Scalable High-Radix Router Microarchitecture Using a Network Switch Organization

JUNG HO AHN and YOUNG HOON SON, Seoul National University
JOHN KIM, Korea Advanced Institute of Science and Technology

As the system size of supercomputers and datacenters increases, cost-efficient networks become critical in achieving good scalability on those systems. *High*-radix routers reduce network cost by lowering the network diameter while providing a high bisection bandwidth and path diversity. The building blocks of these large-scale networks are the routers or the switches and they need to scale accordingly to the increasing port count and increasing pin bandwidth. However, as the port count increases, the high-radix router microarchitecture itself needs to scale efficiently. Hierarchical crossbar switch organization has been proposed where a single large crossbar used for a router switch is partitioned into many small crossbars and overcomes the limitations of conventional router microarchitecture. Although the organization provides high performance, it has limited scalability due to excessive power and area overheads by the wires and intermediate buffers.

In this article, we propose scalable router microarchitectures that leverage a network within the switch design of the high-radix routers themselves. These alternative designs lower the wiring complexity and buffer requirements. For example, when a folded-Clos switch is used instead of the hierarchical crossbar switch for a radix-64 router, it provides up to 73%, 58%, and 87% reduction in area, energy-delay product, and energy-delay-area product, respectively. We also explore more efficient switch designs by exploiting the traffic-pattern characteristics of the *global* network and its impact on the *local* network design within the switch for both folded-Clos and flattened butterfly networks. In particular, we propose a *bilateral butterfly* switch organization that has fewer crossbars and global wires compared to the topology-agnostic folded-Clos switch while achieving better low-load latency and equivalent saturation throughput.

Categories and Subject Descriptors: B.4.3 [**Interconnections (Subsystems)**]: Topology; C.1.2 [**Multiple Data Stream Architectures (Multiprocessors)**]: Interconnection Architectures

General Terms: Design, Performance

Additional Key Words and Phrases: Interconnection architectures, network topology, packet-switching networks

---

This article is an extension of Ahn et al. [2012].
J. Ahn was supported in part by the Center for Integrated Smart Sensors funded by the Ministry of Education, Science and Technology (MEST) of Korea as Global Frontier Project (CISS-2012M3A6A6054193), by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by MEST (2012R1A1B4003447), and by IDEC (EDA Tool). J. Kim was supported by World Class University program under NRF and funded by MEST (project no. R31-30007) and in part by the Ministry of Knowledge Economy, Korea, under the Information Technology Research Center support program supervised by the National IT Industry Promotion Agency (NIPA-2013-H0301-13-1011).
Authors' addresses: J. H. Ahn (corresponding author) and Y. H. Son, Department of Transdisciplinary Studies, Seoul National University, Gwanak-gu, Seoul, Korea; email: ghajh@snu.ac.kr; J. Kim, Division of Web Science and Technology and Department of Computer Science, Korea Advanced Institute of Science and Technology, Daejeon, Korea.

DOI: http://dx.doi.org/10.1145/2512433

ACM Reference Format:
Ahn, J. H., Son, Y. H., and Kim, J. 2013. Scalable high-radix router microarchitecture using a network switch organization. ACM Trans. Architec. Code Optim. 10, 3, Article 17 (September 2013), 25 pages.
DOI: http://dx.doi.org/10.1145/2512433

1. INTRODUCTION

As the number of components to be connected in a large-scale system increases, the interconnection network that connects them becomes increasingly important and can determine the overall performance and cost of the system. These large-scale networks have been traditionally found in high-performance computing, but with the recent emergence of *warehouse-scale* computers [Hoelzle and Barroso 2009] and datacenters that can have up to millions of servers, cost-efficient large-scale networks are also needed in datacenters to provide high performance and energy efficiency [Abts et al. 2010]. Previously, large-scale networks were built with *low*-radix routers where the number of ports per router was relatively small [Agarwal 1991; Dally 1990]. However, with the exponentially increasing pin bandwidth, recent work [Kim et al. 2005a] has shown how it is more cost effective to partition the increasing pin bandwidth into a large number of ports and create *high*-radix networks. The Cray BlackWidow [Abts et al. 2007] is one of the first systems that used high-radix networks as it leveraged radix-64 routers.

Creating a scalable router microarchitecture is one of the design challenges with high-radix routers. The cables [Kim et al. 2008] that connect the routers dominate the cost of these large-scale networks and the off-chip channel bandwidth through these cables often limits the network performance. Thus, the router microarchitecture needs to be scaled properly to ensure that the off-chip channel bandwidth is properly saturated and the internal router microarchitecture does not become the bottleneck. Unfortunately, the conventional router microarchitecture that consists of a single crossbar switch [Dally and Towles 2003] scales poorly with the router radix (k) because the complexity of allocating the packets in input buffers to output ports scales with k^2 and the fanout of the crossbar wires increases linearly with k. To overcome these limitations, a hierarchical crossbar switch organization [Kim et al. 2005a] was proposed and adapted in the YARC router [Scott et al. 2006]. It partitions a single large crossbar into many small crossbars or subswitches and places intermediate buffers at the input and the output of the subswitches. This reduces the number of participants in allocation and the number of output ports an input port can reach from k to p, where p is the radix of a subswitch, so that it is possible to improve the allocation complexity and the energy efficiency of data transfers. However, the YARC router design requires an excessive number of wires as described in Section 3.1 and does not scale efficiently.

We propose alternative and more scalable router microarchitectures using local networks and leverage high-radix topologies, such as folded-Clos [Dally and Towles 2003] and flattened butterfly [Kim et al. 2007b; Ahn et al. 2009], in the switch design of the high-radix routers. We create an on-chip network [Dally and Towles 2001] within a router to construct a *local* network that connects the switch ports of the router, then interconnect these routers together to build a *global* network, introducing a *network within a network*[1] approach to building a scalable switch microarchitecture

[1]Numerous hierarchical networks have been proposed in literature [Dally and Towles 2003; Duato et al. 2002], which can also be regarded as a "network within a network" approach. However, our approach differs as we leverage this idea to create a scalable router microarchitecture.

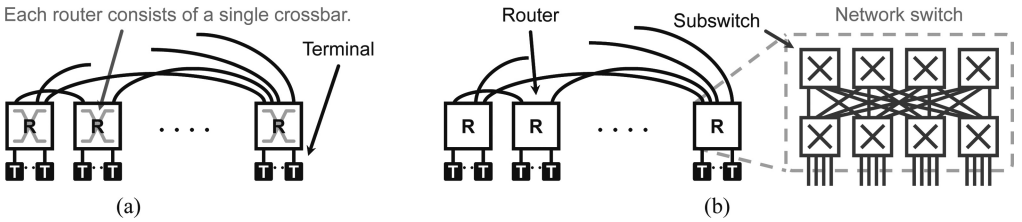


Fig. 1. (a) Conventional crossbar switch-based network and (b) network within a network approach. In (b), we assume that the global network is a 1D flattened butterfly and the local network within the router uses a folded-Clos topology.

(Figure 1(b)). Compared to the hierarchical crossbar design, these network switch organizations require fewer global wires, crossbars, and intermediate buffers and result in improvement of both area and energy efficiency. In addition, since the local network is used within a global network, we leverage the traffic-pattern characteristics of the global network projected on individual routers to further optimize the switch router microarchitecture. For example, load-balancing is often necessary for high performance on any network topology. As a result, we show how load-balancing within the local network is no longer needed, and thus the complexity of the local network within the switch can be further reduced.

Our evaluation shows that the switch design adopting the folded-Clos topology is more area and energy efficient than alternatives, including the hierarchical crossbar design. The design provides up to 73% area reduction compared to the hierarchical crossbar switch while achieving up to 58% and 87% reduction in energy-delay and energy-delay-area products for radix-64 routers. We propose a new topology-aware switch design called *bilateral butterfly* for the local networks in a folded-Clos global network. It consists of fewer crossbars and half the number of global wires compared to the topology-agnostic folded-Clos switch, while achieving better low-load latency and equivalent saturation throughput. We also show that a flattened butterfly switch could be more area and energy efficient than the topology-agnostic folded-Clos switch for a flattened butterfly global network.

This article extends prior work [Ahn et al. 2012] in three main ways. First, additional simulation results have been included to investigate the design space of the folded-Clos global networks utilizing the traffic of both real applications and synthetic patterns (Section 5.2). Second, we have extended our analysis on how the global traffic impacts the internal switch within the flattened butterfly networks in addition to the folded-Clos networks (Section 4.3). Third, the wire and buffer models used in the evaluation of the alternative router microarchitecture were carried out in more detail with SPICE simulations and the details are included (Section 5.1).

The contributions of this article include the following.

—We show that the prior approaches to building a high-radix router using hierarchical organization provide high performance, but suffer high area and low energy efficiency with limited scalability.
—We leverage the concept of *network within a network*, using a local network within a router to create the internal switch microarchitecture and using it within a large-scale, global network.
—We provide a detailed analysis and comparison of alternative approaches and show that the folded-Clos switch designs provide the best area and energy efficiency and are capable of providing fine-grain trade-offs between switch performance and area/energy efficiency.

—We show how the topology-aware switch designs provide superior performance and energy efficiency compared to topology-agnostic designs for folded-Clos and flattened butterfly networks.

2. BACKGROUND

In designing the high-radix routers, it is all important to reduce the die area and to improve the performance and energy efficiency. Performance is the primary design goal of any integrated circuits, and a highly energy-efficient design enables the router to achieve better performance within the power budget of the package. Reduction in the area enables the additional area to be used for more I/Os and other functionalities. For example, in the Cray YARC router [Scott et al. 2006], the SerDes (serializer/deserializer) I/Os occupy approximately 20% of the total area. In the Cray Gemini router [Roweth and Jones 2008], the network interface (NIC) was incorporated into the same router chip. In this section, we review popular high-radix topologies and the technology trends for critical components of high-radix routers, such as I/O ports, buffers, crossbars, route computation units, and allocators.

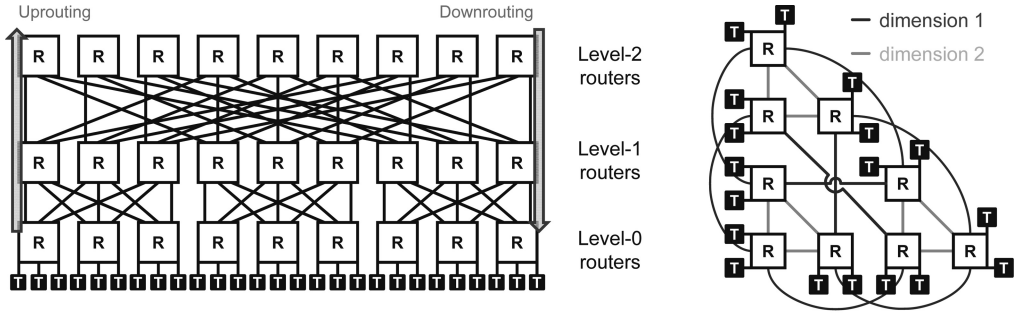
2.1. Technology Trends

Prior work [Kim et al. 2005a] showed the exponential growth in pin bandwidth over the past 30 years. The Cray YARC router published in 2006 has an aggregate off-chip bandwidth of 2.4Tb/s, Rambus demonstrated a 8Tb/s memory system in 2008 [Beyene et al. 2008], and a hub chip in the PERCS interconnect published in 2010 has a 56×56 crossbar with 9Tb/s of raw off-chip bandwidth [Arimilli et al. 2010]. The radix-64 switch fabric reported in 2012 by Satpathy et al. [2012] focused on improving energy efficiency as it achieved 4.5Tb/s throughput while consuming 1.3W. A high-speed serial link design that achieves sub 1pJ/b of energy efficiency over 10Gbps of transfer rate was reported in 2010 [Miura et al. 2010]. Area efficiency of high-speed links has started to gain interest as well [O'Mahony et al. 2010]. Considering the FO4 delay of a future 16nm process projected by a PTM model [Zhao and Cao 2006] and a 4 FO4 design rule [Yang 1998], we expect that the off-chip I/O bandwidth will continue to increase over the next 5 to 10 years. In addition, the recent advances in nanophotonic interconnects are expected to not only provide an increasing bandwidth, but also improve the *efficiency* of signaling so that the impact of I/O signaling on the overall power consumption will be reduced [Miller 2009; Beyene et al. 2008]. In this article, we assume that the bandwidth and energy efficiency of off-chip channels continue to improve and focus on devising scalable router microarchitecture designs.

