2009 18th International Conference on Parallel Architectures and Compilation Techniques

Oblivious Routing in On-Chip Bandwidth-Adaptive Networks

Myong Hyon Cho, Mieszko Lis, Keun Sup Shim, Michel Kinsy, Tina Wen and Srinivas Devadas
Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA
{mhcho, mieszko, ksshim, mkinsy, tinaw, devadas}@mit.edu

***Abstract*—Oblivious routing can be implemented on simple router hardware, but network performance suffers when routes become congested. Adaptive routing attempts to avoid hot spots by re-routing flows, but requires more complex hardware to determine and configure new routing paths. We propose on-chip bandwidth-adaptive networks to mitigate the performance problems of oblivious routing and the complexity issues of adaptive routing.**

In a bandwidth-adaptive network, the bisection bandwidth of a network can adapt to changing network conditions. We describe one implementation of a bandwidth-adaptive network in the form of a two-dimensional mesh with adaptive bidirectional links, where the bandwidth of the link in one direction can be increased at the expense of the other direction. Efficient local intelligence is used to reconfigure each link, and this reconfiguration can be done very rapidly in response to changing traffic demands.

We compare the hardware designs of a unidirectional and bidirectional link and evaluate the performance gains provided by a bandwidth-adaptive network in comparison to a conventional network under uniform and bursty traffic when oblivious routing is used.

I. Introduction

Routers can be generally classified into oblivious and adaptive [1]. In oblivious routing, the path is completely determined by the source and the destination address. Deterministic routing is a subset of oblivious routing, where the same path is always chosen between a source-destination pair. Thanks to its distributed nature where each node can make its routing decisions independent from others, oblivious routing such as dimension-order routing [2] enables simple and fast router designs and is widely adopted in today's on-chip interconnection networks. On the other hand, today's oblivious routing algorithms can have difficulty with certain traffic patterns, especially when bandwidth demands of flows vary with time, because routes are not adjusted for different applications.

In adaptive routing, given a source and a destination address, the path taken by a particular packet is dynamically adjusted depending on, for instance, network congestion. With this dynamic load balancing, adaptive routing can potentially achieve better throughput and latency compared to oblivious routing. However, adaptive routing methods face a difficult challenge in balancing router complexity with the capability to adapt. To achieve the best performance through adaptivity, a router ideally needs global knowledge of the current network status. However, due to router speed and complexity, dynamically obtaining a global and instantaneous view of the network is often impractical. As a result, adaptive routing in practice relies primarily on local knowledge, which limits its effectiveness. If it is necessary to avoid out-of-order packet receipt at the destination, additional mechanisms are required. For example, reorder buffers are required at destination nodes, or packets in transit on the original path of a flow all have to reach the destination before the network is reconfigured and packets are injected into the new path.

We propose *bandwidth-adaptive networks* to mitigate the problems of oblivious routing and avoid the complexity of adaptive routing. In a bandwidth-adaptive network, the bisection bandwidth of a network can adapt to changing network conditions. We describe one implementation of a bandwidth-adaptive network in the form of a two-dimensional mesh with adaptive bidirectional links[1], where the bandwidth of the link in one direction can be increased at the expense of the other direction. Efficient local intelligence is used to appropriately reconfigure each link, and this reconfiguration can be done very rapidly in response to changing traffic demands. Reconfiguration logic compares traffic on either side of a link to determine how to reconfigure each link.

One can think of a bandwidth-adaptive link as a multilane freeway, where a subset or all of the lanes can be set up to carry traffic in either direction. Each lane carries traffic in one particular direction at any point of time, but can be easily switched to carry traffic in the opposite direction depending on the number of cars wishing to travel in each direction. Figure 1 illustrates a scenario where this would be helpful!

We compare the hardware designs of a unidirectional and bidirectional link and argue that the hardware overhead of implementing bidirectionality and reconfiguration is reasonably small. We then evaluate the performance gains provided by a bandwidth-adaptive network in comparison to a conventional network through detailed network simulation of oblivious routing methods under uniform and bursty traffic, and show that the performance gains are significant.

[1]Bidirectional links have been referred to as half-duplex links in router literature.

1089-795X/09 $25.00 © 2009 IEEE
DOI 10.1109/PACT.2009.41

Fig. 1. Motivation for Bandwidth Adaptivity (from www.panoramio.com)

In Section II, we describe a hardware implementation of an adaptive bidirectional link, and compare it with a conventional unidirectional link. In Section III, we describe schemes that determine the configuration of the adaptive link, i.e., decide which direction is preferred and by how much. The frequency of reconfiguration can be varied. Related work is summarized in Section IV. Simulation results comparing oblivious routing on a conventional network against a bandwidth-adaptive network are the subject of Section V. Section VI concludes the paper.

II. Adaptive Bidirectional Link

A. Typical Virtual Channel Router

Although bandwidth adaptivity can be introduced independently of network topology and flow control mechanisms, in the interest of clarity we assume a typical virtual-channel router on a two-dimensional (2-D) mesh network as a baseline.

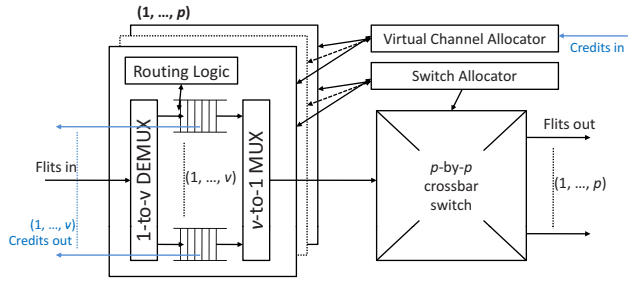

Fig. 2. Typical virtual-channel router architecture with p physical channels and v virtual channels per physical channel.

Figure 2 illustrates a typical virtual-channel router architecture and its operation [3], [4], [5]. As shown in the figure, the datapath of the router consists of buffers and a switch. The input buffers store flits waiting to be forwarded to the next hop; each physical channel often has multiple input buffers, which allows flits to flow as if there were multiple "virtual" channels. When a flit is ready to move, the switch connects an input buffer to an appropriate output channel. To control the datapath, the router also contains three major control modules: a router, a virtual-channel (VC) allocator, and a switch allocator. These control modules determine the next hop, the next virtual channel, and when a switch is available for each packet/flit.

The routing operation comprises four steps: routing (RC), virtual-channel allocation (VA), switch allocation (SA), and switch traversal (ST); these are often implemented as four pipeline stages in modern virtual-channel routers. When a head flit (the first flit of a packet) arrives at an input channel, the router stores the flit in the buffer for the allocated virtual channel and determines the next hop node for the packet (RC stage). Given the next hop, the router then allocates a virtual channel in the next hop (VA stage). The next hop node and virtual channel decision is then used for the remaining flits of the given packet, and the relevant virtual channel is exclusively allocated to that packet until the packet transmission completes. Finally, if the next hop can accept the flit, the flit competes for a switch (SA stage), and moves to the output port (ST stage).

