A Petabit Photonic Packet Switch (P^{3}S)

H. Jonathan Chao, *Kung-Li Deng, and Zhigang Jing

Department of Electrical and Computer Engineering, Polytechnic University, Brooklyn, NY 11201
**General Electric Company, Global Research Center, Niskayuna, NY 12309*

***Abstract*-- This paper presents a new petabit photonic packet switch (P^{3}S) architecture that is highly scalable both in dimension and capacity while maintaining high system performances. Using a new multi-dimensional photonic multiplexing scheme that includes space, time, wavelength, and sub-carrier domains, we propose a photonic switch fabric based on a 3-stage Clos network to provide scalable large-dimension photonic interconnections with nanosecond reconfiguration speed. Packet buffering is implemented electronically at the input and output port controllers, allowing the central photonic switch fabric to transport high-speed optical signals without electrical-to-optical conversion. Optical time division multiplexing (OTDM) technology further scales port speed beyond electronic speed up to 160 Gbits/s to minimize the fiber connections. To solve output contention, we propose a new arbitration scheme, called Frame-based Exhaustive Matching (FEM), using extended frames to aggregate cells from different incoming lines. The extended frame relaxes the stringent arbitration time constraint at a 160 Gbit/s port speed. Based on the FEM scheme in the proposed architecture, a 6400 x 6400 switch with a total capacity of 1.024 petabit/s can be achieved with throughput close to 100% under various traffic conditions.**

***Index terms*—photonic switch, Clos network, scheduling, Optical Time Division Multiplexing (OTDM)**

I. INTRODUCTION

The challenges to build a next-generation packet router with petabit capacity are: 1) building a switching fabric with minimal fiber connections while providing high-speed connectivity among a large number of line cards with fast reconfiguration speed; 2) developing an efficient arbitration scheme to resolve output contention with stringent time requirement while offering high throughput and low latency under various input traffic distributions, and 3) using efficient memory architecture for packet storage and buffer management at a reasonable memory speed. **[1]** Currently, electronic crossbar switches can be implemented on a single chip with a capacity up to 320 Gbits/s at a link rate of 2.5 Gbits/s **[2]**. To build a petabit (10^{6}Gbit/s) router, the dimension of the interconnection may reach 100,000 for port rate at 10 Gbits/s. The total number of interconnections can be estimated as: 100,000 (connections between stages) x 2 (stages) x 4 (number of 2.5-Gbits/s electronic planes to form a 10-Gbits/s port) = 800,000 (interconnections). This large number of fiber interconnections for an electronic switch fabric becomes too complex to implement (or even manage). Thus, we are motivated to look at a photonic switch system as an alternative to building a petabit switch fabric. Several promising photonic technologies have been proposed in the implementation of large-dimension space switches **[3,4]**. The optical micro-electro-mechanical system (optical MEMS) has demonstrated a 1296 x 1296 cross-connection switch with a total capacity of 2.07 petabit/s **[5]**. The drawbacks to these types of switches are the slow reconfiguration speed (~ μs range) due to the moving of mechanics in achieving the switching and its complex control mechanism. For packet switch applications, the switch needs to re-configure its connections on a per-packet basis, usually in the nanosecond time scale. Another photonic approach is to combine tunable lasers with an array-wave guide grating (AWG) device as a space switch **[6]**. The advantage of this approach is the rapid switching speed achieved by tuning the wavelength of the lasers at a nanosecond speed **[7]**. To cascade the AWG-based switches for multi-stage applications, wavelength conversion becomes an essential element since the switching property of the AWG depends on the change of the input wavelengths **[8]**.

