Low Propagation Delay Load-Balanced 4×4 Switch Fabric IC in 0.13-μm CMOS Technology

Ching-Te Chiu, Yu-Hao Hsu, Wei-Chih Lai, Jen-Ming Wu, Shawn S. H. Hsu, Yang-Syu Lin, Fan-Ta Chen, Min-Sheng Kao, and Yar-Sun Hsu

***Abstract*— A load-balanced Birkhoff-von Neumann (LB-BvN) 4 × 4 switch fabric IC is proposed for feedback-based switch systems. This is fabricated in 0.13-μm CMOS technology and the chip area is 1.380 × 1.080 mm^2. The overall data rate of the LB-BvN 4 × 4 switch fabric IC is up to 32 Gb/s (8 Gb/s/channel) with only 0.8 ns propagation delay. The LB-BvN switch is highly recommended for constructing the next-generation terabit switch. In a feedback-based switch system, the long propagation delay of the switch module reduces the system throughput significantly. In this paper, we present a scalable LB-BvN 4 × 4 switch fabric IC directly in the high-speed domain. By observing the deterministic switching pattern of the $N \times N$ LB-BvN switch, we present a low-complexity pattern generator that reduces the PG complexity from $O(N^3)$ to $O(1)$. This technique reduces the propagation delay of the switch module from 30 to 0.8 ns, and also provides 80% area saving and 85% power saving compared to serializer–deserializer interfaces. The proposed LB-BvN 4 × 4 switch fabric IC is suitable for feedback-based switch systems to solve the throughput degradation problem.**

***Index Terms*— Current-mode logic (CML), load-balanced Birkhoff-von Neumann switch, low propagation delay, scalability, serializer–deserializer (SerDes), switch fabric IC.**

I. Introduction

MORE and more computers and commercial devices communicate with each other either by wired or wireless connections, and this revolution has led to increasing data traffic in networks. With the availability of high-speed internet, cloud computing services are being provided to corporate and individual users. These applications, which require high bandwidths, have become more and more popular, and this trend is set to continue. Therefore, to support high-bandwidth traffic, the performance of internet routers and switches should grow drastically.

Most switches currently available in the market are based on the shared memory switch architecture, which is one of the output-buffered switches [1]. In this architecture, packets are stored and forwarded in a common shared memory. As the speed of fiber optics advances, the memory access speed becomes a bottleneck (scalability problem). If the line rate is R, an $N \times N$ common shared memory has to deal with at most $2 \times N \times R$ data rate in the same time. To achieve higher speed, one has to use parallel-buffered switch architectures to obtain the needed speedup. One common approach, known as the input-queued switch architecture, is to have parallel buffers in front of a switch fabric [2].

An input-queued switch, with each input maintaining a single first-in first-out (FIFO) queue, may suffer head-of-line (HOL) blocking problem and then result in degradation in throughput down to 58% [3]. One way to solve this problem is the virtual output queuing (VOQ) technique, which maintains a separate queue for each output at each input. As there are N^2 buffers (memories) at the inputs of an $N \times N$ switch fabric, the key problem of input-queued switches (equipped with VOQs) is to apply a certain matching algorithm to choose at most N of N^2 HOL packets to transmit through the switch fabric [3]–[9]. The maximum weight matching algorithm can find a solution to this problem but, unfortunately, has a complexity of $O(N^{2.5} \log N)$ [10], which makes it difficult to implement in practice.

Several heuristics have been proposed to lower the complexity. The iSLIP has a time complexity of $O(\log N)$ to converge with maximum matching using $2N$ arbiters [11]. The computational complexity of randomized algorithms is $O(\log N)$ at the cost of increasing cell delay [12], [13]. The input-queued switch has a much longer delay than the load-balanced switch when the traffic arrival rate is above 0.9 [14]. Matching algorithms for conflict resolution require extra computation and communication overheads in every time slot, and these overheads result in another scalability issue. Furthermore, matching algorithms cannot guarantee 100% throughput theoretically without a speedup of 2 because

Manuscript received October 10, 2011; revised May 15, 2012; accepted July 21, 2012. Date of publication September 4, 2012; date of current version July 22, 2013. This work was supported in part by the National Science Council, Taiwan, under Contract NSC 97-2221-E-007-112-MY3 and the Advanced Research for Next-Generation Networking and Communications Project 98N2502E.

C.-T. Chiu and W.-C. Lai are with the Department of Computer Science and the Institute of Communications Engineering, National Tsing Hua University, Hsinchu 300, Taiwan (e-mail: ctchiu@cs.nthu.edu.tw; andy751026@hotmail.com).

Y.-H. Hsu is with the Embedded SRAM Library Department, Taiwan Semiconductor Manufacturing Company, Hsinchu 300-77, Taiwan (e-mail: shhsu@ee.nthu.edu.tw).

J.-M. Wu, S. S. H. Hsu, F.-T. Chen, M.-S. Kao, and Y.-S. Hsu are with the Department of Electrical Engineering, National Tsing Hua University, Hsinchu 300, Taiwan (e-mail: jmwu@ee.nthu.edu.tw; shhsu@ee.nthu.edu.tw; fanta524cf@yahoo.com.tw; kaom0711@gmail.com; yshsu@ee.nthu.edu.tw).

Y.-S. Lin is with the High Speed Memory Development Program, Taiwan Semiconductor Manufacturing Company, Hsinchu 300-77, Taiwan (e-mail: yslinze@tsmc.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2012.2212618

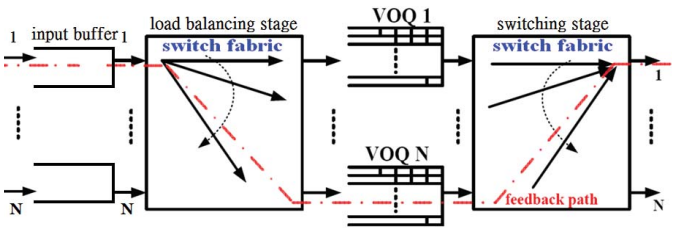


Fig. 1. Concept of the LB-BvN switch system architecture.

the use of a maximal matching algorithm, such as PIM [4] and SLIP [2], can only achieve about 50% throughput [1]. Although heuristic scheduling algorithms require speedup to achieve higher throughput, on-chip speedup could be inexpensive using parallel processing with the advance of semiconductor technologies. Heuristic scheduling algorithms are adopted in building switch fabrics [13], [14].

A breakthrough in high-speed switch architectures to overcome the problem of conflict resolution is the load-balanced Birkhoff-von Neumann (LB-BvN) switch [15]–[21], which is presented in Fig. 1. It was proposed to resolve the memory access conflict and to yield 100% throughput without extra computation and communication overheads. The LB-BvN switch consists of two-stage switch fabrics and one-stage parallel buffers (equipped with VOQs) between them. The first stage performs load balancing for the incoming traffic, so that the traffic arrives at the second stage uniformly. The second stage performs BvN switching on the uniform traffic [1], [15]. Since the connection patterns in both stages are periodic and deterministic, there is no need to find a matching result in every time slot. Therefore, the switch can expand its port number without the limitation of computation complexity. This high scalability is one of the most significant features of the LB-BvN switch. Compared to the input-queued switches, LB-BvN switches have drastic delay (which increases linearly with N) at low to medium traffic loads. This drawback affects its applications in latency-sensitive systems such as high-performance computing or high-frequency financial trading.

With this important feature (100% throughput), the LB-BvN switch is one of the best architectures for implementing the next-generation terabit switch. However, one drawback of the LB-BvN switch is the out-of-sequence issue. Packets in the flow of the same input to the same output might be out of order due to multiple paths created by the load-balancing stage.

Recently, much research has focused on resolving the out-of-sequence issue in this two-stage switch architecture [16]–[18]. The uniform frame spreading (UFS) scheme [22] adds VOQs at the inputs of the whole switch and operates the system in frames. Packets destined for the same output are stored in the same VOQ. Once a VOQ has more packets than the number of input/output ports, that VOQ is called a full-framed VOQ. At the beginning of a frame, a full-framed VOQ is selected and transmitted to the second stage. If there is no full-framed VOQ, then nothing is transmitted. A full-framed VOQ reserves a frame (of time slots) and transmits its packets consecutively in that frame. Though the frame-based scheme is shown to achieve 100% throughput, the packet delay is large even in light traffic. This is known as the starvation problem, as it takes time to accumulate packets for a full-framed VOQ.

The concept of feedback-based path in this two-stage system is introduced in mailbox algorithm [16]. The key idea of the mailbox switch is to use a set of symmetric connection patterns to create a feedback path in the two-stage switches. Through the feedback path, the packet departure times in the central VOQs are delivered back to the input buffers in front of the first-stage switch. With the information of packet departure times, the mailbox switch can schedule packets so that they depart in the order of their arrivals.

The contention and reservation (CR) switch has been proposed to solve the long delay of the UFS switch under light traffic [17]. They adopt the mailbox switch mechanism under light traffic. If there is no full-framed VOQ, then HOL packets are transmitted instead. As shown in [17] and [18], the average delay of the CR switch is low for light traffic but still large for heavy traffic. It is because there are few collisions under light traffic. A packet, upon its arrival, is transmitted immediately to the central buffer as an HOL packet. As for the approach of adding reorder buffers at the output, in general, mechanisms for ensuring in-order packet delivery tend to penalize the packet delay performance more than throughput. If resequencing buffers are used for solving the missequencing problem, packets suffer from the additional resequencing delay. It is because packets of the same flow experience different delays at different middle-stage ports [18].

Feedback-based scheduling switch [18] further extends this idea to create different feedback paths in the two stages. These feedback-based methods use a set of connection patterns to create a feedback path for packet departure time. With the information of packet departure time, the system can schedule packets so that they depart in the order of their arrivals.

In the design and implementation of a switch fabric IC, a digital signal processing (DSP) switch core with the serializer–deserializer (SerDes) with 8 B/10 B CODEC interfaces is commonly used, as shown in Fig. 2(a). The SerDes interfaces help reduce the pin counts of chips. However, in a feedback-based two-stage switch system, a long propagation delay in the feedback path makes the system throughput to decrease significantly [23], [24]. high-order switch fabric is usually constructed from lower order switches. The effect of propagation delay in the SerDes interface and DSP core becomes worse when a switch fabric scales up. To show this effect, the analysis of throughput degradation versus feedback path propagation delay is provided in this paper.