B. Bidirectional Links

In the typical virtual-channel router shown in Figure 2, each output channel is connected to an input buffer in an adjacent router by a unidirectional link; the maximum bandwidth is related to the number of physical wires that constitute the link. In an on-chip 2-D mesh with nearest neighbor connections there will always be two links in close proximity to each other, delivering packets in opposite directions.

We propose to merge the two links between a pair of network nodes into a set of bidirectional links, each of which can be configured to deliver packets in either direction, increasing the bandwidth in one direction at the expense of the other. The links can be are driven from two different sources, with local arbitration logic and tristate buffers ensuring that both do not simultaneously drive the same wire.

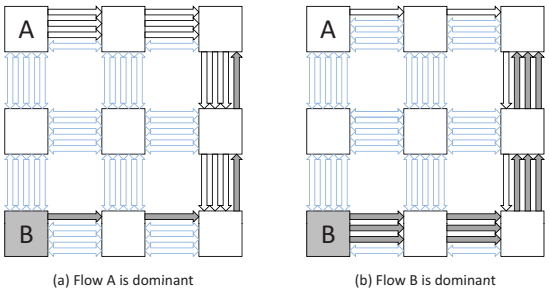

Fig. 3. Adaptivity of a mesh network with bidirectional links

Figure 3 illustrates the adaptivity of a mesh network using bidirectional links. Flow A is generated at the upper left corner and goes to the bottom right corner, while flow B is generated at the bottom left corner and ends at the upper right corner. When one flow becomes dominant, bidirectional links change their directions in order to achieve maximal total throughput. In this way, the network capacity for each flow can be adjusted taking into account flow burstiness without changing routes.

Figure 4 shows a bidirectional link connecting two network nodes (for clarity, only one bidirectional link is shown between the nodes, but multiple bidirectional links can be used to connect the nodes if desired). The bidirectional link can be

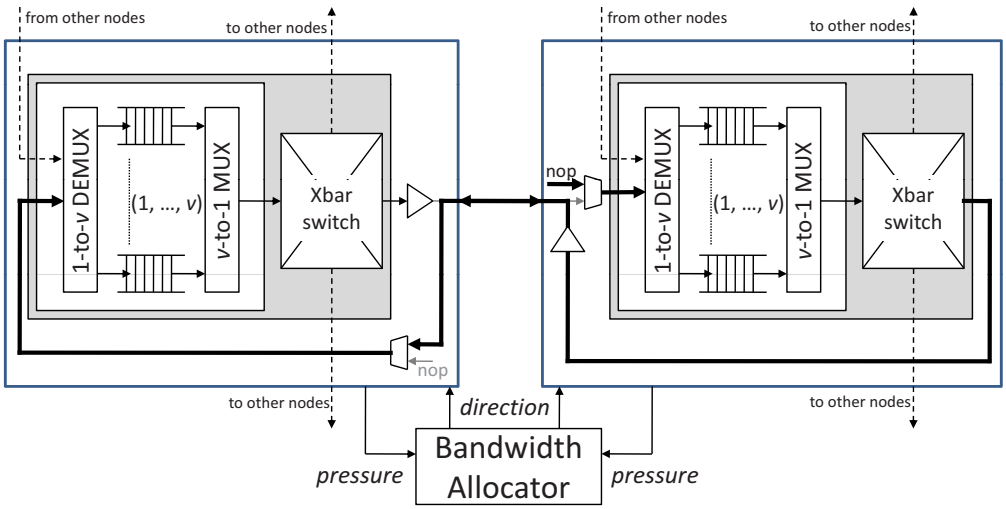

Fig. 4. Connection between two network nodes through a bidirectional link (configured going left).

regarded as a bus with two read ports and two write ports that are interdependent. A bandwidth arbiter governs the direction of a bidirectional link based on *pressure* (see Section III) from each node, a value reflecting how much bandwidth a node requires to send flits to the other node. Bold arrows in Figure 4 illustrate a case when flits are delivered from right to left; a tri-state buffer in the left node prevents the output of its crossbar switch from driving the bidirectional link, and the right node does not receive flits as the input is being multiplexed. If the link is configured to be in the opposite way, only the left node will drive the link and only the right node will receive flits.

Router logic invalidates the input channel at the driving node so that only the other node will read from the link. The switching of tri-state buffers can be done faster than other pipeline stages in the router so that we can change the direction without dead cycles in which no flits can move in any direction. Note that if a dead cycle is required in a particular implementation, we can minimize performance loss by switching directions relatively infrequently. We discuss this tradeoff in Section V.

Long wires in on-chip networks require repeaters. In this paper we are focused on a nearest-neighbor mesh network. As can be seen in Figure 4, only a short section of the link is bidirectional. Tri-state buffers are placed immediately to either side of the bidirectional section. This will be true of links connecting to the top and bottom network nodes as well. Therefore, the bi-directional sections do not need repeaters. If a bi-directional link is used to connect faraway nodes in a different network topology, a pair of repeaters with enable signals will be required in place of a conventional repeater on a unidirectional link.

C. Router Architecture with Bidirectional Links

Figure 5 illustrates a network node with b bidirectional links, where each link has a bandwidth of one flit per router cycle; gray blocks highlight modules modified from the baseline

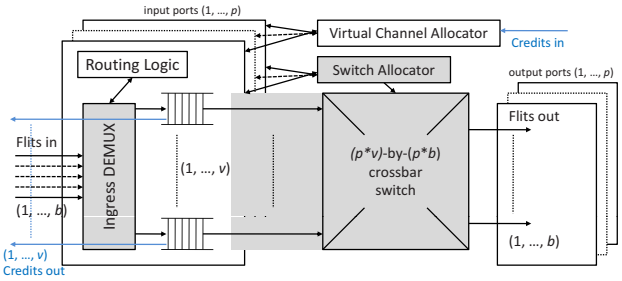

Fig. 5. Network node architecture with u unidirectional links and b bidirectional links between each of p neighbor nodes and itself.

architecture shown in Figure 2. Adjacent nodes are connected via p ports (for the 2-D mesh we consider here, $p = 4$ at most). At each port, b input channels and b output channels share the b bidirectional links via tri-state buffers: if a given link is configured to be ingressive, its input channel is connected to the link while the output channel is disconnected, and vice versa (the output channels are not shown in the figure).