We previously proposed a single-stage photonic switch fabric with capacity up to several terabits **[9,10]**. In this paper, we propose a new petabit photonic packet switch (P^{3}S) architecture that is suitable to be the platform for building a next-generation IP router. The switching fabric utilizes multi-dimensional photonic technologies to achieve highly scalable, rapidly reconfigurable interconnections in a multi-stage Clos network. Time dimension multiplexing is added using optical time division multiplexing (OTDM) technology **[11-13]** to increase the port speed up to hundreds of gigabits per second. This greatly reduces the number of fiber interconnections by aggregating more capacity on a single fiber. For a system with 160-Gbit/s OTDM port rate, the total number of fiber connections inside a 3-stage photonic switch fabric can be reduced to 12k. The reduction is contributed by increasing the bandwidth of fiber interconnections from 2.5 Gbits/s to 160 Gbits/s. Compared to 800k interconnections in the electronics approach, the proposed photonic switch fabric shows excellent scalability for petabit capacity.

Up to now, only a few researchers have attempted to find good matching schemes for the multi-stage Clos-network switch **[14]**. It is very challenging to find an efficient and fast scheduling scheme to provide high throughput, no-starvation, acceptable delay, and fairness performance under various traffic conditions in the multi-stage bufferless Clos-network switch architecture. The scheduling scheme must be able to scale gracefully with a large switch capacity because the time

This work was supported in part by the National Science Foundation under Grant 9814856 and Grant 9906673 and in part by the New York State Center for Advanced Technology in Telecommunications.

0-7803-7753-2/03/$17.00 (C) 2003 IEEE

for resolving output contention becomes more constrained with the increase of switch size and port speed. This paper presents a new scheduling scheme, called Frame-based Exhaustive Matching (FEM), as a solution to these requirements. To relax the strict arbitration time constraint, FEM operates based on a frame of *r* cells (*r*>1). Multiple cells from different input lines are aggregated to form a photonic frame, a switching data unit in the photonic switch fabric. The FEM scheme is an extension of the EDRRM scheme **[15]** by including the notion of frames to relax the arbitration time constraint, and slightly modifying the EDRRM to further improve the throughput under various traffic conditions.

II. System Architecture

A. System Architecture

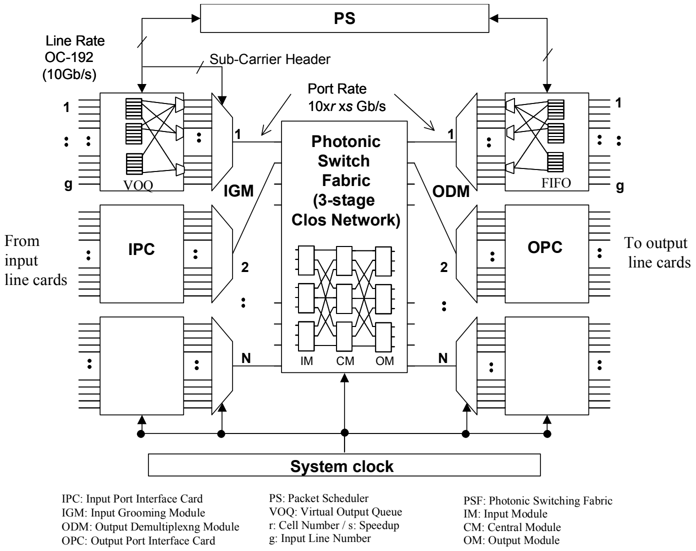


Figure 1. System architecture of a petabit photonic switch (P^3S)

Figure 1 shows the system architecture of a petabit photonic packet switch (P^3S). The basic building modules include the input and output port controllers (IPC and OPC), input grooming and output demultiplexing modules (IGMs and ODMs), photonic switch fabric (PSF), centralized packet scheduler (PS), and a system clock distribution unit. The incoming and outgoing line rates are assumed to be 10 Gbits/s. All incoming lines will first be terminated at line cards, where packets are segmented into cells (fixed length data units) and stored in a memory. The packet headers are extracted for IP address lookup, classification, and other functions such as traffic shaping/policing. All cell buffering is implemented electronically at the IPCs and OPCs, leaving the central PSF bufferless, i.e., no photonic buffering required in the system. The port speed can be equal to or greater than (with a speedup) *r* times the line rate, where *r* is the grooming factor. Virtual output queues (VOQs) at IPCs, together with the PS, provide contention resolution and hence achieve high performances in the packet switch operation. At the input end, the majority of incoming packets are stored in the ingress line cards, where packets are segmented and stored in its VOQs. Packets destined for the same output port are stored in the same VOQ. VOQs implemented at the IPC serve as the mirror of the VOQ memory structure in the line cards. As long as they can keep the cells flowing between the line card and IPC, the size of the VOQs at the IPC can be considerably smaller than its mirror part in the input line cards. Buffers at the OPC are used to store cells before they are sent out to the destined output line cards. A large buffer with a virtual input queue (VIQ) structure implemented in the line card is used to store the egress cells from the PSF and re-assemble them into packets.