The motivation for our work is to present an LB-BvN switch architecture for reducing the long propagation delay in a feedback-based LB-BvN switch system [25]. The overall architecture of an LB-BvN 4×4 switch is shown in Fig. 2(b). Based on the characteristics of deterministic and periodic connection patterns in an LB-BvN switch, we implement the switch directly in the high-speed domain instead of a DSP core. The current-mode logic D-type flip-flops (CML DFFs) and CML multiplexers (CML MUXes) are adopted to achieve higher operating speed. By operating the switch system directly using high-speed circuits, the SerDes interfaces,

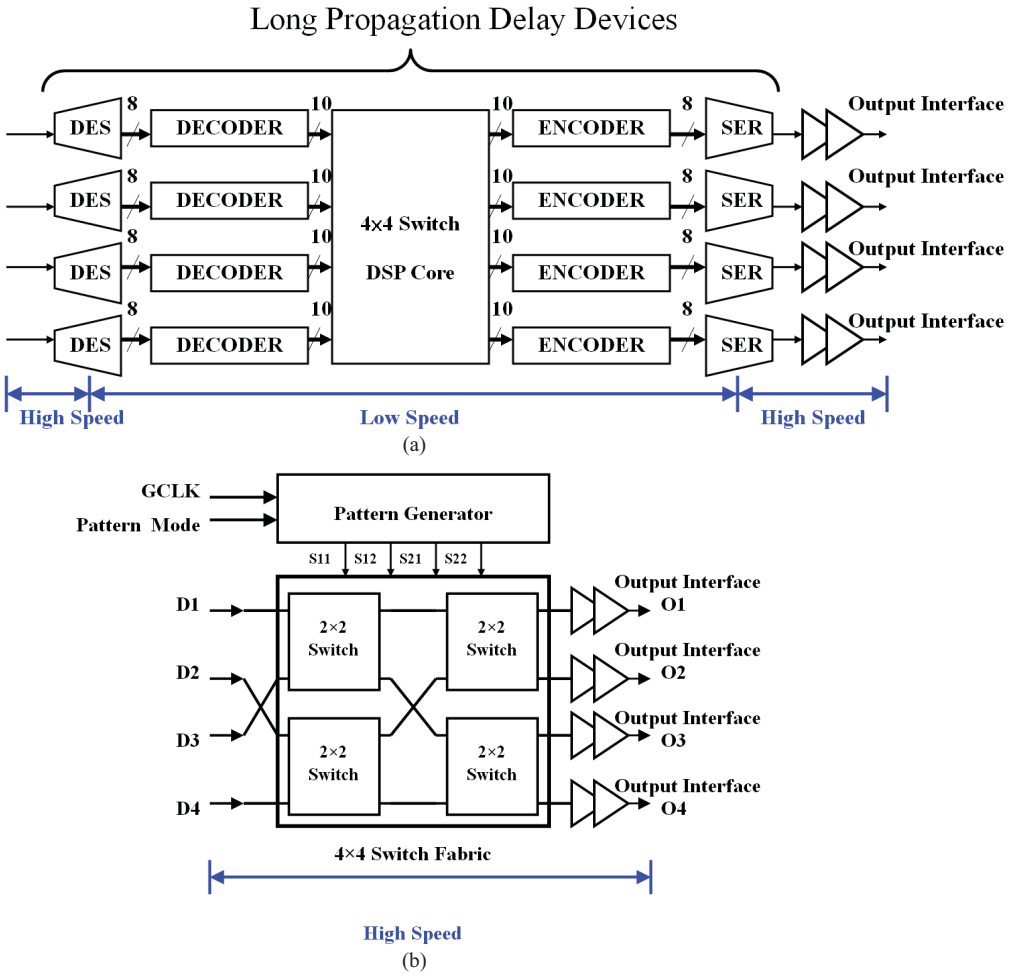


Fig. 2. (a) Conventional 4 × 4 DSP switch architecture. (b) Proposed LB-BvN 4 × 4 switch architecture.

which convert high-speed serial stream to low-speed parallel data for DSP core, can be saved in the design. In this paper, a stacked current source and symmetric topology in CML DFFs are applied to ensure high-speed performance, low propagation delay, and area saving. The pMOS active load and active back-end termination are also adopted to increase the speed of the circuit. This is the first LB-BvN 4 × 4 switch fabric IC without SerDes interfaces that can operate directly in high-speed domain. The overall data switching rate of the LB-BvN 4 × 4 switch fabric IC is up to 32 Gb/s (8 Gb/s/channel). With only 0.8 ns propagation delay, the proposed LB-BvN 4 × 4 switch fabric IC is designed for feedback-based switch systems. A high order LB-BvN $N \times N$ switch can be constructed by using the LB-BvN 4 × 4 switch modules recursively. The contributions of this paper are as follows:

1) design and implementation of a low propagation delay 4 × 4 switch fabric IC for a feedback-based switch system to improve throughput;
2) operating the switch core at the high-speed domain to remove the latency and power consumption overheads of the serial-to-parallel (S/P) conversion circuits in SerDes;
3) design of the switch pattern generator by using only two DFFs to replace the $O(N^3)$ matching algorithm.

The rest of this paper is organized as follows. In Section II, design motivation and the analysis of propagation delay and throughput are given. Then, we introduce two special connection patterns for creating feedback path in Section III. The overall system architecture design of the LB-BvN switch and the details of circuit design techniques to boost switching speed are presented in Section IV. We also present the concept of parallel distribution of the pattern generation block into each basis 4 × 4 switch, in which only two DFFs are used. In Section V, the measurement results and the system simulation results of the proposed LB-BvN 4 × 4 switch fabric IC are provided. Finally, we give a short conclusion in Section VI.

II. Propagation Delay and Throughput in a Feedback-Based Switch System

The performance of feedback-based two-stage switch system depends on the round-trip time (RTT) of transmitting packet departure information. The system architecture of an LB-BvN $N \times N$ ($N = 16$) switch constructed by 4×4 switches with feedback path is shown in Fig. 3(a). It includes a load-balancing stage (the first stage switch), VOQs, and a switching stage (the second stage switch). Especially, SerDes interfaces are commonly inserted in commercial switch fabric IC to reduce pin counts and the routing complexity. The SerDes are the blue blocks shown in Fig. 3(a).

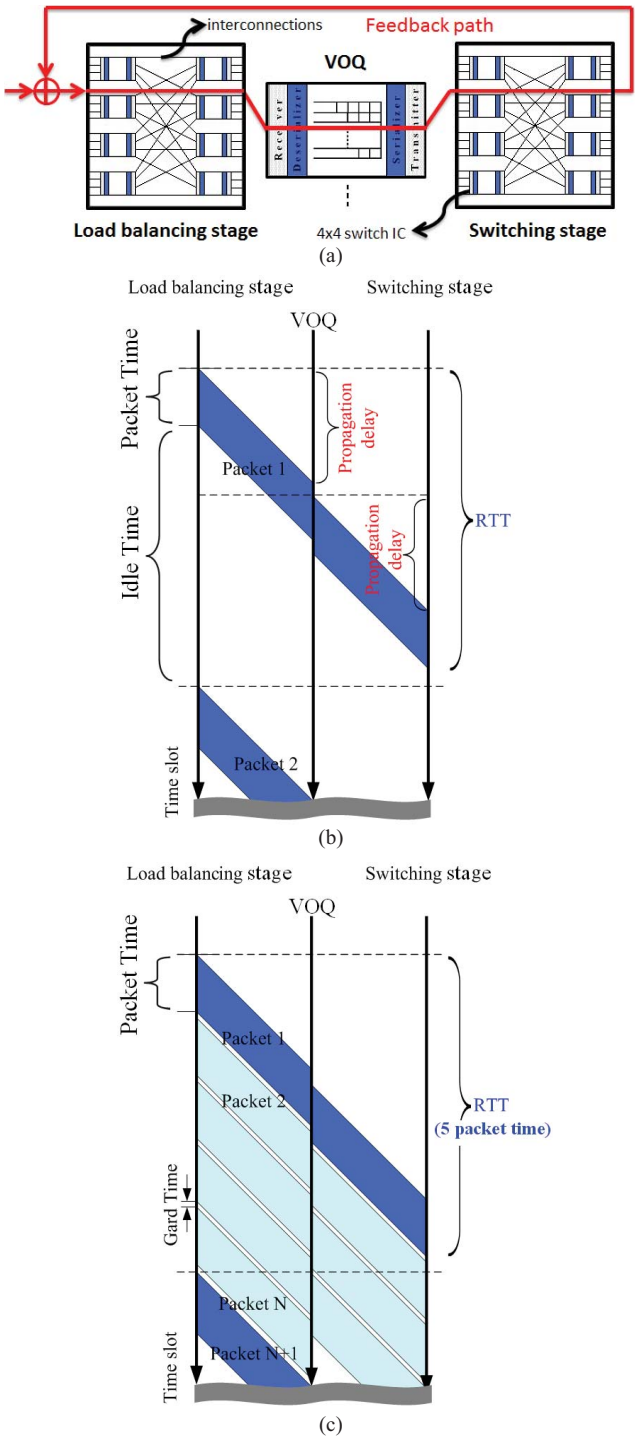


Fig. 3. (a) Feedback-based LB-BvN switch with SerDes and transceiver interfaces. (b) RTT of the feedback path (without pipelining). (c) Pipelined version RTT of the feedback path.

The SerDes includes S/P, parallel to serial (P/S) and 8 B/10 B encoder and decoder (CODEC) functions that are operated at the low-speed domain. The transceivers (dotted grey blocks) operated at the high-speed domain are added for high-speed interfaces. A receiver contains the amplifier, equalizer, and clock data recovery (CDR) blocks for resynchronization. The transmitter includes the equalization and pre-emphasis functions. Since transceivers are operated in the high-speed domain, their delay is much smaller than that of the SerDes. When building high-radix switches, we assume that all the line cards (VOQs are in the line cards) and switch cards are placed in the same chassis or are very close. The delay from the switch card to the line card is through backplane or short cables. So we neglect the transceiver and trace delay for simplicity in the following analysis and focus on the delay through the two load balanced switches.

Fig. 3(b) shows the RTT of the feedback path which includes the packet time, propagation delay from the load-balancing stage to the VOQs (physical delay), delay in the VOQs (physical delay), and propagation delay from the VOQs to the switching stage (physical delay). A longer RTT means that input packets have to wait longer for information from the middle-stage VOQs to keep packets in sequence. After scaling up, the feedback-based system might degrade the system throughput since the next packet has to wait for the departure information of the previous packets in the VOQs. When the propagation delay in the two-stage switch is longer than the packet time, the system throughput rate decreases as the RTT increases.

Taking an LB-BvN 16×16 switch system as an example, we have four sets of SerDes interfaces in the switches and one set in the VOQs that are included in the feedback path. Most of the SerDes components add the 8 B/10 B function to generate d.c.-balanced data stream for the ease of clock recovery. Since the 8 B/10 B encoder/decoder operates at the parallel data domain and runs at a much lower clock so the 8 B/10 B encoder/decoder function adds extra latency to the SerDes. The SerDes latency includes two parts: 1) the latency of the SerDes tree and 2) the 8 B/10 B encoder and decoder latency.

Taking the 10-Gb/s Ethernet (10 GbE) transceiver design in [26] for example, it contains four-channel 3.125 Gb/s data with multiplexing index N of 8. The SerDes runs at 1.6 GHz clock with one unit interval (UI) equal to 0.625 ns. The SerDes front-end takes 1 UI and the multiplexing and demultiplexing in the SerDes tree estimates another N UI each. Therefore the data path latency of the SerDes is on the order of $2N+1$. The 8 B/10B encoder/decoder runs at a slower clock of 160 MHz, whose clock period Tclkp is 10 UI (6.25 ns). The latency of the 8 B/10 B encoder/decoder takes more than 14 Tclkp (3 Tclkp for encoder, 5 Tclkp for decoder, and 6 Tclkp for alignment and other functions) [27]. The delay of SerDes including 8 B/10 B is approximately $(14\times6.25+(2N+1)\times0.625)=98.1253$ ns. The standard allows a combined transceiver (TX + RX) latency of 2048 bit times (in terms of the bitrate in serial link) [26]. Each pair of SerDes interface with 8 B/10 B encoder/decoder operated in the low-speed domain contributes about 100 ns propagation delay [23], [24], [26], [28]. Then it results in over 500 ns RTT in the system without considering the delay in the VOQs. A longer RTT means that input packets have to wait longer for information from the middle-stage VOQs to keep packets in sequence.