We parametrize architectures with and without bidirectional links by the number of unidirectional links u and the number of bidirectional links b; in this scheme, the conventional router architecture in Figure 2 has $u = 1$ and $b = 0$. We will compare configurations with the same bisection bandwidth. A router with $u = 0$ and $b = 2$ has the same bisection bandwidth as $u = 1$ and $b = 0$. In general, we may have hybrid architectures with some of the links bidirectional and some undirectional (that is, $u > 0$ and $b > 0$). A (u,b) router with bidirectional links will be compared to a conventional router with $u + b/2$ unidirectional links in each direction; this will be denoted as $(u + b/2, 0)$.

We assume, as in typical routers, that at most one flit from each virtual channel can be transferred in a given cycle – if there are v virtual channels in the router, then at most v flits can be transferred in one cycle regardless of the bandwidth

available. In a (u,b) router, if i out of b bidirectional links are configured to be ingressive at a router node, the node can receive up to $u+i$ flits per cycle from the node across the link and send out up to $(u+b-i)$ flits to the other node. Since each incoming flit will go to a different virtual channel queue,[2] the ingress demultiplexer in Figure 5 can be implemented with b instances of a v-to-1 demultiplexer with tri-state buffers at the outputs; no additional arbitration is necessary between demultiplexers because only one of their outputs will drive the input of each virtual channel.

In a bidirectional router architecture, the egress link can be configured to exceed one flit per cycle; consequently, the crossbar switch must be able to consider flits from more than one virtual channel from the same node. In the architecture described so far, the output of each virtual channel is directly connected to the switch and competes for an outgoing link. However, one can use a hierarchical solution where the v virtual channels are multiplexed to a smaller number of switch inputs. The Intel Teraflops has a direct connection of virtual channels to the switch [6]. Most routers have v-to-1 multiplexers that select one virtual channel from each port for each link prior to the crossbar.

In addition, the crossbar switch must now be able to drive all $p\cdot(u+b)$ outgoing links when every bidirectional link is configured as egressive, and there are u unidirectional links. Consequently, the router requires a $p\cdot v$-by-$p\cdot(u+b)$ crossbar switch, compared to a $p\cdot v$-by-$p\cdot(u+b/2)$ switch of a conventional $(u+b/2,0)$ router that has the same bisection bandwidth; this larger switch is the most significant hardware cost of bidirectional router architecture. If the v virtual channels are multiplexed to reduce the number of inputs of the switch, the number of inputs to the crossbar should be at least equal to the maximum number of outputs in order to fully utilize the bisection bandwidth. In this case, we have a $p\cdot(u+b/2)$-by-$p\cdot(u+b/2)$ crossbar in the $(u+b/2,0)$ case. In the (u,b) router, we will need a $p\cdot(u+b)$-by-$p\cdot(u+b)$ crossbar. The v virtual channels at each port will be multiplexed into $(u+b)$ inputs to the crossbar.

To evaluate the flexibility and effectiveness of bidirectional links, we compare, in Section V, the performance of bidirectional routers with $(u,b)=(0,2)$ and $(u,b)=(0,4)$ against unidirectional routers with $(u,b)=(1,0)$ and $(u,b)=(2,0)$, which, respectively, have the same total bandwidth as the bidirectional routers. We also consider a hybrid architecture with $(u,b)=(1,2)$ which has the same total bandwidth as the $(u,b)=(2,0)$ and $(u,b)=(0,4)$ configurations. Table I summarizes the sizes of hardware components of unidirectional, bidirectional and hybrid router architectures assuming four virtual channels per ingress port (i.e., $v=4$). There are two cases considered. The numbers in bold correspond to the case where all virtual channels compete for the switch. The numbers in plain text correspond to the case where virtual channels are multiplexed before the switch so the number of inputs to the switch is restricted by the bisection bandwidth. While switch allocation logic grows as the size of crossbar switch increases and bidirectional routers incur the additional cost of the bandwidth allocation logic shown in Figure 4, these are insignificant compared to the increased size of the demultiplexer and crossbar. In our simulation experiments we have compare the configurations in bold, as well as the ones in plain text.

Architecture	Ingress Demux	Xbar Switch
(u, b) = (1, 0)	one 1-to-4 demux	4-by-4 or **16-by-4**
(u, b) = (0, 2)	two 1-to-4 demuxes	8-by-8 or **16-by-8**
(u, b) = (2, 0)	two 1-to-4 demuxes	8-by-8 or **16-by-8**
(u, b) = (0, 4)	four 1-to-4 demuxes	16-by-16 or **16-by-16**
(u, b) = (1, 2)	three 1-to-4 demuxes	12-by-12 or **16-by-12**

TABLE I
THE SUMMARY OF DIFFERENCES IN HARDWARE COMPONENTS BETWEEN 4-VC ROUTER ARCHITECTURES

When virtual channels directly compete for the crossbar, the number of the crossbar input ports remains the same in both the unidirectional case and the bidirectional case. The number of crossbar output ports is the only factor increasing the crossbar size in bidirectional routers $(u,b)=(0,4)$ and $(1,2)$ when compared with the unidirectional $(2,0)$ case; this increase is size is roughly equal to the ratio of the output ports. Considering that a 32×32 crossbar takes approximately 30% of the gate count of a switch [7] with much of the actual area being accounted for by queue memory and wiring which is not part of the gate count, we estimate that a $1.5\times$ increase in crossbar size for the $(1,2)$ case will increase the area of the node by $<15\%$. If the queues are smaller, then this number will be larger. Similar numbers are reported in [8].

There is another way to compare the crossbars in the unidirectional and bidirectional cases. It is well known that the size of a $n\times n$ crossbar increases as n^2 (e.g., [9]). We can think of n as $p\cdot(u+b/2)\cdot w$, where w is the bit-width for the unidirectional case. If a bidirectional router's crossbar is $1.5\times$ larger, then one can create an equivalent-size unidirectional crossbar with the same number of links but $\sqrt{1.5}\times$ bit-width, assuming zero buffer sizes. In reality, the buffers will increase by $\sqrt{1.5}=1.22\times$ due to the bit-width increase, and so the equivalent-size undirectional crossbar will have a bit-width that is approximately $1.15\times$ of the bidirectional crossbar, assuming typical buffer sizes. This implies the performance of this crossbar in a network will be $1.15\times$ the baseline unidirectional case. As can be seen in Section V, the bidirectional link architecture results in greater gains in performance.

III. BANDWIDTH ALLOCATION IN BIDIRECTIONAL LINKS

Bidirectional links contain a bandwidth arbiter (see Figure 4) which governs the direction of the bidirectional links connecting a pair of nodes and attempts to maximize the connection throughput. Keys to our approach are the locality and simplicity of this logic: the arbiter makes its decisions based on very simple information local to the nodes it connects.