B. Data Packet Flow

Figure 2 shows how packets flow across the system. At the input, variable-length IP packets are first segmented into cells with a fixed length of 64 bytes, suitable to accommodate the shortest IP packet (40 bytes). At each IPC, a total of *g* input lines at 10 Gbits/s enter the system and terminate at the IPC. Cells designated to the same output port of the PSF are stored in the same VOQ. In order to reduce the memory speed, each VOQ has a parallel memory structure to accommodate the *r* cells that eventually form a photonic frame. Each cell, before entering the IGM for compression, is scrambled to guarantee sufficient transitions in the data bits for optical receiver and clock recovery. In the IGM, these cells are compressed at the optical time domain to form a time-interleaved OTDM frame at *r*x10 Gbits/s. Due to the compression factor in the IGM, the photonic TDM frame with *r* cells occupies the same time interval as each of the incoming cells (compressed OTDM photonic frame size = electronics cell size =*T*). Let us call *T* to be the time slot and *T* = 51.2 ns for 10 Gbits/s line rate. The detailed data structures at each stage of the router are shown in Figure 3. Guardband of ~5% of the frame length is added at the head and tail of the frame to compensate for the relative misalignment of the photonic frames after passing through the PSF and to cover the switching transitions of optical devices.

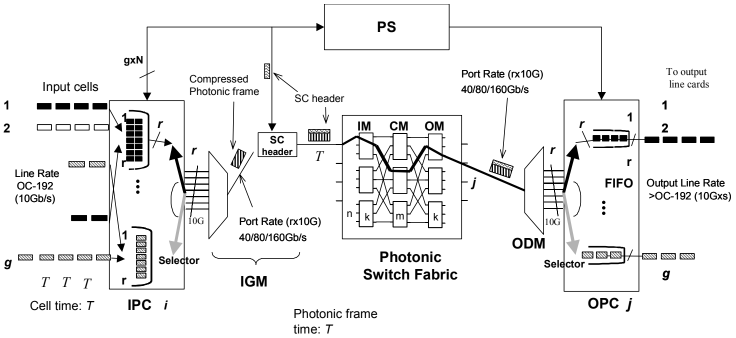


Figure 2. Data packet flow

The data structure at each stage of the switch is illustrated in Figure 3. Prior to the data payload, two header fields that contain output line number (OL) in the destined output port and the input line number (IL) of the switch are added to each incoming cell (see Figure 3 (a)). The OL is used to deliver cells to the destined output lines when the photonic frame (*r* cells) arrives at the OPC. A validity bit is inserted at the beginning of the cell to indicate if the cell is valid or not. The overhead bits introduced by OL and IL can be calculated as $log_2(g)$ and $log_2(g\text{x}N)$, respectively, where g is the total number of input (or output) lines at each IPC (or OPC). For a petabit system with N=6400 and g =16, the cell header length is 21 bits (1+4+16). Bits in each cell are compressed and time-interleaved using OTDM techniques in the IGM to form the photonic frames that are ready to transmit through the PSF (see Figure 3(b)). Each photonic frame also carries with an out-of-band sub-carrier (SC) header. Using the OTDM photonic frame as its carrier, the SC header is amplitude-modulated on the OTDM photonic frame at a much lower sub-carrier frequency. The estimated raw bandwidth required for the SC header is about 600 MHz. Standard multi-level coding schemes can be applied to further compress the SC bandwidth to 80 MHz or less, allowing the SC header to be carried around the DC frequency. The first field in the SC header is a flag containing a specific pattern for frame delineation since the OTDM photonic frames carrying the SC header are not precisely repeating in the time domain. The payload is 8B/10B coded for finding the flag correctly. Three fields are attached to the SC header to provide the routing information at each stage of the PSF. The three fields include CM, OM, and OPC numbers with $log_2(m)$, $log_2(k)$, and $log_2(n)$ bits of information, where m and k are the numbers of CM and OM, and n is the total number of outputs at each OM. At the beginning of the frame, a validity bit is added to indicate if the frame contains valid cells.

