Title

A New Approach to Switch Fabrics based on Mini-Router Grids and Output Queueing

Authors

Karadeniz, T.
Dabirmoghaddam, A.
Goren, Y.
et al.

Publication Date

2015-02-16

Peer reviewed

A New Approach to Switch Fabrics based on Mini-Router Grids and Output Queueing

Turhan Karadeniz† Ali Dabirmoghaddam† Yusuf Goren‡ J.J. Garcia-Luna-Aceves†

†Department of Computer Engineering ‡Department of Mathematics
University of California, Santa Cruz
Santa Cruz, CA, 95064
Email: †{tkaradeniz, alid, jj}@soe.ucsc.edu ‡ygoren@ucsc.edu

***Abstract*—A number of switch fabric architectures based on mini-router grids (MRG) have been proposed as a replacement of buses for system-on-chip communication, as well as a replacement of crossbars for network routers. The rationale for using MRGs in switch fabrics is that they provide high delivery ratios, low latencies, high degree of parallelism and pipelining, load balancing properties, and sub-quadratic cost growth for their implementation. The traditional approaches to switch fabrics are based on input queuing (IQ) or virtual output queueing (VOQ), because output queuing (OQ) solutions to date are unscalable and expensive due to the speedup problem. However, we show that the speedup problem introduced by OQ can be bounded by 3 by using MRGs.**

We present the design of a switch fabric based on OQ MRGs that offers high delivery ratios, smaller queue sizes, and QoS guarantees. Queueing and scheduling are distributed over the MRs, where each MR is a pipestage, thus allowing MRGs to provide high throughput by nature. We present the first in-depth analytical model of switch fabric architectures based on OQ MRG, and support our model with register-transfer level (RTL) simulations in SystemC. The analytical and simulation results are shown to have close correlation over a range of design parameters and evaluation metrics.

I. Introduction

Whether it is used in routers for computer communications or in the system-on-chip domain as a replacement of buses, switch fabric design is an important part of the effort carried out for packet-switched information transmission paradigm. The switch fabric is the hardware component that acts as the intermediate connection point of ingress ports to egress ports in the aforementioned and other similar information systems.

From the computer networks perspective, the routers require ever increasing need for scalable switch fabrics with high delivery ratios and low latencies. From the system-on-chip (SoC) perspective, using packet switched communication remedies the difficulty of interconnecting components with heterogeneous interfaces, thus abstracting data from the signaling interface, while offering high degree of parallelism and pipelining, which results in greater throughput.

A switch fabric consists of queueing memories and memory controllers for temporary storage of the packets, scheduling unit(s) to avoid contention, and other computational components that facilitate these tasks. Switch fabrics need to support high delivery ratios, in order not to become the bottleneck in the communication themselves, and therefore their design remains to be an open research problem.

In Section 2, we outline the related literature. In Section 3, we describe a novel switch fabric architecture. In Section 4 we describe its analytical model and in Section 5, correlate the model to the simulation results. In Section 6 and 7, we discuss a possible placement and comparison to other switches, respectively. Finally, Section 8 concludes the paper.

II. Related Work

The switch fabric is one of the most important building blocks of communication related hardware, and various architectures have been proposed for both packet-switched computer networks and system-on-chip communication. Most of these architectures require a combination of queueing units, scheduling units and various other hardware components.

The main design challenges for a switch fabric include delivery ratio, bandwidth, latency, scheduling algorithms, interfacing, and routing algorithms as required in grid-based solutions. Several switch fabric architectures have been proposed, including the crossbar, shared-bus and shared-memory switches, which deal with these design challenges in various ways. The crossbar switch is the dominant architecture in today's high-performance switches, due to a number of reasons: crossbar switch is more scalable than the shared-bus and shared-memory; this is due to the limitations in bus transfer bandwidth and memory access bandwidth, respectively. Crossbar switch [1] provides point-to-point connections and non-blocking properties, as well as supporting multiple simultaneous transactions, increasing the bandwidth and speed of the router. However, their cost grows quadratically with the number of ports, since they require internal crosspoints and queues for every input/output port pair.

A number of architectures based on virtual output queues (VOQ) were proposed in order to remediate head-of-line (HOL) blocking of FIFO scheduling, including PIM, RRM, iSLIP [2][3][4], claiming a theoretical delivery ratio of 100%. However, VOQs do not scale optimally as the number of ports is increased, and therefore they are impractical. Another proposal, the load-balancing switch [5], claims greater scalability, by eliminating the need for a scheduler at the cost of duplicating the packets. Output queueing (OQ) provides high performance, due to the fact the ingress packets are forwarded to the output queues without any delays, and as a result they achieve a theoretical delivery ratio of 100%. However, OQ suffers from the speedup problem, which means if there are N contending packets, they need to be written to the queue in the same time cycle requiring a frequency speedup of N. Prabhakar and McKeown [6] propose a combined input and output queued (CIOQ) architecture, which bounds speedup by 4.