An 8 B/10 B CODEC is primarily used to support ac-coupled links, particularly in optical systems. It can also be used as a runlength-limited (RLL) code, but all modern systems such as 10 GbE, 16 GFC, and 25 GbE move away

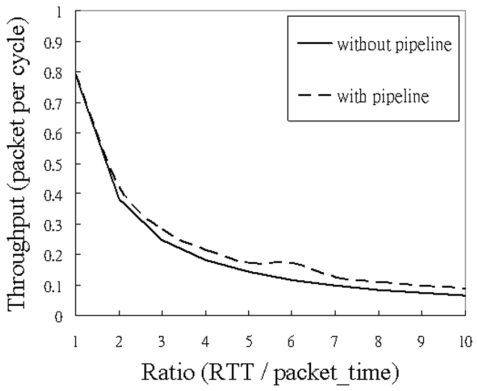


Fig. 4. 100 × 100 switch throughput degradation simulations of mailbox algorithm with $\delta = 3$.

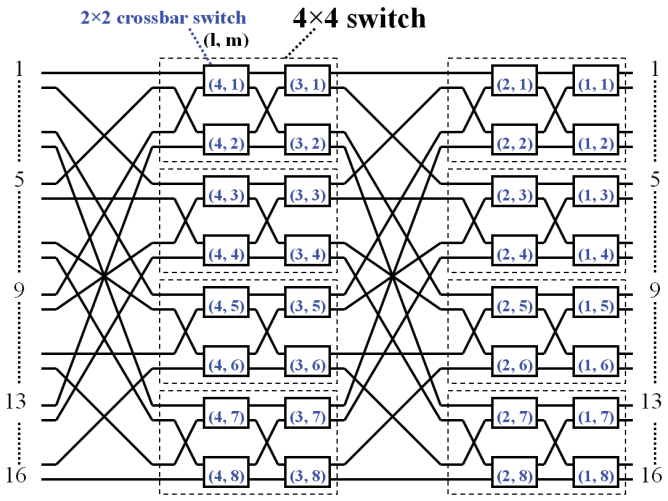


Fig. 5. 16×16 switch fabric constructed using eight perfect shuffle-connected 4 × 4 switch modules [16].

from it due to too much overhead. It is theoretically possible to get much better latency using the SerDes approach with DSP in a custom ASIC. High latency in SerDes of 100 ns arises from the use of off-the-shelf chips with legacy protocols such as XAUI with 8 B/10 B [26].

We adopt the mailbox algorithm [16] in our simulation to show the throughput degradation. The simulation setups are as follows: 1) switch size 100 × 100; simulation cycle count or packet time slots per input port 10 000; 2) packet arrival rate 1.0 (one packet generated per cycle for each input port); and 3) the cell index (δ) of virtual waiting time that keeps middle stage packets in order fixed as 3 to achieve the maximum throughput [16].

Fig. 4 shows the simulation results for the throughput rate versus the RTT. In most current literature on two-stage feedback switches [15]–[20], only throughput and delay versus traffic loads or maximum throughput versus switch size is provided. We assume that there is no propagation delay in the ideal case. The ideal feedback-based system without any propagation delay only needs one packet time to send the packet as well as the information (RTT = 1). Other simulation results with different propagation delays (RTT from 1 packet time to 10 packet time) are also demonstrated in Fig. 4 (solid line). The throughput degrades from 79.1% to 6.5% when the RTT ratio increases from 1 to 10 because several time slots are needed to send one packet as well as the middle-stage VOQs information.

To solve this problem, one straightforward strategy is to fill up the idle time [shown on the left side of Fig. 3(b)] in the RTT by sending more packets [as shown in Fig. 3(c)]. When the RTT is r times longer than the packet transmission time, we send r packets to fill up the idle time. We call this the pipeline method, as shown by the dashed line in Fig. 4. The improvement of the pipeline method is limited. In the mailbox switch architecture, N FIFOs are placed in front of the first-stage switch as input buffers. Since FIFOs have the first-come-first-serve property, an HOL packet cannot be transmitted to the central VOQs because of collision under medium or heavy traffic. Under these traffic conditions, every input port can send only one packet on average to a specific destination port, even for the pipeline case, because the FIFO input buffers block multiple packets in the line card.

Adding VOQs at the input ports improves the throughput significantly in pipelined transmission. We run the simulation of adding VOQ with size of three packets at the input ports for pipelined transmission under traffic injection rate 1. Under pipelined transmission, the throughput of the VOQ based (versus the FIFO-based) approach under RTT 2, 4, 8, and 10 are 82.3%(41.97%), 66.5%(21.63%), 37.3%(11.08%) and 29.9%(8.96%), respectively. The throughput of the VOQ-based approach is larger than that of the FIFO-based approach. When the packet size in the VOQ increases to 8, the throughput can still be above 75.5% for RTT = 10. The memory buffers needed in VOQ and FIFO are O(N^2) and O(N), respectively.

The throughput degradation problem becomes severe when the switch system scales up. We use a 4×4 switch as a module to construct a high-order switch, as shown in Fig. 5. From cost and flexibility considerations, the approach of using discrete small switch chips on a PCB board is commonly adopted. An LB-BvN $N \times N$ switch needs $((N/4) \times \log_4 N)$ 4×4 switches. In a feedback-based switch system, $2 \times (N/4) \times \log_4 N$ SerDes are needed, and it means that $2 \times \log_4 N$ SerDes propagation delay will be introduced to the feedback-based system. The throughput simulation after scaling up is shown in Fig. 6, which demonstrates the throughput degradation problem when the propagation delay is 1, 2, or 4 times that of the packet transmission time. The throughput degradation results are compared with those of the ideal case (without any propagation delay). It is important to reduce the propagation delay especially in a feedback-based system for high-radix switches.

III. Connection Patterns for Switch Systems

In this section, we review two special series of patterns, named symmetric time division multiplexing (STDM) patterns [16] and staggered symmetric patterns [18], which can be applied in the feedback-based LB-BvN $N \times N$ switch system. In an LB-BvN $N \times N$ switch, we can use the special set of the STDM connection patterns to replace the matching algorithms. As shown in Fig. 7(a), the input i is connected to the middle

stage j on the tth time slot by the following rule:

$$(i + j) \mod N = (t + 1) \mod N. \quad (1)$$

A switch fabric that implements the connection patterns in (1) is called an STDM switch because input i is connected to output j if and only if input j is connected to output i. The connection patterns of the STDM change with time t, and the connection patterns for the two-stage switches at four different time slots for a 4×4 switch are shown in Fig. 7(a). The input ports are linked back to the same output ports through two-stage switch connection patterns for sending feedback information.

An LB-BvN $N \times N$ switch is constructed by using $((N/2) \times \log_2 N)$ 2×2 switches. The STDM connection pattern of each 2×2 switch depends on its position in the LB-BvN $N \times N$ switch module and the current time slot. The position of a 2×2 switch is defined by the column index l and row index m (l, m) as shown in Fig. 5. The column index l of each 2×2 switch is defined from right to left as $1, 2, \ldots, \log_2 N$, and the row index m is defined from top to bottom as $1, 2, \ldots, N/2$. According to [16], the STDM connection pattern of the mth switch of the lth stage at time t for a 2×2 switch can be determined by

$$\Psi(l, m, t) = \left\lfloor \frac{\left(t - \Phi(l, m)\right) \mod 2^l}{2^{l-1}} \right\rfloor \quad (2)$$

where

$$\Phi(l, m) = \left((m - 1) \mod 2^{l-1}\right) + 1. \quad (3)$$

We set the bar connection pattern if (2) equals zero, and set the cross-connection pattern otherwise. Table I shows the STDM connection pattern for each 2×2 switch in our proposed 4×4 switch module.

The 4×4 STDM switch, shown in Fig. 8(a), is built by cascading four 2×2 switches with banyan network. The sequence of four numbers on the top of each 2×2 switch represents the four connection patterns, which are shown in Fig. 7(a). The "1" denotes cross-connection and the "0" denotes bar connection of a 2×2 switch. Each connection pattern in Fig. 7(a) can be realized by the corresponding connection patterns in the 2×2 switches in the banyan network. In particular, the connection patterns in each 2×2 switch of Fig. 8(a) represent the third of the connection patterns ($t = 4n+3$) in the 4×4 STDM switch.

The staggered symmetric connection patterns can be applied to the feedback-based switch system using staggered ordering algorithm [18]. The input i, connected to the middle stage j, and the middle stage j, connected to the output k, on the tth time slot follow the following two rules [18]:

$$j = (i + t) \mod N \quad (4)$$

$$k = (j + N - 1 - t) \mod N. \quad (5)$$

The feedback-based scheduling approach proposed by Bing *et al.* [18] can resolve the long propagation delay problem of the UFS and CR switches under heavy traffic. This approach uses VOQs as input buffers and adopts pipelined transmission, small middle-stage VOQs, and staggered symmetry connection patterns. Information in central VOQs is fed back to ensure in-order packet delivery. The simulation results show that this feedback-based scheduling outperforms the CR and UFS switches for delay and throughput [18]. Our implemented switch can be configured to generate the staggered patterns for this feedback-based switch. The staggered symmetric 4×4 connection patterns are demonstrated in Fig. 7(b). The staggered symmetric connection patterns provide another kind of feedback path that is different from STDM connection patterns. According to [18], the feedback information from VOQs to output ports (line cards) is passed through the previous cycle switching pattern. The line cards use the information, just renewed, to decide which packets can be sent to the VOQs. For example, ordering information through the $t = 4n + 1$ staggered symmetric connection pattern of the switching stage to the output ports helps the next cycle $(t = 4n + 2)$ input ports send packets to the VOQs. It can be applied in systems using staggered ordering algorithms [18]. However, these connection patterns are different in the load-balancing stage and switching stage, so reconfigurable design is required. The pattern generator design details are provided in the next section. The staggered symmetric connection patterns in each 2×2 switch are shown in Fig. 8(b) and (c). Table I shows the staggered symmetric connection patterns for each 2×2 switch and the relations between the STDM connection patterns and the staggered symmetric connection patterns.

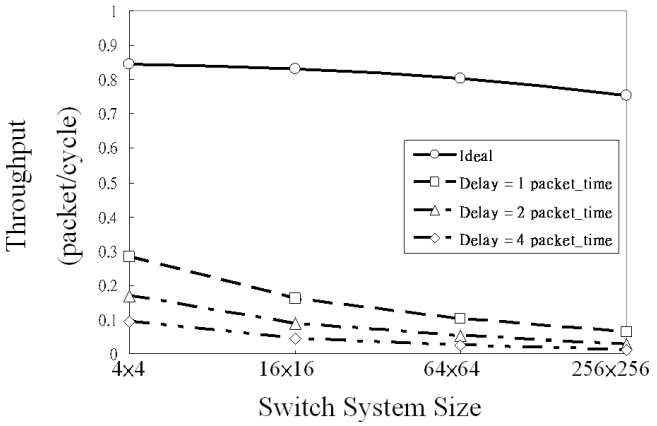


Fig. 6. Throughput degradation simulations of mailbox algorithm with $\delta = 3$ in different switch sizes.