[2]Recall that once a virtual channel is allocated to a packet at the previous node, other packets cannot use the virtual channel until the current packet completes transmission.

Each network node tells the arbiter of a given bidirectional links how much *pressure* it wishes to exert on the link; this pressure indicates how much of the available link bandwidth the node expects to be able to use in the next cycle. In our design, each node counts the number of flits ready to be sent out on a given link (i.e., at the head of some virtual channel queue), and sends this as the pressure for that link. The arbiter then configures the links so that the ratio of bandwidths in the two directions approximates the pressure ratio, additionally ensuring that the bandwidth granted does not exceed the free space in the destination node. Consequently, if traffic is heavier in one direction than in the other, more bandwidth will be allocated to that direction.

The arbitration logic considers only the next-hop nodes of the flits at the front of the virtual channel queues and the available buffer space in the destination queues, both of which are local to the two relevant nodes and easy to compute. The arbitration logic itself consists of threshold comparisons and is also negligible in cost.

When each packet consists of one flit, the pressure as defined above exactly reflects the traffic that can be transmitted on the link; it becomes approximate when there are multiple flits per packet, since some of the destination queues with available space may be in the middle of receiving packets and may have been assigned to flows different from the flits about to be transmitted. Although more complex and accurate definitions of pressure are possible, our experience thus far is that this simple logic performs well in practice.

In some cases we may not want arbitration to take place in every cycle; for example, implementations which require a dead cycle after each link direction switch will perform poorly if switching takes place too often. On the other hand, switching too infrequently reduces the adaptivity of the bidirectional network, potentially limiting the benefits for quickly changing traffic and possibly requiring more complex arbitration logic. We explore this tradeoff in Section V.

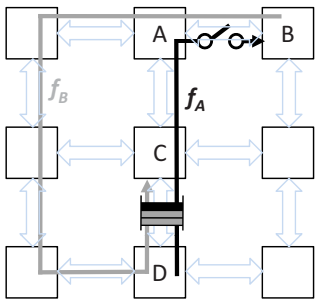

Fig. 6. Deadlock on deadlock-free routes due to bidirectional links

When analyzing link bandwidth allocation and routing in a bidirectional adaptive network, we must take care to avoid additional deadlock due to bidirectional links, which may arise in some routing schemes. Consider, for example, the situation shown in Figure 6: a flow f_B travels from node B to node C via node A, and all links connecting A with B are configured in the direction $B \rightarrow A$. Now, if another, smaller flow f_A starts at D and heads for B, it may not exert enough pressure on the $A \rightarrow B$ link to overcome that of f_B, and, with no bandwidth allocated there, may be blocked. The flits of f_A will thus eventually fill the buffers along its path, which might prevent other flows, including f_B, from proceeding: in the figure, f_B shares buffering resources with f_A between nodes C and D, and deadlock results. Note that the deadlock arises only because the bidirectional nature of the link between A and B can cause the connection $A \rightarrow B$ to disappear; since the routes of f_A and f_B obey the west-first turn model [10], the deadlock does not arise in the absence of bidirectional links. One easy way to avoid deadlock is to require, in the definition of pressure, that some bandwidth is always available in a given direction if some flits are waiting to be sent in that direction. For example, if there are four bidirectional links and there are eight flits waiting to travel in one direction and one in the opposite direction, we will assign three links to the first direction and one to the opposite direction.

IV. Related Work

A. Routing Techniques

A basic deterministic routing method is dimension ordered routing (DOR) [2] which becomes XY routing in a 2-D mesh. ROMM [11] and Valiant [12] are classic oblivious routing algorithms, which are randomized in order to achieve better load distribution. In o1turn [13], Seo *et al* show that simply balancing traffic between XY and YX routing can guarantee provable worst-case throughput. A weighted ordered toggle (WOT) algorithm that assumes 2 or more virtual channels [14] assigns XY and YX routes to source-destination pairs in a way that reduces the maximum network load for a given traffic pattern. While we have focused on dimension-ordered routing in this paper due to its speed and simplicity, other methods can be used in conjunction with bandwidth adaptivity.

Classic adaptive routing schemes include the turn routing methods [10] and odd even routing [15]. These are general schemes that allow packets to take different paths through the network while ensuring deadlock freedom but do not specify the mechanism by which a particular path is selected. An adaptive routing policy determines what path a packet takes based on network congestion.

Adaptive routing policies can be classified as either congestion-oblivious or congestion-aware, based on whether they take output link demand into account [8]. Some examples of congestion-oblivious routing strategies are random [16], zigzag [17] and no-turn [10]. Congestion-aware routing policies use various metrics to determine congestion. For example, Dally and Aoki [18] favor the port with the largest number of available virtual channels, and give results that have better performance than congestion-oblivious algorithms. In [19] a scheme that switches between deterministic and adaptive modes depending on the application is presented, where local FIFO information is used to adapt routes. Buffer availability at adjacent routers has been used as a congestion metric [20], as well as output queue length [21], [22]. These

routing algorithms all rely on local congestion indicators. Regional Congestion Awareness (RCA) [8] is an adaptive routing approach that propagates congestion information across the network in a scalable manner, improving the ability of adaptive routers to spread network load.

We have used oblivious routing methods in this paper, and therefore the hardware requirements are smaller than for conventional adaptive routing methods. The router only has to support DOR, and we have used simple, local congestion metrics to determine how best to configure each link. Rather than making decisions on a per-packet basis, our network makes decisions on a per-link basis.

B. Router Designs

Dally's virtual channels [23] allocate buffer space for virtual channels in a decoupled way from bandwidth allocation. Many designs of virtual channel routers have been proposed (e.g., [5], [24], [25], [26]). Our virtual channel router design is modified to enable adaptive bidirectional links in the network.

Router designs with bidirectional or half-duplex links have been proposed. For example, Ariadne [27], the Intel Cavallino [28] and NetworkDesignFrame [29] use half-duplex links, with the Cavallino using simultaneous bidirectional signalling. The MIT J-Machine [30] has bidirectional links where flits waiting on either side are sequentially transferred. Our architecture differs from previous architectures in the fine-grained adaptive control of multiple channels based on the amount of waiting data that can be accepted by the destination.

A recent paper also proposes reconfigurable bidirectional links [31]. Our work was carried out independently ([32] is an earlier version of this paper), and has significant differences with the BiNOC architecture of [31]. We use pressure-based control as opposed to request-based control in BiNOC which incurs direction-switching delays. BiNOC does not incorporate multiple virtual channels, and does not discuss any additional deadlock possibilities (cf. Figure 6). These possibilities are precluded by our definition of pressure.