In deep submicron technologies, the performance of on-chip global wires, not transistors, has become a major limiting factor on IC design [Ho et al. 2001]. There is a large design space on semiglobal and global wires for different design goals, such as delay, throughput density, and energy efficiency [Zhang et al. 2009]. Since throughput density and energy efficiency are important for on-chip datapath wires in high-radix routers, we assume that these wires are RC repeated and have the minimum pitch, similar to the assumptions made by Balfour and Dally [2006] and by Joshi et al. [2009] for their on-chip network designs.

2.2. Topology

Conventional low-radix topologies, such as 2D/3D mesh or torus topologies, are not appropriate for fully exploiting the benefits of high-radix routers. Instead, topologies tailored to high-radix routers, such as folded-Clos [Dally and Towles 2003] or flattened butterfly [Kim et al. 2007b], are necessary. A folded-Clos network is a multilevel network made by folding a Clos network that consists of an odd number of stages and

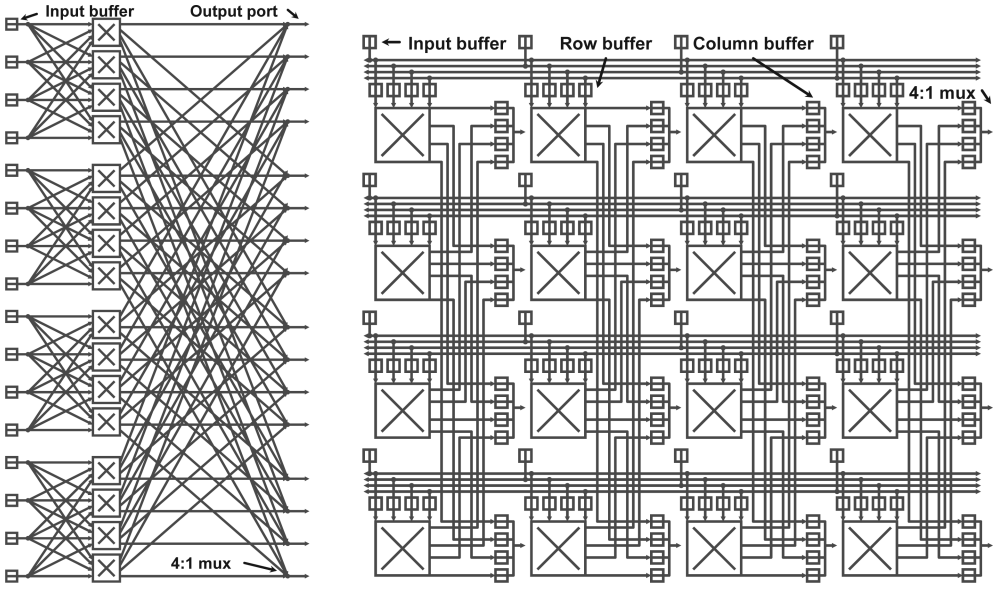


(a) folded Clos (b) flattened butterfly (HyperX)

Fig. 2. (a) An exemplar 3-level folded-Clos network connecting 27 terminal nodes using radix-6 routers. Each router is connected to lower-level routers through downlinks and to upper-level routers through uplinks. Note that the top-level routers do not have uplinks connected. (b) An exemplar flattened butterfly (HyperX) topology with ($L = 2, S = 3, K = 1, T = 2$).

then by fusing input and output switch pairs that collocate. A symmetric 3-stage Clos network can be represented by three tuples (m, n, r), where m, n, and r are the number of middle-stage router, input ports per input router, and the number of input routers, respectively. The Clos network has a path diversity m and is noninterfering [Dally and Towles 2003] if $m \geq n$. By folding, the 3-stage Clos network becomes an (nr)-radix 2-level folded-Clos network where each port becomes bidirectional by having an input and an output channel, a middle-stage router becomes a top-level router with radix r, and a pair of an input and an output router becomes a bottom-level router with radix $(n + m)$. Throughout the article, we assume that each router or switch has the same number of input and output channels and we call it a radix of the router or switch. We also interchange the terms channel and link. A folded-Clos network with more than 2 levels can be constructed by composing each top-level router using a folded-Clos network recursively. Figure 2(a) shows an exemplar 3-level folded-Clos network connecting 27 terminal nodes using radix-6 routers. The links in a folded Clos consist of uplinks and downlinks. Links used to send data from Level i to Level $i + 1$ routers are called uplinks while downlinks send data from Level i to Level $i - 1$ routers. Routing in a folded Clos consists of two phases: uprouting, and then, downrouting. In uprouting, the packets are routed to the nearest common ancestor between the source and destination using uplinks. Once the packet reaches this intermediate router, the packet is downrouted to its destination.

The flattened butterfly topology [Kim et al. 2007b] can also exploit high-radix routers. An m-ary n-flat network is derived from an m-ary n-fly butterfly network where all the routers in each row of the butterfly network are *flattened* into a single router and the same connections are maintained. HyperX [Ahn et al. 2009] extends the flattened butterfly topology and we use the notation provided by the HyperX framework in this article. A regular HyperX is an L-dimensional direct network where the size of each dimension is S and the routers in each dimension are fully connected. Each router in the HyperX is connected to T terminals and the bandwidth of links between the routers is K times the bandwidth of links to terminals. Using the (L, S, K, T) notation, an m-ary n-flat network can be represented as an $(n - 1, m, 1, m)$ HyperX network. It connects TS^L terminals with $\frac{KS^{L+1}}{2}$ channels crossing its bisection. Therefore, the ratio of bisection bandwidth to aggregate terminal bandwidth $\beta = \frac{KS}{2T}$ is 0.5 for the m-ary n-flat network. Figure 2(b) shows an exemplar 2D HyperX network composed of radix-6



(a) block diagram (b) layout

Fig. 3. (a) A logical block diagram and (b) a layout of a radix-16 hierarchical crossbar switch design. Subswitch size p is 4.

routers with (2, 3, 1, 2). In this article, HyperX and flattened butterfly terms are used interchangeably.

3. MICROARCHITECTURES FOR HIGH-RADIX ROUTERS

In this section, we first describe the previously proposed hierarchical switch organization for high-radix routers [Kim et al. 2005a; Scott et al. 2006] and its limitations. Then, we describe an alternative switch design that uses a network, and thus creates a network within a network organization. Throughout this work, we define a *local* network as the network used for a switch within a single router and a *global* network as the system-level network that interconnects the routers together.

3.1. Hierarchical Switch Organization

The hierarchical router organization [Kim et al. 2005a] partitions a single $k \times k$ switch into $(\frac{k}{p})^2$ $p \times p$ subswitches where k is the router radix and p is the subswitch crossbar radix. The hierarchical router organization introduces $2\frac{k^2}{p}$ intermediate buffers (or subswitch buffers) at both the input and the output of the subswitches. The intermediate buffers decouple the input and the output arbitration and avoid global arbitration by localizing the arbitration. Figure 3(a) shows a logical block diagram of a hierarchical crossbar switch organization with $k = 16$ and $p = 4$. Its corresponding layout example is shown in Figure 3(b).

The hierarchical crossbar switch consists of k row buses or horizontal broadcast channels and $\frac{k}{2} \times \frac{k}{p}$ vertical, column channels since all output ports in a column are fully connected to all column buffers in the same column. These horizontal and vertical channels require global wires, whose area scales at the rate of $k \times \frac{k}{2} \times \frac{k}{p} = \frac{k^3}{2p}$. Note that the total area of all the subswitch crossbars scales at the rate of k^2 as there are $(k/p)^2$ subswitches and each $p \times p$ subswitch area is proportional to p^2. We set $p = \sqrt{k}$

to minimize the aggregate fanout[2] and to minimize the dynamic energy of a packet that traverses the switch [Sutherland et al. 1999]. As a result, the area of horizontal and vertical wires dominates the switch area with a large k, which scales at the rate of $k^{2.5}$. The number of subswitch buffers also scales superlinearly at the rate of $\frac{k^2}{p} = k^{1.5}$. These all lead to an inefficient area and power scaling of the hierarchical switches as k increases. The large amount of speedup also leads to inefficient scaling, as shown in Figure 3(b). Thus, we explore an alternative organization using different multistage networks as the switch organization to provide a scalable switch design.

3.2. Network within a Network Switch Architecture

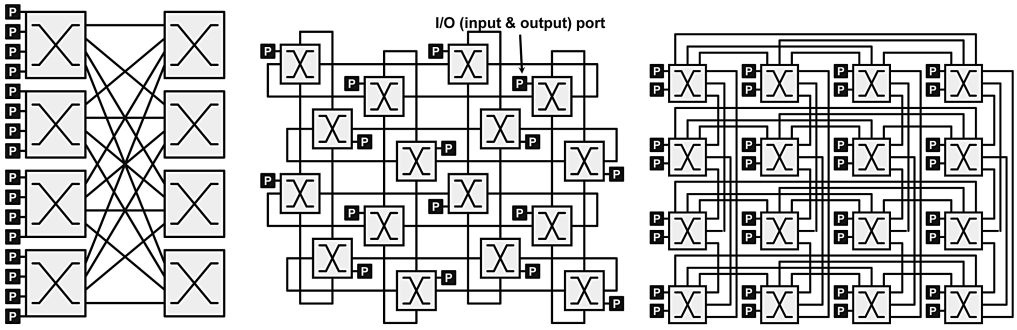
We explore the switch microarchitecture designs based on network topologies and create a network within a network switch architecture, or a network switch architecture, where a switch is composed of multiple subswitches connected through internal point-to-point channels. Each subswitch is input buffered so that arbitration can be localized to a subswitch, similar to the hierarchical switch organization. For the local network, we focus on three different topologies: folded Clos [Dally and Towles 2003], 2D torus [Dally and Towles 2003], and 2D HyperX [Kim et al. 2007a; Ahn et al. 2009]. Other topologies, such as a 2D mesh topology or a tree-based topology, can be used for the local network. However, these topologies either introduce nonuniformity (i.e., 2D mesh) or a single-point bandwidth bottleneck (i.e., tree) and are not suitable as the local network. The off-chip bandwidth is an expensive resource in large-scale systems and they need to be fully utilized. Thus, the local network cannot become the bottleneck.

We set all switch organizations to support the same aggregate off-chip channel bandwidth for a given router radix k and use the following metrics for comparison: the number of buffers and crosspoints in a switch, the aggregate fanout, and the total switch area. Each crosspoint in a crossbar and each buffer are implemented by transistors that consume leakage power regardless of switch activity, and thus more crosspoints and buffers result in higher static power. The aggregate fanout of the datapath a packet traverses in a router impacts dynamic energy as higher dynamic energy is consumed with higher fanout.

