDP	CEP	–	–	i′ T		27			14	33,35	34		31	33
34	CDP	CEP	CEP′	–	i′ T		28			16	33			31	
35	CDP	CEP	CDP′	–	i′ T		29			18	35	36			33
36	CDP	CEP	CDP′	CEP′	i′ T		30			20	35				34
37	CDP	CEP	CDP	–	i T		37	38		22	37	37		37	37
38	CDP	CEP	CDP	CEP	i T		37			24	38	38		38	38

Table 1: State Transition Table

of $H_{i'}$, which is not displayed in Table 1 because it is not associated with RVC i.

Table 1 is used below to show that a packet P, which is injected on virtual circuit V, reaches the destination of circuit V, and the host interface at the destination associates packet P with circuit V.

Theorem 5: Whenever packet P is mapped, its input RVC identifies virtual circuit V.

Proof: The proof is by induction on the distance from the virtual circuit's source.

H_i and T_*		T_* only		U_X		map_i	
-	empty	CDP′	CDP on RVC i′ ≠ i	-	no unmapped packet	T	RVC i is mapped to an output RVC k, which also means that N_X may store mapped data packets from RVC i.
CDP	CDP on RVC i	CEP′	CEP on RVC i′ ≠ i	i	one unmapped packet on RVC i is in N_X (note: N_X may hold at most one unmapped packet)	F	RVC i is unmapped, which means any mapped packets in N_X did not arrive on RVC i.
				i′	the unmapped packet is on some RVC i′ ≠ i		

Table 2: State Notation for Table 1

Basis: The theorem is true at the injecting host interface because the host transmits only one CEP for the circuit until it is disestablished.

Induction Step: Assume the theorem is true for P at switch or host S_n (n hops from the source). We now show the theorem is also true at the host or switch that is $n+1$ hops from the source (at host/switch S_{n+1}). Suppose packet P is about to transmit to host/switch S_{n+1}'s primary BVC on RVC k. Since P is about to transmit, P must be a mapped data packet at host/switch S_n. Therefore, the previous control packet on RVC k must have been a CEP that identified V.

If S_{n+1} is a host instead of a switch, then the host will correctly associate packet P with virtual circuit V. However, if S_{n+1} is a switch, then there may be a danger that the CEP will be deleted before it reaches the head of H_k. Only if the CEP reaches the head of H_k will the switch S_{n+1} record the circuit identification information for V in the IMT entry for RVC k (algorithm line 21). We will now prove that the danger is unwarranted. The CEP will successfully reach the head of H_k at switch S_{n+1}.

Switch S_{n+1} may delete a CEP if a CDP arrives from S_n or is generated locally. In our scenario, a CDP may not arrive at S_{n+1} from S_n, since packet P is about to transmit from S_n. It is impossible for a CDP to transmit ahead of a mapped data packet such as P (see arbitration priority rules in lines 1–3 of Figure 4). Therefore, only a CDP that is generated locally at S_{n+1} could trigger the deletion of the CEP.

From Table 1, we can identify all the state transitions in which a CEP is deleted upon introduction of a CDP. That is, we can identify all present states in the table such that the column D transition takes the switch to a next state that has one less CEP than the present state. For some of those states (10, 11, and 14 through 21), the CEP that identifies virtual circuit V is already at the head of the logical queue (i.e. H_k for switch S_{n+1}), hence the IMT entry identifies V. Even if that CEP is deleted, the information is available when packet P arrives at switch S_{n+1}. The remaining qualifying states are 24, 25, 31, and 33 through 36. In those states, a CDP is present in the buffer. According to line 27 of Figure 4, the presence of the CDP prevents the introduction of a locally-generated CDP. Therefore we conclude that it is not possible for a locally-generated CDP to cancel the CEP before the IMT is written to identify circuit V. □

Theorem 6: If packet P is a diverted packet, then P's header identifies circuit V.

Proof: By theorem 5, at the time P is diverted, the IMT entry identifies circuit V. The procedure for diversion attaches that information to packet P's header. After diversion, the packet header is not altered until P is delivered. Therefore, packet P is always associated with circuit V. □

6. Implementation Issues

Efficient implementation of the DVC algorithms presented earlier requires specialized hardware support. To demonstrate the feasibility of using DVCs, this section presents key components of a possible implementation. For cost-effective high performance, new high-speed hardware is introduced only where needed to support frequently executed performance-critical operations. Less expensive (in terms of chip area) hardware components can be used to perform the remaining operations which are less frequent. To illustrate the costs of the proposed hardware support for DVCs, we consider a realistic example switch configuration in Section 6.4. The example demonstrates that the additional hardware components needed to implement the DVC algorithms have modest cost and are practical to implement with current technology.

6.1. Switch Organization

An $n \times n$ switch can be implemented as a single-chip device. To perform the relatively complex yet infrequent circuit manipulation operations of the DVC algorithm, the switch can include a simple processor core on-chip. The processor can also be used for other purposes, for example executing higher-level routing protocols that identify low latency paths [33]. Compared to specialized logic, using a programmable processor core may lower design time (processor cores are readily available as standard cells for chip designs), and it provides maximum flexibility for implementing updates to change, extend, or customize functionality. The processor needs on-chip memory for use as an instruction store. It needs only a small amount of on-chip memory to store data since the DVC algorithms are control-intensive rather than data-intensive. An example 32-bit microprocessor with embedded 512 kilobytes of flash memory and 32 kilobytes of SRAM can be implemented with 1.65 million transistors [27]. Since current widely-used chips consist of hundreds of millions of transistors, the simple processor core and instruction memory require only a very small fraction of the silicon area of a modern chip.

At each input port of the switch, there is a set of packet buffers. Each input port X has primary buffer N_X, diversion buffer D_X, and control packet buffer T_*^X. The input buffers are connected to output ports through an $n \times n$ crossbar. Each input port also has an Input Mapping Table (IMT) that records RVC mappings (Figure 5).

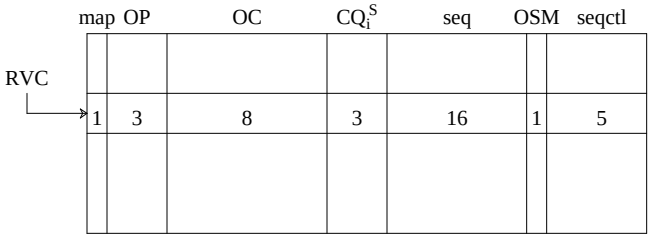


Figure 5: Input Mapping Table (one at each input port). Field widths in bits are shown.

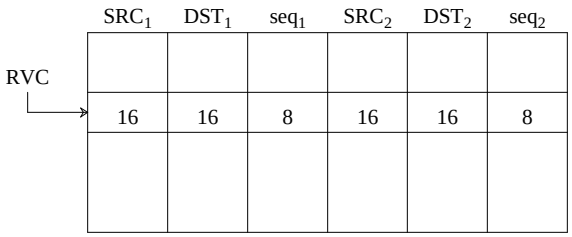


Figure 6: Circuit Information Table (one for each input port). Holds information about mapped and torn DVCs, and CEP information from H_i.