V. Results and Comparisons

A. Experimental Setup

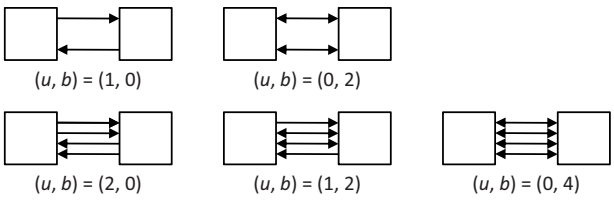

Fig. 7. Link configurations used in the experiments. The configurations on each row have the same bisection bandwidth.

A cycle-accurate network simulator was used to model the bidirectional router architectures with different combinations of unidirectional (u) and bidirectional (b) links in each connection (see Figure 7 and Table II for details). To evaluate performance under general traffic patterns, we employed a set of standard synthetic traffic patterns (*transpose*, *bit-complement*, *shuffle*, and *uniform-random*) both without burstiness and with a Markov Modulated Process (MMP) bursty traffic model. For the evaluation of performance under real-world applications, we profiled the network load of an H.264 decoder and employed the traffic pattern on the unidirectional and the bidirectional networks. We also examined several frequencies of bandwidth allocation to estimate the impact on architectures where a dead cycle is required to switch the link direction.

Characteristic	Configuration
Topology	8x8 2D MESH
Link configuration	(u, b) = (1,0), (0,2) (2,0), (1,2), (0,4)
Routing	DOR-XY and DOR-YX
VC output multiplexing	None, Matching maximum bandwidth
Per-hop latency	1 cycle
Virtual channels per port	4
Flit buffers per VC	4
Average packet length (flits)	8
Traffic workload	transpose, bit-complement, shuffle, uniform-random profiled H.264 decoder
Burstiness model	Markov modulated process
Warmup cycles	20,000
Analyzed cycles	100,000

TABLE II
Summary of network configuration

Although the bidirectional routing technique applies to various oblivious routing algorithms, we have, for evaluation purposes, focused on Dimension Ordered Routing (DOR), the most widely implemented oblivious routing method. While our experiments included both DOR-XY and DOR-YX routing, we did not see significant differences in the results, and consequently report only DOR-XY results. In all of our experiments, the router was configured for four virtual channels per ingress port under a dynamic virtual channel allocation regimen. The effect of multiplexing virtual channels in front of the crossbar switches was also examined.

B. Non-bursty Synthetic Traffic

Figure 8 shows the throughput in the unidirectional and bidirectional networks under non-bursty traffic. When traffic is consistent, the improvement offered by bidirectional links depends on how symmetric the flows are. On the one extreme, *bit-complement*, which in steady state is entirely symmetric when routed using DOR and results in equal traffic in each direction on any link, shows no improvement; on the other extreme, in *transpose*, packets move in only one direction over any given link, and bidirectional links improve throughput twofold. *Shuffle* lies between the two extremes, with the bidirectional network outperforming the unidirectional solution by 60% when total bandwidth is equal.

Uniformly random traffic is also symmetric when averaged over a period of time. For very short periods of time, however, the symmetry is imperfect, allowing the bidirectional network