The block diagrams of the alternative local networks are shown in Figure 4. For the folded 2D torus organization (Figure 4(b)), we assume a $\sqrt{k}$-ary 2-cube network with $\sqrt{k} \times \sqrt{k} = k$ subswitches. The network has $4\sqrt{k}$ channels crossing its bisection, so the bandwidth of each internal channel must be $\frac{k}{4\sqrt{k}} = \frac{\sqrt{k}}{4}$ times higher than the bandwidth of each external channel connected to an I/O port. We assume that each subswitch dedicates a port to an I/O port and a single logical internal channel consists of $\frac{\sqrt{k}}{4}$ physical channels, thus, the radix of a subswitch becomes $\frac{\sqrt{k}}{4} \times 4 + 1 = \sqrt{k} + 1$. We choose Valiant's routing algorithm [Valiant 1982], which gives $\sqrt{k}$ as the average hop count. As a result, the folded 2D torus switch has $k\sqrt{k}$ subswitch buffers, the aggregate fanout is proportional to k, and both switch area and the number of crosspoints are proportional to k^2.

For the folded-Clos switch, we assume that it is an $(m, \frac{k}{r}, r)$ Clos network and $m = \frac{k}{r}$ (Section 5.3 discusses the folded-Clos switches with a higher m). This 2-level local network consists of r radix-$\frac{2k}{r}$ bottom-level subswitches (which are connected to the

[2]In this article, we define fanout as the number of output ports that an input port has to drive in a component, such as a crossbar, a mux, or a bus. When a switch consists of multiple components, we add the fanout values of all the components that a packet passes in the switch and call it an aggregate fanout. In the hierarchical switch, the broadcast bus delivers the packet to $\frac{k}{p}$ row buffers, a $p \times p$ subswitch, and a column multiplexer so that the aggregate fanout is $\frac{k}{p} + p + 1$.



(a) folded Clos (b) folded 2D torus (c) 2D HyperX

Fig. 4. (a) Folded-Clos, (b) folded 2D torus, and (c) 2D HyperX switch organization. **P** represents an I/O port and each crossbar represents a subswitch.

Table I. Router Microarchitecture Complexity Analysis

	Subswitch Buffers	Aggregate Fanout	Total Crosspoints	Switch Area
Canonical Crossbar	N/A	$k-1$	k^2	k^2
Hierarchical Crossbar	$2\frac{k^2}{p}$	$\frac{k}{p}+p+1$	k^2	$\frac{k}{p}(k^2+2p^2)$
Folded-Clos	$2k$	$\frac{3k}{r}+r-2$	$(rk+\frac{4k^2}{r})$	$\frac{3}{2}k(\frac{4k}{r}+k+r)$
Folded 2D Torus	$k^{\frac{3}{2}}$	$\frac{3}{4}k+\frac{5}{4}k^{\frac{1}{2}}-1$	$k(2k^{\frac{1}{2}}+\frac{3}{4}k)$	$\frac{9}{4}k(k^{\frac{1}{2}}+1)^2$
2D HyperX	$4k^{\frac{2}{3}}(k^{\frac{1}{3}}-1)$	$25(k^{\frac{1}{3}}-1)$	$(5(k^{\frac{1}{3}}-1)k^{\frac{1}{3}})^2$	$(\frac{3}{2}(5k^{\frac{1}{3}}-4)k^{\frac{1}{3}}+k)^2$

k is the radix of the switches, p is the radix of the subswitch crossbars in a hierarchical crossbar switch, and r is the radix of the top-level subswitches in a folded-Clos switch.

I/O ports) and $\frac{k}{r}$ radix-r top-level subswitches, so there are $rk+\frac{4k^2}{r}$ crosspoints and the aggregate fanout is up to $\frac{3k}{r}+r-2$. We place the top-level subswitches in the middle and half of the bottom-level subswitches on each side to provide a better aspect ratio, as shown in Figure 8(b). The switch area is dominated by global wires that are used for channels connecting subswitches, which scales at the rate of k^2.

As for the 2D HyperX, we assume $K=2$ and $S=T$, so that a channel between subswitches has twice the bandwidth of a channel to an I/O port. This switch is a ($L=2$, $S=k^{\frac{1}{3}}$, $K=2$, $T=k^{\frac{1}{3}}$) HyperX network. The subswitch radix is $5k^{\frac{1}{3}}-4$ and there are $k^{\frac{2}{3}}$ subswitches, so the local network has about $4k$ subswitch buffers and $25k^{\frac{4}{3}}$ crosspoints. We assume a minimal or Valiant's routing algorithm for the 2D HyperX switch so that the worst-case aggregate fanout becomes $25(k^{\frac{1}{3}}-1)$. The switch area is dominated by global wires as well so that it scales at the rate of k^2.

We summarize the complexity of the canonical crossbar, hierarchical crossbar, folded-Clos, folded 2D torus, and 2D HyperX switch organizations in Table I. The switch area is normalized to the square of a channel pitch, and includes the area of buffers, crossbars, and the global wires used for the channels between subswitches. Even though a channel of a high-radix router is narrower than that of a low-radix router, we assume that each channel is wide enough so that the control logic of the router represents a very small fraction of the total area and ignore its area [Wang et al. 2002]. Figure 5 shows the relative subswitch buffers, aggregate fanout, total crosspoints, and switch area of the switch organizations as the switch radix is increased from 8 to 128. The results are normalized to the values of the hierarchical crossbar. We set $p=\sqrt{k}$ and $r=2\sqrt{k}$ to minimize aggregate fanout and switch area.

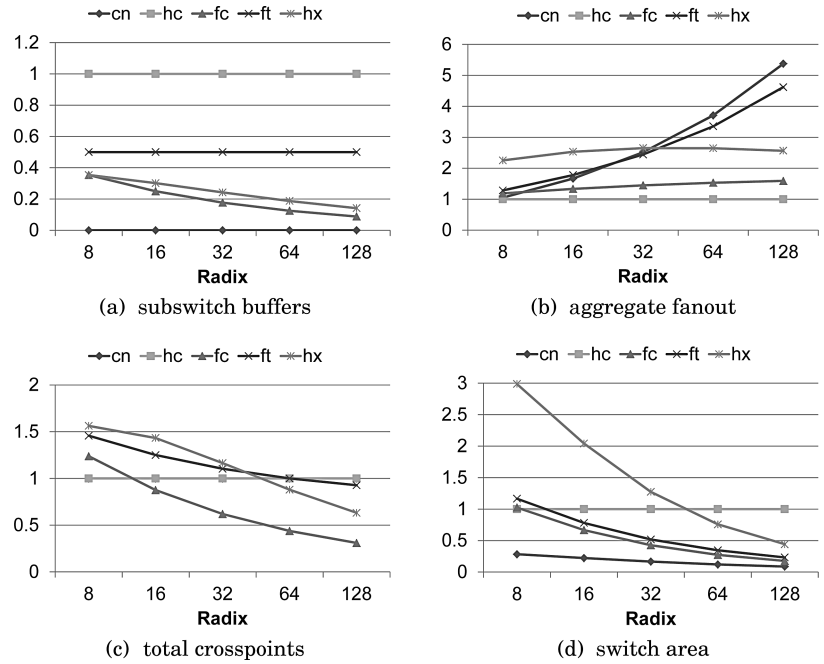


Fig. 5. The relative (a) switch buffers; (b) aggregate fanout; (c) total crosspoints; and (d) switch area of the five switch organizations in Table I. cn, hc, fc, ft, and hx stand for canonical crossbar, hierarchical crossbar, folded-Clos, folded 2D torus, and 2D HyperX switches. The hierarchical crossbar switch design is used as a baseline.

The network switch architectures, especially the folded-Clos switch designs, scale better than the canonical crossbar and the hierarchical crossbar switches in terms of power and energy efficiency as the switch radix increases. All the network switch organizations require fewer subswitch buffers than the hierarchical crossbar switch, and also have fewer crosspoints as the switch radix is 64 or higher. The results show that the switches adopting torus, folded-Clos, or HyperX topologies consume less static power than the hierarchical crossbar organization on high-radix designs. These network switches require no broadcast channels through global wires and result in a decrease of relative switch area as the switch radix is increased. These switches have higher aggregate fanout values than the hierarchical crossbar switch, consuming more dynamic power. However, static power consumption is more important for high-radix routers since most of the crosspoint transistors do not drive wires. For example, in a canonical crossbar, only k out of k^2 crosspoints are active and the remaining $k^2 - k$ ones consume static power only, evidenced by the YARC design where static power dominates dynamic power [Scott et al. 2006]. The torus switch scales worse than the folded-Clos switch in all aspects because it has higher hop count and requires wider channels between subswitches, which are implemented by multiple physical channels, to provide the same bisection bandwidth. Table I shows that the HyperX switch has a lower exponent value than the folded-Clos switch on the aggregate fanout and the number of crosspoints. Therefore, the 2D HyperX switch has better scalability, but its large coefficients make the folded-Clos switch become more area and power efficient on the radices we consider. The performance of these alternative switch networks is compared in Section 5.

4. EXPLOITING GLOBAL NETWORK TRAFFIC CHARACTERISTICS TO ROUTER DESIGNS

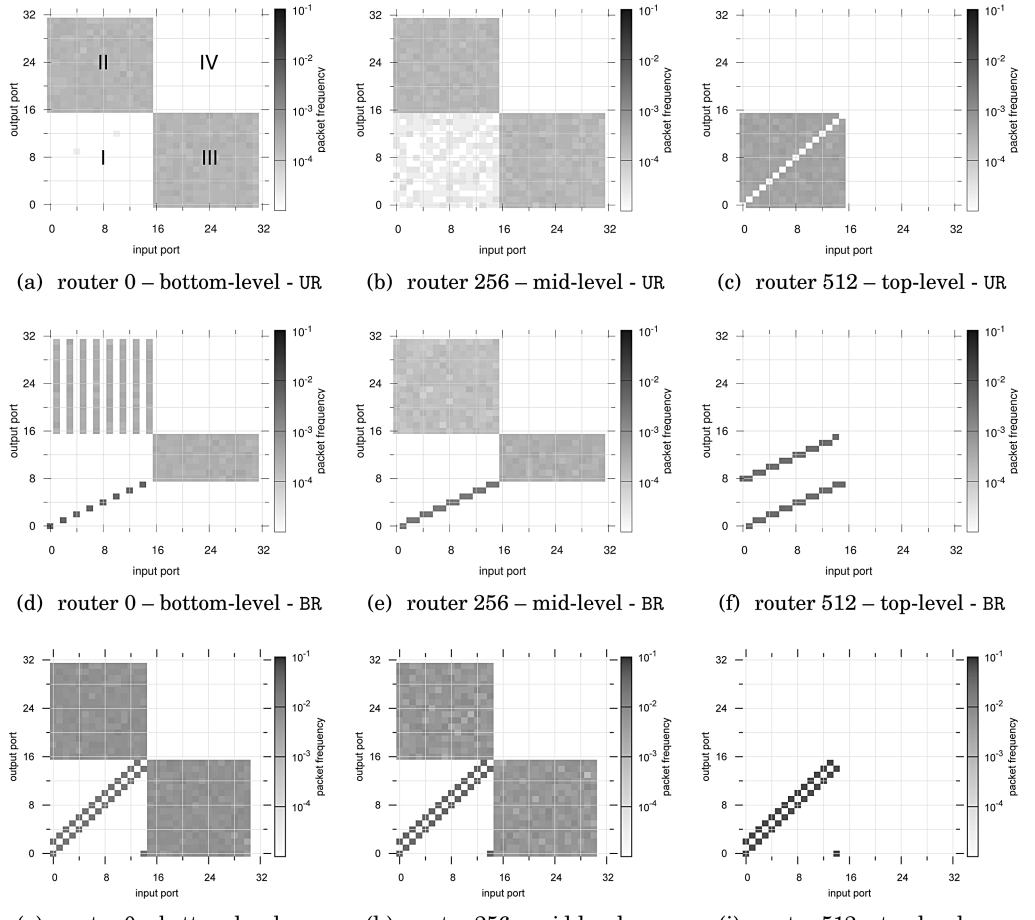
In the previous section, we identified the limitations of single-stage and hierarchical crossbar switch organizations and explored alternative router switch microarchitectures using a network organization. Our analysis showed that a switch design based on the folded-Clos network scales better in terms of energy efficiency and die area, compared to other alternatives, as the router radix increases. In this section, we further reduce the cost of a high-radix switch microarchitecture by exploiting the traffic characteristics of the global network and its impact on the local network. We make the observation that the traffic of the global network, based on its topology and routing, can be exploited to further reduce the cost of the switch microarchitecture. We first describe the impact of the global network traffic on the local network and then describe *bilateral butterfly*, a novel switch organization that exploits the characteristics of the global network traffic for high-radix routers.