When a data packet arrives to an input port, its RVC identifier is used to access an IMT entry. The entry's ''map'' field indicates whether the input RVC is mapped to an output RVC (i.e., whether a circuit is established). If so, the ''OP'' field specifies the output port to forward the arriving packet, and the ''OC'' field specifies the output RVC value which replaces the value in the packet's header. The input RVC value is saved in N_X with the data packet and is discarded when the packet is dequeued upon transmission. The remaining fields of the IMT are used to keep track of control packet buffering (Section 6.2) and FIFO sequencing (Section 6.3).

Each output port has an Output Mapping Table (OMT). The OMT is indexed by output RVC value and indicates the status of the output RVC. Each OMT entry has 3 bits. The ''map'' bit is set when the output RVC becomes mapped to an input RVC. It is cleared when a CDP transmits to the next switch using that RVC, releasing the output RVC for use by a new circuit. The ''victim'' bit is set when the circuit mapped to the output RVC is chosen as a victim for teardown. It too is cleared when the CDP transmits to the next switch. At that point, the RVC can be mapped to an input RVC that has a CEP waiting to establish a circuit. The ''active'' bit indicates whether the RVC has been used by any recently transmitted packet. The switch uses the ''active'' bit to choose a victim. In particular, the ''clock'' algorithm [10], which approximates LRU, can be used to select a victim.

The IMT and OMT are accessed for each packet and must therefore be stored in high-speed low density on-chip memory (SRAM). Other information is accessed only when packets are diverted or when circuits are manipulated and can therefore be stored in lower speed dense memory, most likely DRAM, which could be embedded on-chip [29, 5]. This memory will contain tables for routing CEPs, tables that store DVC identification and routing information (these tables are updated upon DVC establishment and read when DVCs are rerouted), and tables used to implement the H_i control buffers. The storage requirements for this lower speed dense memory are derived in the following subsection which focuses on

the handling of control packets.

6.2. Control Buffer Implementation

At each input port, the algorithm in Section 4 requires the ability to store up to four control packets (CEPs, CDPs) for one RVC (in H_i and T_*) and up to two control packets for the rest of the RVCs (in H_j). The cost of this storage can be minimized by splitting the storage between high-speed SRAM and slower denser memory such as DRAM, and by using efficient encoding to minimize the memory needed for CDPs. The storage for each RVC is split into two components: CQ_i^S — a frequently-accessed component stored as part of the IMT in dedicated SRAM at each port, and CQ_i^{CEP} — an infrequently-accessed component stored in the lower speed dense memory available on the switch. The CQ_i^S component consists of a compact representation of the state of the logical queue of control packets for input RVC i. From Table 1, the logical queue of control packets for an RVC i can at any time be in one of eight states: empty, CEP, CDP, CEP CDP, CDP CEP, CEP CDP CEP, CDP CEP CDP, CDP CEP CDP CEP. The CQ_i^S field of the IMT specifies the current state as a 3-bit value. Since CDPs carry no information other than the RVC value, the 3-bit state encoding represents all the CDPs in the queue.

The infrequently-accessed CQ_i^{CEP} component is maintained in the Circuit Information Table (CIT) (Figure 6) which is implemented using dense memory such as embedded DRAM. For each input RVC, there is an entry in the table with space to record the information of two CEPs (DVC source, destination, and sequencing information). The first set of fields (SRC_1, DST_1, and seq_1) reflects the current status of the DVC and is set from the contents of the CEP that established the current DVC. The second set of fields (SRC_2, DST_2, and seq_2), if valid, corresponds to the DVC that will replace the current DVC. This second set of fields is the storage of CQ_i^{CEP} — the CEP that is part of H_i (Section 4.3). Storage for information from two CEPs is needed since the switch may contain data packets on the previous DVC when the CEP for a new DVC arrives. If one or more of these data packets needs to be diverted, the information for the previous DVC will still be needed.

Finally, a dedicated buffer for each input port provides storage for the one CEP that is part of T_*.

Performance can be optimized by adding a small, fast write buffer for arriving CEPs. The switch can access the buffer as a cache to forward CEPs quickly without incurring the DRAM access time for writing to the CIT. A write buffer only benefits CEPs that arrive to empty H_i queues. This should be the common case since control packets present only a light load with reasonable traffic patterns.

The switch must transmit the control packets and data packets in a logical queue in FIFO order, even though the packets are stored physically at the switch in various memory buffers (i.e., the primary

buffer for data packets, and the control buffers described above for control packets). Each buffer preserves the partial order of the packets it stores from the logical queue. The partial order of mapped data packets of a logical queue is preserved because the primary buffer maintains in FIFO order its data packets that are destined to a single output port. The partial order of control packets in the logical queue is recorded and maintained by using the IMT CQ_i^S field. Although these buffers preserve the correct partial order of transmission, a mechanism is needed to control the total interleaved order of transmissions from these buffers.

The total order of a logical queue is preserved by transmitting all of its mapped data packets before any of its control packets. The mapped data packets must precede the control packets because a CDP that precedes a mapped data packet would delete the current RVC mapping, and a preceding CEP would establish a new mapping before the data packet is transmitted. Transmission of mapped data packets ahead of control packets is enforced through the use of the ‘‘seqctl’’ field in the IMT (Figure 5). The seqctl field has initial value zero and records the number of mapped data packets that are present in the logical queue for an RVC. Control packets in the logical queue are allowed to transmit only when the value in ‘‘seqctl’’ is zero. Mapped data packets can transmit any time. The ‘‘seqctl’’ value is incremented when a mapped data packet arrives at a switch on the RVC or when an unmapped data packet on the RVC becomes mapped as a result of transmitting a CEP. The ‘‘seqctl’’ field is decremented when a data packet that previously arrived on the RVC is either transmitted to the next switch along its circuit or is diverted.

6.3. Sequencing and Diversion

As explained in Section 4, to support FIFO delivery, one sequence number per input RVC is maintained in the IMT. The IMT entry ‘‘seq’’ (sequence number) field (Figure 5) is incremented whenever a packet is transmitted whose header has no sequence number field (Figure 1). The index used to access the IMT entry is the value of the input RVC on which the packet arrived at the switch. As discussed in Section 6.1, this value is saved in the main data packet buffer N_x. Whenever a packet with a sequence number is transmitted from the primary buffer, that sequence number replaces the value in the IMT ‘‘seq’’ field.

After a data packet is diverted, the next data packet that is transmitted normally on the same circuit is stamped with a sequence number (Section 4.1). The IMT entry’s single-bit ‘‘OSM’’ (Out of Sequence Mode) field is used as a flag to determine whether a data packet that is transmitted normally should be stamped. The flag is set when a data packet is diverted. Before a data packet begins transmission to the

next switch, the OSM flag for the packet's input RVC is read. If the flag is set, then it is cleared and the sequence number is added to the data packet header on-the-fly as it is transmitted to the next switch. With virtual cut-through, a packet is transmitted only if the next switch has sufficient buffer space for a maximum size packet. Hence, lengthening the packet, following diversion, as it is transmitted, will not cause the buffer at the next switch to overflow.