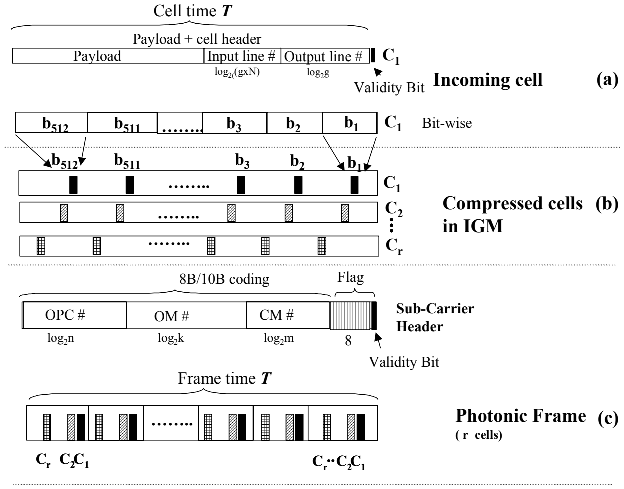


Figure 3 Data structure of (a) incoming cell, (b) compressed cells and (c) photonic frame

At each stage of the photonic switching fabric, the corresponding sub-carrier header is extracted and processed to control the switching fabric. Since the PS has already resolved the contention, the photonic frame is able to find a path by selecting the proper output links at each stage in the switching fabric. Once the photonic frame arrives successfully at the designated output port, it is demultiplexed and converted back to *r* cells at 10 Gbits/s in the electronic domain. The OPC then forwards the cells to their corresponding line cards based on the output line numbers (OLs).

C. Input Port Controller (IPC)

As shown in Figure 4, the input port controller (IPC) aggregates g input lines, each at 10 Gbits/s. For each input line, a memory space (input line memory) is used to store the incoming cells before they are granted to transmit. The input line memory consists of r frame memories. Within each frame memory, there are N VOQs, where N is the total number of output ports in the PSF, each corresponding to an output port of the PSF. The indices of VOQ(i,j,h) are denoted based on the following order: i is the input port number, j is the output port number, and h the input line memory number ($1 \leq i, j \leq N$, $1 \leq h \leq g$). Selectors select one of N VOQs of a specific input line memory that has been granted transmission at the current time slot (T). Cells in the r frame memories of the selected VOQ are read simultaneously to the IGM, where they are time-division multiplexed in the optical domain.

An example of how cells flow through the IPC is also illustrated in Figure 4. In this case, packets A, B, and C from input lines 1 and g, respectively, are destined for the same output port of the PSF (port 1). Packets A and B are heading toward output line 1 while packet C is headed toward output line g at the same OPC. Upon arriving at the IPC, each packet, already segmented into a number of fixed-size cells, is stored in the corresponding input line memory. In this example, packet A is segmented into 24 cells (cells A_0 to A_{23}), packet B contains 16 cells (cells B_0 to B_{15}), and packet C has 7 cells (cells C_0 to C_6). All incoming cells are stored in the r frame memories in a round-robin manner.