4.1. Network Traffic Patterns Projected on High-Radix Routers

We first examine the traffic patterns projected on the high-radix routers in a folded-Clos global network. In Figure 6, we show the heatmap of traffic on a noninterfering folded-Clos network. The results are shown for a 3-level 4096-node network with radix-32 routers ($k = 32$), but the observations made in this section can be generalized to other folded-Clos configurations. The heatmap shows the frequencies of packets moving from input ports (x-axis) to output ports (y-axis) at routers in different levels of a folded-Clos network on the `UR` (uniform-random), `BR` (bit-rotation), and `mg` (multigrid from NAS Parallel Benchmark [Jin et al. 1999]) traffic patterns. On each router, the ports 0 to $\frac{k}{2} - 1$ are connected to lower-level routers (or terminal nodes at level-0 routers) while ports $\frac{k}{2}$ to $k-1$ are connected to upper-level routers. The communication characteristics can be partitioned into four quadrants, as shown in Figure 6(a). Each quadrant represents different types of traffic. Quadrant II represents uprouting traffic, where the packets are routed to the upper-level routers, while quadrant III is downrouting traffic, where the packets are routed to the lower-level routers, in the folded-Clos network. Quadrant I represents traffic "turning" in the network, that is, a packet arrives at the nearest common ancestor between the source and the destination and switches from uprouting to downrouting. However, quadrant IV corresponds to the packets that would switch from downrouting to uprouting. This is not possible in the folded-Clos network regardless of the traffic pattern because a packet that starts downrouting will continue to be downrouted until it reaches its destination. The heatmaps reflect this by showing that quadrant IV is not used in any routers.

Folded-Clos networks use random routing or its variants [Kim et al. 2006a] as a packet chooses one of the nearest common ancestors between its source and destination with equal probability[3]. Because of this load-balancing in the global network, the uprouting within a switch is load-balanced and results in uniformly utilizing the output ports (quadrant II). Similarly, regardless of the traffic pattern, the randomization in the uprouting leads to the input ports used in downrouting to be utilized with equal probability (quadrant III). However, the characteristics of the quadrant I traffic depend on the traffic patterns of the network. On `UR`, most packets reach top-level routers and fewer packets "turn" in middle- and bottom-level routers. Thus, quadrant I is more heavily utilized in upper-level routers. On `BR`, the bit-rotating traffic pattern of the network is reflected only in quadrant I, as shown in Figure 6(d,e,f).

[3]In some systems where ordering is required, deterministic routing might be used, even for uprouting. However, some form of load-balancing is used to spread traffic, for example, Cray BlackWidow system used hashing based on the address for traffic that needs to be routed deterministically [Scott et al. 2006].



(a) router 0 – bottom-level - UR (b) router 256 – mid-level - UR (c) router 512 – top-level - UR

(d) router 0 – bottom-level - BR (e) router 256 – mid-level - BR (f) router 512 – top-level - BR

(g) router 0 – bottom-level - mg (h) router 256 – mid-level - mg (i) router 512 – top-level - mg

Fig. 6. 2D heatmaps showing the frequencies of packets moving from the input ports (in the x-axis) to the output ports (in the y-axis) at (a,d,g) a bottom-level, (b,e,h) a mid-level, and (c,f,i) a top-level radix-32 router of a 4096-node folded-Clos network. (a,b,c), (d,e,f), and (g,h,i) are on the UR, BR, and mg traffic, respectively.

The router switch needs to provide high performance for traffic "turning" (quadrant I traffic) while others (quadrant II and quadrant III traffic) are load-balanced by the global network. The *local* network also does not need to load-balance the traffic. In comparison, based on the global network, the quadrant IV traffic is nonexistent. We take advantage of these observations to further optimize the switch microarchitecture and propose the bilateral butterfly switch microarchitecture in Section 4.2.

Figure 7 shows the heatmaps of the routers in a 8-ary 4-flat flattened butterfly network on the UR, BC (bit complement), and mg traffic. The network can be described as ($L = 3, S = 8, K = 1, T = 8$) using the HyperX notation. There are four port groups: 8 ports connected to the terminals and 8 ports for each of the three dimensions. These 4 groups of ports on both input and output ports form 16 tiles that have different utilization characteristics. Similar to the folded-Clos network, the utilization is dependent on the traffic pattern. In addition, if load-balanced routing is not used (e.g., minimal routing), the utilization differs compared to Valiant's routing, which randomizes all traffic patterns. The characteristics of the tiles in the diagonal depend on the traffic

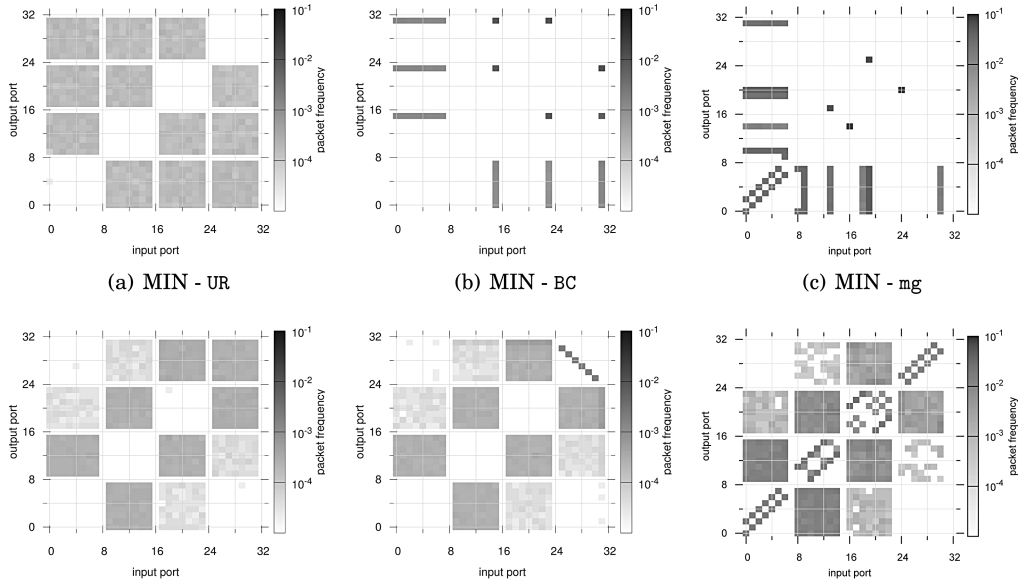


Fig. 7. 2D heatmaps showing the frequencies of packets moving from the input ports (in the x-axis) to the output ports (in the y-axis) at a radix-32 router of a 4096-node flattened butterfly network on the UR, BC, and mg traffic adopting (a,b,c) a minimal and (d,e,f) a Valiant's routing algorithms.

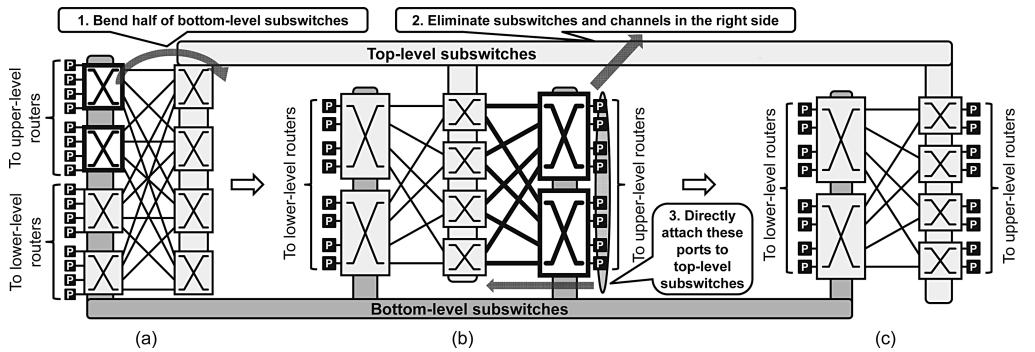


Fig. 8. Bottom-level subswitches connected to upper-level routers and global wires connected to these subswitches (highlighted with thick lines) in a folded-Clos switch are eliminated forming a bilateral butterfly network. All ports are bidirectional.

patterns of the network. In the following section, we describe switch microarchitecture that exploits these observations to further reduce the cost of a high-radix router switch.

4.2. Bilateral Butterfly Switch Architecture

The observations described in Section 4.1 enable further optimization for the folded-Clos switch design used in the folded-Clos global network. Figure 8 illustrates how some of the subswitches in the folded-Clos switch architecture can be removed. Figure 8(a) shows a folded-Clos switch to be transformed, which is the same as Figure 4(a). We assume that the lower half of the router ports are connected to lower-level routers and the upper half are connected to upper-level routers in the global network. We first bend the upper half of the bottom-level subswitches in the router such that the

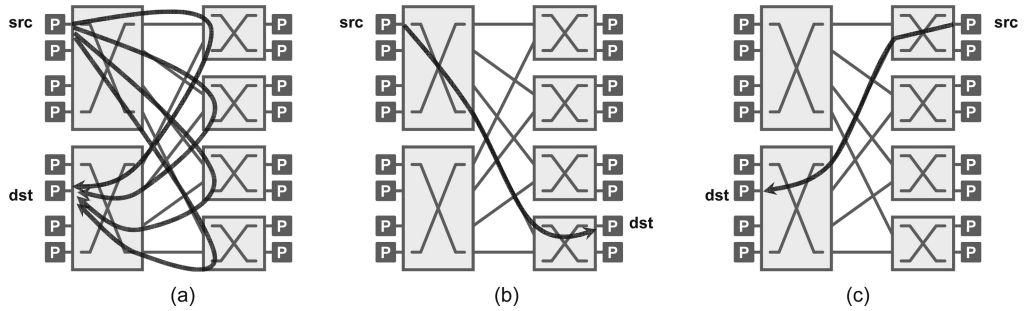


Fig. 9. Routing example of the bilateral butterfly for packets (a) turning; (b) uprouting; and (c) downrouting.

top-level subswitches are located in the middle (Figure 8(b)). Then we eliminate the bottom-level subswitches and attach intersubswitch channels in the rightmost column and connect the ports to the upper-level routers on the right side directly to the top-level subswitches. This forms a *bilateral butterfly* switch (Figure 8(c)) where the switch resembles a butterfly switch but all ports are bidirectional. The middle stages of a Clos network are needed for load-balancing, but with *global* load-balancing achieved from the global network, the load-balancing within the *local* network is no longer needed.