Recently, there have been a number of proposals on Mini-Router Grid (MRG) based switch fabric architectures for system-on-chip communication as a replacement of bus (a paradigm also referred to as network-on-chip), as well as for

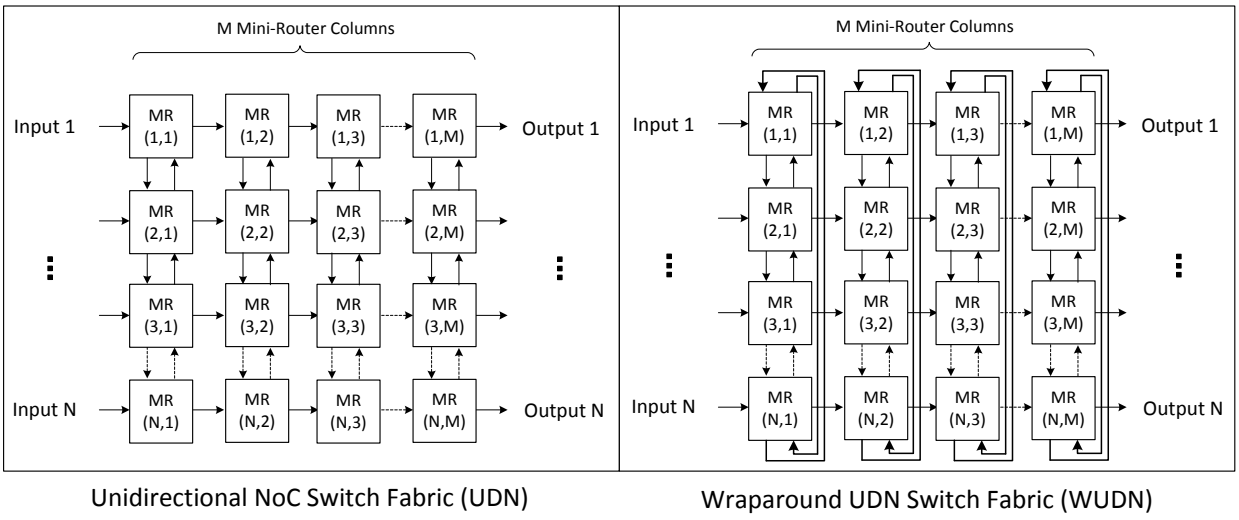

Fig. 1: UDN / WUDN Architectures

routers used in computer communications as a replacement of crossbars, including [7], [8], [9], [10].

Goossens et al [7] and Karadeniz et al [8] propose the UDN switch fabric, its functional model and the hardware design respectively, with the premise that they provide i) high throughput due to their pipelined nature, ii) low latencies, iii) ability to decouple switch size from cost growth, and iv) load balancing properties. UDN is in the form of an $N \times M$ MRG, where the input ports are placed on the West of the grid, and output ports are placed on the East, as shown in Fig. 1. These proposals achieve sub-quadratic cost growth by decoupling the number of ports from the switch cost, at the expense of performance. Moreover, they introduce load-balancing without duplicating the packets, which in turn improves the delivery ratio and latency. Goossens et al [7] describe the aforementioned architecture, presents some limited analytical modeling and functional level simulations and Karadeniz et al [8] proposes a feasible hardware implementation and cost analysis; however, both lack an in-depth analytical model correlating the design parameters to each other, as well as to the performance metrics.

In this paper, we extend the MRG based switch fabric proposal to replace VOQs by output queues (OQ), show that we are able to bound the speedup by 3 and to decrease latency by adding wraparound connections between North and South MR rows. Furthermore, we present the first in-depth analytical model of switch fabric architectures based on OQ MRG and validate our model with register-transfer level (RTL) simulations in the synthesizable subset of SystemC by showing that the analytical and simulation results have close correlation over a range of design parameters.

III. Mini-Router Grid Based Switch Fabric Architecture

A. Architecture Design Parameters

In this paper, we propose an output-queued, wraparound version of UDN (WUDN), where the North ports of the North-most row are connected to South ports of the South-most row, reducing the maximum vertical packet transmission from N hops to $N/2$ hops, and increase the load balancing and traffic uniformity further. The comparison of the UDN and WUDN architectures are presented in Fig. 1.

The size of the WUDN switch is defined by the 2-tuple (N, M), where N denotes the number of input-output (I/O) ports. M denotes the number of MR columns. The number of MRs in the WUDN switch is equal to $N \times M$, and they all have 3 I/O ports. Hardware design constraints and the modulo based routing algorithm (Subsection III-D) inflict some restrictions to the (N, M) values:

- $N \in \mathbb{N}$, and $N \geq 2$, trivial restriction that ensures the number of ports is greater than or equal to 2.
- $M \leq N$ and $M = 2^{\mathrm{m}}$, where $m \in \mathbb{N}_0$: The restrictions on M are inflicted by the routing algorithm involving Modulo M operations (See Section III-D). Because Modulo M operation requires division in case $M \neq 2^m$, but it is a simple bit-selection operation in case $M = 2^m$, we can avoid the extra cycles caused by the division operation by applying this restriction.
- $M \mid N$: This ensures that the load is distributed uniformly on M columns, in order to avoid congestion due to structural non-uniformity.

The WUDN switch architecture has the wormhole switching mode, buffered flow control implemented by output queueing scheme, Modulo XY Algorithm allows deterministic-uniform routing within the MRG, and incremental routing path decision. For simplicity, we use fixed-length cells.