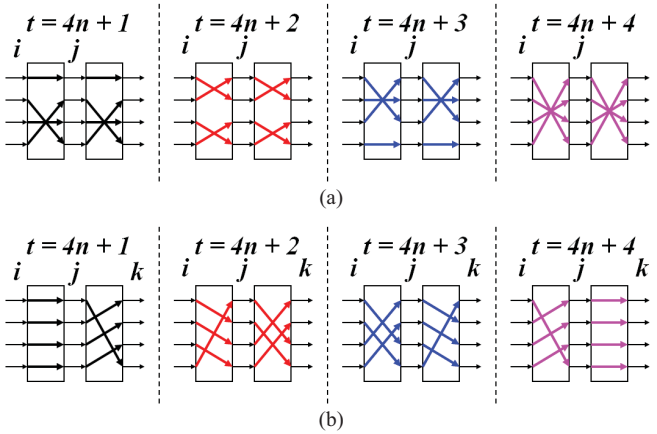


Fig. 7. (a) Example of two-stage 4×4 STDM connection patterns [16]. (b) Example of two-stage 4×4 staggered symmetric connection patterns [18].

TABLE I
4 × 4 SWITCH SELECT SIGNAL COMPARISON

Load balancing stage		
	STDM connection patterns [16]	Staggered symmetric connection patterns [18]
S11	0101	0101
S12	0101	0101
S21	0011	0011
S22	1001	**0110**

Switching stage		
	STDM connection patterns [16]	Staggered symmetric connection patterns [18]
S11	0101	**1010**
S12	0101	**1010**
S21	0011	**1100**
S22	1001	**0110**

IV. PROPOSED LB-BVN SWITCH ARCHITECTURE AND CIRCUITS DESIGN TECHNIQUE

To solve the long-standing RTT issue in the feedback-based LB-BvN switch system, we propose the switch IC architecture directly in high-speed domain without SerDes interfaces. The equalization, error correction, and resynchronization functions are handled by the transmitter, receiver, and CDR. Our design is limited to a fully synchronous system where one global clock drives all switches. Under an asynchronous system, CDRs can be adopted to recover the data in each channel. Then, a a global clock is used to resample all the data in different channels before sending to the switch fabric. The system does not go through traditional DSP system design flow [see Fig. 2(a)] because the connection patterns in LB-BvN are periodic and deterministic. The overall architecture of the LB-BvN 4 × 4 switch fabric is shown in Fig. 2(b). It contains a switch core, a pattern generator, and input/output interfaces.

In the following, we introduce the design of an LB-BvN $N \times N$ switch at the system level. Obviously, one can implement an LB-BvN $N \times N$ switch directly with a specific number N. But for flexibility and ease of VLSI design complexity, one strategy is to design a smaller switch to be the basic unit, and then construct an $N \times N$ switch with these basic units by perfect shuffle connection [16]. For example, a 16×16 switch is constructed by 32 2×2 switches as shown in Fig. 5. In the following subsection, we show how to decompose a series of $N \times N$ connection patterns into those small 2×2 switches.

The overall architecture of an LB-BvN 4×4 switch is shown in Fig. 2(b). The pattern generator, the 4×4 switch built by four 2×2 switches, and the input/output interfaces are the key blocks. The pattern generator provides the deterministic connection patterns for LB-BvN switches. The 2×2 switch realizes the crossbar function which is controlled by the connection pattern. The output interfaces provide high current-driving efficiency for back-end termination.

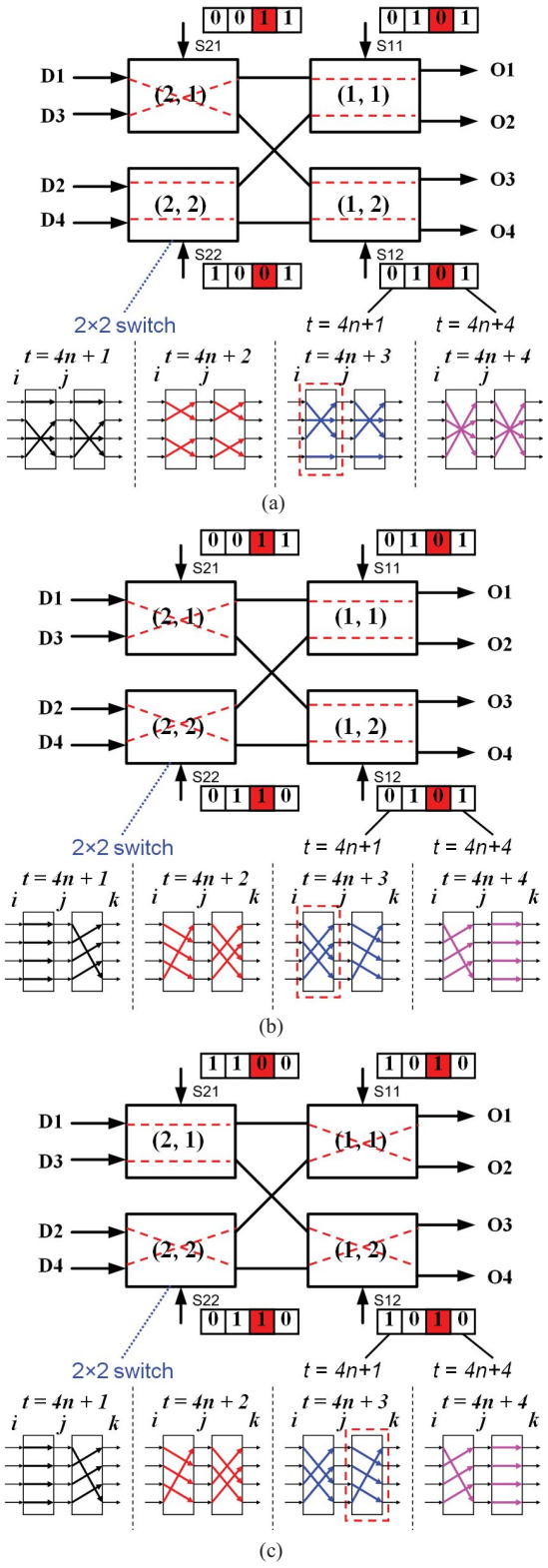


Fig. 8. Block diagram of LB-BvN 4 × 4 switch fabric basic IC with (a) STDM connection patterns, (b) staggered symmetric connection patterns for load-balancing stage, and (c) staggered symmetric connection patterns for switching stage.

A. Pattern Generator Block Design

There are three methods to implement the pattern generation block:

1) mapping directly from mathematical equations;
2) using shift registers to memorize all the states;
3) using a clock divider with a phase shifter.

Method 1 directly implements methods 2 and 3, which deal with power-of-two modulus divisions and many time-consuming arithmetic operations, such as the addition and subtraction. In method 2, equations are expanded in advance and then all states have to be memorized in considerable registers. Actually, connection patterns expressed by 2) are periodic. After observing the behavior of all states expanded by 2), the third method is proposed and only two DFFs and three MUXs are used for constructing an LB-BvN 4×4 switch module.

In the LB-BvN 4×4 switch [Fig. 8(a)–(c)], we can index the four 2×2 switches using the stage index l and switch index m. For example, the index (l, m) of the upright 2×2 switch is $(1, 1)$. From Table I, we find that the STDM connection patterns ($S11$, $S12$, $S21$, and $S22$) are the same for the load-balancing stage and for the switching stage. The switching rate of ($S11$, $S12$) are twice that of ($S21$, $S22$). The $S22$ is the circular shift version of $S21$. So the pattern generator can be built by dividing a clock signal (GCLK) and generating different phase signals by a phase shifter. The implementation of 4×4 STDM connection patterns is by using two DFFs as shown in Fig. 9(a), and the two-bit reconfigurable selecting signal for the three MUXes (the pattern mode) needs to be set at "00." These two DFFs can be built by traditional true single-phase clocked DFFs or CML DFFs, depending on the system switching speed. The 4×4 staggered symmetric connection patterns also can be generated by the same circuit as shown in Fig. 9(a). The relations ($S11$, $S12$, $S21$, and $S22$) between STDM patterns and staggered symmetric patterns are signal inversions (the bold ones in Table I). If we take the load-balancing stage for example, $S11$, $S12$, and $S22$ of the staggered symmetric patterns are equal to the STDM ones. On the other hand, the $S21$ of the staggered symmetric patterns is obtained from inverting the $S21$ of the STDM pattern. In Fig. 9(a), the staggered symmetric patterns can be got by setting the pattern mode to "10" for the load-balancing stage and "11" for the switching stage. Furthermore, if we want to construct higher order switches, such as a 16×16 or 64×64 switch, the same method can be applied.

We compare the pattern generator only to our prior work since this pattern generator is used to generate switch connection pattern and not the conventional pseudo-random binary sequence (PRBS) generator. The switch pattern generator is the key component to control the connection between the input and output ports. We are the first and the only ones to realize the load-balanced BvN switch pattern generator in hardware [24]. From the surveyed literature, there are no other LB-BvN switch pattern generators available except ours. We make significant improvement in the function, hardware complexity, and power compared to the switch pattern generator of the mathematical model as well as our previous work. The architecture of our previous pattern generator is shown in Fig. 9(b) [24]. In this implementation, there are 6 DFFs and 12 MUXs for generating the STDM pattern only. In this paper, the switch pattern generator shown in Fig. 9(a) uses only two DFFs, three MUXs, and three inverters to generate both STDM and staggered symmetric patterns. Table II shows the comparison between these two architectures. The new

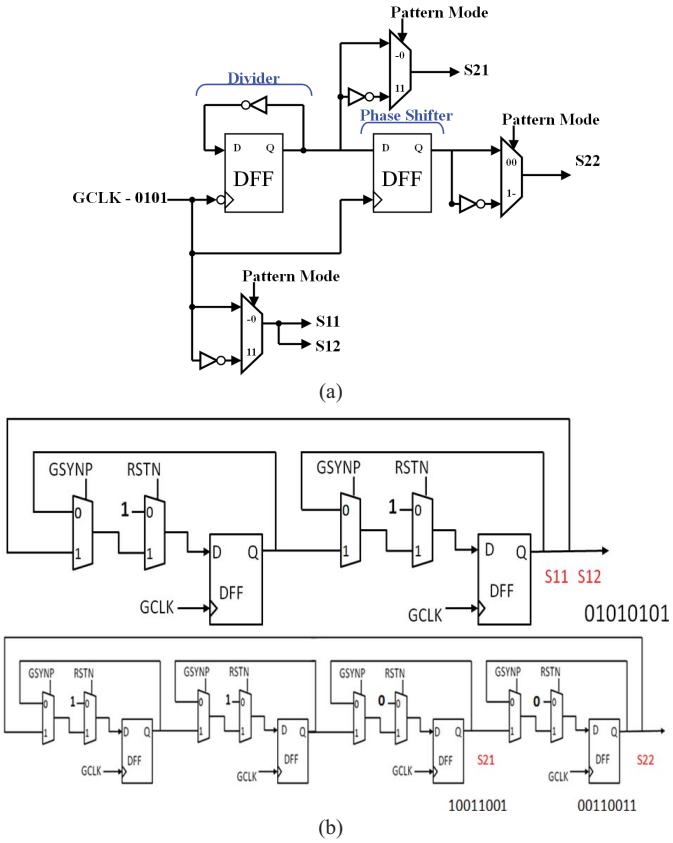


Fig. 9. (a) Proposed pattern generator block and select signals ("-" means that the bit can be either "1" or "0"). (b) Pattern generator block in type-II switch [24].