Depending on the router radix, one-fourth or one-third of crossbar crosspoints are saved on each router. But more importantly, half of intersubswitch datapath wires are eliminated, providing significant energy and static power savings compared to the already efficient folded-Clos switch. The bilateral butterfly switch provides sufficient path diversity for packets that "turn" in the router (Figure 9(a)), but provides no path diversity to packets passing through the router. However, based on the traffic characteristics described in Section 4.1, no further load-balancing is needed for packets that keep moving upwards or downwards because the routing algorithms of the global folded-Clos networks already load-balance both traffic types projected on the routers. Thus, deterministic routing is sufficient (Figure 9(b,c)). For the turning packets, the same path diversity as the folded-Clos router is still available with half the number of bottom-level subswitches because the routing algorithms of the global networks do not allow the packets from the upper-level routers to turn instead of pass through. Packets entering the bilateral butterfly switch experience fewer aggregate fanout on average and this results in a lower dynamic power. The bilateral butterfly switches can also be used as top-level routers by not utilizing ports dedicated to upper-level routers or by pairing an input port at one side with an output port on the other side.

4.3. Switch Architecture for Flattened Butterfly Networks

The traffic patterns projected on the routers in a flattened butterfly network are different from those in a folded-Clos network. A folded-Clos network has the same saturation throughput regardless of traffic patterns. However, that of a flattened butterfly network depends on traffic patterns so that the network has higher saturation throughput on benign traffic, such as a uniform-random traffic pattern, compared to adversarial traffic patterns. The flattened butterfly networks typically exploit adaptive routing algorithms, such as UGAL [Singh 2005], Adaptive Clos [Kim et al. 2007b], and DAL [Ahn et al. 2009], to make the saturation throughput on adversarial traffic patterns be half of that on the uniform-random traffic by misrouting packets and turning the traffic into two separate uniform random patterns. Hereafter, we focus on the uniform random pattern projected on the routers in a flattened butterfly network.

The way of assigning the external ports of a router to its subswitches determines the characteristics of the traffic between the subswitches within the router of a flattened

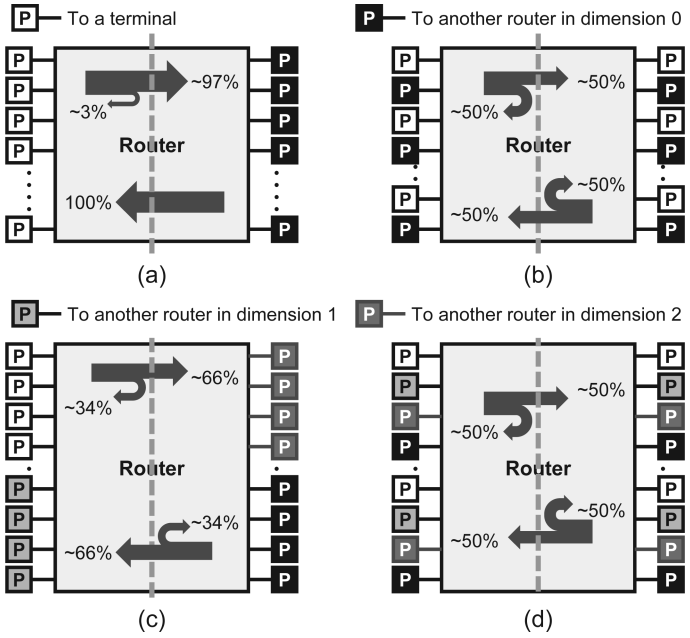


Fig. 10. The ratio of the packets crossing the bisection of a router to the ones injected to the router in (a,b) a 1D flattened butterfly network and (c,d) a 3D flattened butterfly network. The ports connected to other routers belonging to the same dimension are tied together in (a,c), while they are interleaved such that the ones to different dimensions and terminals are grouped in (b,d).

butterfly network (Figure 10). We assume a minimal routing algorithm for the global flattened butterfly networks that have $\beta = 0.5$ and use radix-64 routers. For a 1D flattened butterfly network, if we connect the ports at one side of a router to terminals (terminal ports) and the ones at the other side to other routers (router ports), most of the packets injected to the router cross its bisection (Figure 10(a)). However, if we interleave the terminal ports and the router ports such that both are evenly distributed at both sides, only half of the injected packets cross the bisection (Figure 10(b)). Interleaving the terminal and router ports halves the number of packets crossing the router bisection for a 3D flattened butterfly network as well (Figure 10(c,d)). The probability distribution of choosing the next dimension for a packet in a certain dimension depends on the topological distance between the source and destination of the packet and the routing algorithm of the network, and therefore it is not uniformly distributed. Once the next dimension is determined, the packet proceeds to any node in the dimension with equal probability because the traffic pattern is uniform random. As a result, it is possible to use flattened butterfly switches with $\beta = 0.5$ for global flattened butterfly networks with $\beta = 0.5$. The evaluation in Section 5.4 shows that this flattened butterfly switch design provides more area- and energy-efficient router microarchitectures than the alternatives.

5. EVALUATION

We evaluated the performance and efficiency of the proposed switch organizations using various traffic patterns on single-router and global network configurations. Single-router results show that the folded-Clos switch design performs better than other network switch designs, but still performs worse than the hierarchical crossbar design. The performance of the folded-Clos switch was further improved by adding more top-level subswitches, providing input speedup on bottom-level subswitches, or both. This

fine-tuning made the folded-Clos switch perform comparable to the hierarchical crossbar design and more energy efficient. The bilateral butterfly switch and the HyperX switch with less bisection bandwidth further improved the performance and energy-efficiency of the global networks. We used a 22nm process technology for quantitative evaluations, but the observations made here can be applied to other future high-radix network designs as well.

5.1. Experimental Setup

We developed an event-driven cycle accurate simulator which supports single-stage crossbars, hierarchical crossbars, and folded torus for local networks and folded-Clos and flattened butterfly for both local and global networks. The folded-Clos networks use an adaptive routing algorithm that randomly picks the two nearest common ancestors between the source and destination of a packet and chooses one with less congestion as an intermediate node, which is the same as the *greedy_r*(2) algorithm in Kim et al. [2006a]. The flattened butterfly and folded 2D torus networks use Valiant's routing [Valiant 1982] or minimal routing algorithm [Ahn et al. 2009]. A packet that arrives at a router proceeds through the steps of route computation, virtual channel allocation, switch allocation, and switch traversal. If the router has a network switch, it performs both the routing of the global and the local network during the route computation step so that the subswitches within the router only have subswitch allocation and traversal steps. Routers and subswitches are input buffered and each buffer has 4 Virtual Channels (VCs). iSLIP [McKeown 1999] separable allocation method is used for VC and switch allocation. We assumed virtual cut-through flow control, which was also used in the YARC router [Scott et al. 2006]. We warmed up the network before measuring the performance of the networks. Unless mentioned otherwise, switch speedup was not used. Single-flit packets were used by default.

We applied the 22nm predictive technology model (PTM [Zhao and Cao 2006]) for transistors in SRAMs, crossbars, and interconnection wires. We extrapolated the wiring design rules of the 65nm PTM and assumed that the wires used to connect subswitches and compose crossbars have 0.4um pitch, 0.35um thickness, 0.15um height, 0.002Ω/□, and 0.3fF/um. We targeted 5 GHz operating frequency for buffers, crossbars, and global wires. The bandwidth of each channel was 320Gbps. Even though it was infeasible to operate the crossbar in cn at 5 GHz, we assumed that it can be pipelined and has the same zero-load latency as hc. SPICE simulation results showed that 240fJ was needed to send a bit over a 1-cycle distance (2.5mm in 5 GHz), 7fJ for reading or writing a bit on a 128kb input port buffer and an 8kb subswitch buffer, and 30fJ per radix for sending a bit through a crossbar. As for static power, 20fJ was needed for a 1-cycle distance global wire, 4fJ per radix for each crosspoint per cycle, 11.2mW and 0.7mW on an input port buffer and a subswitch buffer, and 1pJ for transferring a bit off-chip. We modeled multiplexer-based crossbars for subswitches, which were shown to be more area and power efficient than matrix crossbars [Shamim 2009]. 5 GHz operating frequency was achievable for both 8-port and 16-port subswitches, while both types have 2 cycles of latency between an input and an output port. Propagation delay was 2 cycles for internal channels between subswitches, except for hc, where it became 4 cycles for horizontal broadcast channels due to the relatively large area of hc, as shown in Table II. These values are similar to the ones Joshi et al. [2009] estimated on a 22nm process. The size of a flit was assumed to be 128bits, which translates to 2data path cycles (128b/320Gbps = 0.4ns). We assume that the control path, such as switch allocators for the 8-port and 16-port crossbars, is not in a critical path, as 2 datapath cycles are enough for the allocation [Passas et al. 2012].

We used both synthetic traffic patterns and the network message traces from NAS Parallel Benchmarks (NPB [Jin et al. 1999]) to evaluate the performance of

Table II. Router Design Complexity Analysis on a Radix-64 Router

		Subswitch Buffers	Aggregate Fanout	Total Crosspoints	Switch Area
Canonical Crossbar		N/A	63	4,096	4,096
Hierarchical Crossbar		1,024	17	4,096	33,792
Folded-Clos	fc	128	26	1,664	9,216
	fc5	160	27	2,032	13,680
	fc6	192	28	2,400	19,008
	fc5-isu2	160	35	3,376	20,496
	fc6-isu2	192	36	3,872	24,640
Folded 2D Torus		512	45∼75	4,096	11,664
2D HyperX		192		3,600	25,600

$n = 4$, $p = 8$, $r = 16$, while m is varied.

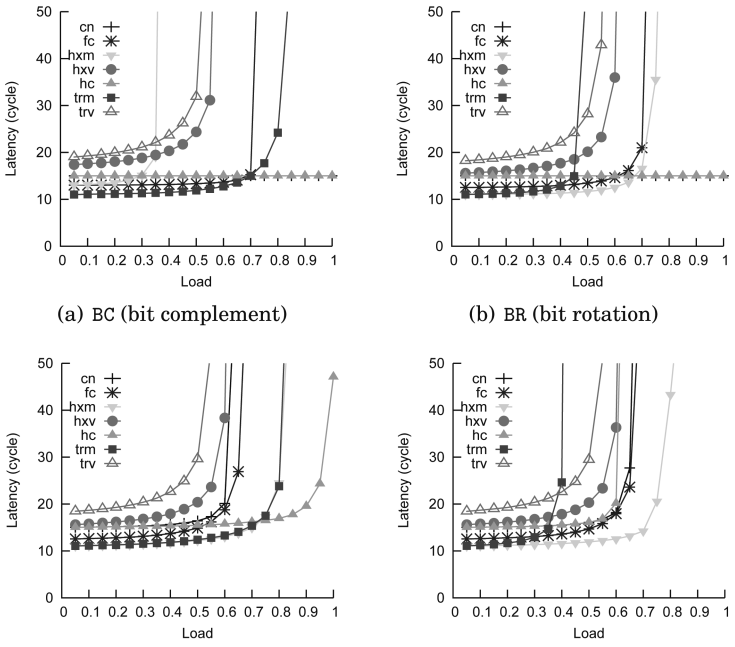
network-switch organizations. We used various synthetic traffic patterns, including two bit permutation patterns (BC and BR) and two random traffic patterns (UR and TR). BC is a bit-complement traffic pattern, BR is a bit-rotation traffic pattern, UR is a uniform-random traffic, and TR is a transpose-random traffic, where a node in a row A of a square matrix of network terminals is equally likely to send packets to nodes in a column A of the square matrix. Note that TR is one of the worst-case traffic patterns for the hierarchical crossbar [Kim et al. 2005a]. We also tested other traffic, but they showed similar performance trends to BC, BR, UR, and TR patterns. Their results were not included due to space limitation, except in Section 5.3. We ran the OpenMP implementations of NPB, collected the message passing traces, and fed them into the simulator. We used the problem size of class B. As for the application network message traces, we divided a long message into multiple packets, each of which has a header flit and 64B data flits. We simulated 2 million packets for each trace.