The data delivery through the MRG is pipelined due to the point-to-point nature of MRs. This restricts the critical path to the control logic in a single MR and improves the scalability and throughput. WUDN is a deadlock-free since it is unidirectional.

By introducing output queuing (OQ) scheme to MRGs, we show that the speedup problem can be bounded by 3 and therefore be feasibly implemented in hardware. Our architecture offers a theoretical throughput of 100%, smaller queue sizes, and QoS guarantees, while not suffering from unscalability due to speedup.

The cost of WUDN is a function of N and M, and it is directly proportional to the number of MRs. As N is increased, more MR columns would be required to support the traffic; however the architecture permits M to be decoupled from N, trading performance for lower cost. For constant performance, M is a function of N.

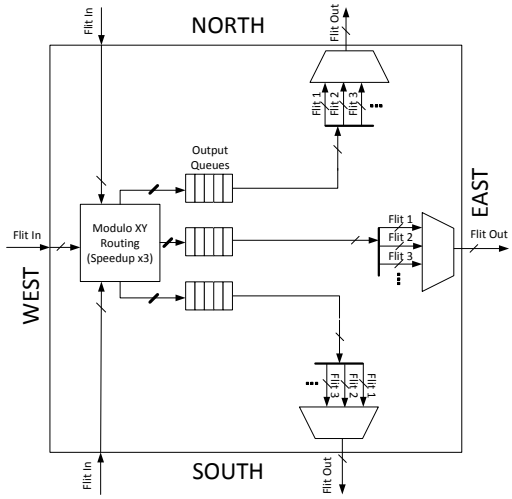

Fig. 2: 3 I/O Port WUDN MR, Top-Level Block Diagram

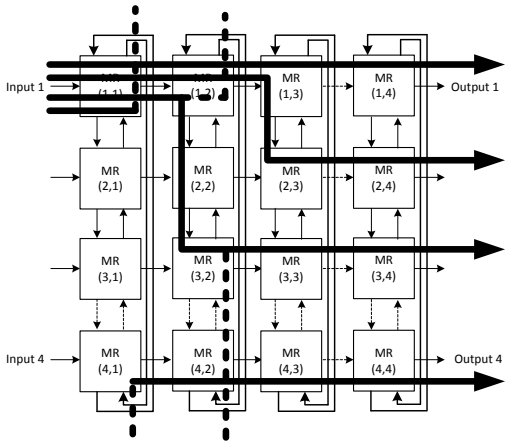

Fig. 3: Modulo XY Routing for WUDN

B. WUDN 3 I/O Port Mini-Router

The block diagram for a 3 I/O Port WUDN MR is given in Fig. 2. There are 3 input ports, West, North and South; and there are 3 output ports, East, North and South. The output queues and memory controllers are placed at each output port, whereas the next-hop logic and demultiplexers are placed at the input ports. Packets are transmitted in equally size flits in between MRs.

C. Memory Organization for Output Queues

MRs require a maximum speedup of 3, which is the number of MR I/O ports. In order to overcome the speedup problem that affects only the write operations, we use a dual clock-domain memory controller and dual-port queues.

First, the next-hop port for the ingress packets are computed, which generates the signal to demultiplex the packet into the appropriate output queue controller; then, within the memory controller for write operation, contending packets are multiplexed and serially written to the queueing memory that is on the higher frequency clock domain. We use a Round Robin (RR) based scheduling algorithm to multiplex and serialize the packets into the output queues. The priorities of the three input ports are shifted in a circular fashion, and the present contending packets are serialized according to these priorities. The output queue control is implemented in a circular fashion.

D. Modulo XY Routing in the Mini-Router Grid

We present Modulo XY Routing algorithm for WUDN, a special case of Manhattan routing, which makes a balanced distribution of the vertical traffic over the MR columns by using the modulo operation. The algorithm is applied to each packet at each MR, thus implementing an incremental routing scheme. Fig. 3 exemplifies 4 packet transmissions from I_1 to O_{1-4}. The packets are routed on different MR columns per I/O port pairs, distributing the load on the switch. A turning point MR (MR_{TP}) denotes an MR in which a horizontal flow is directed into a vertical one ($MR_{TP,H\to V}$), or vice versa ($MR_{TP,V\to H}$).

The algorithm incorporates some computational operations including modulo, absolute value, sign function, subtraction, and comparison. With the restrictions we have defined in Subsection III-A, the modulo operation is reduced to a simple bit select, and the remaining operations can be computed in fewer number of operations, using 2's complement algebra.

IV. Analytical Model

In this section, we present a mathematical analysis of WUDN switch fabric that provides useful insights into evaluation of the performance and scalability of the proposed architecture.

In this section, we use notation R_{xy} to denote the MR at coordinate (x, y) on the grid. I_s and O_t are used when referring to input port s and output port t, respectively. Further, $s \rightsquigarrow t$ is to denote the path between I_s and O_t found by the modulo algorithm and $|s - t|$ denotes the lattice distance between I_s and O_t.

A. Analysis of the Modulo Routing Algorithm

We begin with characterization of the routing algorithm. Modulo routes exhibit some useful properties that make the analysis of the system easier. In the sequel, we review some of these properties.