TABLE II
SWITCH PATTERN GENERATOR COMPARISON

	This paper	Type-II [24]
DFF/MUX/Inverter no.	2/3/3	6/12/0
Area (μm $\times$ μm)	230 $\times$ 115	500 $\times$ 150
Total power consumption	11.5 mW	30.6 mW

design reduces by about 65% the area and by 63% the power consumption.

B. 2×2 Switch and CML MUX/CML DFF Design

As shown in Fig. 10, the 2×2 switch consists of two CML MUXes, and two CML DFFs to realize the crossbar function. Traditionally, CMOS MUXes are applied in many digital designs or analog designs below 1 GHz. CML MUXes are adopted in our design to achieve 8 Gb/s line rate operating speed because the CML MUX outperforms the traditional CMOS MUX in both rise time and fall time.

The other component is the CML DFF, which is composed of two D-type latches. In the traditional CML DFF circuit design, each CML latch consists of an input tracking pair, which is utilized to track the input data signal while the clock transistor pair switches the current to the left branch, and a cross-coupled regenerative pair (also called the holding pair), which is utilized to hold the data while the current is switched to the right branch. A few drawbacks exist in this

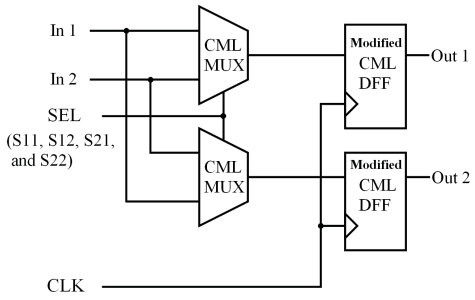


Fig. 10. Block diagram of a 2 × 2 switch.

circuit. Especially, two inherently different branches, tracking and holding, share the same current source, which tends to tie up the bias condition of these two circuits. At high data rates, the parasitic capacitances of the tracking pair transistors degrade the required minimum small-signal gain for proper tracking operation. Therefore, the tail current source must be sufficiently high to achieve a wider range of linearity and a larger transconductance. On the other hand, the holding pair does not need a large bias current at ultrahigh frequencies [29].

To solve these problems, a traditional CML DFF is modified so that the tracking sides in the two latches share a single current source and the holding sides share another current source as shows in Fig. 11(a). With this modification, the DFF becomes more symmetric and thus results in a lower level of switching noise at 10 Gb/s data rate [30] [as shown in Fig. 11(b)]. In addition, each of the tail current sources in MUXes and DFFs is replaced by a stacked current source, which consists of two cascaded nMOS transistors [31]. The upper transistor is a low-threshold voltage device and the bottom one is a regular-threshold voltage device. This configuration results in a flat current source characteristic shown in Fig. 11(c), since output resistance increases from r_o to $2r_o + g_m r_o^2$ as shown below

$$R_o = r_o + r_o(1 + g_m r_o) = 2r_o + g_m r_o^2. \tag{6}$$

From our simulations, a single CML MUX/DFF consumes approximately 72% more power than a conventional CMOS MUX/DFF at 10 Gb/s. The clock-to-output delay of a single CML DFF is half of that of a conventional CMOS DFF. The 4 × 4 switch and the pattern generator are built from CML MUX/DFFs, which contribute 78.6% of the total power. The remaining power is consumed by the CML back-end termination design described below.

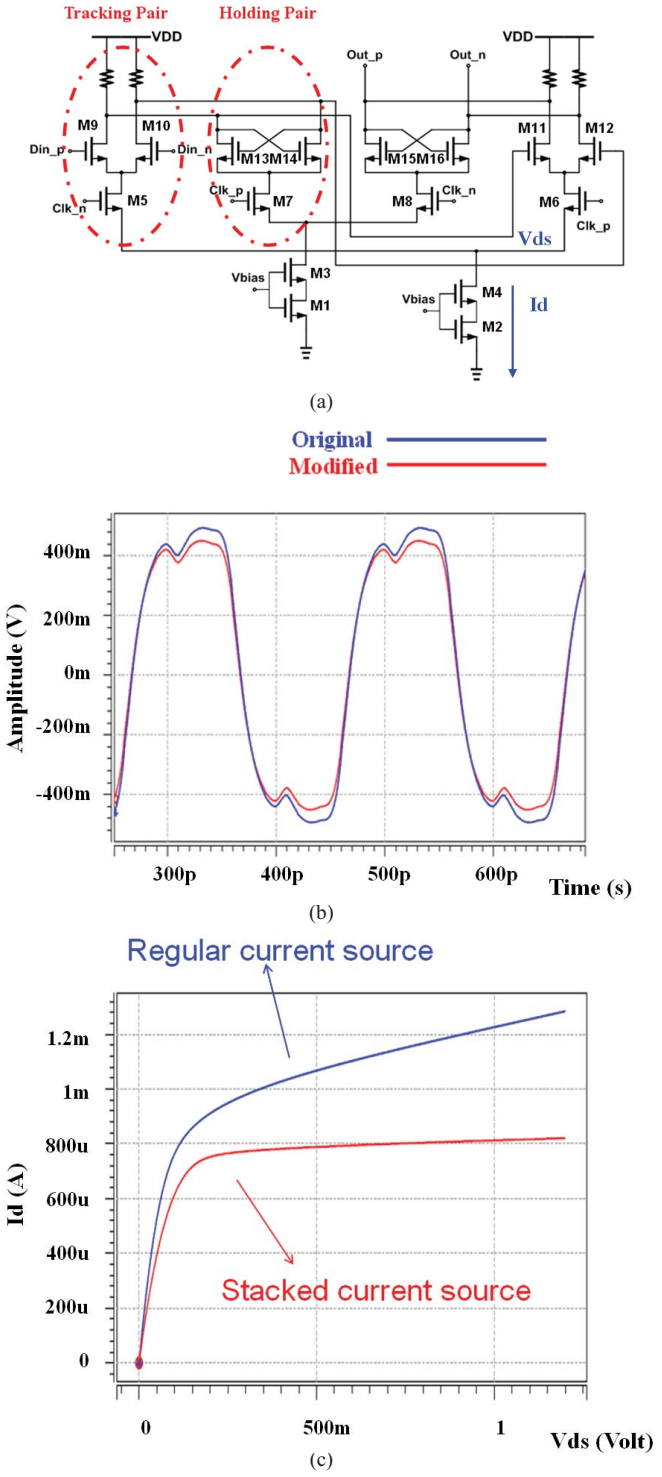


Fig. 11. (a) Modified DFF design. (b) Stable logic level characteristic. (c) Flat current source characteristic.

C. Back-End Termination Design

The CML output interface, which consists of two-stage CML buffers, is shown in Fig. 12(a). In the first stage, we use our patented pMOS active load inductive peaking technique [32], [33] to improve the operating speed. In the second stage, we propose the active back-end termination for impedance-matching the 50-Ω load.

The traditional CML output interface uses resistor loads between supply voltage (VDD) and output pairs. To improve the operating bandwidth, one can choose the on-chip inductors to replace resistors. However, on-chip inductors occupy the largest chip area and introduce significant parasitic capacitance. In our design, we use the pMOS active load inductive peaking technique [32], [33] in the first stage of the CML output interface to enhance the bandwidth. It includes active inductors formed by pMOS transistors (M9–M10) that act as active resistors connected to nMOS transistors load (M7–M8). They act as the on-chip inductors to employ inductive peaking.

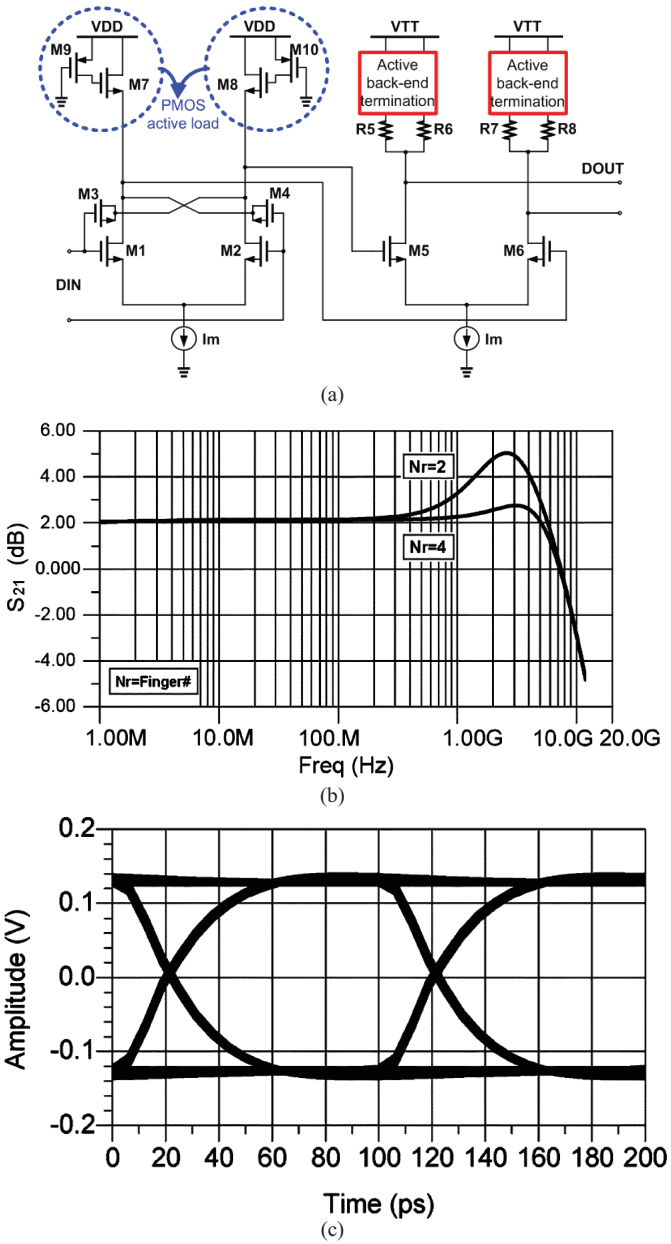


Fig. 12. (a) CML output interface: active load inductive peaking and active back-end termination. (b) *S*-parameter S_{21} plot of CML output interface. (c) Simulation eye diagram of CML output interface.

Compared to on-chip inductors, active inductors require much lower chip area and consume less power to achieve the same frequency response. We also incorporate negative Miller capacitance (M3–M4) to meet the high-speed requirement.

With the increasing operation speed of communication networks, signal reflection is getting worse because of the impedance mismatch, which impacts the performance of the transmission. To resolve this problem, some circuit designs use passive back-end termination but it costs 50% modulation current. On the other hand, some other circuit designs use ac-coupled back-termination. However, it is very difficult to design a high-quality capacitor in chip process and it also occupies a large chip area. We adopt the active back-end termination technique in the last stage of CML output interface

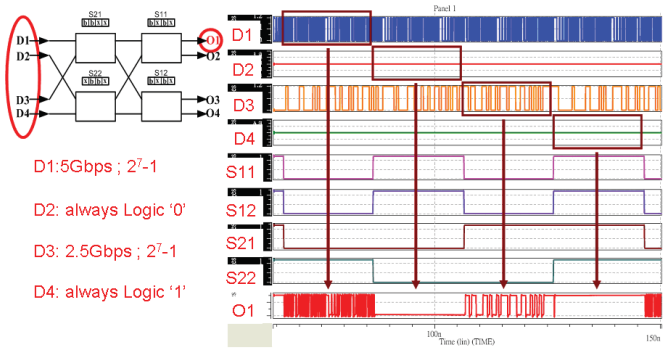


Fig. 13. Post-simulation waveform of one channel output at 5 Gb/s.