Packet diversion requires forwarding a timed-out packet to an output other than the one associated with the logical queue storing the packet. Hence, when a packet times out, its request from the crossbar arbiter is modified to include access to the switch output that was its original destination as well as access to the switch output(s) on the diversion path(s). If access to the diversion network is granted first, the packet's RVC field is changed to indicate use of the diversion BVC, and DVC information from the IMT and CIT is added to the header.

6.4. Hardware Costs for Example Switch Configuration

To illustrate the costs of the hardware required to implement the DVC algorithms, we consider the hardware additions for an example switch configuration. We take the example of an 8×8 switch which supports 32 RVCs per port. We assume a maximum packet data payload of 32 bytes with packet header of 1 byte for regular data packets or 8 bytes for diverted packets and CEPs.

With these values, the packet buffers D_X and T_*^X respectively use 40 bytes and 8 bytes of fast SRAM per input port. The IMT with 32 entries at each input port requires 148 bytes of SRAM, and the OMT at each output port contributes 12 bytes. Thus, for DVCs the switch needs an additional 208 bytes of SRAM per port, or 1664 bytes of SRAM for the entire 8×8 switch.

As described in Section 6.2, the Circuit Information Table records source ID, destination ID, and sequencing information for two CEPs for each RVC. To minimize silicon area, this infrequently accessed table can be maintained in dense memory such as embedded DRAM instead of faster SRAM. In our example switch configuration, the Circuit Information Table consists of 320 bytes of DRAM per port for a total of 2560 bytes for an 8×8 switch.

Compared to packet switching, DVCs require additional control logic to execute the infrequent, complex circuit manipulation operations. As described earlier, these operations can be performed by specialized logic or can be implemented using a small, fully programmable processor core that occupies a small fraction of the silicon area on a typical chip.

7. Performance Evaluation

A key advantage of DVCs is the potential for reducing overall network contention by establishing circuits or rerouting existing circuits onto low latency paths. Any routing algorithm can be used to select the circuit paths, since the underlying DVC mechanism does not constrain the choice of routes. In particular, high-level global knowledge of the expected traffic patterns can be used to minimize contention in a DVC network instead of relying on simple algorithmic routing. In most systems, traffic patterns change dynamically, and circuits require time to adjust their paths to compensate. DVCs are intended to support adaptation for traffic patterns that do not change so frequently that the costs of circuit establishment and teardown become prohibitive. The performance of a system with DVCs with shifting traffic patterns will highly depend on the particular routing and rerouting algorithms that are used. Since the focus of this paper is on the DVC mechanism itself, which could be used with a variety of routing algorithms, we leave the question of performance under shifting traffic patterns as a potential topic for future investigations.

For a first-order evaluation of the performance *potential* for DVCs with adaptive routing, we consider simpler limit cases with stable traffic patterns and circuit placements. This allows us to identify an upper limit on the performance benefits that can be achieved by allowing full freedom for circuit path selection. Since the DVC mechanism can divert packets from circuit paths in order to avoid deadlocks, we also evaluate the impact of packet diversion in the presence of good circuit placement. Packet diversion has the potential to help performance by providing multi-path routing, but it may also hurt performance by diverting packets away from good paths and onto congested paths. A packet is diverted if it is blocked at the head of the queue at a switch and the timeout expires. A short timeout interval may cause packets to be diverted unnecessarily, whereas a long timeout interval may increase the time required to resolve deadlocks. To evaluate the sensitivity of performance to the timeout parameter, our evaluation includes simulations for a variety of timeout values.

As we describe below, our experimental results demonstrate that good path selection, along with packet diversion, has significant advantages compared to fixed packet routing, providing higher throughput and better fairness. In most cases this approach can achieve close to the theoretical maximum throughput. Thus, there is limited potential to achieve further improvements by using an alternative approach like fully-adaptive packet switching, and such benefits may also disappear if the packets have small data payloads leading to significant waste of link bandwidth consumed by the larger packet headers required with packet switching instead of DVCs. The results also show that packet diversion is effective for removing deadlocks and is rare enough that the benefits of good path selection are preserved.

However, in most of the experiments the use of multiple paths because of packet diversion does not improve performance significantly. Finally, our results show that although the optimal setting of the timeout parameter is traffic pattern-dependent, in most cases throughput and latency are reasonably insensitive to the timeout parameter value. While the optimal timeout value may depend on packet size, this is not likely to be a large obstacle for DVC systems, which favor the use of small packets. DVC networks use virtual cut-through which limits packet size (packets must fit in the switching element buffer). This is in contrast to wormhole networks which may require alternative deadlock detection mechanisms to avoid the sensitivity of timeouts to message length [32]. Also, in parallel applications most payloads are likely to be small (acknowledgments, synchronization, cache blocks). Furthermore, in order to limit the amount of data to be retransmitted when an error occurs there is reason to keep the packet size small. For these reasons, packet sizes will vary over a narrow range, allowing the selection of a timeout value that is close to optimal.

7.1. Experimental Setup

We consider Uniform, Transpose and Bit-Reversal traffic patterns. In all cases, we precompute the paths used by DVCs in order to simulate ideal conditions where circuits have settled into a steady-state, low contention configuration. We compare the performance of the resulting configuration against a packet switched network using Dimension Order Routing (DOR), which is known to perform well for Uniform and poorly for Transpose and Bit-Reversal patterns [22]. To emulate the behavior of a sophisticated routing algorithm that uses global knowledge of the network to select good paths for circuits, in the experiments the routes for DVCs are precomputed using a simple heuristic Bellman-Ford minimum cost path algorithm. The cost of each link is set to an estimate of the delay experienced by packets waiting to use the link. The delay is estimated by modeling the link as a Geo(N)/D/1 queueing system fed by packet arrivals from all DVCs whose routes include the link. Details of this route precomputation procedure are described elsewhere [37].

For our simulation experiments, packet size is 32 phits, and switches have DAMQ primary input buffers. For DVC simulations, there is also a diversion buffer of capacity 32 phits. The results for DOR and routed DVCs are shown for equal total input buffer capacity, which for DVCs is the sum of the primary input buffer capacity plus 32 phits for the diversion buffer. The interval between packet creations has a geometric distribution. At each switch, the crossbar arbitration policy gives priority to the packet that has resided longest at the switch. The performance metrics are the average latency, the average node throughput and the normalized throughput. Packet latency is defined as the number of cycles from when the first byte of a packet is generated at a sender to when it leaves the network.

Normalized throughput expresses throughput as a fraction of the network bisection bandwidth.

7.2. Transpose Traffic Pattern

The transpose traffic pattern sends all packets across the diagonal of the mesh; packets from the source at row i column j are sent to the destination at row j column i. Nodes along the diagonal only forward packets; they do not produce or consume packets. For this traffic pattern, Figure 7 shows the latency versus throughput for an 8×8 mesh, with input buffer capacity 64, 96, and 288 phits.