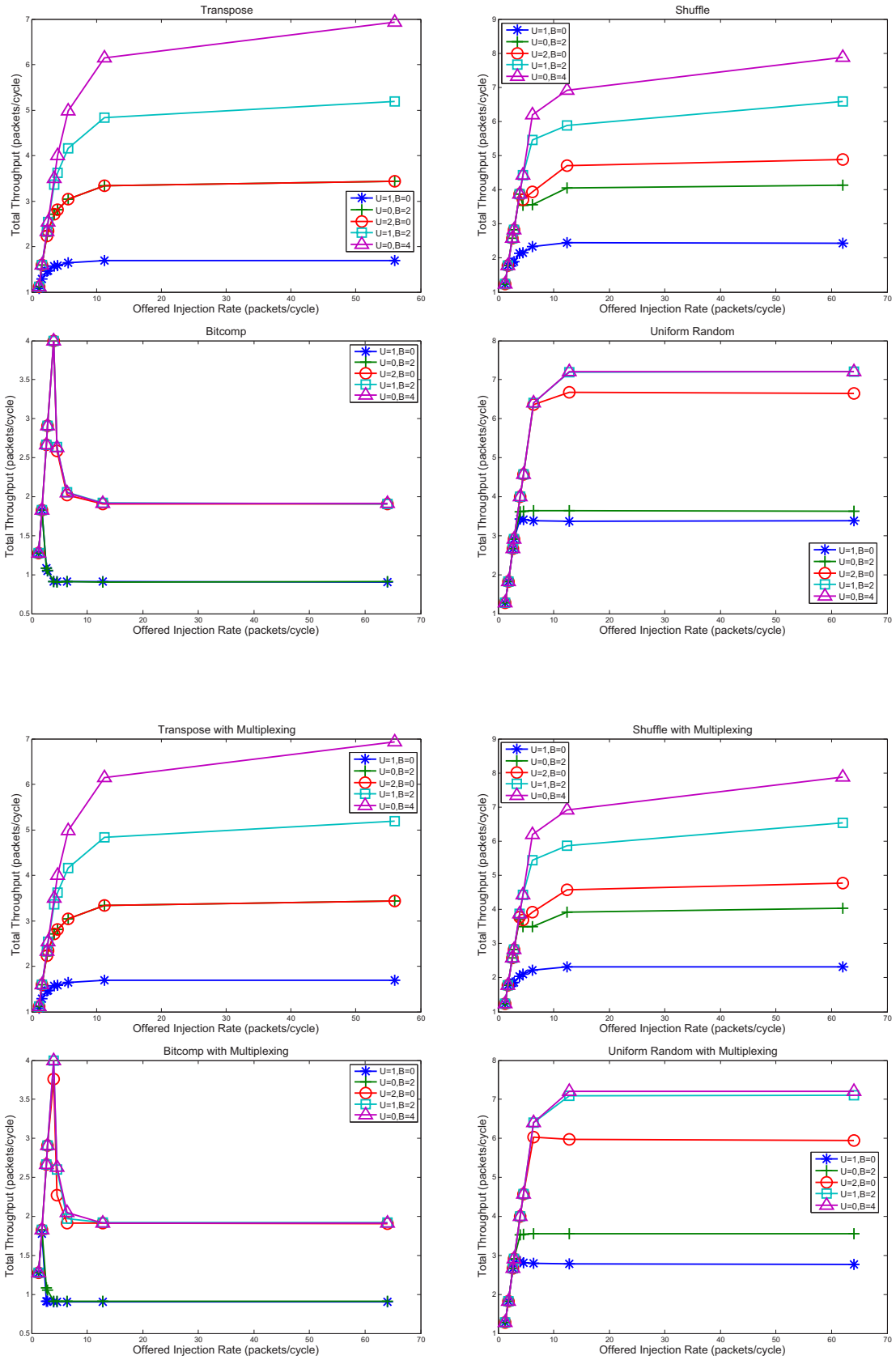

Fig. 8. Throughput under non-bursty traffic

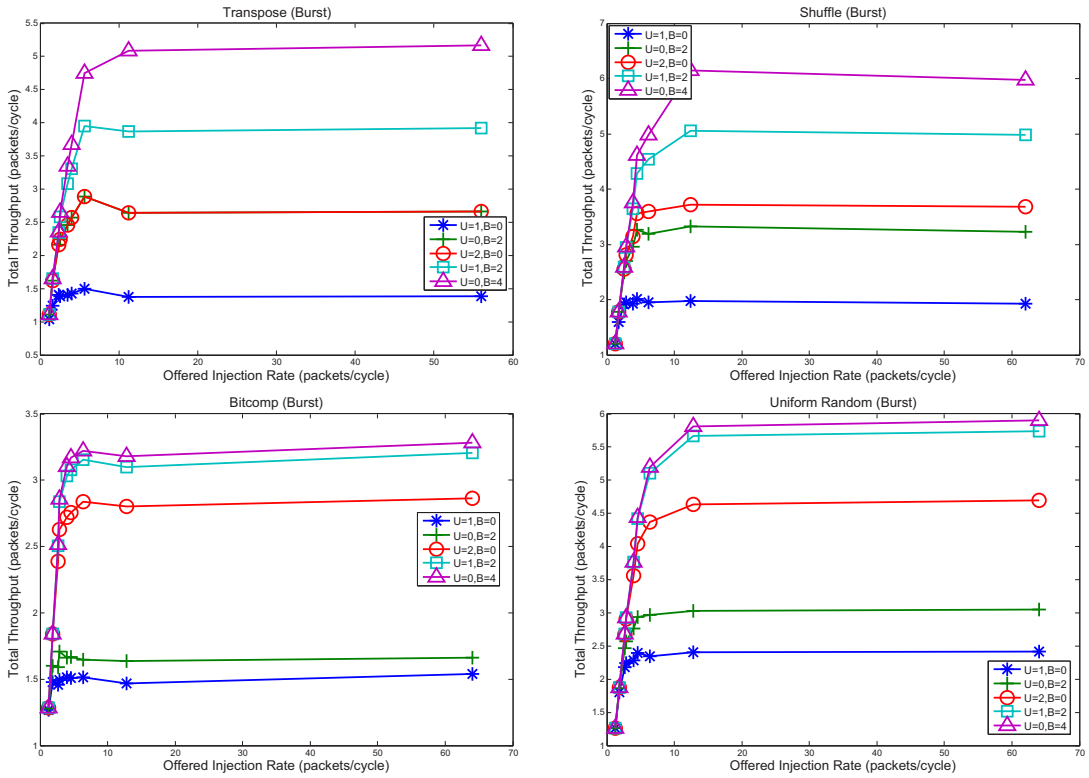

Fig. 9. Throughput under bursty traffic

to track the traffic shifts as they happen and outperform the unidirectional network throughput by up to 8% without multiplexing virtual channel outputs.

C. Non-bursty Synthetic Traffic with Multiplexed VC Outputs

If the outputs of virtual channels are multiplexed, the number of inputs to the crossbar switch can be significantly reduced, especially in unidirectional networks. However, the use of multiplexers can limit the flexibility of switch allocation because less virtual channels can compete for the switch at a given cycle.

This limited flexibility does not significantly affect performance of *bit-complement*, *transpose* and *shuffle* because packet flow at each network node is in steady-state under these traffic patterns. If packet flow is in steady-state, each port at each network node has the same inflows and outflows of flits, which are bounded by the maximum outgoing bandwidth. Therefore, multiplexing corresponding to the maximum outgoing bandwidth does not affect throughput because we need not connect more virtual channels to the switch than the number of multiplexer outputs.

On the other hand, if the congestion at each link is not in steady-state as in the *uniform-random* example, each port sees a temporal mismatch between inflows and outflows of flits. If all virtual channels can compete for the switch without multiplexers, flits in ingress queues can be quickly pulled out as soon as the link to the next hop becomes less congested. The results show the unidirectional networks have 10% less throughput under *uniform-random* when multiplexers are used, as they cannot pull out congested flits as fast as networks without multiplexers. Bidirectional networks have more multiplexer outputs than unidirectional networks because their maximum outgoing bandwidth is greater than unidirectional networks. Therefore, the size of crossbar switches of bidirectional networks increases, but they can send out more flits in congested ports than unidirectional networks. Consequently, the bidirectional networks outperform the unidirectional network throughput by up to 20% under *uniform-random* when virtual channel outputs are multiplexed as shown in Figure 8.

D. Bursty Synthetic Traffic

The temporary nature of bursty traffic allows the bidirectional network to adjust the direction of each link to favor whichever direction is prevalent at the time, and results in throughput improvements across all traffic patterns (see Figure 9). With bursty traffic, even *bit-complement*, for which the bidirectional network does not win over the unidirectional case without burstiness, shows a 20% improvement in total throughput because its symmetry is broken over short periods of time by the bursts. For the same reason, *shuffle* and *uniform-random* outperform the unidirectional network by 66% and 26% respectively, compared to 60% and 8% in non-bursty mode. Finally, *transpose* performance is the same as for the non-bursty case, because the traffic, if any, still only flows in one direction and requires no changes in link direction after the initial adaptation.

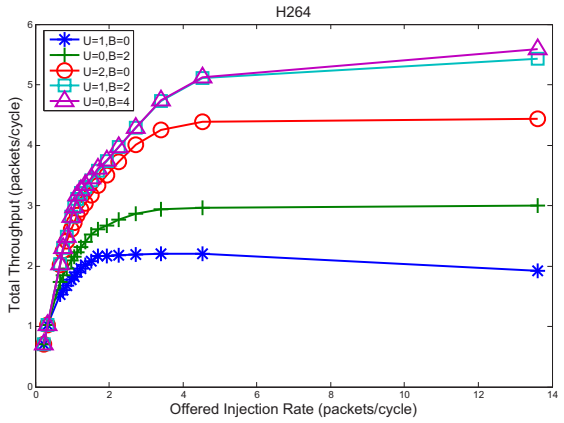

Fig. 10. Throughput under traffic of an H.264 decoder application

These results were obtained with virtual channels directly competing for the crossbar. We have simulated these examples with multiplexed VC outputs and the results have the same trends as in Figure 9, and therefore are not shown here.