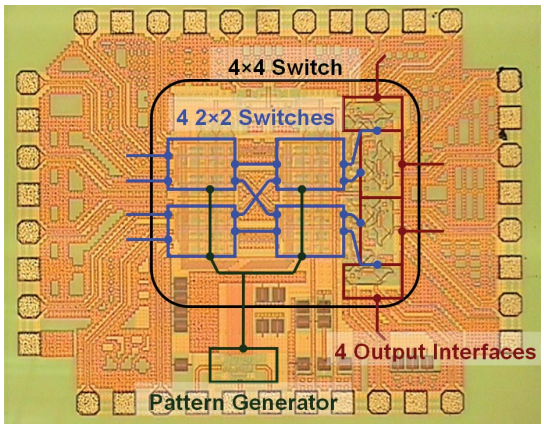


Fig. 14. Die photo of the LB-BvN 4 × 4 switch fabric IC.

[see Fig. 12(a)] to match the 50-Ω load environment. The *S*-parameter S_{21} plot and simulation eye diagram of CML output interface are shown in Fig. 12(b) and (c). This scheme provides high current driving efficiency than passive back-end termination. As compared to ac-coupled active back-end termination, it occupies less chip area because it needs no on-chip capacitor. The CML output interface is suitable for 10-Gb/s design to drive the 50-Ω load environment, so we adopt it in our design.

V. Experimental Results and System Throughput Performance Analysis

A. Experimental Results

The LB-BvN 4 × 4 switch fabric IC has been implemented in 0.13-μm CMOS technology. In Fig. 13, the post-simulation waveform of the connection patterns and one channel output at 5 Gb/s are shown. For the ease of demonstration, we set the input port D1 with 5 Gb/s pseudo random binary sequence content and input port D3 with 2.5 Gb/s PRBS content; we also set the input port D2 with logic "0" and the input port D4 with logic "1."

The total area including PADs is 1.380 × 1.080 mm^2. Fig. 14 shows the chip microphotograph of the 4 × 4 switch. The configuration printed circuit board (PCB) is shown in Fig. 15. The four-layer PCB is fabricated with Nelco 4000-13. From the measurement result, the total power is 134 mW, which is almost 20% and 15% of that in previous works as

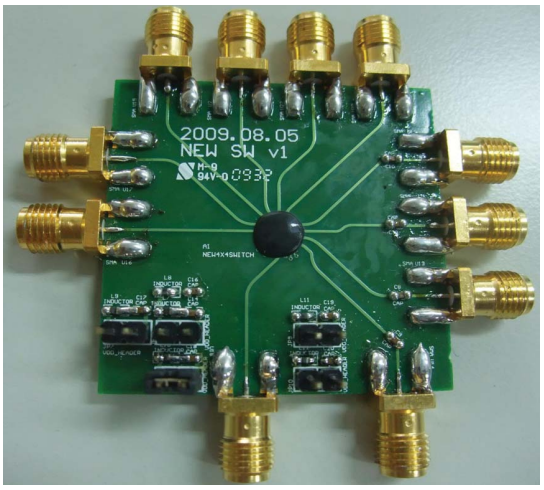


Fig. 15. PCB of the LB-BvN 4 × 4 switch fabric IC.

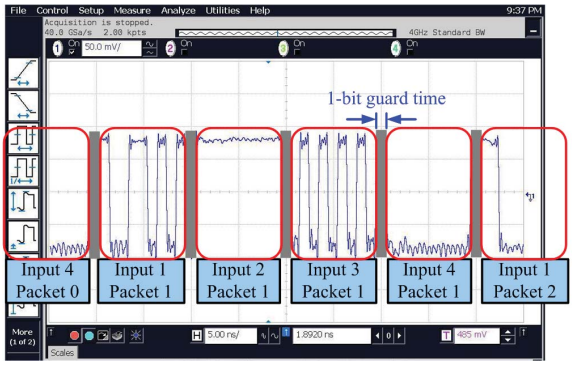
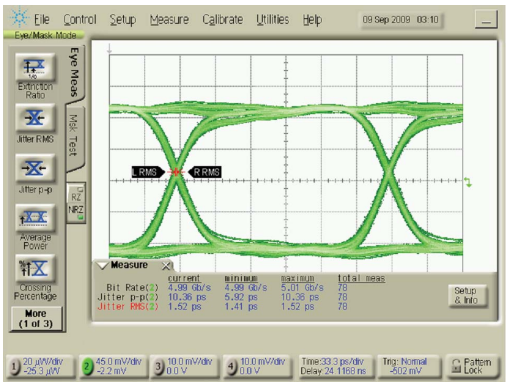


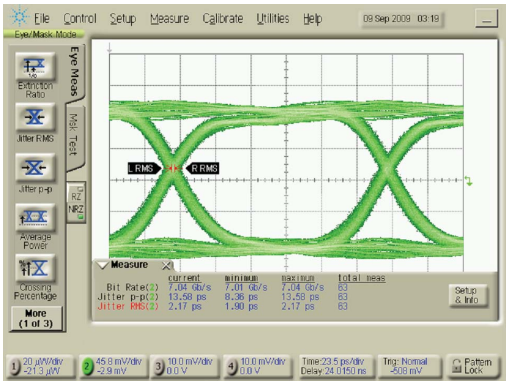
Fig. 16. One of the measurement output waveforms.

shown in Table III. With the same circuit function as the post-simulation, one of the measurement output waveforms is presented in Fig. 16 for comparison. Packets are evenly switched to each output port with 1-bit guard time. We also test the chip with different data rates: 5, 7, and 8 Gb/s, and different PRBS patterns. Eye diagrams of one channel output for different data rates are shown in Fig. 17(a)–(c), respectively. The peak-to-peak jitter ($\text{Jitter}_{\text{pp}}$) is 10 ps at 5 Gb/s, 14 ps at 7 Gb/s, and 20 ps at 8 Gb/s.

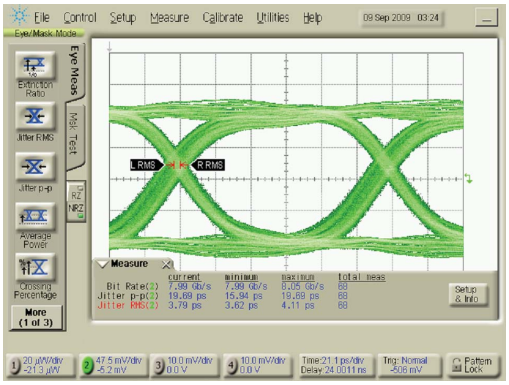
The data switching rate of the proposed LB-BvN 4 × 4 switch is 32 Gb/s (8 Gb/s/channel). Table III shows the comparison with our previous works and other types of switches. The type-I switch in [23] and the type-II switch in [24] are implemented with different SerDes interface structures. Compared to the type-II 4 × 4 switch, the overall propagation delay is reduced from 30 to 0.8 ns by removing the SerDes part and operating the switch on the high-speed domain. The 0.8 ns delay includes the delay from the switch and CML drivers. The delay of the 2 × 2 switch and the CML output interface are 0.28 and 0.24 ns, respectively, under the data rate of 8 Gb/s. The main power saving over type-II switch is from two parts. One is from removing the four channels of SerDes (560 mW) and one PLL (60 mW), which contributes 620 mW power (72.9% of the total power). The other is from removing the 8 B/10 B CODEC, which contributes 125 mW



(a)



(b)



(c)

Fig. 17. PCB measured eye diagrams at different specifications. (a) $\text{Jitter}_{\text{pp}}$ = 10 ps at 5 Gb/s. (b) $\text{Jitter}_{\text{pp}}$ = 14 ps at 7 Gb/s. (c) $\text{Jitter}_{\text{pp}}$ = 20 ps at 8 Gb/s.

power (14.7% of the total power). This proposed LB-BvN 4 × 4 switch fabric IC outperforms the results of our previous works.

We compare our design with the shared memory switch Fulcrum FM6000 [34] and the input-queued switch [14]. The Fulcrum FM6000 data center switch can support 36 × 36 switching with per port up to 10 Gb/s and total 720 Gb/s bandwidth [34]. The latency is from 300 ns to multiple microseconds depending on the switch nodes. When the switch size scales up, the speed and latency become challenging issues in the shared memory switch design. The input-queued 32 × 32 switch [14] adopts the 1SLIP algorithm. The implementation has 3.125 Gb/s data rate per port and has a virtually full throughput for realistic VOQ sizes. However, the input-queued switch has a much longer cell delay than the load balanced

TABLE III
SWITCHES COMPARISON

	This paper	Type-II [24]	Type-I [23]	Input-queued [14]	Shared memory [34]
Switch size	4×4	4×4	8×8	32×32	36×36
Technology	0.13 μm	0.13 μm	0.18 μm	0.13 μm	65 nm
Supply voltage	1.2 V	1.2 V	1.8 V	1.2 V	0.6 V
Maximum speed/Ch.	9 Gb/s	8.8 Gb/s	3.2 Gb/s	3.125 Gb/s	10 Gb/s
Total data rate	32 Gb/s	28 Gb/s	25.6 Gb/s	100 Gb/s	360 Gb/s
Jitter_{pp}	20 ps	18 ps	21 ps	NA	NA
Chip area (including PADs)	1.380×1.080 1.49 mm^2	3.000×2.480 7.44 mm^2	3.650×3.570 13.03 mm^2	1.75×1.75 3 mm^2 (w/o pads)	NA mm^2
Total power	105 mW (switch core) 29 mW (driver) 134 mW	260 mW (switch core) 590 mW (4SerDes + PLL) 850 mW	230 mW (switch core) 500 mW (SerDes + PLL) 730 mW	NA NA NA	NA 1 W/10G port NA
Propagation delay	0.8 ns	29.5 ns	50 ns	NA	300 ns to sub-microsecond

TABLE IV
PROPOSED SWITCH ON TECHNOLOGY/SIZE SCALING

Switch size	4×4	4×4	4×4	32×32
Technology	0.13 μm	90 nm	40 nm	0.13 μm
Supply voltage	1.2 V	1.2 V	1.2 V	1.2 V
Data rate/Ch.	8 Gb/s	10 Gb/s	20–25 Gb/s	8 Gb/s
Switching rate	32 Gb/s	40 Gb/s	80–100 Gb/s	256 Gb/s
Jitter_{pp}	20 ps	25 ps	0.63 ps (20 Gb/s)	25 ps
Chip area (without PADs)	0.347 mm^2	0.265 mm^2	0.0304 mm^2	9.4 mm^2
Power consumption	134 mW	95 mW	137 mW	2.8 W
Propagation delay	0.8 ns	0.5 ns	0.088 ns (20 Gb/s)	1.7 ns

switch under heavy traffic, especially when the arrival rate is over 0.9 [14].

We implement our proposed 4×4 switch in 130-, 90-, and 40-nm CMOS technology. The implementation results such as area, propagation delay, and operation speed in terms of various technologies are shown in Table II Under 40-nm CMOS technology, the post-simulations show that the maximum speed per channel in the 4×4 switch is 25 Gb/s with a total data rate of 100 Gb/s. The area and propagation delay of the 4×4 switch are reduced significantly in 40-nm technology. Using more advanced CMOS technology or InP double-heterojunction bipolar transistor (DHBT) technology, a line rate of 40–100 Gb/s is possible to achieve.