5.2. Comparing the Performance of High-Radix Router Microarchitectures

We set the bisection bandwidth of each high-radix router equal to its aggregate off-chip incoming channel bandwidth and compared the performance of the router microarchitectures. The performance comparison showed that all the organizations suffered the Head-of-Line (HoL) blocking problem [Karol et al. 1987], but the network switch designs were more often susceptible to the HoL blocking problem.

We evaluated the following router switch microarchitecture organizations: canonical crossbar (cn), folded-Clos (fc), HyperX with Valiant's routing (hxv), HyperX with minimal routing (hxm), folded torus with Valiant's routing (trv), folded torus with minimal routing (trm), and hierarchical crossbar (YARC) switches (hc). Figure 11 shows the average latency of injected packets over offered loads of radix-64 switches on BC, BR, UR, and TR patterns. The subswitch radix in hc (p) is 8 and the top-level subswitch radix in fc (r) is 16. Crossbars have no input speedup. All switch designs used 4 VCs, where hxv and trv used half of them to send packets to random intermediate subswitches and the others to deliver them to destinations ports. hxm chose the VC channel based on the number of hops until the destination.

The canonical crossbar design performs ideally on any permutation pattern, such as BC and BR, but it experiences HoL blocking on random traffic similar to other network switch designs and its achieved throughput saturates around 66%. fc, trv, trm, hxv, and hxm all suffer the HoL blocking problem. Under low loads, fc has a lower average latency than hc because fc is smaller so it has lower global wire propagation delays. The average latency of hxm is the lowest, except for the BC traffic pattern, because there are packets with minimum paths whose hop counts are lower than 3. However, it



(a) BC (bit complement) (b) BR (bit rotation)

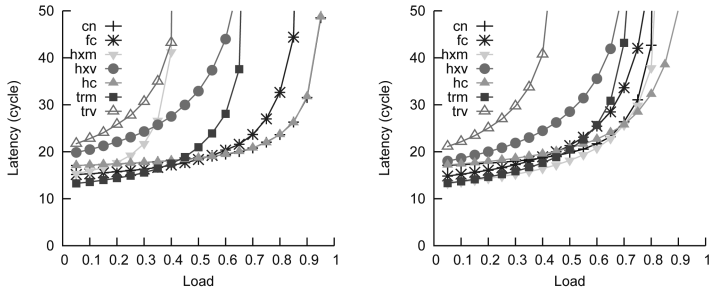
(c) UR (uniform random) (d) TR (transpose random)

Fig. 11. Load-latency graphs of radix-64 switches on (a) BC; (b) BR; (c) UR; and (d) TR traffic patterns. cn, fc, hxv, hxm, trv, trm, and hc stand for canonical crossbar, folded-Clos, HyperX with Valiant's routing, HyperX with minimal-adaptive routing, folded torus with Valiant's routing, folded torus with minimal routing, and hierarchical crossbar switches. There was no input speedup on crossbars.

saturates quickly on adversarial traffic compared to hxv, so adaptive routing algorithms combining the benefits of both can be beneficial, such as UGAL [Singh 2005], Adaptive Clos [Kim et al. 2007b], and DAL [Ahn et al. 2009]. hc performs almost ideally on BC, BR, and UR, even without the crossbar input speedup owing to its property of distributing loads from a port to 8 subswitches on random traffic [Kim et al. 2005a]. As a result, the average load in a subswitch is 0.125, even with full loads from the ports due to this effective output speedup through broadcast. This load distribution does not happen at the TR traffic pattern and hc suffers the same HoL blocking problem (Figure 11(d)).

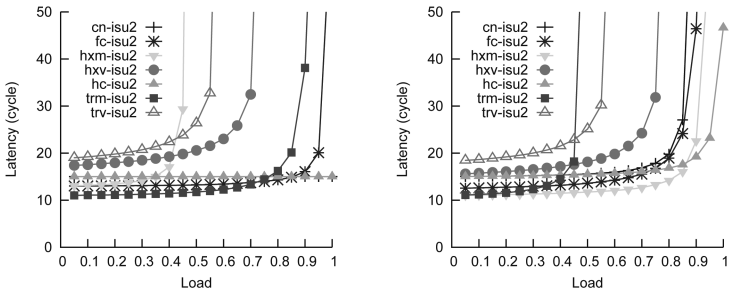
A larger packet size leads to higher saturation throughput of folded-Clos, folded 2D torus, and 2D HyperX switches in general because it effectively increases the number of switch allocation steps. Figure 12 shows the load-latency plot of radix-64 switches for bimodal traffic pattern where half of the traffic is 1-flit packets and the other half is 5-flit packets. The bimodal traffic is representative of request-reply traffic. On both BC and UR, the performance of the folded-Clos switches is enhanced so that their saturation throughputs get closer to those of hc. Other traffic patterns show similar trends but are omitted due to space limitation.

Another way of improving the saturation throughput of the switch designs is through speeding up the input ports of the subswitch crossbars. Load-latency graphs in Figure 13 show that the performance of folded-Clos and 2D HyperX switches improves noticeably but is still outperformed by the hierarchical crossbar switches. Considering the high area and energy-efficiency overhead imposed by speeding up the crossbars, crossbar speedup should be applied carefully, which is further explored in the following subsection.



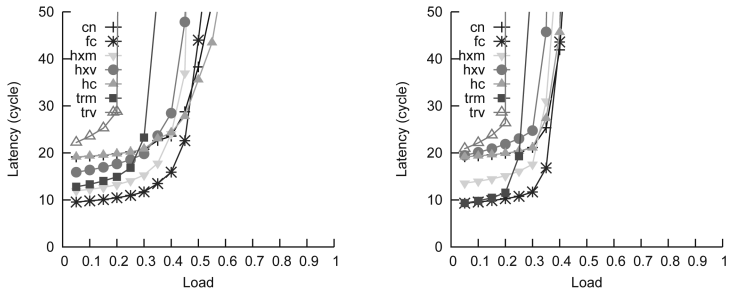
(a) BC (bit complement) (b) UR (uniform random)

Fig. 12. Load-latency graphs of radix-64 switches with bimodal traffic consisting of 1-flit and 5-flit packets on (a) BC and (b) UR traffic patterns.



(a) BC (bit complement) (b) TR (transpose random)

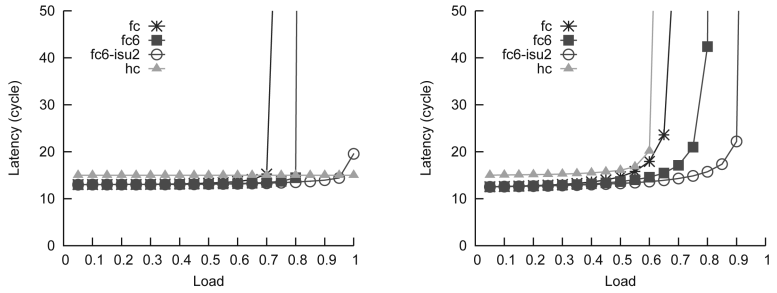
Fig. 13. Load-latency graphs of radix-64 switches on (a) BC and (b) TR traffic patterns. -isu2 stands for the input speedup of 2 for each crossbar.



(a) mg (b) lu

Fig. 14. Load-latency graphs of radix-64 switches on (a) mg and (b) lu NPB message passing traces.

Figure 14 compares the performance of the router designs on the NPB message passing traces. Larger packet sizes, time-varying nature of traffic, and the existence of hotspots cause the difference in the load-latency graphs between the application traces and the synthetic patterns, but both share similar trends. The folded-Clos switches have the lowest latency under low loads and their saturation throughputs are comparable to those of the hierarchical crossbar switches on both mg and lu. lu also has hotspots such that even the throughput of the flattened butterfly router with Valiant routing saturates before the load reaches 40% of its bisection bandwidth.



(a) BC (bit complement) (b) TR (transpose random)

Fig. 15. Load-latency graphs of radix-64 switches with more top-level subswitches and selective input speedup on (a) BC and (b) TR traffic patterns.

5.3. Exploring the Design Space of the Folded-Clos Switch Architecture

The folded-Clos switch design outperforms the 2D HyperX design, but the performance does not match that of the hierarchical crossbar design. Since fc is smaller and dissipates less static power than hc with comparable dynamic energy consumption, additional performance-cost trade-off can be made with fc to increase its performance.

We achieved this trade-off by putting more top-level subswitches (increasing m) and speeding up bottom-level subswitches for the folded-Clos switch designs. We have assumed $m = n$ but with $m > n$, an output speedup of $\frac{m}{n}$ in the input-stage subswitches is provided such that $\frac{n}{m}$ becomes the maximum load of a top-level subswitch. Since top-level subswitches are chosen randomly, fc with a higher m experiences adversarial traffic less frequently than hc. Further performance improvement is available by speeding up bottom-level subswitches, whose speedup overhead is lower than top-level ones since bottom-level subswitches typically have a lower radix. Figure 15 shows the performance improvement on both BC and TR patterns, where fc6-isu2 performs comparable to hc on BC and even fc6 outperforms hc on TR. fc$m$ stands for a folded-Clos switch with m top-level subswitches and fc$m$-isu2 is fc$m$ with an input speedup of 2 only to bottom-level subswitches. Table II compares their design complexity. The complexity of the folded-Clos switch increases as more speedup techniques are applied, but even with fc6-isu2, its complexity is still lower than that of hc.

In Figure 16, we compared the energy consumption, Energy-Delay Product (EDP), and Energy-Delay-Area Product (EDAP) of hc, fc, fc$m$, and fc$m$-isu2. The energy consumption, EDP, and EDAP values are normalized to those of hc. Results show that fc consumes the lowest energy, but also performs the worst. It has the best EDAP as well (lower is better for the energy consumption, EDP, and EDAP metrics). For example, on TR, fc consumes 1.9× lower energy and shows 7.6× better EDAP over hc. Either fc5-isu2 or fc6-isu2 has the best energy-delay product, among which fc6-isu2 has 2.4× lower EDP than hc on TR and has 1.2× lower EDP than fc on BR. Among the NPB message passing traces, bt, cg, and mg show similar trends. The difference in performance between yc and fc is reduced compared to the synthetic patterns because of the larger packet size, which was also observed in Figure 12. sp and especially lu have hotspots and therefore their saturation throughputs are lower than others. For both synthetic patterns and application traces, folded-Clos switch designs provide better energy efficiency and energy-delay product than the hierarchical crossbar switch design. We also evaluated other traffic patterns, such as bit reverse, tornado, and 100RP [Dally and Towles 2003] where 100RP is the average of 100 random permutation traffic patterns. These traffic patterns show similar trends to UR, BC, BR, and TR. Note that some NPB

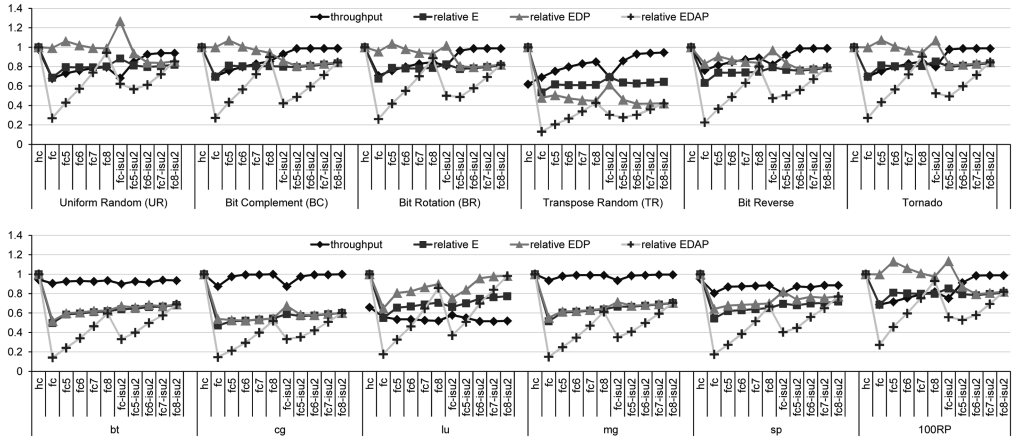
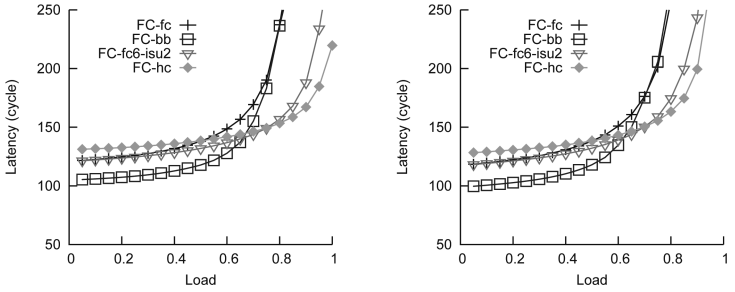


Fig. 16. Achieved throughput, relative energy, relative Energy-Delay Product (EDP), and relative Energy-Delay-Area Product (EDAP) of 11 switch designs on UR, BC, BR, TR, bit-reverse, tornado, bt, cg, lu, mg, sp, and 100RP patterns. All relative values are normalized to those of hc.