Property 1. *For any given source-destination pair, modulo routing finds the shortest path between the two end.*

Proof: The modulo routing is a greedy forwarding scheme in a taxicab geometry in which at each hop, the lattice distance between the packet and destination is decremented by one. This is, indeed, the maximum the grid structure can offer and thus, guarantees that the routing is performed over the shortest path. Such a path, however, is not necessarily unique. ■

An immediate deduction from Property 1 is that the length of the modulo route between an arbitrary choice of I_s and O_t is indeed $|s - t|$, that is the lattice distance between source and destination. On a $N \times M$ grid and with a uniform and independent selection of source and destination, every path is comprised of an exact number of M horizontal and an average of $N/4$ vertical transmissions. The latter can be inferred noting the fact that $|s - t|_v$ takes on values between 0 and $N/2$ with equal probability. Therefore, the average path length is $M + N/4$ hops. We shall use this fact later on when proving Property 2.

Property 2. *Under a uniform and independent selection of source-destination pairs, modulo routing maintains a uniform distribution of traffic across all MRs throughout the grid.*

Proof: We quantify the fraction of transmissions contributed by each MR in the grid. In a general case and on a $N \times M$ grid, the modulo algorithm can generate a total of N^2 deterministic paths connecting every source to every destination port. Assuming that source-destination pairs are chosen uniformly and independently, then the choice of every path is also uniform; that is, all paths are equally utilized.

Suppose that every source deterministically generates a packet for every destination. From Property 1, we recall that a modulo path comprises an average of $M+N/4$ hops. Thus, the total number of transmissions, T, to be made over the entire grid is given by

$$T = N^2\Big(M + \frac{N}{4}\Big)\,. \tag{1}$$

Consider the x^{th} row, denoted by R_{x*}, on the grid. We enumerate the total number of horizontal transmissions over R_{x*}. Note that such transmissions are either from the traffic originated from I_x or destined at O_x. We consider each case separately.

1) Traffic originated from I_x (but not destined at O_x): There are N packets coming out of I_x, each of which takes between 0 to $M-1$ horizontal hops on R_{x*} with equal probability before getting to the turning point. That makes a total of $N(M-1)/2$ horizontal transmissions on R_{x*}.
2) Traffic destined at O_x: There are N such packets each taking between 1 to M horizontal hops on R_{x*} with equal probability. This adds up to a total of $N(M+1)/2$ horizontal transmissions on R_{x*}.

The total number of horizontal transmissions across R_{x*} is thus given by

$$T_h(x) = NM\,. \tag{2}$$

In order to enumerate the vertical transmissions across row R_{x*}, we perform a transformation on the grid. Let us assume that the grid is horizontally contracted such that all MRs in every row are consolidated into a single MR. This transforms the original $N \times M$ grid into a $N \times 1$ layout. This transformation is safe, in the sense that it does not affect the number of vertical transmissions. In fact, all vertical transmissions that were supposed to be carried out across R_{x*} will now take place on $R_x \equiv R_{x1}$. Therefore, the number of vertical transmissions across row R_{x*} is now equal to the number of individual paths intersecting R_x.

We take advantage of the symmetry of vertical connections on the grid to enumerate all such paths. To that end, note that R_x is used on every $s \rightsquigarrow t$ where $|s-x|+|x-t| = |s-t|$. By Property 1, $|s-t| \leq 1+N/2$. However, where $|s-t| = 1+N/2$, there are exactly two such paths between s and t, only one of which uses R_x if $s \neq x$. In that case, the grid uses both paths alternatively to maintain a fair balance of traffic over all MRs. In our analysis, we count such cases as *half-paths*. It is easy to observe that there are exactly N half-paths that cross through any given R_x. Offsetting all such paths that are double-counted, we obtain the following equation for the total number of paths intersecting R_{x*}

$$T_v(x) = 2\sum_{i=0}^{\frac{N}{2}-1}\left(\frac{N}{2} - i\right) - \frac{N}{2} = \frac{N^2}{4}\,, \tag{3}$$

which is indeed equivalent to the number of vertical transmissions carried out across R_{x*}.

From Equations (1), (2) and (3), we realize that R_{x*} handles $1/N$ of the total transmissions. For uniformly and independently chosen I_s and O_t, choice of the turning point column (*i.e.*, $(I_s + O_t)\%M$) is also uniform (see [11] for the actual theorem and a rigorous proof). This implies that every column also carries equal share of the traffic. Therefore, the total number of horizontal transmissions, $\bar{T}_h$, and vertical transmissions, $\bar{T}_v$, contributed by each MR is

$$\bar{T}_h = N\,, \qquad \bar{T}_v = \frac{N^2}{4M}\,. \tag{4}$$

Overall, each MR transmits a total of $N+N^2/4M$ packets, which is in fact $1/MN$ of the total traffic given by Equation (1). ∎

In the following, we use the foregoing discussion to obtain a detailed characterization of the dynamics of the proposed architecture.