Analog circuits such as synchronization and error correction are needed in building a realistic large switch system. This part of circuit may account for as much as 40% of the total power consumption of typical SerDes macros with synchronization and error correction in 130-nm CMOS technology. These circuits are needed for both the digital domain and our proposed analog domain switches. This power consumption is unavoidable for building high-speed switches. The conversion between serial and parallel is needed only in digital domain switches and this circuit accounts for 10% –15% power of the SerDes with synchronization and error correction [35], [36]. Although the conversion circuit consumes less power than the synchronization and error correction, a few watts of power can be saved by removing the conversion circuit at every port while building a large switch system.

To show the scalability of the load balanced switch, we have implemented a 32×32 load-balanced switch that is recursively constructed from the 4×4 and 8×8 switches [16]. The 8×8 switch is constructed from the 4×4 and 2×2 switches. As shown in the Fig. 10, every connection port has only two fixed loads. This makes the loading consideration of the design easier, and the speed deterioration is limited when high-radix switches are constructed. The post-simulation shows that design could run at 8 Gb/s per channel and the total data rate is 256 Gb/s. The area of the implemented 32×32 switch core is $2.5912 \times 3.628 = 9.4$ mm^2 and the consumed power is 2.8 W. When an advanced 40-nm CMOS technology is adopted, the area can be reduced and the speed can be increased significantly.

B. System Throughput With Measured Propagation Delay

In the following, we show the system throughput performance of our proposed LB-BvN 4×4 switch fabric IC. There are more than 160 bits in a TCP/IP IPv4 or IPv6 packet. If we assume that the switch system is under a line rate of 8 Gb/s (8 Gb/s/channel), it means that the time to transmit one packet (one packet time) is about 20 ns. According to Table III, the propagation delay of our proposed LB-BvN 4×4 switch fabric IC is 0.8 ns. Our previous works with SerDes interfaces with 8 B/10 B encoder/decoder (type-I [23] and type-II [24]) are 50 and 30 ns, respectively. Part of the propagation delay is from the delay in the SerDes. For example, the delay of the type-II 4×4 switch [23] is 30 ns. The SerDes is around 21.4 ns, which contributes 71.4% of the total delay. The other contributors are the 4×4 switch and the CML output interface, which contribute 28.6% delay. Assume the transceiver delay is 0.75 ns. In a 16×16 switch system, RTT of these designs are 26.2, 273.75 [23], and 173.75 ns [24]. When the propagation delay is up to 100 ns for each pair SerDes interface in [26] and [28], it needs at least 523.75 ns RTT to send a

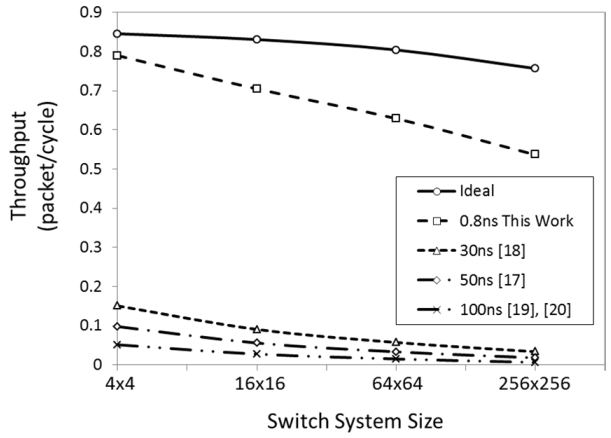


Fig. 18. Maximum throughput simulation based on measured propagation delay for different switch sizes.

TABLE V
THROUGHPUT DEGRADATION RATIO (RTT/PACKET_TIME)

Reference	This paper	[24]	[23]	[26], [28]
Propagation delay	0.8 ns	30 ns	50 ns	100 ns
Packet (160 bits)				
RTT/packet_Time	1.31	8.6875	13.6875	26.1875
Maximum throughput	0.691	0.091	0.0562	0.028
Flit (40 bits)				
RTT/packet_Time	2.24	31.75	51.75	101.75
Maximum throughput	0.421	0.0223	0.013	0.00615

packet. Under these simulation setups, the 16 × 16 switch system using the proposed LB-BvN 4 × 4 switch fabric IC modules still can maintain reasonable throughput (69.1%) as shown in Table V, while the others suffer from the throughput degradation problem (under 10%). When we further divide a packet into many smaller flits (most switch systems are based on the flits instead of packets), the throughput degradation becomes more severe. Based on the simulation results, the throughput degradation ratio of different propagation delays for packet (160 bits) and flit (40 bits) versions are shown in Table V.

The RTT is the critical parameter that impacts the throughput of a feedback-based load-balanced switch. These simulation results are close to realistic expectations in a real large-scale switch system. Reducing the switch propagation delay is one direct way to reduce the RTT of feedback-based load-balanced switches. By removing the S/P conversion and switching directly on the high-speed analog domain, we reduce the switch propagation delay significantly. For example, reducing the propagation delay from 100 to 50, 30, and 0.8 ns, the RTT/packet ratio of a 16 × 16 switch is reduced from 26.1875 to 13.6875, 8.6875, and 1.31 and the throughput increases from 2.8% to 5.62%, 9.1%, and 69.1% as in Table IV. This demonstrates that our throughput is much better than that of the others.

Fig. 18 shows the throughput simulation based on the measured propagation delay (see Table III) when the LB-BvN switch scales from 4 × 4 to 256 × 256. The system throughput of switch system using the proposed LB-BvN 4 × 4 switch fabric IC modules can achieve about 80% compared to a 15% throughput of a switch with SerDes [24]. When the switch system scales up to 256 × 256, the system throughput can still maintain about 53.7%. However, the throughput reduces to 0.673% when the switch ports are 256 with 100-ns propagation delay. The ideal case (without any propagation delay) is also shown in Fig. 18 for comparison.

VI. CONCLUSION

In feedback-based LB-BvN switch systems, the propagation delay of the whole system becomes an important issue because the packets in the switch system have to wait for feedback information for ordering. The system throughput degrades when the propagation delay gets longer, especially in systems with numerous input/output ports. In this paper, an LB-BvN 4 × 4 switch fabric IC was proposed for a feedback-based switch system to solve the throughput degradation problem; the system throughput performance analysis was also provided to support our idea. This design was fabricated in 0.13-μm CMOS technology and the chip area was 1.380×1.080 mm^2. The overall data switching rate of the LB-BvN 4 × 4 switch fabric IC was up to 32 Gb/s (8 Gb/s/channel) with only 0.8-ns propagation delay.

In the proposed circuit designs, the pattern generator for switch connections was implemented by using only two DFFs to replace the $O(N^3)$ matching algorithm. The high-speed CML DFFs with stacked current source and symmetric topology were adopted to replace the low-speed DSP core for implementing switch function. pMOS' active load and active back-end termination were used for CML output interfaces. This reduced the propagation delay of the switch module from 30 to 0.8 ns, and provided 80% area saving and 85% power saving, compared to our previous work with SerDes interfaces. The system throughput performance analysis based on the measured propagation delay also demonstrated that the switch system built by the proposed LB-BvN 4 × 4 switches outperforms other switch systems with SerDes interfaces in overall system throughput. The system throughput achieves 80% under 1.0 packet injection for each input port, and the proposed system performance is comparable to ideal cases. The proposed LB-BvN 4 × 4 switch fabric IC is highly recommended for feedback-based switch systems to solve the throughput degradation problem.

REFERENCES

[1] C. S. Chang and D. S. Lee, *Principles, Architectures and Mathematical Theories of High Performance Switches*. Beijing, China: National Tsinghua Univ. Press, May 2008.
[2] N. McKeown, "Scheduling algorithms for input-queued cell switches," Ph.D. thesis, Dept. Eng. Elect. Eng. Comput. Sci., Univ. California, Berkeley, 1995.
[3] M. J. Karol, M. G. Hluchyj, and S. P. Morgan, "Input versus output queueing on a space-division packet switch," *IEEE Trans. Commun.*, vol. 35, no. 12, pp. 1347–1356, Dec. 1987.
[4] T. Anderson, S. Owicki, J. Saxe, and C. Thacker, "High-speed switch scheduling for local-area networks," *ACM Trans. Comput. Syst.*, vol. 11, no. 4, pp. 319–352, 1993.
[5] Y. Tamir and H. C. Chi, "Symmetric crossbar arbiters for VLSI communication switches," *IEEE Trans. Parallel Dist. Syst.*, vol. 4, no. 1, pp. 13–27, Aug. 1993.

[6] N. McKeown, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," *IEEE Trans. Commun.*, vol. 47, no. 8, pp. 1260–1267, Aug. 1999.

[7] A. Mekkittikul and N. McKeown, "A practical scheduling algorithm to achieve 100% throughput in input-queued switches," in *Proc. IEEE INFOCOM 17th Ann. Joint Conf. Comput. Commun. Soc.*, Mar.–Apr. 1998, pp. 792–799.

[8] J. Dai and B. Prabhakar, "The throughput of data switches with and without speedup," in *Proc. IEEE INFOCOM 19th Ann. Joint Conf. Comput. Commun. Soc.*, 2000, pp. 556–564.

[9] Y. Li, S. Panwar, and H. J. Chao, "On the performance of a dual round-robin switch," in *Proc. IEEE INFOCOM 20th Ann. Joint Conf. Comput. Commun. Soc.*, vol. 3. 2001, pp. 1688–1697.

[10] H. N. Gabow and R. E. Tarjan, "Faster scaling algorithms for network problems," *SIAM J. Comput.*, vol. 18, no. 5, pp. 1013–1036, 1989.

[11] N. McKeown, "The iSLIP scheduling algorithm for input-queued switches," *IEEE/ACM Trans. Netw.*, vol. 7, no. 2, pp. 188–201, Apr. 1999.

[12] L. Tassiulas, "Linear complexity algorithms for maximum throughput in radio networks and input-queued switches," in *Proc. IEEE INFOCOM 17th Ann. Joint Conf. Comput. Commun. Soc.*, Mar. 1998, pp. 533–539.

[13] P. Giaccone, D. Shah, and B. Prabhakar, "An implementable parallel scheduler for input-queued switches," *IEEE Micro*, vol. 22, no. 1, pp. 19–25, Jan. 2002.

[14] N. Chrysos and G. Dimitrakopoulos, "Practical high-throughput crossbar scheduling," *IEEE Micro*, vol. 29, no. 4, pp. 22–35, Jul.-Aug. 2009.

[15] C. S. Chang, D. S. Lee, and Y. S. Jou, "Load balanced Birkhoff-von Neumann switches, part I: One-stage buffering," *Comput. Commun.*, vol. 25, no. 6, pp. 611–622, 2002.

[16] C. S. Chang, D. S. Lee, and Y. J. Shih, "Mailbox switch: A scalable two-stage switch architecture for conflict resolution of ordered packets," *IEEE Trans. Commun.*, vol. 56, no. 1, pp. 136–149, Jan. 2008.

[17] C. L. Yu, C. S. Chang, and D. S. Lee, "CR switch: A load-balanced switch with contention and reservation," *IEEE Trans. Netw.*, vol. 17, no. 5, pp. 1659–1671, Oct. 2009.

[18] B. Hu and K. L. Yeung, "Feedback-based scheduling for load-balanced two-stage switches," *IEEE Trans. Netw.*, vol. 18, no. 4, pp. 1077–1090, Aug. 2010.