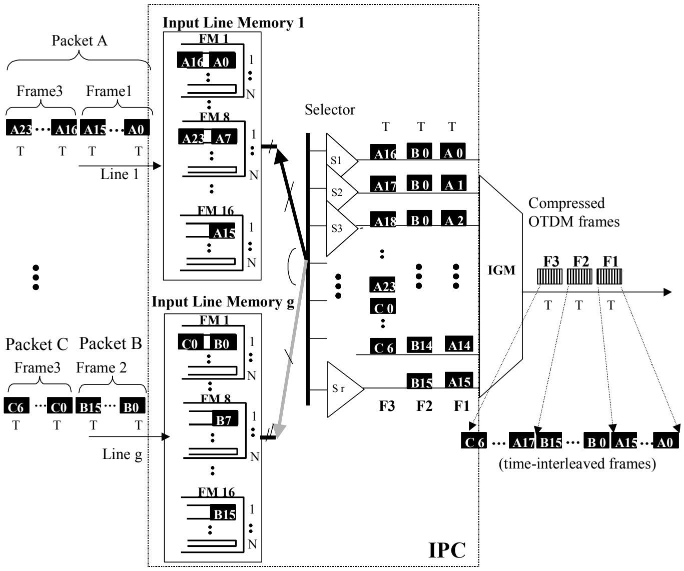


Figure 4 IPC architecture and an example of illustrating how frames are formed in the IPC

As soon as a cell arrives at the input line memory, a request is sent to the packet scheduler that tracks of all the incoming cells. The scheduler, based on a new frame-based exhaustive matching scheme, sends back the grant signal if the transmission has been granted. As a result, 16 cells (A_0 to A_{15}) from input line memory 1 are selected at the first time slot ($T \sim$ 50 ns) to form frame number 1, followed by 16 cells (B_0 to B_{15}) from input line memory g selected at the next time slot to form frame number 2. In this example, the remaining 8 cells (A_{16} to A_{23}) will be aggregated with another 7 cells from C packet (C_0 to C_6) to form frame number 3.

D. Output Port Controller (OPC)

The output port controller (OPC) is shown in Figure 5. The main function of the OPC is to demultiplex cells in the frames to proper output lines using *g* parallel FIFOs in each of the output line memories. Selectors select the output line memory according to the output line number (OL) of the cell header at each time slot. The FIFO size can be kept sufficiently small if the data rate to each output line is much larger than the line rate 10 Gbit/s. There are *N*x*g* virtual input queues (VIQs) at the output line cards used for packet assembly.

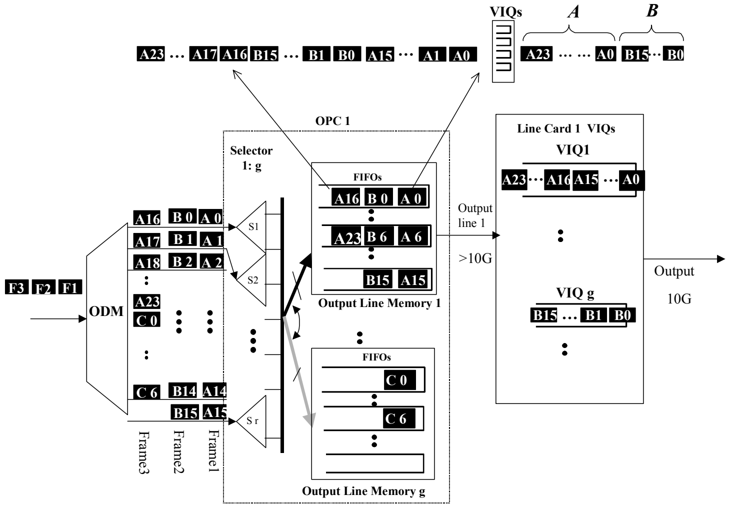


Figure 5 OPC architecture and an example of illustrating how packets are demultiplexed at the OPC and reassembled at the egress line cards

Following the above example, Figure 5 also shows how packets *A, B,* and *C* are as they are demultiplexed at the OPC and reassembled at the egress line cards. Photonic frames containing the compressed cells are demultiplexed in the ODM and sent into *r* parallel inputs to the selector array. According to the cell header, A_0 to A_{15} go to the 16 FIFOs located in output line memory 1 at the first time slot. At the next time slot, B_0 to B_{15} are sent to the same 16 FIFOs in output line memory 1. At the next time slot, photonic frame number 3 arrives at the OPC. The remaining part of packet *A* is sent to input line memory 1, while cells from packet *C* go to output line memory *g*. These cells are then read out from the FIFOs to the designated output line (output line 1 in this case) at a speed larger than 10 Gbits/s. The VIQs at the line card are used to reassemble packet *A*, *B* and *C*.