These results show that for all levels of buffer capacity, the maximum throughput achieved with DVCs is about twice that achieved with DOR. With DOR, only the horizontal incoming links of the switches along the diagonal are utilized. With routed DVCs, both the vertical and horizontal incoming links of the diagonal are utilized with approximately equal traffic loads assigned to each link. The results also show that the impact of increasing the buffer capacity is higher latency, not higher throughput. Throughput does not increase for either DOR or routed DVCs because it is limited by the bandwidth of saturated links along the mesh diagonal. Finally, the results show that using DVCs with long timeouts for packet diversion results in higher maximum throughput than using short timeouts. Short timeouts increase the frequency of packet diversions, which for the transpose traffic pattern occur before the packet crosses the diagonal, the congested point in the network. Since diverted packets use DOR, they can only use the horizontal incoming links of switches on the diagonal. Hence packet diversions shift traffic from the vertical to the horizontal incoming links at the diagonal. These traffic imbalances reduce performance, thus in this case longer timeouts which minimize packet diversions are better.

Figure 8 shows the fraction of traffic diverted versus normalized throughput for the DVC network with input buffer capacity of 96 phits. For low and medium network loads, as the load increases, the fraction of diverted packets increases. However, past a certain point, the fraction of diverted packets decreases as the load increases. The reason for this is that at these high loads the increase in network throughput is mostly for the subset of circuits using low-contention routes. Other circuits and their diversion paths are saturated and their throughput does not increase as the applied load increases. For the low-contention circuits no diversion occurs so more packets get through the network without a corresponding increase in the number of diverted packets.

The performance of a distributed application is often limited by the performance of its slowest member rather than by the aggregate throughput available to the application. For example, an application whose nodes communicate via the transpose traffic pattern may occasionally need to synchronize to ensure that all sending nodes are in a known state. If some flow is particularly slow, progress of nodes

associated with this flow will be impeded and the progress of all other nodes will be throttled upon synchronization. Hence, it is useful to evaluate the fairness of the system by comparing the throughputs achieved by individual senders.

Figure 9 shows the raw (not normalized) throughput achieved by each source node in the 8×8 mesh using the transpose traffic pattern. The throughputs from each sender are displayed, sorted to be monotonic (the first eight sources are along the diagonal and do not generate packets). Throughputs for routed DVCs and DOR are displayed as separate curves. Since fairness in a network decreases as the load increases, comparison of the fairness of the two policies should be done at the same average node throughput. In Figure 9, average node throughput for DOR is at its maximum, 0.233, and average node throughput for routed DVCs is 0.242. Since unfairness increases with average node throughput, the result in Figure 9 is biased slightly in favor of DOR, yet the routed DVCs achieve far greater uniformity of sender throughput than does DOR. As we increase applied load further, the routed DVCs policy also becomes unfair, but only at much higher levels of average node throughput than can be achieved with DOR. This is demonstrated in Figure 10, which shows throughput fairness at saturation, in which each source constantly tries to inject packets.

7.3. Bit-Reversal Traffic Pattern

The bit-reversal traffic pattern sends messages from each source $x_{n-1}x_{n-2}\cdots x_0 y_{n-1}y_{n-2}\cdots y_0$ to destination $y_0 y_1 \cdots y_{n-1} x_0 x_1 \cdots x_{n-1}$. Figure 11 shows latency versus throughput on an 8×8 mesh, for total input buffer capacity 64, 96 and 288 phits. The reported throughput is normalized to the bisection bandwidth, the upper bound on throughput for the bit-reversal traffic pattern.

The results for bit-reversal show that, as with transpose traffic, routed DVCs significantly outperforms DOR and there is no advantage to increasing the buffer size. Unlike with transpose traffic the results with bit-reversal traffic are nearly independent of the diversion timeout value. With bit-reversal traffic, diverted packets do not necessarily follow poor paths that increase congestion.

Figure 12 shows the fraction of traffic diverted versus throughput with total input buffer capacity 96 phits. These results show, as with transpose, that increasing timeout values greatly reduce the fraction of traffic diverted. Since diverted packets are handled less efficiently than packets on DVCs, these results and the transpose traffic results indicate that long timeout values should be used.

7.4. Uniform Traffic Pattern

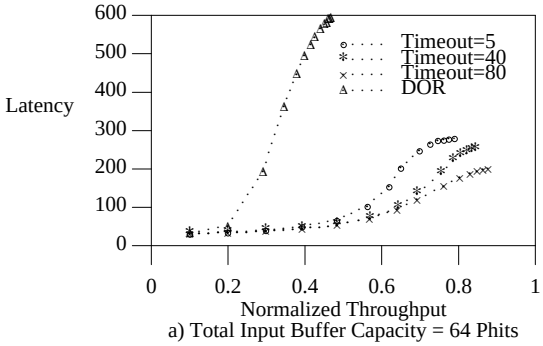
For uniform traffic in a square mesh, DOR distributes load evenly across the links that comprise the network bisection and thus should perform well. In contrast, some adaptive routing schemes tend to steer more traffic toward the center of the network, causing congestion [30].

Figure 13 shows latency versus throughput for uniform traffic with total input buffer capacity 64 and 288 phits. The results show that the performance of routed DVCs is close to that of DOR. Unlike transpose and bit-reversal traffic, with uniform traffic the use of larger buffers improves performance. Increasing the buffer capacity increases the number of flows that can be represented at any instant by the packets that are present at a switch. Larger buffers are therefore more helpful for uniform traffic with $O(N^2)$ flows than for the previous traffic patterns which have only $O(N)$ flows (one from each source node). Performance also improves with the use of smaller timeout values which effectively increase the useful buffer capacity by enabling more packets to take advantage of the 32 phit diversion buffers. With large buffers (288 phits), routed DVCs and DOR have nearly identical performance.

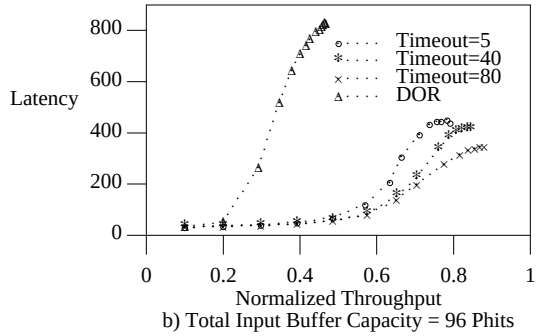
For routed DVCs with short timeouts, as the applied load increases beyond saturation the network throughput decreases slightly. This may occur because congestion in the primary virtual network causes a larger number of packets to enter the diversion virtual network which has limited buffering and therefore limited throughput.