[19] J. J. Jaramillo, F. Milan, and R. Srikant, "Padded frames: A novel algorithm for stable scheduling in load-balanced switches," in *Proc. Inform. Sci. Syst. 40th Ann. Conf.*, Mar. 2006, pp. 1732–1737.

[20] I. Keslassy and N. McKeown, "Maintaining packet order in two-stage switches," in *Proc. IEEE INFOCOM 21st Ann. Joint Conf. Comput. Commun. Soc.*, vol. 2. May 2002, pp. 1032–1041.

[21] Y. Shen, S. Jiang, S. S. Panwar, and H. J. Chao, "Byte-focal: A practical load balanced switch," in *Proc. IEEE High Perform. Switch. Rout.*, 2005, pp. 6–12.

[22] I. Keslassy, S.-T. Chuang, K. Yu, D. Miller, M. Horowitz, O. Solgaard, and N. McKeown, "Scaling internet routers using optics," in *Proc. Conf. Appl. Technol. Archit. Protocols Comput. Commun.*, Karlsruhe, Germany, Aug. 2003, pp. 189–200.

[23] C. T. Chiu, Y. H. Hsu, M. S. Kao, H. C. Tzeng, M. C. Du, P. L. Yang, M. H. Lu, F. T. Chen, H. Y. Lin, J. M. Wu, S. S. H. Hsu, and Y. S. Hsu, "A scalable load balanced Birkhoff-von Neumann symmetric TDM switch IC for high-speed networking applications," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2007, pp. 2754–2757.

[24] Y. H. Hsu, M. H. Lu, P. L. Yang, F. T. Chen, Y. H. Li, M. S. Kao, C. H. Lin, C. T. Chiu, J. M. Wu, S. H. Hsu, and Y. S. Hsu, "A 28Gb/s 4×4 switch with low jitter SerDes using area-saving RF model in 0.13μm CMOS technology," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2008, pp. 3086–3089.

[25] Y. H. Hsu, Y. S. Lin, P. L. Yang, C. T. Chiu, J. M. Wu, S. H. Hsu, F. T. Chen, M. S. Kao, and Y. S. Hsu, "A 32Gb/s low propagation delay 4×4 switch IC for feedback-based system in 0.13μm CMOS technology," in *Proc. IEEE Int. Symp. Circuits Syst.*, May 2010, pp. 581–584.

[26] *Texas Instruments 10 Gigabit (XAUI) Ethernet Transceivers Datasheet*. (2009) [Online]. Available: http://focus.ti.com/lit/ds/symlink/tlk3138.pdf

[27] A. X. Widmer and P. A. Franaszek, "A DC-balanced, partitioned-block, 8B/10B transmission code," *IBM J. Res. Devel.*, vol. 27, no. 5, pp. 440–451, 1983.

[28] *XILINX RocketIO^{TM} Transceiver User Guide*. (2007) [Online]. Available: http://www.xilinx.com/support/documentation/user_guides/ug024.pdf

[29] P. Heydari and R. Mohanavelu, "Design of ultrahigh speed low-voltage CMOS CML buffers, and latches," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 12, no. 10, pp. 1081–1093, Oct. 2004.

[30] T. Otsuji, M. Yoneyama, K. Murata, and E. Sano, "A super-dynamic flip-flop circuit for broadband applications up to 24 Gb/s utilizing production-level 0.2-μm GaAs MESFETS," *IEEE J. Solid-State Circuits*, vol. 32, no. 9, pp. 1357–1362, Sep. 1997.

[31] H. D. Wohlmuth and D. Kehrer, "A low power 13-Gb/s 27-1 pseudo random bit sequence generator IC in 120 nm bulk CMOS," in *Proc. IEEE Symp. Integr. Circuits Syst. Des.*, Pernambuco, Brazil, Sep. 2004, pp. 233–236.

[32] M. S. Kao, C. H. Jen, J. M. Wu, C. T. Chiu, and S. S. H. Hsu, "Transmission circuit for use in input/output interface," U.S. Patent 7 443 210, Oct. 2008.

[33] M. S. Kao, J. M. Wu, C. H. Lin, F. T. Chen, C. T. Chiu, and S. S. H. Hsu, "A 10-Gb/s CML I/O circuit for backplane interconnection in 0.18-μm CMOS technology," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 17, no. 5, pp. 688–696, May 2009.

[34] *Data Center CEE/DCB Switch Chip Family, FM 6000*. (2010) [Online]. Available: http://www.fulcrummicro.com/product/FM6000_Product_Brief.pdf

[35] R. Reutemann, M. Ruegg, F. Keyser, J. Bergkvist, D. Dreps, T. Toifl, and M. Schmatz, "A 4.5 mW/Gb/s 6.4 Gb/s 22+1-lane source synchronous receiver core with optional cleanup PLL in 65 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 45, no. 12, pp. 2850–2861, Dec. 2010.

[36] K. Fukuda, H. Yamashita, G. Ono, R. Nemoto, E. Suzuki, N. Masuda, T. Takemoto, F. Yuki, and T. Saito, "A 12.3-mW 12.5-Gb/s complete transceiver in 65-nm CMOS process," *IEEE J. Solid-State Circuits*, vol. 45, no. 12, pp. 2838–2849, Dec. 2010.

Ching-Te Chiu received the B.S. and M.S. degrees from National Taiwan University, Taipei, Taiwan, and the Ph.D. degree from the University of Maryland, College Park, all in electrical engineering.

She was an Associate Professor with National Chung Cheng University, Chia-Yi, Taiwan. She is currently with the Computer Science Department and Institute of Communications Engineering, National Tsing Hua University, Hsinchu, Taiwan, as an Associate Professor. Her current research interests include high-speed SerDes design, multichip interconnect, fault tolerance for network-on-chip, high-dynamic range image and video processing, high-definition television video decoder chip design, and the SONET/SDH mapper and framer IC design.

Dr. Chiu was a recipient of the First Prize, the Best Advisor Award, and the Best Innovation Award from the Golden Silicon Award in 2006. She is a TC Member of the Nanoelectronics and Gigascale Systems Group, the IEEE Circuits and Systems Society and the Design and Implementation of Signal Processing Systems Group, the IEEE Signal Processing Society. She is the Program Chair of the first IEEE Signal Processing Society Summer School, Hsinchu, in 2011. She was a Technical Staff Member with AT&T, Murray Hill, NJ, and at Lucent Technologies, Murray Hill, and with Agere Systems, Santa Clara, CA.

Yu-Hao Hsu received the B.S. degree in electrical engineering and the Ph.D. degree in communications engineering from National Tsing Hua University, Hsinchu, Taiwan, in 2002 and 2010, respectively.

He is currently a Principle Engineer with the Memory Design Program, Taiwan Semiconductor Manufacturing Company Limited, Hsinchu. His current research interests include high-speed switch architecture design, high SERDES interface design, and SRAM compiler design.

Wei-Chih Lai received the B.S. degree from National Central University, Jhongli, Taiwan, and the M.S. degree from National Tsing Hua University, Hsinchu, Taiwan, in 2009 and 2011, respectively, both in computer science.

He is currently with United Microelectronics Corporation, Taipei, Taiwan. His current research interests include high-speed switch and data center networks.

Jen-Ming Wu received the B.S. degree from National Taiwan University, Taipei, Taiwan, the M.S. degree from Polytechnic Institute, New York University, New York, and the Ph.D. degree from the University of Southern California, Los Angeles, in 1988, 1991, and 1998, respectively, all in electrical engineering.

He was with Sun Microsystems Inc., Sunnyvale, CA, from 1998 to 2003, as a Technical Staff Member. Since 2003, he has been with the Institute of Communications Engineering, Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan, where he is currently an Associate Professor. He has been involved in research on various fields of electrical engineering, including signal processing for communications, wireless communication transceiver integrated circuit (IC) design, high-speed interface IC design, and microprocessor architectures. He has authored or co-authored more than 60 technical papers in IEEE journals and conferences. His current research interests include high-speed interface technologies, multiple-input and multiple-output (MIMO) signal processing, MIMO cognitive radios, and wireless applications for healthcare monitoring.

Shawn S. H. Hsu (M'04) was born in Tainan, Taiwan. He received the B.S. degree from National Tsing Hua University, Hsinchu, Taiwan, in 1992, and the M.S. and Ph.D. degrees from the University of Michigan, Ann Arbor, in 1997 and 2003, respectively.

He is currently a Professor with the Department of Electrical Engineering and Institute of Electronics Engineering, National Tsing Hua University. He is also involved in research on design and modeling of high-frequency transistors and interconnects. His current research interests include the design of monolithic microwave integrated circuits and radio frequency interface chips using Si/III-V-based devices for low-noise, high-linearity, and high-efficiency system-on-chip applications.

Prof. Hsu was a recipient of the Junior Faculty Research Award from National Tsing Hua University in 2007 and the Outstanding Young Electrical Engineer Award from the Chinese Institute of Electrical Engineering in 2009. He has been a Technical Program Committee Member of SSDM from 2008 to 2011 and the IEEE A-SSCC since 2008.

Yang-Syu Lin received the M.S. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 2008.

He is currently a Senior Engineer with the Memory Design Program, Taiwan Semiconductor Manufacturing Company Limited, Hsinchu. His current research interests include high-speed switch architecture design, high-speed SerDes interface design, and high-speed SRAM macro design.

Fan-Ta Chen was born in Hsinchu, Taiwan, on May 24, 1983. He received the B.S.E.E. degree from Yuan Ze University, Jhongli, Taiwan, in 2005, and the M.S.C.E. degree from the University of National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 2007, where he is currently pursuing the Ph.D. degree with NTHU.

His current research interests include phased locked loops, and clock and data recovery for high-speed and low-power SerDes circuit design.

Min-Sheng Kao received the M.S. degree in electrical engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, and the Ph.D. degree in communications engineering from National Tsing Hua University, Hsinchu, Taiwan, in 1999 and 2011, respectively.

He was with the Optical Communication and Optical Display Division, Industrial Technology Research Institute, Hsinchu, from 2000 to 2004. He is currently the Director of APAC marketing and product application engineering with Mindspeed Technologies. He holds eight U.S. patents. His current research interests include analog front-end circuits for both wireless and wireline communications.

Yar-Sun Hsu received the B.S. and M.S. degrees in electronics engineering from National Chiao Tung University, Hsinchu, Taiwan, and the Ph.D. degree from Rensselaer Polytechnic Institute, Troy, NY.

He was with General Electric Company, New York, NY, for three years before joining IBM T.J. Watson Research Center, Yorktown Heights, NY, as a Research Staff Member. He was involved in research on computer architectures, parallel and distributed systems, parallel file systems, interconnection networks, and VLSI design. He was the Manager of the system department involved in research and design of the IBM Scalable Power Parallel System, the base machine used for IBM's Deep Blue Program, in 1988. He also led his group working on cache coherence protocol for multiprocessor systems, performance evaluation and visualization for scalable parallel systems, and scalable parallel inputs and outputs. He has been with the Department of Electrical Engineering, National Tsing Hua University, since 2002, where he is currently a Professor.

Dr. Hsu was a recipient of the IBM Outstanding Technical Achievement Award, three IBM Invention Plateau Awards, two IBM Supplemental Invention Awards for top-rated patents, three IBM Research Division Technical Achievement Awards, the Best System Paper Award at the ACM SIGMETRICS Conference in 2000, the Best Paper Award at the International Computer Symposium in 2004, and Outstanding Teaching Award from National Tsing Hua University in 2006 and 2009, respectively.