(a) BC (bit complement) (b) UR (uniform random)

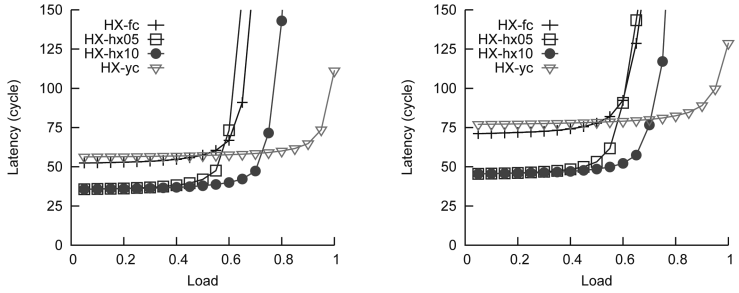
Fig. 17. Load-latency graph of folded-Clos networks with 4,096 nodes connected by radix-32 switches on (a) BC and (b) UR traffic patterns. bb stands for a bilateral butterfly switch.

applications, such as ft and is, consist mostly of an all-to-all communication pattern, whose performance resembles 100RP.

5.4. Comparing the Performance of Network Switch Designs in Global Networks

We evaluated the performance of the network within network switch designs within a global network. First, the folded-Clos topology is used as the 4,096-node global network and is simulated with the following radix-32 switch designs: fc, fc6-isu2, hc, and the bilateral butterfly (bb). 4 VCs are used for all the configurations and the port-to-port latency between two routers is assumed to be 10 cycles, which includes serialization/deserialization, synchronization, and link propagation delays.

Figure 17 shows that hc performs well. fc6-isu2 performs comparable to hc on traffic patterns including BC and UR. These results show that the cost- and energy-efficient folded-Clos switch designs perform well on the global network traffic patterns as well as the single-router traffic patterns. Considering that they scale better than the hierarchical crossbar switches and perform comparably, or even better, by adding more top-level subswitches and speeding up bottom-level subswitches, the folded-Clos switch designs are compelling router microarchitecture options.



(a) 1,024 HyperX network (b) 1,728 HyperX network

Fig. 18. Load-latency graph of HyperX global networks on the UR traffic pattern with (a) 1,024 nodes connected by radix-64 switches and (b) 1,728 nodes connected by radix-36 switches.

Without speedup, bb performs better than fc. Even with an input speedup of 2 for crossbars, bb performs better on most cases, especially when loads are low, since a packet traverses one fewer internal channels and subswitches per router in a bilateral butterfly router. Compared to the folded-Clos switch design, bb reduces the switch area by 25% and the switch power by 10% on UR when neither switch has input speedup. These results support that further cost- and energy-efficiency optimizations are possible for the routers composing high-radix folded-Clos networks.

We verified the observations with regard to global HyperX networks in Section 4.1 by applying uniform-random traffic to 1,024-node and 1,728-node networks. In Figure 18, we used the following 4 switch designs, namely, fc, hx05, hx10, and yc, where hx05 stands for a HyperX switch with $\beta = 0.5$ and hx10 is a HyperX switch with $\beta = 1.0$. The 1,024-node network uses $(1, 32, 1, 32)$ global 1D HyperX topology, $(2, 4, 1, 4)$ for hx05, $(2, 4, 2, 4)$ for hx10, and 4 VCs. The 1,728-node network uses $(2, 12, 1, 12)$ global 2D HyperX topology, $(1, 6, 1, 6)$ for hx05, $(1, 6, 2, 6)$ for hx10, $p = 6$ for hc, $r = 12$ for fc, and 6 VCs. All use minimal routing. For both 1,024-node and 1,728-node HyperX networks, hx05 performs similar to fc even if hx05 has half the bisection bandwidth of fc. Both hx05 and hx10 lead to smaller low-load latencies because a packet traverses fewer subswitches per router on average, similar to the case of bb. The area of the radix-64 hx05 router for the 1,024-node version is much smaller than that of the hx10 router and even 8% smaller than that of the area-efficient fc router. These facts reinforce the argument that it is possible to explore more efficient switch designs by exploiting the traffic-pattern characteristics of the global network and its impact on the local network design within the switch.

6. RELATED WORK

Kim et al. [2005a] showed that increasing the router radix and creating a high-radix router is more cost effective than increasing the bandwidth per port. They proposed a hierarchical router organization for high-radix routers and it was implemented in the Cray YARC router [Scott et al. 2006]. However, the hierarchical organization did not consider how the microarchitecture can exploit the global network and it has scalability limitations. In addition to electrical signaling, different technologies have been recently proposed for high-radix switches, including ones using proximity communication [Eberle et al. 2008] and optics [Binkert et al. 2011]. Mora et al. [2006] also proposed a high-radix switch organization to reduce the impact of head-of-line blocking. These studies are orthogonal to our work as we explore alternative switch microarchitectures.

To properly exploit high-radix routers, an appropriate topology is needed. The folded Clos [Dally and Towles 2003] is one such topology and the Cray BlackWidow system

Table III. Qualitative Comparison of On-Chip Networks (NoC) and this Work

	On-Chip Networks	Network Switch Microarchitectures
Terminal nodes	CPU, caches, memories, etc.	I/O ports
Bandwidth	high local bandwidth required to exploit locality on-chip	high bisection bandwidth required to provide a non-interfering network
Traffic pattern	adversarial traffic pattern can occur, based on node communication pattern	influenced by global network
Area	area dominated by end nodes	entire chip dedicated to local network

employed a variant of the folded-Clos topology [Scott et al. 2006]. Cost-efficient high-radix topologies [Ahn et al. 2009; Kim et al. 2007b, 2008] have been recently proposed to further reduce network cost. These topologies require a load-balancing routing algorithm [Singh 2005; Ahn et al. 2009; Jiang et al. 2009] to properly exploit the path diversity. In this work, we exploited the load-balancing of these high-radix topologies to optimize the switch design of a high-radix router. These load-balancing algorithms are often nondeterministic, leading to a packet reordering problem. Packet order can be guaranteed using a sliding window protocol [Peterson and Davie 2007] that reorders packets at the destination nodes using sequence numbers. In contrast, deterministic routing can be used and other techniques can be leveraged for load-balancing. For example, the BlackWidow system uses deterministic routing to maintain ordering for requests while hashing the requests across the different paths, based on the address. These methods can be applied to our network-switch organizations.

Many different hierarchical networks [Duato et al. 2002; Dally and Towles 2003] have been proposed in the literature and our work can be viewed as a hierarchical approach as well. For example, the recently proposed Dragonfly topology [Kim et al. 2009] also provides a hierarchical network. However, our work differs from prior approaches as we used a local network to create a scalable high-radix switch and the local network is created within the constraint of being placed inside a single chip. The techniques presented in Section 4.2 are not feasible for conventional, hierarchical networks. The Cray T3D router [Dally and Towles 2003] used a similar approach of creating a network of switches in the router design. Because of the technological constraints, a single switch was partitioned into three subswitches, one for each x, y, and z dimension. However, this organization can severely limit the network throughput because of the limited connection between subswitches. Choi and Pirtkston [1997] also explored these decoupled crossbar designs to support fully adaptive routers with a deadlock recovery feature.

A significant amount of research has been done recently in on-chip networks [Dally and Towles 2001; Nicopoulos et al. 2010]. We also leverage on-chip networks to build a switch. However, the architecture for a scalable high-radix switch differs from on-chip networks for multicore processors because of the different constraints on the type of terminal nodes, bandwidth requirement, traffic patterns, and switch area, as summarized in Table III. For example, a rotary router [Abad et al. 2007] replaced a crossbar into multiple smaller building blocks connected together to constitute a ring. It can be regarded as an example of the network-switch designs generalized in this article. Because the bisection bandwidth of a ring does not scale with increasing the number of nodes, it is more suitable as an NoC microarchitecture. There are several proposals to lower the complexity of crossbars within routers by exploiting the topology and routing characteristics of on-chip networks. Kim et al. [2005b] split a 5×5 crossbar into smaller components, demultiplexers, and a decomposed crossbar, for a mesh or torus network, a Row-Column Decoupled Router [Kim et al. 2006b] further saves area by substituting the decomposed crossbar into lower-radix ones, and CHIPPER [Fallin

et al. 2011] replaces a crossbar into a permutation network to lower the cost of a bufferless deflection router.

Recently, there have been proposals to implement high-radix switches using a single-stage network for on-chip networks [Satpathy et al. 2012; Passas et al. 2012]. To scale a high-radix crossbar, the swizzle switch network [Satpathy et al. 2012] used SRAM circuit techniques to scale to a 64×64 crossbar while radix-128 crossbar was designed using bit-slicing [Passas et al. 2012]. Both of the proposals improve the performance and efficiency of canonical crossbars and some of the ideas presented in these works can be leveraged in our network within network switch design—for example, the techniques can be used in the design of the local switches. However, these works targeted on-chip networks which have lower bandwidth (under 5Tb/s), compared with the bandwidth required in the high-radix routers we explore in this work, for example, a recent high-radix switch provided approximately 10Tb/s [Arimilli et al. 2010] while we assumed switches that required 20Tb/s or more of bandwidth.

7. CONCLUSION

This article approached the problem of building a high-radix router switch by treating it as a network design. We explored alternative network topologies, such as folded Clos, 2D torus, and flattened butterfly, for switch microarchitecture and compared them with the state-of-the-art hierarchical crossbar architecture. We showed that our network within a network design approach scales better than the hierarchical crossbar architecture on area and power consumption. Among the network switch designs, the folded-Clos switch design has the smallest area and dissipates the lowest power. The folded-Clos switch also enables fine-grain trade-offs between switch performance and cost/energy efficiency by adjusting the number of top-level subswitches and speeding up bottom-level subswitches. Compared to the hierarchical crossbar design, a radix-64 folded-Clos switch without speedup consumes 35% and 53% less energy on uniform-random and transpose-random traffic. A folded-Clos switch with two more top-level subswitches is 1.2 and 2.4× better in energy-delay product on uniform-random and transpose-random traffic compared with the hierarchical crossbar. To further optimize the local network, we exploit the traffic pattern of the global network and propose a *bilateral butterfly* organization that removes up to 33% of crosspoints in a folded Clos. Thus, we achieve a more area- and energy-efficient switch design for a folded-Clos network. As for a flattened butterfly global network, a flattened butterfly switch design, which is not optimal in general, provides better area and energy efficiency than a folded-Clos switch design with the same radix. In this work, we focused mainly on optimizing the local network organization based on the global network characteristics, but it remains to be seen if both the local and global network can be jointly optimized to reduce overall network cost.