B. Detailed Characterization of Traffic Distribution for MRs

Inside an individual MR, let us denote with P_E, P_N and P_S the probabilities of sending a packet on the east, north and south ports, respectively. Using Property 2, we note that these probabilities are identical for all MRs throughout the grid. From Equation (4) and the facts that

$$\frac{\bar{T}_h}{\bar{T}_v} = \frac{P_E}{P_N + P_S}, P_N = P_S \tag{5}$$

we readily obtain that

$$P_E = \frac{4}{\beta+4}\,, \quad P_N = P_S = \frac{1}{2}\Big(\frac{\beta}{\beta+4}\Big)\,, \tag{6}$$

where $\beta := N/M$ is the grid's aspect ratio.

As seen, these probabilities depend on number of inputs N and number of layers M only through β. In fact, grids with similar aspect ratio exhibit similar behavior in terms of traffic distribution. We call such grids *isomorphic* and shall further investigate on their properties later in Section V.

A second important observation from Equation (6) is that when $\beta = 8$, we have $P_E = P_N = P_S$. In that case, all three queues are equally utilized and there appears a balanced distribution of traffic across all directions on the grid.

C. Incoming Traffic Rate and Stability

For a switch with infinite capacity queues, using Equation (6), we find the maximum rate of incoming traffic (denoted by λ) under which the switch would be stable. Let us denote with λ^* the total input rate to an arbitrary edge MR, R_{x1}. In the steady state, the output rate is equal to the input rate. Thus,

$$\lambda^* = \lambda + 2 \times \frac{1}{2}\Big(\frac{\beta}{\beta+4}\Big)\lambda^* = \lambda\left(1 + \frac{\beta}{4}\right). \tag{7}$$

The stability condition requires to maintain a utilization factor $\rho := \lambda^*/\mu$ of less than one. Therefore, the sufficient condition for stability is to maintain

$$\lambda < \frac{4\,\mu}{\beta+4} = \lambda_{max}\,. \tag{8}$$

The service rate, μ, is deterministic and is equal to 3, since MR is able to transmit 3 packets every round. Thus, the

maximum tolerable input rate is 2.4 when $\beta = 1$ (*i.e.*, perfect square grids). The switch would only be able to handle lower rates as β increases. Of course, with finite capacity queues, irrespective of the queue size, the switch would always remain stable; yet certain fraction of the incoming traffic would be blocked and the system would inevitably be subject to packet losses should the system is loaded with a rate higher than λ_{max}.

D. State Distribution and Optimal Queue Size

The incoming traffic being a memory-less process, a WUDN switch can effectively be modeled as a Jackson network of $M^X/D^Y/1/N$ queues[1]. Due to the architectural symmetry and traffic uniformity (Property 2), most of the analysis of such a complex system can be well reduced to the analysis of a single $M/D/1/N$ queue.

Even in finite capacity systems, a delivery ratio of close to 1 is often desired. In order to obtain insights into finding the optimal queue size, we concentrate on the analysis of the more generic case of $M/D/1$ with infinite capacity and characterize the state distribution. The state distribution, $P(i)$, for $M/D/1$ queues can be computed according to Fry's derivation [12] when $n = 1$, as follows:

$$P(i) = (1-\rho)\sum_{j=0}^{i} e^{\rho j}\left(\frac{(-\rho j)^{i-j}}{(i-j)!} - \frac{(-\rho j)^{i-j-1}}{(i-j-1)!}\right). \quad (9)$$

Using that, we readily compute the cumulative state distribution, $S(i)$, as follows.

$$S(i) = \sum_{j=0}^{i} P(j) = (1-\rho)\sum_{j=0}^{i} \frac{\big(\rho(j-i)\big)^j}{j!} \cdot e^{\rho(i-j)}. \quad (10)$$

In Fig. 5, the dashed lines illustrate $S(i)$, the probability of having at most i packets in each MR. In the actual $M/D/1/N$ architecture, there is a chance of packet loss due to blocking at the queues. By numerical evaluation of $S(i)$ for a given utilization factor, we are able to pinpoint the optimal queue size by looking at the k^{th}-percentile on the cumulative distribution, when a steady-state delivery ratio of $k\%$ is expected.

V. The Model & Simulations

A. RTL & Functional Models

The switch fabric model and simulations are implemented in SystemC [13]. SystemC is an event-driven simulation library for C++ that imitates Hardware Description Languages (HDL) by simulating concurrent events. SystemC provides a synthesizable subset to approach RTL design languages like Verilog or VHDL further, compared to functional models.

We implement most of our switch fabric components, including *queue memory controllers*, *Modulo XY Routing unit*, and other *forwarding related circuitry* in synthesizable SystemC, whereas *clock generation*,*Source* and *Sink* modules are functional models. The Source module is instantiated *N* times, once per ingress port, and it generates random traffic according to a Poisson distribution. Similarly, the Sink module is instantiated *N* times, once per egress port, and it is responsible for performance reporting.

[1] $M^X/D^Y/1/N$ is Kendall's notation describing memory-less batch arrival and deterministic batch service distributions, single server, with queue size of $N-1$. Here M and N are part of a standard notation and should not be confused with what we have used for specifying grid dimensions.