E. Traffic of an H.264 Decoder Application

As illustrated in the example of *transpose* and *bit-complement*, bidirectional networks can significantly improve network performance when network flows are not symmetric. As opposed to the synthetic traffic such as *bit-complement*, the traffic patterns in many real applications are not symmetric as data is processed by a sequence of modules. Therefore, bidirectional networks are expected to have significant performance improvement with many real applications. Figure 10 illustrates the performance of the bidirectional and the unidirectional networks under traffic patterns profiled from an H.264 decoder application, where the bidirectional networks outperforms unidirectional networks up to 35%. The results correspond to the case where virtual channels directly compete for the crossbar, and is virtually identical to the results with VC multiplexing.

F. Link Arbitration Frequency

So far, our results have assumed that the bandwidth arbiter may alter the direction of every link on every cycle. While we believe this is realistic, we also considered the possibility that switching directions might require a dead cycle, in which case changing too often could limit the throughput up to

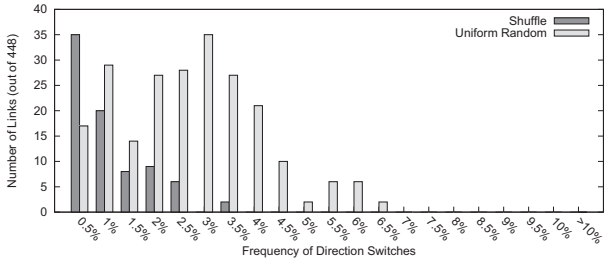

Fig. 11. The frequency of direction changes on bidirectional links

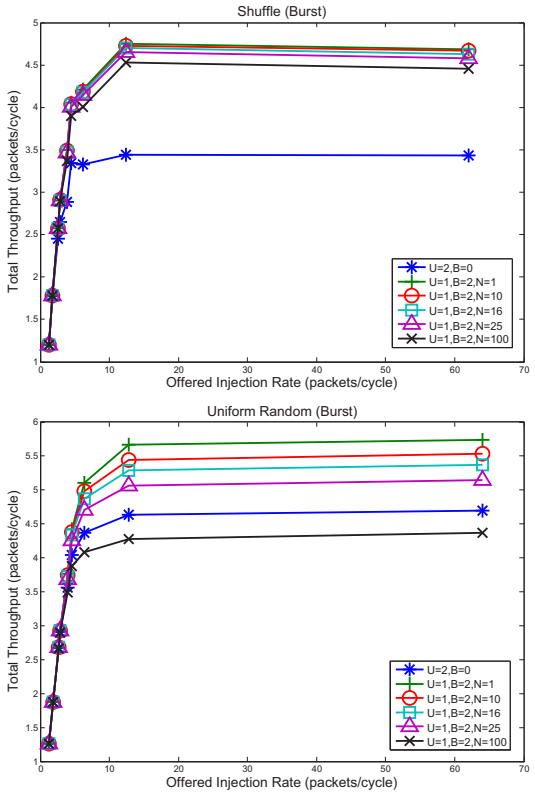

Fig. 12. Bursty traffic throughput with different link arbitration periods (N)

50% in the worst case. We therefore reduced the arbitration frequency and examined the tradeoff between switching every N cycles (thereby lessening the impact of a dead cycle to $\frac{1}{N+1}$) and limiting the network's adaptivity to rapid changes in traffic patterns. The results in this section illustrate the relevant tradeoffs.

Figure 11 shows how often each bidirectional link actually changes its direction under bursty *shuffle* and *uniform-random* traffic: the x-axis shows how frequently links directions change and the y-axis shows how many links switch that often. For example, under *shuffle* traffic, about 8% of bidirectional links change their direction less than once every two hundred cycles. Traffic exhibiting the *uniform-random* pattern, in comparison, is more symmetric than *shuffle*, and so the link directions change more often.

The observation that no link changes its direction more frequently than once in ten cycles led us to investigate how infrequent the link switches could be without significantly affecting performance. In Figure 12 we compare the performance of the bidirectional network under different link arbitration frequencies; as expected, throughput decreases when the links are allowed to switch less often.

Even with a switching period as large as 100 cycles, the bidirectional network still significantly outperforms the unidirectional design under many loads (e.g., by more than 20% for *shuffle*). In the case of *uniform-random*, however, the bidirectional network performance trails the unidirectional de-

sign when switching is infrequent. This is because, when each link arbitration decision lasts 100 cycles, any temporary benefit from asymmetric bandwidth allocation is nullified by changes in traffic patterns, and, instead of improving throughput, the asymmetric allocations only serve to throttle down the total throughput compared to the unidirectional router.

Infrequent link switching, therefore, demands a more sophisticated link bandwidth arbiter that bases its decisions on the pressures observed over a period of time rather than on instantaneous measurements. For *uniform-random*, for example, the symmetry of uniform random traffic over time would cause the link bandwidths to be allocated evenly by such an arbiter, allowing it to match the performance of the unidirectional network.

VI. Conclusions

We have proposed the notion of bandwidth-adaptive networks in this paper, given one concrete example of bidirectional links in a 2-D mesh, and evaluated it. Adaptivity is controlled by local pressure that is easily computed. While more comprehensive evaluation should be performed, adaptive bidirectional links provide better performance under both uniform and bursty traffic for the tested benchmarks.

We have focused on a mesh; however, adaptive bidirectional links can clearly be used in other network topologies. In adaptive routing decisions are made on a per-packet basis at each switch. In bandwidth-adaptive networks, decisions are made on a per-link basis. We believe this difference makes bandwidth-adaptivity more amenable to local decision making, though more rigorous analysis is required.

References

[1] L. M. Ni and P. K. McKinley, "A Survey of Wormhole Routing Techniques in Direct Networks," *Computer*, vol. 26, no. 2, pp. 62–76, 1993.

[2] W. J. Dally and C. L. Seitz, "Deadlock-Free Message Routing in Multiprocessor Interconnection Networks," *IEEE Trans. Computers*, vol. 36, no. 5, pp. 547–553, 1987.

[3] W. J. Dally and B. Towles, *Principles and Practices of Interconnection Networks*. Morgan Kaufmann, 2003.

[4] L.-S. Peh and W. J. Dally, "A Delay Model and Speculative Architecture for Pipelined Routers," in *Proc. International Symposium on High-Performance Computer Architecture (HPCA)*, Jan. 2001, pp. 255–266.

[5] R. D. Mullins, A. F. West, and S. W. Moore, "Low-latency virtual-channel routers for on-chip networks," in *Proc. of the 31st Annual Intl. Symp. on Computer Architecture (ISCA)*, 2004, pp. 188–197.