7.5. The Impact of Network Size

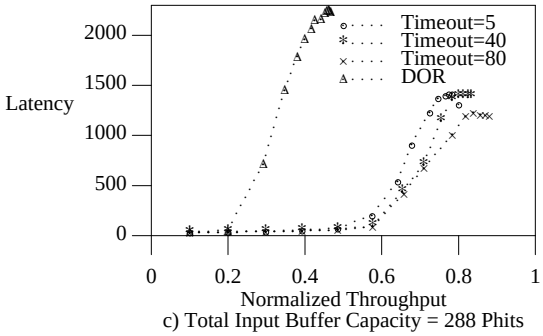
For a larger mesh network of size 16×16 and the same traffic patterns as used previously, Figures 14, 15 and 16 show latency versus throughput, and Figures 17, 18 and 19 show the fraction of traffic diverted versus normalized throughput. For non-uniform traffic, the results show that routed DVCs significantly outperform DOR, but the performance difference is smaller than on the 8×8 mesh. With the larger network, packets travel longer distances, and there are more opportunities for delays and deadlocks. Hence, the fraction of packets diverted tends to be larger than on the 8×8 mesh, resulting in more of the traffic facing congestion as with DOR.



a) Total Input Buffer Capacity = 64 Phits



b) Total Input Buffer Capacity = 96 Phits



c) Total Input Buffer Capacity = 288 Phits

Figure 7: Transpose traffic: Latency vs. Normalized Throughput.

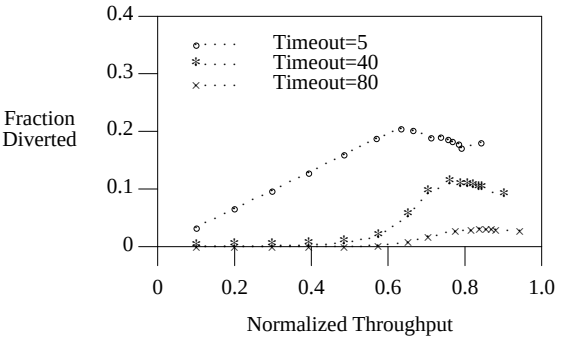


Figure 8: Transpose traffic: Fraction of Traffic Diverted vs. Normalized Throughput. Total input buffer capacity = 96 phits.

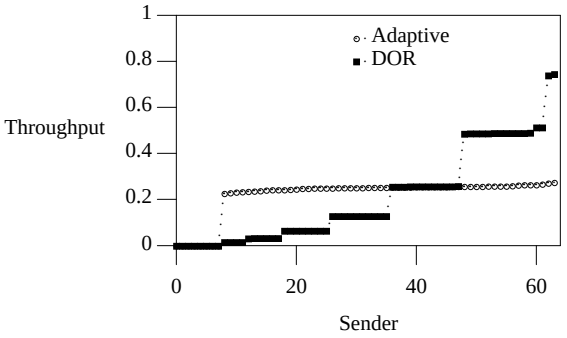


Figure 9: Transpose traffic: Throughput fairness. Throughput vs. sender, sorted. Total input buffer capacity = 64 phits. Average node throughput = 0.241 for DOR (48.2% normalized), 0.242 for routed DVCs (48.4% normalized).

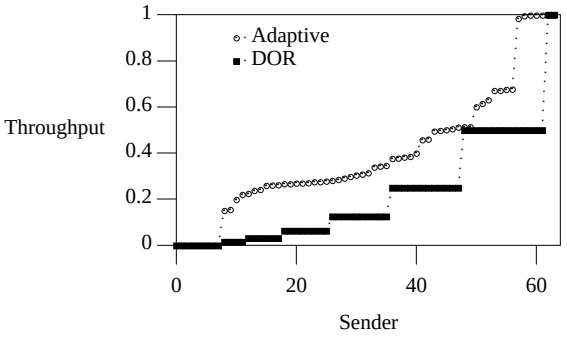
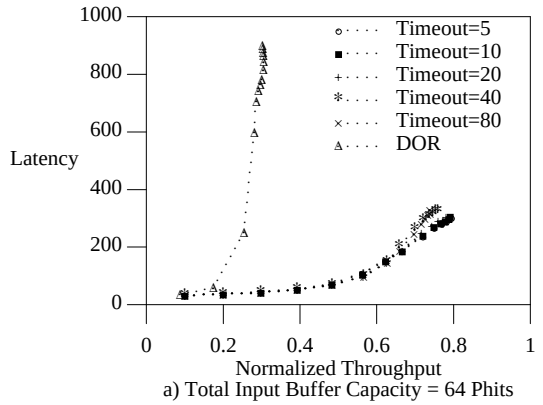
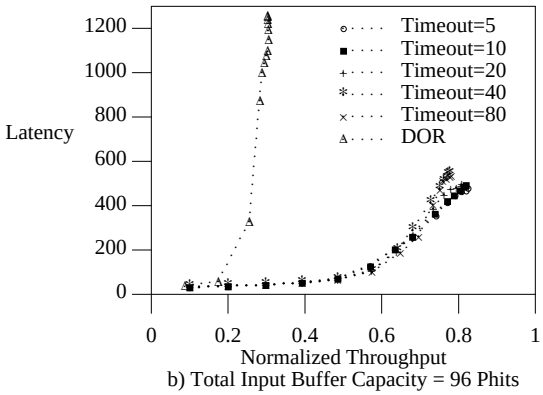


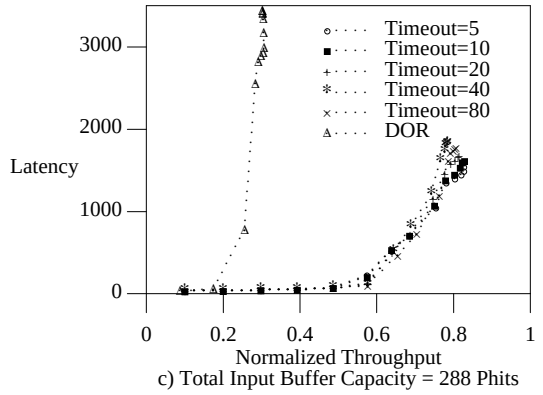
Figure 10: Transpose traffic: Throughput fairness at saturation. Throughput vs. Sender, sorted. Total input buffer capacity = 288 phits. Average node throughput = 0.24 for DOR (48% normalized), 0.47 for routed DVCs (94% normalized)



a) Total Input Buffer Capacity = 64 Phits



b) Total Input Buffer Capacity = 96 Phits



c) Total Input Buffer Capacity = 288 Phits

Figure 11: Bit-Reversal traffic: Latency vs. Normalized Throughput.