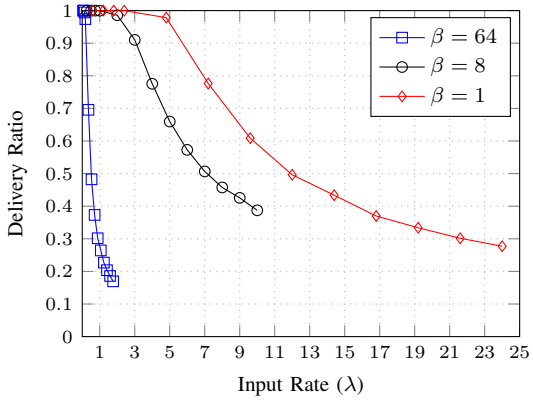

Fig. 4: Delivery Ratio Vs. Input Rate

B. Simulation Parameters & Performance Metrics

The input parameters to our simulation experiments are switch size N, number of MR layers M, maximum queue size QS, incoming traffic rate λ, and the simulation runtime duration T.

The performance metrics of interest are as follows:

- Delivery Ratio (DR): Ratio of the packets successfully delivered to the destination ports. We analyze this metric under variable input rate and queue size;
- Local traffic distribution: Fraction of packets being sent over each output port within MRs. For this measure, we only evaluate P_E, probability of transmitting over east port. P_S and P_N provide no further insight and can be computed accordingly;
- Global distribution of traffic across the entire MR grid.

C. Delivery Ratio under Variable Traffic Rates

Our first experiment examines the impact of increasing traffic rate on the overall delivery ratio. We simulate the performance of a switch with 64 input ports, where the number of layers M are $\{1, 8, 64\}$. For this analysis, we use Equation (10) to anticipate and employ a queue capacity for which a delivery ratio of close to 1 is achieved when the system is highly loaded. Note that the optimal queue size also depends on the grid's aspect ratio as seen in Section IV. Therefore, the queue sizes used on each of the three individual grid layouts are different.

Fig. 4 shows how the delivery ratio decays with an increasing load of traffic over the x-axis. λ_{max} for $\beta = \{1, 8, 64\}$ are $\{2.4, 1, 0.176\}$, respectively. As clearly seen, a delivery ratio of 1 is obtained in all cases when $\lambda < \lambda_{max}$. Recall from Section IV that this corresponds to a utilization factor (ρ) of 1. The utilization factor, in fact, determines the fraction of time that the system is busy (*i.e.*, non-empty MRs). Having $\rho \geq 1$ results in instability (infinite queue size and waiting time) in infinite capacity systems and packet drops in finite capacity queues. This behavior is accurately captured both by model and simulations. Even though reporting results for $\rho \geq 1$ might seem futile, it's actually useful to take into account how the system would react to bursty traffic or unexpected high loads of short durations.

D. Delivery Ratio as a Function of Queue Size

We study the impact of changing the queue size on the overall delivery ratio. In every experiment, the queue size is fixed and identical for all MRs throughout the grid. Also, the

queue sizes for north, east and south ports are the same and are equal to one-third of the total MR capacity. This, in fact, justifies the reason why we have used queue sizes that are multiples of 3.

In order to conduct a fair experiment, we have used a 64×8 grid to maintain an aspect ratio (β) of 8 for which all queues are equally utilized.

Fig. 5 demonstrates how increasing the system capacity enhances overall delivery ratio. The solid curves show the simulation results, while dotted lines correspond to theoretical values obtained by Equation (10). As seen, the behavior of the system is accurately predicted by the model.

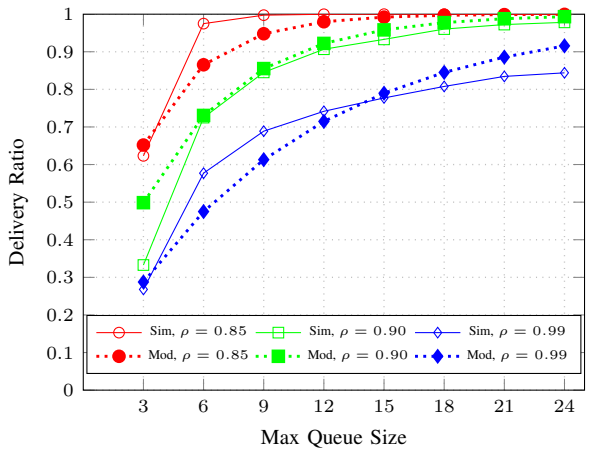

Fig. 5: Delivery Ratio vs. Max Queue Size

E. Impact of Grid's Layout on Local Traffic Distribution

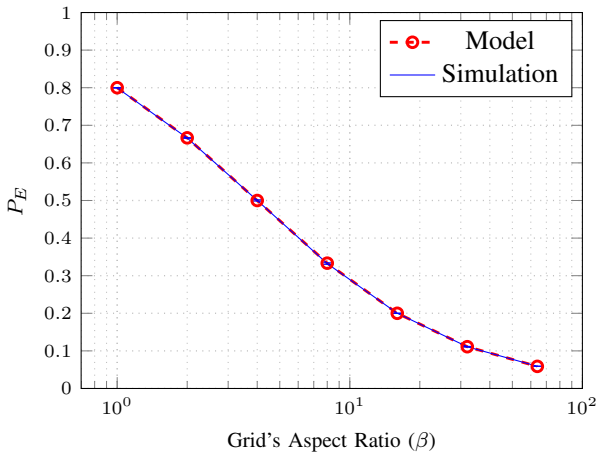

Fig. 6: Probability of sending on east vs. grid's aspect ratio

In this subsection, we verify the property captured by Equation (6) that the within-MR distribution of traffic across all ports depends on N and M only through β. More precisely, grids with similar aspect ratios (isomorphic grids) result in similar traffic distributions across their output ports.