III. PACKET SCHEDULING FOR 3-STAGE CLOS-NETWORK SWITCHES

For a matched input-output pair in a 3-stage switch, there can be multiple possible paths (determined by the number of center switch modules). Having to choose a center switch module that reduces internal blocking to increase the throughput further complicates the scheduling complexity. Here, we attempt to solve the problem in two phases. In the first phase, we use a new matching scheme to find the matching between inputs and outputs. In the second phase, we use a parallel matching scheme to find the paths for the matched input-output pairs. Both schemes are described in this section.

A. FEM Packet Scheduler

Let us consider an *N*×*N* photonic switch fabric (PSF) with *g* input lines multiplexed at each IPC as shown in Figure 1 in Section II. Here we propose a new matching scheme, called frame-based exhaustive matching (FEM), to match the input and output ports of the PSF.

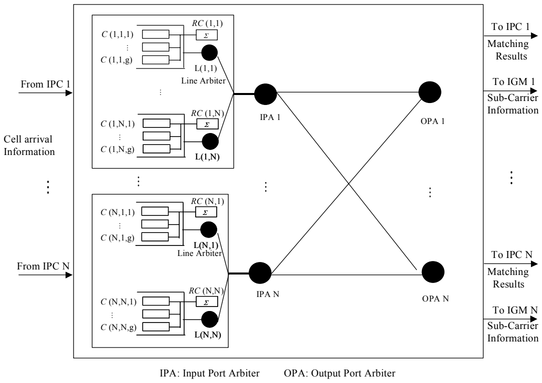


Figure 6. The schematic of the FEM packet scheduler

The FEM scheduler consists of *N* input port arbiters (IPA) and *N* output port arbiters (OPA) as shown in Figure 6. As discussed in Section II, each IPC has *N* VOQs, each corresponding to an OPC. Each *VOQ(i,j)* accepting cells from g input line consists of *g VOQ(i,j,h),* where *i* denotes the IPC number*, j* denotes the OPC number*,* and *h* denotes the input line number. A counter *C(i,j,h)* is used to record the number of cells in the corresponding *VOQ(i,j,h).* An *RC(i,j)* is used to record the total number of cells in each *VOQ(i,j).* So, $RC(i,j) = \sum_{h=1}^{g} C(i,j,h)$. Each input port arbiter, corresponding to each input port of the PSF, maintains *N* line arbiters *L(i,j)*s. When the IPC *i* is matched with an OPC, say *j*, the line arbiter *L(i,j)* then chooses cells from *g* input lines, based on the value of each *C(i,j,h*), to form a frame.

The FEM scheduler operates as follows:

- At each time slot, each IPC sends cell arrival information, including the input line number and the destined OPCnumber, to the FEM scheduler. The total number of bits from each IPC to the FEM is $g\times\log_2(N)$. The FEM scheduler then updates the corresponding counters *C(i,j,h)* and *RC(i,j)*.
- Based on the counter *RC(i,j)* value, indicating the queue length of *VOQ(i,j),* input port arbiters and output port arbiters finish their matching based on the packet scheduling scheme described in Section III.B.
- Each line arbiter *L(i,j)* chooses the cells from *VOQ(i,j,h)* in a round-robin manner until a frame is formed or partially formed when all g *VOQ(i,j,h)* become empty. The pointer of the line arbiter *L(i,j)* is updated to one location beyond the current chosen position.