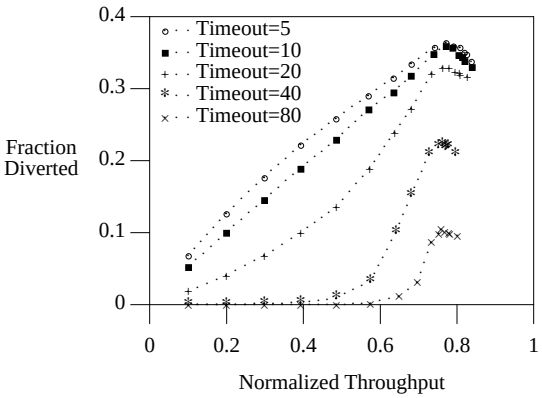
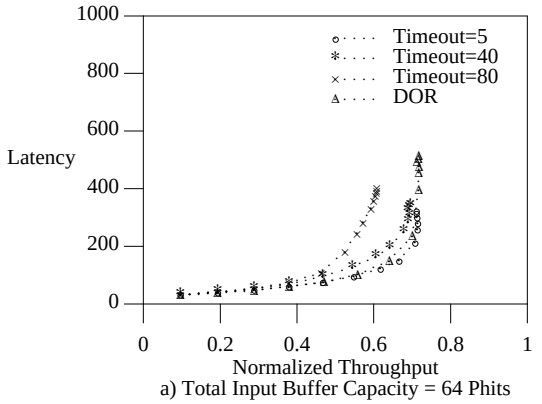
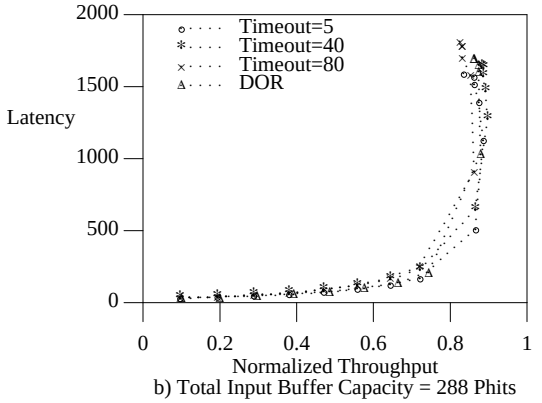


Figure 12: Bit-Reversal traffic: Fraction of Traffic Diverted vs. Normalized Throughput. Total input buffer capacity = 96 phits.



a) Total Input Buffer Capacity = 64 Phits



b) Total Input Buffer Capacity = 288 Phits

Figure 13: Uniform traffic: Latency vs. Normalized Throughput.

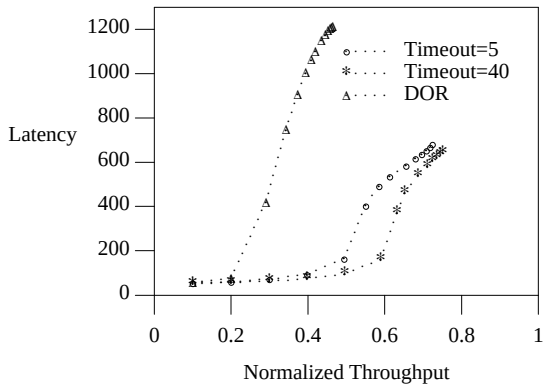


Figure 14: 16×16 Transpose traffic: Latency vs. Normalized Throughput. Total input buffer capacity = 64 phits.

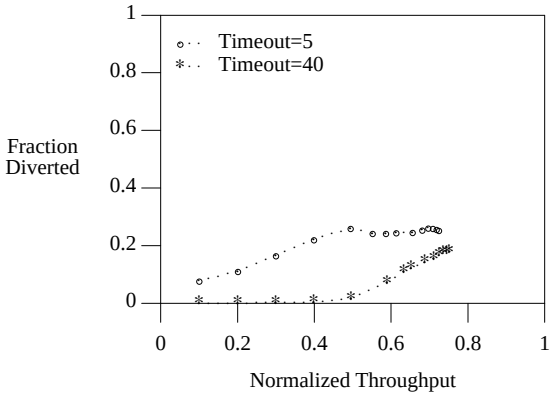


Figure 17: 16×16 Transpose traffic: Fraction of Traffic Diverted vs. Normalized Throughput. Total input buffer capacity = 64 phits.

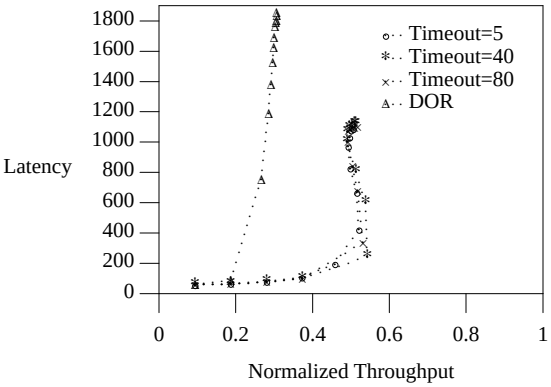


Figure 15: 16×16 Bit-Reversal traffic: Latency vs. Normalized Throughput. Total input buffer capacity = 64 phits.

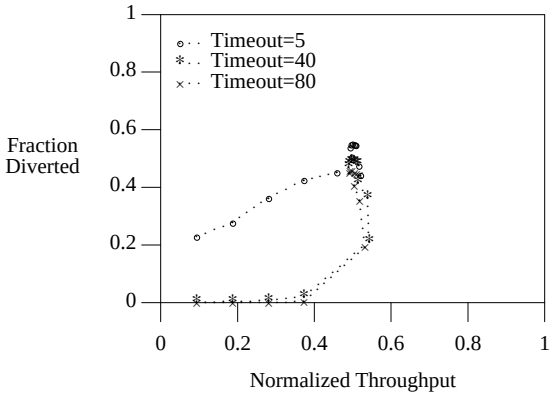


Figure 18: 16×16 Bit-Reversal traffic: Fraction of Traffic Diverted vs. Normalized Throughput. Total buffer capacity = 64 phits.

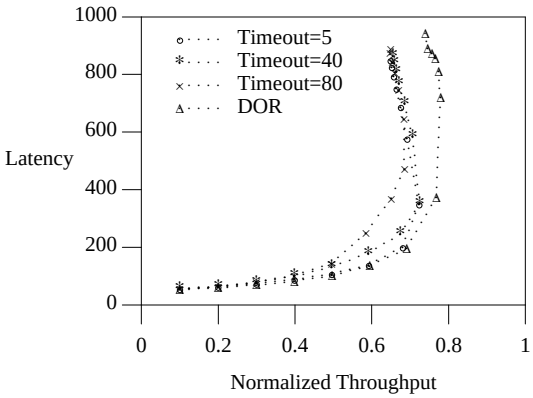


Figure 16: 16×16 Uniform traffic: Latency vs. Normalized Throughput. Total input buffer capacity = 64 phits.

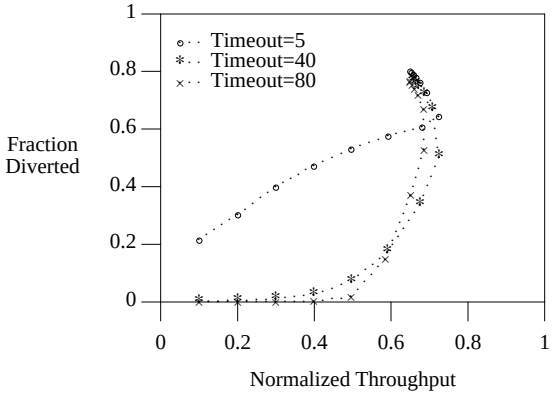


Figure 19: 16×16 Uniform traffic: Fraction of Traffic Diverted vs. Normalized Throughput. Total input buffer capacity = 64 phits.

8. Related Work

In this section we describe related work on connection-based routing for multicomputers or high-end cluster networks, and existing techniques for virtual circuits to support adaptive routing and circuit rerouting. We compare these approaches to our proposed Dynamic Virtual Circuits.