For this experiment, we generate multiple grid instances of different sizes that have identical aspect ratios. For every layout, we calculate the fraction of packets transmitted through the east port (*i.e.*, P_E). Fig. 6 demonstrates that the simulation results (solid line) accurately support our theoretical analyses (dashed line). The simulation results, in fact, are the average

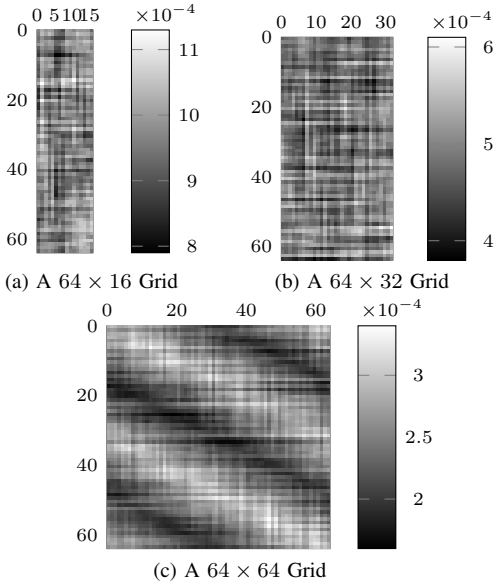

Fig. 7: Heatmaps illustrating uniform distribution of the the traffic across the grid

values for different isomorphic grids. The standard deviations, shown as error bars, are so small that are barely visible on the graph.

According to Fig. 6, the more the aspect ratio is, the lighter becomes the eastbound traffic. This behavior is intuitive noting the fact that for a fixed M (say, $M = 1$), a larger β corresponds to having a larger number of sources (N). In that case, higher proportion of traffic should be carried through a fixed number of columns (M), which results in a lower P_E. This also points out to the fact that isomorphic grids provide identical distributions of traffic across output ports.

F. Global Distribution of Traffic Across the Grid

We provide experimental results to demonstrate the uniformity of the traffic across the entire grid. This is, indeed, an attribute of the XY modulo routing that we discussed earlier as Property 2 of the routing algorithm.

We run extensive simulations on three different grid layouts and quantify the portion of traffic handled by each MR. We illustrate the traffic through individual MRs in the grid using heatmaps in Fig. 7. Lighter colors represent higher load intensity.

As illustrated in Fig. 7, the load range is very tight all over the grid, and the varying colors on the heatmap only point to very small variances. Therefore, in all scenarios, the traffic distribution is very balanced and the grid allows fair routing. A slight pattern can be observed in Fig. 7c, but the variance between the limiting values within the colormap is so low that the pattern is virtually negligible.

VI. Placement Considerations

The block diagram of WUDN Fig. 1 might incorrectly imply that the wraparound links connecting North ports of the North-most MR row to South ports of the South-most MR row are implemented as long buses, which in return would inflict asymmetric delays between the MRs, requiring the clocking to be slowed down. Moreover, having a long bus proportional to N would render the switch fabric unscalable due to the propagation delay of O(N). On the contrary, WUDN can be

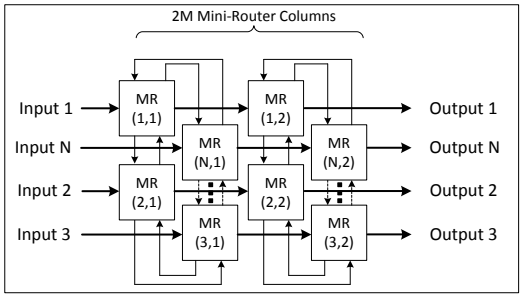

Fig. 8: Placement of WUDN

placed on the chip in such a way that the wraparound links are not longer than the other vertical or horizontal links, preventing such issues from arising.

The idea behind WUDN placement is "folding" the columns in the mesh such that each column has the "U" shape, as opposed to "I". This is shown in Figure 8, where wraparound links are not a function of switch size N anymore. Since the topology and the architecture remain the same, the descriptions and analysis in the previous sections hold.

Please note that, with this approach the area of the switch fabric remains the same; the height is halved and the width is doubled. After the placement, WUDN has longer horizontal links that connect every other column, but the total link length throughout the switch fabric remains the same. In addition, to our advantage, the length of any link between the MRs is now decoupled from the switch size N (as well as, number of layers M), to yield a scalable and low delay architecture.

VII. Comparison & Discussion

Crossbar, in terms of many performance aspects, resemble MRG based UDN and WUDN, supporting point-to-point connections and non-blocking properties that allows multiple simultaneous transactions. However, their cost grows quadratically with the number of ports, since they require internal crosspoints and queueing memories for every input/output port pair. MRG based switches improve upon crossbar switches for their ability to lower switch cost growth from $O(N^2)$ at the expense of performance. This is due to their ability to add or remove MR columns (or layers) as necessary.