ACKNOWLEDGMENT

The authors would like to thank Sungwoo Choo of NHN for his efforts on router modeling.

REFERENCES

Abad, P., Puente, V., Prieto, P., and Gregorio, J. A. 2007. Rotary router: An efficient architecture for CMP interconnection networks. In *Proceedings of the 34th International Symposium on Computer Architecture*.

Abts, D., Bataineh, A., Scott, S., Faanes, G., Schwarzmeier, J., Lundberg, E., Johnson, T., Bye, M., and Schwoerer, G. 2007. The cray black widow: A highly scalable vector multiprocessor. In *Proceedings of the ACM/IEEE Conference on Supercomputing*.

Abts, D., Marty, M. R., Wells, P. M., Klausler, P., and Liu, H. 2010. Energy proportional datacenter networks. In *Proceedings of the 37th International Symposium on Computer Architecture*.

Agarwal, A. 1991. Limits on interconnection network performance. *IEEE Trans. Parallel Distrib. Syst. 2,* 4.

AHN, J. H., BINKERT, N., DAVIS, A., MCLAREN, M., AND SCHREIBER, R. S. 2009. HyperX: Topology, routing, and packaging of efficient large-scale networks. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*.

AHN, J. H., CHOO, S., AND KIM, J. 2012. Network within a network approach to create a scalable high-radix router microarchitecture. In *Proceedings of the 18th International Symposium on High Performance Computer Architecture*.

ARIMILLI, B., ARIMILLI, R., CHUNG, V., CLARK, S., DENZEL, W., DRERUP, B., HOEFLER, T., JOYNER, J., LEWIS, J., LI, J., NI, N., AND RAJAMONY, R. 2010. The percsh high-performance interconnect. In *Proceedings of the 18th IEEE Symposium on High Performance Interconnects*.

BALFOUR, J. AND DALLY, W. J. 2006. Design tradeoffs for tiled CMP on-chip networks. In *Proceedings of the 20th International Conference on Supercomputing*.

BEYENE, W. T., MADDEN, C., KIM, N., LEE, H.-C., PEREGO, R., SECKER, D., YUAN, C., VAIDYANATH, A., AND CHANG, K. 2008. Design and analysis of a tb/sec memory system. In *Electrical Performance of Electronic Packaging*.

BINKERT, N., DAVIS, A., JOUPPI, N. P., MCLAREN, M., MURALIMANOHAR, N., SCHREIBER, R., AND AHN, J. H. 2011. The role of optics in future high radix switch design. In *Proceedings of the 38th International Symposium on Computer Architecture*.

CHOI, Y. AND PIRTKSTON, T. M. 1997. Crossbar analysis for optimal deadlock recovery router architecture. In *Proceedings of the 10th International Parallel Processing Symposium*.

DALLY, W. J. 1990. Performance Analysis of k-ary n-cube interconnection networks. *IEEE Trans. Comput. 39,* 6.

DALLY, W. J. AND TOWLES, B. 2001. Route packets, not wires: On-chip interconnection networks. In *Proceedings of the 38th Design Automation Conference*.

DALLY, W. J. AND TOWLES, B. 2003. *Principles and Practices of Interconnection Networks*. Morgan Kaufmann, San Fransisco, CA.

DUATO, J., YALAMANCHILI, S., AND LIONEL, N. 2002. *Interconnection Networks: An Engineering Approach*. Morgan Kaufmann, San Fransisco, CA.

EBERLE, H., GARCIA, P. J., FLICH, J., DUATO, J., DROST, R., GURA, N., HOPKINS, D., AND OLESINSKI, W. 2008. High-radix crossbar switches enabled by proximity communication. In *Proceedings of the ACM/IEEE Conference on Supercomputing*.

FALLIN, C., CRAIK, C., AND MUTLU, O. 2011. CHIPPER: A low-complexity bufferless deflection router. In *Proceedings of the 17th International Symposium on High Performance Computer Architecture*.

HO, R., MAI, K. W., AND HOROWITZ, M. A. 2001. The future of wires. *Proc. IEEE 89,* 4.

HOELZLE, U. AND BARROSO, L. A. 2009. *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines* 1st Ed. Morgan and Claypool.

JIANG, N., KIM, J., AND DALLY, W. J. 2009. Indirect adaptive routing on large scale interconnection networks. In *Proceedings of the 36th International Symposium on Computer Architecture*.

JIN, H., FRUMKIN, M. A., AND YAN, J. C. 1999. The OpenMP implementation of naas parallel benchmarks and its performance. Tech. rep. NAS-99-011, NASA Ames Research Center.

JOSHI, A., BATTEN, C., KWON, Y.-J., BEAMER, S., SHAMIM, I., ASANOVIC, K., AND STOJANOVIC, V. 2009. Silicon-photonic clos networks for global on-chip communication. In *Proceedings of the 3rd International Symposium on Networks-on-Chip*.

KAROL, M. J., HLUCHYJ, M. G., AND MORGAN, S. P. 1987. Input versus output queueing on a space-division packet switch. *IEEE Trans. Comm. 35,* 12.

KIM, J., BALFOUR, J., AND ABTS, D. 2007a. Flattened butterfly topology for on-chip networks. In *Proceedings of the 40th IEEE/ACM International Symposium on Microarchitecture*.

KIM, J., DALLY, W. J., AND ABTS, D. 2006a. Adaptive routing in high-radix clos network. In *Proceedings of the ACM/IEEE Conference on Supercomputing*.

KIM, J., DALLY, W. J., AND ABTS, D. 2007b. Flattened butterfly: A cost-efficient topology for high-radix networks. In *Proceedings of the 34th International Symposium on Computer Architecture*.

KIM, J., DALLY, W. J., SCOTT, S., AND ABTS, D. 2008. Technology-driven, highly-scalable dragonfly topology. In *Proceedings of the 35th International Symposium on Computer Architecture*.

KIM, J., DALLY, W. J., SCOTT, S., AND ABTS, D. 2009. Cost-efficient dragonfly topology for large-scale systems. *IEEE Micro 29,* 1.

KIM, J., DALLY, W. J., TOWLES, B., AND GUPTA, A. K. 2005a. Microarchitecture of a high-radix router. In *Proceedings of the 32nd International Symposium on Computer Architecture*.

KIM, J., NICOPOULOS, C., PARK, D., NARAYANAN, V., YOUSIF, M. S., AND DAS, C. R. 2006b. A gracefully degrading and energy-efficient modular router architecture for on-chip networks. In *Proceedings of the 33rd International Symposium on Computer Architecture*.

KIM, J., PARK, D., THEOCHARIDES, T., VIJAYKRISHNAN, N., AND DAS, C. R. 2005b. A low latency router supporting adaptivity for on-chip interconnects. In *Proceedings of the 42nd Design Automation Conference*.

MCKEOWN, N. 1999. The islip scheduling algorithm for input-queued switches. *IEEE/ACM Trans. Netw. 7,* 2.

MILLER, D. A.B. 2009. Device requirements for optical interconnects to silicon chips. *Proc. IEEE 97,* 7.

MIURA, N., KASUGA, K., SAITO, M., AND KURODA, T. 2010. An 8tb/s 1pj/b 0.8mm2/tb/s QDR inductive-coupling interface between 65nm CMOS GPU and 0.1um DRAM. In *Proceedings of the 57th International Solid-State Circuits Conference*.

MORA, G., FLICH, J., DUATO, J., LOPEZ, P., BAYDAL, E., AND LYSNE, O. 2006. Towards an efficient switch architecture for high-radix switches. In *Proceedings of the ACM/IEEE Symposium on Architecture for Networking and Communications Systems*.

NICOPOULOS, C., NARAYANAN, V., AND DAS, C. R. 2010. *Network-on-Chip Architectures*. Springer.

O'MAHONY, F., KENNEDY, J., JAUSSI, J. E., BALAMURUGAN, G., MANSURI, M., ROBERTS, C., SHEKHAR, S., MOONEY, R., AND CASPER, B. 2010. A 47x10gb/s 1.4mw/(gb/s) parallel interface in 45nm CMOS. In *Proceedings of the 57th International Solid-State Circuits Conference*.

PASSAS, G., KATEVENIS, M., AND PNEVMATIKATOS, D. N. 2012. Crossbar NoCs are scalable beyond 100 nodes. *IEEE Trans. Comput.-Aid. Des. Integr. Circ. Syst. 31,* 4.

PETERSON, L. L. AND DAVIE, B. S. 2007. *Computer Networks: A Systems Approach*. Morgan Kaufmann, San Fransisco, CA.

ROWETH, D. AND JONES, T. 2008. QsNetIII an adaptively routed network for high performance computing. In *Proceedings of the 16th IEEE Symposium on High Performance Interconnects*.

SATPATHY, S., SEWELL, MANVILLE, T., CHEN, Y.-P., DRESLINSKI, R., SYLVESTER, D., MUDGE, T., AND BLAAUW, D. 2012. A 4.5tb/s 3.4tb/s/w 64×64 switch fabric with self-updating least-recently-granted priority and quality-of-service arbitration in 45nm CMOS. In *Proceedings of the 59th International Solid-State Circuits Conference*.

SCOTT, S., ABTS, D., KIM, J., DALLY, A. J. 2006. The blackwidow high-radix clos network. In *Proceedings of the 33rd International Symposium on Computer Architecture*.

SHAMIM, I. 2009. Energy efficient links and routers for multi-processor computer systems. Master's thesis. Massachusetts Institute of Technology, Cambridge, MA. http://dspace.mit.edu/handle/1721.1/54650.

SINGH, A. 2005. Load-balanced routing in interconnection networks. Ph.D. dissertation, Stanford University. http://pdf.aminer.org/000/338/323/balanced_routing.pdf.

SUTHERLAND, I. E., SPROULL, R. F., AND HARRIS, D. F. 1999. *Logical Effort: Designing Fast CMOS Circuits*. Morgan Kaufmann, San Fransisco, CA.

VALIANT, L. G. 1982. A scheme for fast parallel communication. *SIAM J. Comput. 11,* 2.

WANG, H.-S., PEH, L.-S., AND MALIK, S. 2002. A power model for routers: Modeling alpha 21364 and infiniband routers. In *Proceedings of the 10th IEEE Symposium on High Performance Interconnects*.

YANG, C.-K. K. 1998. Design of high-speed serial links in CMOS. Ph.D. dissertation, Stanford University. ftp://reports.stanford.edu/pub/cstr/reports/csl/tr/98/775/CSL-TR-98-775.pdf.

ZHANG, Y., HU, X., DEUTSCH, A., ENGIN, A. E., BUCKWALTER, J. F., AND CHENG, C.-K. 2009. Prediction of high-performance on-chip global interconnection. In *Proceedings of the 11th International Workshop on System Level Interconnect Prediction*.

ZHAO, W. AND CAO, Y. 2006. New generation of predictive technology model for sub-45nm design exploration. In *Proceedings of the 7th International Symposium on Quality Electronic Design*.

Received October 2012; revised January 2013; accepted March 2013