An approach to connection-based routing that combines static virtual circuits (for traffic exhibiting locality) and conventional wormhole routing (for other traffic) was proposed by Dao, Yalamanchili, and Duato [13]. In this proposal, a source node tries to establish a static virtual circuit by injecting into the network a circuit establishment probe. The probe is adaptively routed toward the destination; along the way the probe may backtrack as it searches for a path with free resources on each link. If the probe backtracks to the source node, the source either re-injects the probe for another try or gives up on establishing a circuit, in which case it uses conventional wormhole routing for traffic it sends to the destination. Existing circuits are not torn down to free resources for new circuits. Nor are circuits rerouted to adjust to congestion or faults.

A different approach that relaxes the static nature of virtual circuits was proposed by Hsu and Banerjee [23]. They proposed adding logic to support on each switch a small number of virtual circuits, called *cached circuits*. A cached circuit relaxes the static restrictions of virtual circuits; a cached circuit may be torn down in order to free up resources at an intermediate switch. The resources may be needed, for example, to establish a new cached circuit. The intermediate switch selects an existing cached circuit to be a victim, and it sends a request to the source node of the victim circuit to tear it down. The packets in transit from the victim's source must progress past the intermediate switch before the resources held by the victim circuit can be released. Therefore, new circuit establishment may be blocked for an extended period while packets on the victim circuit are being flushed out. In addition, the packets being flushed out are forced to take the existing path of the victim circuit, which may no longer be a desirable path because of the possible prior onset of congestion or faults.

Virtual circuits are used in Asynchronous Transfer Mode (ATM) [14]. ATM supports two types of connections: virtual circuits (VCs) and virtual paths (VPs). A VC is composed of a sequence of VPs from source to destination. Each VP can support 2^{16} VCs. A VC can be rerouted to improve quality of service via a round trip exchange of control messages between the source and destination [9]. A VP that is used by many VCs can be rerouted when a link on its route fails. Rerouting a VP is transparent to the VCs that use it. A VP can be rerouted onto a backup route that is either pre-defined or is selected after a failure is detected [25, 21]. VP rerouting is accomplished through a round trip exchange of control messages on the backup path between the VP endpoints [25]. Alternatives to end-to-end VP rerouting include rerouting

only the portion of the VP between two switches that are adjacent to the failure, and rerouting the portion of a VP from a switch that is upstream from the failure and the VP destination. These strategies differ in the time required to reroute a VP and in the spare bandwidth that is needed to guarantee that all VPs can be rerouted and meet their bandwidth requirements after any single failure [28, 3].

Our Dynamic Virtual Circuits (DVCs) [34] proposal differs from the above proposals by allowing virtual channel resources to be quickly reallocated through local operations at a switch, avoiding the long delays of schemes that require interactions with faraway nodes before releasing local resources. Resource reallocation avoids blocking circuit establishment, and it enables adaptive circuit rerouting.

9. Conclusion

In this paper, we presented the algorithms and hardware/firmware architecture for Dynamic Virtual Circuits (DVCs), a novel technique that enables multicomputer and cluster interconnection networks to combine the benefits of connection-based routing and adaptive routing. In particular, DVCs reduce link bandwidth overheads and packet processing delays at switches compared to pure packet switching networks. A Dynamic Virtual Circuit can be established on any path from source to destination without the possibility of deadlocks involving packet buffer resources or virtual circuit manipulation operations. Unlike prior approaches, DVCs can be rerouted dynamically through the use of fast operations at any switch along a virtual circuit's path without requiring costly delays for coordination with remote nodes. It is practical to implement the DVC mechanism, which has only modest hardware requirements and applies to networks with arbitrary topologies. Emerging interconnect standards such as InfiniBand could be extended to support DVCs, which fit well with the semantics of InfiniBand end-to-end queue pairs connections and could be used to improve transmission efficiency for InfiniBand network fabrics.

To guarantee deadlock-freedom in DVC networks, our solution decouples data packet routing and circuit manipulation operations. To enable data packets to use unconstrained virtual circuit paths, we leverage the existing approach from packet switching networks of allowing data packets that encounter buffer dependency cycles to transition to a dependency cycle-free virtual network, the *diversion network*. To avoid deadlocks involving circuit manipulation operations, we present a new approach, based on an analysis of control packet arrival sequences, which guarantees that control packets cannot experience blocking across switches. We use an elaboration of the state space of a switch to develop correctness arguments showing that the DVC algorithms ensure the network is deadlock-free and that data packets are delivered to correct destinations. Our performance evaluation results show that with virtual circuits, global routing optimization is possible and provides performance superior to fixed routing. Furthermore,

the results show that the use of deadlock-free escape paths is sufficiently infrequent to preserve the bandwidth efficiencies of the DVC mechanism.

For future investigation, an interesting and important question is how DVCs behave with shifting traffic patterns. In particular, there are many alternatives for choosing when and how to reroute existing circuits. Other future investigations could focus on extending DVCs to provide advanced functionalities including improved fault tolerance and multicast capabilities.

REFERENCES

1. *InfiniBand Architecture Specification, Volume 1, Release 1.0,* InfiniBand Trade Organization www.infinibandta.org (October 2000).
2. *TOP500 Supercomputer Sites,* www.top500.org (November 2005).
3. J. Anderson, B. T. Doshi, S. Dravida, and P. Harshavardhana, ‘‘Fast Restoration of ATM Networks,’’ *IEEE Journal on Selected Areas in Communications* **12**(1), pp. 128-138 (January 1994).
4. K. V. Anjan and T. M. Pinkston, ‘‘An efficient, fully adaptive deadlock recovery scheme: DISHA,’’ *Proceedings 22nd Annual International Symposium on Computer Architecture*, Santa Margherita Ligure, Italy, pp. 201-10 (22-24 June 1995).
5. J. E. Barth, Jr., J. H. Dreibelbis, E. A. Nelson, D. L. Anand, G. Pomichter, P. Jakobsen, M. R. Nelms, J. Leach, and G. M. Belansek, ‘‘Embedded DRAM design and architecture for the IBM 0.11-um ASIC offering,’’ *IBM Journal of Research and Development* **46**(6) (November 2002).
6. N. J. Boden, D. Cohen, R. E. Felderman, A. E. Kulawik, C. L. Seitz, J. N. Seizovic, and W.-K. Su, ‘‘Myrinet -- a Gigabit-per-Second Local Area Network,’’ *IEEE Micro* **15**(1), pp. 29-36 (February 1995).
7. K. Bolding, M. Fulgham, and L. Snyder, ‘‘The case for chaotic adaptive routing,’’ *IEEE Transactions on Computers* **46**(12), pp. 1281-1292 (December 1997).
8. S. Borkar, R. Cohn, G. Cox, T. Gross, H. T. Kung, M. Lam, M. Levine, B. Moore, W. Moore, C. Peterson, J. Susman, J. Sutton, J. Urbanski, and J. Webb, ‘‘Supporting Systolic and Memory Communication in iWarp,’’ *17th Annual International Symposium on Computer Architecture*, Seattle, Washington, pp. 70-81 (May 28-31, 1990).
9. R. Cohen, ‘‘Smooth Intentional Rerouting and its Applications in ATM Networks,’’ *IEEE INFOCOM’94*, Toronto, pp. 1490-1497 (June 1994).
10. F. J. Corbato, ‘‘A Paging Experiment with the MULTICS System,’’ Project MAC Memo MAC-M-384, MIT, Cambridge, MA (July 1968).
11. W. J. Dally and C. L. Seitz, ‘‘Deadlock-Free Message Routing in Multiprocessor Interconnection Networks,’’ *IEEE Transactions on Computers* **C-36**(5), pp. 547-553 (May 1987).
12. W. J. Dally and H. Aoki, ‘‘Deadlock-Free Adaptive Routing in Multicomputer Networks Using Virtual Channels,’’ *IEEE Transactions on Parallel and Distributed Systems* **4**(4), pp. 466-475 (April 1993).
13. B. V. Dao, S. Yalamanchili, and J. Duato, ‘‘Architectural support for reducing communication overhead in multiprocessor interconnection networks,’’ *Third International Symposium on High-Performance Computer Architecture*, San Antonio, TX, pp. 343-52 (1-5 Feb. 1997).
14. M. De Prycker, *Asynchronous Transfer Mode: Solution for Broadband ISDN (3rd edition),* Prentice Hall, New York (1996).