UDN inflicts increased congestion in the central rows, since there is more traffic towards the center due to its lack of uniformity. WUDN improves upon the uniformity by adding the wraparound links that balances the traffic over all of the rows, resulting in less congestion. Moreover, the wraparound links decrease the vertical latency by a factor of 2.

It should be mentioned that WUDN is more cost efficient for larger switch sizes. For example, in the case of a switch with 4 ports, it would be more efficient to remedy a speedup of 4 instead of implementing it using 16 3-port mini-routers, which is clearly inefficient. However, as the switch size increases, the speedup cannot be remedied by a single switch, and WUDN becomes a viable option.

Another important point that the reader should note is that we chose to carry out in-depth analysis and simulations assuming uniform all-to-all traffic. [7] reports good performance for unbalanced and bursty traffic patterns, and our proposal with wraparound links and OQ would only improve those results further.

WUDN has queuing memory at every output port for each MR; however, observing Fig. 5 (where values are provided in terms of the total queueing memory size per MR) it's possible to deduce that the WUDN requires very small queue sizes, and thus that this does not in fact issue a threat on the cost.

VIII. Conclusion

In this paper, we have proposed a novel switch fabric based on OQ MRGs, which offers promising delivery ratios, small queue sizes, and low latencies. Moreover, we showed that the speedup problem introduced by OQ can be bounded by 3 by using MRGs. We have presented the first in-depth analytical model of switch fabric architectures based on OQ MRs, where we correlate design parameters to several performance metrics, such as the maximum supportable input rate λ_{max}, the optimal queue size, and local and global distributions of traffic. We have supported and validated our model with RTL simulations and showed that the simulation results closely match the analytical model. Finally, we have shown that WUDN does not inflict additional delays over UDN due to the wraparound links, by describing a feasible placement on-chip.

Power profiling and comparison to other architectures in terms of power consumption remain as future work. Analytical modeling under unbalanced traffic patterns would also be beneficial to this work.

References

[1] N. McKeown, M. Izzard, A. Mekkittikul, W. Ellersick, and M. Horowitz, *The Tiny Tera: A Packet Switch Core*. 1996.

[2] T. E. Anderson, S. S. Owicki, J. B. Saxe, and C. P. Thacker, "High-speed switch scheduling for local-area networks," *ACM Transactions on Computer Systems*, vol. 11, pp. 319–352, Nov. 1993.

[3] N. McKeown and T. E. Anderson, "A quantitative comparison of iterative scheduling algorithms for input-queued switches," *COMPUTER NETWORKS AND ISDN SYSTEMS*, vol. 30, pp. 2309—2326, 1998.

[4] N. McKeown, "The iSLIP scheduling algorithm for input-queued switches," *IEEE/ACM Transactions on Networking*, vol. 7, pp. 188–201, Apr. 1999.

[5] I. Keslassy, C.-S. Chang, N. McKeown, and D.-S. Lee, "Optimal load-balancing," in *Proceedings IEEE INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies*, vol. 3, pp. 1712– 1722 vol. 3, IEEE, Mar. 2005.

[6] B. Prabhakar and N. McKeown, "On the speedup required for combined input- and output-queued switching," *Automatica*, vol. 35, pp. 1909–1920, Dec. 1999.

[7] K. Goossens, L. Mhamdi, and I. Senin, "Internet-router buffered crossbars based on networks on chip," in *Digital System Design, Architectures, Methods and Tools, 2009. DSD '09. 12th Euromicro Conference on*, pp. 365 –374, Aug. 2009.

[8] T. Karadeniz, L. Mhamdi, K. Goossens, and J. Garcia-Luna-Aceves, "Hardware design and implementation of a network-on-chip based load balancing switch fabric," in *2012 International Conference on Reconfigurable Computing and FPGAs (ReConFig)*, pp. 1 –7, Dec. 2012.

[9] D. Wiklund, A. Ehliar, and D. Liu, "Design of an internet core router using the SoCBUS network on chip," in *Signals, Circuits and Systems, 2005. ISSCS 2005. International Symposium on*, vol. 2, pp. 513 – 516 Vol. 2, July 2005.

[10] G. Luo-Feng, D. Gao-ming, Z. Duo-Li, G. Ming-Lun, H. Ning, and S. Yu-Kun, "Design and performance evaluation of a 2D-mesh network on chip prototype using FPGA," in *Circuits and Systems, 2008. APCCAS 2008. IEEE Asia Pacific Conference on*, pp. 1264 –1267, Dec. 2008.

[11] P. Scozzafava, "Uniform Distribution and Sum Modulo m of Independent Random Variables," *Statistics & Probability Letters*, vol. 18, no. 4, pp. 313–314, 1993.

[12] T. C. Fry, *Probability and its engineering uses*. Van Nostrand New York, 1928.

[13] O. S. Initiative, "IEEE standard SystemC language reference manual," *IEEE Computer Society*, pp. 1666–2005, 2006.