[6] Y. Hoskote, S. Vangal, A. Singh, N. Borkar, and S. Borkar, "A 5-GHz Mesh Interconnect for a Teraflops Processor," *IEEE Micro*, vol. 27, no. 5, pp. 51–61, Sept/Oct 2007.

[7] M. Katevenis, G. Passas, D. Simos, I. Papaefstathiou, and N. Chrysos, "Variable packet size buffered crossbar (CICQ) switches," in *2004 IEEE International Conference on Communications*, vol. 2, Jun. 2004, pp. 1090–1096.

[8] P. Gratz, B. Grot, and S. W. Keckler, "Regional Congestion Awareness for Load Balance in Networks-on-Chip," in *In Proc. of the 14th Int. Symp. on High-Performance Computer Architecture (HPCA)*, Feb. 2008, pp. 203–214.

[9] T. Wu, C. Y. Tsui, and M. Hamdi, "CMOS Crossbar," in *Proceedings of the* 14^{th} *IEEE Symposium on High Performance Chips (Hot-Chips 2002)*, August 2002.

[10] C. J. Glass and L. M. Ni, "The turn model for adaptive routing," *J. ACM*, vol. 41, no. 5, pp. 874–902, 1994.

[11] T. Nesson and S. L. Johnsson, "ROMM routing on mesh and torus networks," in *Proc. 7th Annual ACM Symposium on Parallel Algorithms and Architectures SPAA'95*, 1995, pp. 275–287.

[12] L. G. Valiant and G. J. Brebner, "Universal schemes for parallel communication," in *STOC '81: Proceedings of the thirteenth annual ACM symposium on Theory of computing*, 1981, pp. 263–277.

[13] D. Seo, A. Ali, W.-T. Lim, N. Rafique, and M. Thottethodi, "Near-Optimal Worst-Case Throughput Routing for Two-Dimensional Mesh Networks," in *Proceedings of the 32nd Annual International Symposium on Computer Architecture (ISCA 2005)*, 2005, pp. 432–443.

[14] R. Gindin, I. Cidon, and I. Keidar, "NoC-Based FPGA: Architecture and Routing," in *First International Symposium on Networks-on-Chips (NOCS 2007)*, 2007, pp. 253–264.

[15] G.-M. Chiu, "The Odd-Even Turn Model for Adaptive Routing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 11, no. 7, pp. 729–738, 2000.

[16] W.-C. Feng and K. G. Shin, "Impact of Selection Functions on Routing Algorithm Performance in Multicomputer Networks," in *In Proc. of the Int. Conf. on Supercomputing*, 1997, pp. 132–139.

[17] H. Badr and S. Podar, "An Optimal Shortest-Path Routing Policy for Network Computers with Regular Mesh-Connected Topologies," *IEEE Transactions on Computers*, vol. 38, no. 10, pp. 1362–1371, 1989.

[18] W. J. Dally and H. Aoki, "Deadlock-free adaptive routing in multicomputer networks using virtual channels," *IEEE Transactions on Parallel and Distributed Systems*, vol. 04, no. 4, pp. 466–475, 1993.

[19] J. Hu and R. Marculescu, "DyAD: Smart Routing for Networks on Chip," in *Design Automation Conference*, Jun. 2004.

[20] H. J. Kim, D. Park, T. Theocharides, C. Das, and V. Narayanan, "A Low Latency Router Supporting Adaptivity for On-Chip Interconnects," in *Proceedings of Design Automation Conference*, June 2005, pp. 559–564.

[21] A. Singh, W. J. Dally, A. K. Gupta, and B. Towles, "GOAL: a load-balanced adaptive routing algorithm for torus networks," *SIGARCH Comput. Archit. News*, vol. 31, no. 2, pp. 194–205, 2003.

[22] A. Singh, W. J. Dally, B. Towles, and A. K. Gupta, "Globally Adaptive Load-Balanced Routing on Tori," *IEEE Comput. Archit. Lett.*, vol. 3, no. 1, 2004.

[23] W. Dally, "Virtual-Channel Flow Control," *IEEE Transactions on Parallel and Distributed Systems*, vol. 03, no. 2, pp. 194–205, 1992.

[24] T. Bjerregaard and J. Sparsø, "Virtual channel designs for guaranteeing bandwidth in asynchronous network-on-chip," in *Proceedings of the IEEE Norchip Conference (NORCHIP 2004)*. IEEE, 2004.

[25] N. K. Kavaldjiev, G. J. M. Smit, and P. G. Jansen, "A virtual channel router for on-chip networks," in *IEEE Int. SOC Conf., Santa Clara, California*. IEEE Computer Society Press, Sep. 2004, pp. 289–293. [Online]. Available: http://eprints.eemcs.utwente.nl/775/

[26] C. A. Nicopoulos, D. Park, J. Kim, N. Vijaykrishnan, M. S. Yousif, and C. R. Das, "ViChaR: A dynamic virtual channel regulator for network-on-chip routers," in *Proc. of the 39th Annual Intl. Symp. on Microarchitecture (MICRO)*, 2006.

[27] J. D. Allen, P. T. Gaughan, D. E. Schimmel, and S. Yalamanchili, "Ariadne—an adaptive router for fault-tolerant multicomputers," *SIGARCH Comput. Archit. News*, vol. 22, no. 2, 1994.

[28] J. Carbonaro and S. Verhoorn, "Cavallino: The Teraflops Router and NIC," in *Proceedings of the Fourth Symp. High-Performance Interconnects (Hot Interconnects 4)*, August 1996.

[29] W. J. Dally and P. Song, "Design of a self-timed VLSI multicomputer communication controller," in *Proceedings of International Conference on Computer Design (ICCD-87)*, 1987, pp. 230–234.

[30] P. R. Nuth and W. J. Dally, "The J-machine Network," in *Proceedings of the IEEE Conference on Computer Design: VLSI in Computers and Processors*, 1992, pp. 420–423.

[31] Y.-C. Lan, S.-H. Lo, Y.-C. Lin, Y.-H. Hu, and S.-J. Chen, "BiNoC: A Bidirectional NoC Architecture with Dynamic Self-Reconfigurable Channel," in *Proceedings of the* 3^{rd} *ACM/IEEE International Symposium on Networks-on-Chip*, May 2009.

[32] M. H. Cho, M. Lis, K. S. Shim, M. Kinsy, T. Wen, and S. Devadas, "Oblivious Routing in On-Chip Bandwidth-Adaptive Networks," Massachusetts Institute of Technology, Tech. Rep. CSAIL-TR-2009-011 (http://hdl.handle.net/1721.1/44958), Mar. 2009.