15. J. Duato, ‘‘A Necessary And Sufficient Condition For Deadlock-Free Adaptive Routing In Wormhole Networks.,’’ *IEEE Transactions on Parallel and Distributed Systems* **6**(10), pp. 1055-1067 (October 1995).

16. J. Duato, ‘‘A necessary and sufficient condition for deadlock-free routing in cut-through and store-and-forward networks,’’ *IEEE Transactions on Parallel and Distributed Systems* **7**(8), pp. 841-54. (August 1996).

17. J. Duato, ‘‘Deadlock avoidance and adaptive routing in interconnection networks,’’ *Proceedings of the Sixth Euromicro Workshop on Parallel and Distributed Processing*, Madrid, Spain, pp. 359-364 (21-23 Jan. 1998).

18. C. Eddington, ‘‘InfiniBridge: An InfiniBand Channel Adapter with Integrated Switch,’’ *IEEE Micro* **22**(2), pp. 48-56 (Mar/Apr 2002).

19. J. Flinch, M. P. Malumbres, P. Lopez, and J. Duato, ‘‘Performance Evaluation of a New Routing Strategy for Irregular Networks with Source Routing,’’ *14th Int’l Conf on Supercomputing*, pp. 34-43 (2000).

20. M. Galles, ‘‘Spider: A High-Speed Network Interconnect,’’ *IEEE Micro* **17**(1), pp. 34-39 (January/February 1997).

21. A. Gersht and A. Shulman, ‘‘Architecture for Restorable Call Allocation and Fast VP Restoration in Mesh ATM Networks,’’ *IEEE Transactions on Communications* **47**(3), pp. 397-403 (March 1999).

22. C. J. Glass and L. M. Ni, ‘‘The Turn Model for Adaptive Routing,’’ *Journal of the Association for Computing Machinery* **41**(5), pp. 874-902 (September 1994).

23. J.-M. Hsu and P. Banerjee, ‘‘Hardware Support for Message Routing in a Distributed Memory Multicomputer,’’ *1990 International Conference on Parallel Processing*, St. Charles, IL (August 1990).

24. 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), pp. 451-464 (July 1992).

25. R. Kawamura and H. Ohta, ‘‘Architectures for ATM Network Survivability and Their Field Deployment,’’ *IEEE Communications Magazine* **37**(8), pp. 88-94 (August 1999).

26. J. H. Kim, Z. Liu, and A. A. Chien, ‘‘Compressionless Routing: A Framework For Adaptive and Fault-Tolerant Routing,’’ *IEEE Transactions on Parallel and Distributed Systems* **8**(3), pp. 229-244 (March 1997).

27. C. Kuo, M. Weidner, T. Toms, H. Choe, K.-M. Chang, A. Harwood, J. Jelemensky, and P. Smith, ‘‘A 512-kb Flash EEPROM Embedded in a 32-b Microcontroller,’’ *IEEE Journal of Solid-State Circuits* **27**(4) (April 1992).

28. K. Murakami and H. S. Kim, ‘‘Comparative Study on Restoration Schemes of Survivable ATM Networks,’’ *IEEE INFOCOM’97*, Kobe, Japan, pp. 345-352 (April 1997).

29. M. Ohmacht, D. Hoenicke, R. Haring, and A. Gara, ‘‘The eDRAM based L3-Cache of the BlueGene/L Supercomputer Processor Node,’’ *16th Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)* (October 2004).

30. M. J. Pertel, ‘‘A Critique of Adaptive Routing,’’ Computer Science Technical Report 92-06, California Institute of Technology, Pasadena, CA (June 1992).

31. F. Petrini, W.-C. Feng, A. Hoisie, S. Coll, and E. Frachtenberg, ‘‘The Quadrics Network: High-Performance Clustering Technology,’’ *IEEE Micro* **22**(1), pp. 46-57 (February 2002).

32. J.-M. M. Rubio, P. López, and J. Duato, ‘‘FC3D: Flow Control-Based Distributed Deadlock Detection Mechanism for True Fully Adaptive Routing in Wormhole Networks,’’ *IEEE*

Transactions on Parallel and Distributed Systems **14**(8) (August 2003).

33. W. D. Tajibnapis, ‘‘A Correctness Proof of a Topology Information Maintenance Protocol for a Distributed Computer Network,’’ *Communications of the ACM* **20**(7), pp. 477-485 (July 1977).
34. Y. Tamir and Y. F. Turner, ‘‘High-Performance Adaptive Routing in Multicomputers Using Dynamic Virtual Circuits,’’ *6th Distributed Memory Computing Conference*, Portland, OR, pp. 404-411 (April 1991).
35. Y. Tamir and G. L. Frazier, ‘‘Dynamically-Allocated Multi-Queue Buffers for VLSI Communication Switches,’’ *IEEE Transactions on Computers* **41**(6), pp. 725-737 (June 1992).
36. Y. F. Turner and Y. Tamir, ‘‘Connection-Based Adaptive Routing Using Dynamic Virtual Circuits,’’ *International Conference on Parallel and Distributed Computing and Systems*, Las Vegas, NV, pp. 379-384 (October 1998).
37. Y. F. Turner, ‘‘High-Performance Adaptive Routing in Multicomputers Using Dynamic Virtual Circuits,’’ Ph.D. Dissertation, Computer Science Department Technical Report #TR050022, University of California, Los Angeles, CA (July 2005).
38. L. Tymes, ‘‘Routing and flow control in TYMNET,’’ *IEEE Transactions on Communication* **COM-29**, pp. 392-398 (1981).
39. A. K. Venkatramani, T. M. Pinkston, and J. Duato, ‘‘Generalized theory for deadlock-free adaptive wormhole routing and its application to Disha Concurrent,’’ *The 10th International Parallel Processing Symposium*, Honolulu, HI, pp. 815-21 (15-19 April 1996).