B. *Frame-Based Exhaustive Matching (FEM) Algorithm*

The FEM scheme is an extension of the exhaustive dual round-robin matching (EDRRM) scheme **[15]**. Most maximal-sized matching schemes, such as iSLIP **[16]** and DRRM **[17]**, suffer from the problem of throughput degradation under unbalanced traffic distribution. The EDRRM scheme improves throughput by maintaining the existing matched pairs between the inputs and outputs so that the number of unmatched inputs and outputs is drastically reduced (especially at high load), thus reducing the inefficiency caused by not being able to find matches among those unmatched inputs and outputs. To relax the stringent arbitration time constraint, the switching unit for the FEM scheme is extended from a cell to a frame of *r* cells (r>1), contributed by *g* input lines in the same group. The FEM also modifies the EDRRM slightly to further improve throughput. One of the major problems of the exhaustive matching is that it may cause starvation in some inputs. One way to overcome this problem is to set a timer for each head-of-line frame. When the timer expires, the request from the "expired" frame has the highest preference to be granted. The FEM consists of three steps and is described below.

○ **Step 1: Request**

Each unmatched input (including the currently matched input but whose matched VOQ's queue length is less than a threshold *r*) sends a request to every output port arbiter for which it has a queued cell in the corresponding VOQ. The request is set as high priority if the queue length is above (or equal to) the threshold *r*; otherwise, the request is set as low priority. Each matched input (with matched VOQ's queue length greater than or equal to the threshold *r*) only sends a request to its matched output. The request is set as high priority.

○ **Step 2: Grant**

▪ If an output port arbiter receives one or more high-priority requests, it grants the one that appears next in a fixed round-robin schedule starting from the current position of the high-priority pointer. If there are no high-priority requests, the output port arbiter grants one low-priority request in a fixed round-robin schedule starting from the current position of the low-priority pointer. The output port arbiter notifies each requesting input whether or not its request is granted. The reason to give higher priority to those filled frames is to increase the frame utilization and, thus, the system throughput.

▪ Update of the pointer of the output port arbiter. The pointer of the output port arbiter moves to the selected input position only if this grant is accepted by the input port arbiter in Step 3. Otherwise, the pointer remains unchanged.

○ **Step 3: Accept**

▪ If an input receives one or more high-priority grants, it accepts one that appears next in a fixed round-robin schedule starting from the current position of the high-priority pointer. If there are no high-priority grants, the input port arbiter accepts one low-priority grant that appears next in a fixed round-robin schedule starting from the current position of the low-priority pointer.

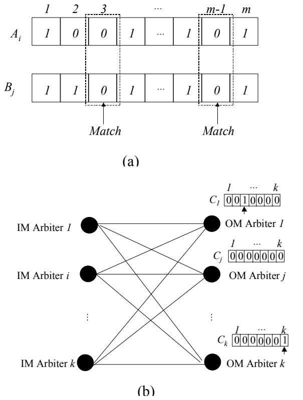
▪ Update of the pointer of input port arbiter: The pointer of the input port arbiter moves to the accepted output position.

C. *Parallel Matching Scheme*

After input-output matching is completed in phase 1, parallel matching is used to find a routing path for each matched input-output pair through the 3-stage bufferless Clos-network switch. Finding a routing path is equivalent to finding an available CM (out of *m* CMs) for the matched input-output pair. When $m\geq 2n-1$, the 3-stage Clos-Network is a non-blocking circuit switch. Although it was theoretically proven that a 3-Stage bufferless Clos-network switch is rearrangeablly non-blocking when $m\geq n$, the already known rearrangement scheme is impractical to implement in the high-speed switch due to their prohibited high time complexity. To achieve high matching efficiency with low computing complexity, a parallel matching scheme is adopted **[18]**. Let us label the set of output links from each IM_i by A_i ($1 \leq i \leq k$) and the set of input links into each OM_j by B_j ($1 \leq j \leq k$). Each A_i and B_j contains exactly *m* elements denoting the status of each physical link as shown in Figure 7(a) ("0" means that this link has not been matched; "1" means that this link has been matched). In order to find a path between IM_i and OM_j, one just needs to find an available CM in the center stage by finding a vertical pair of zeros in the A_i and B_j. For instance, Figure 7(a) shows two paths available between A_i and B_j.

Ideally, we would like to divide each time slot (T) into *k* mini-slots. In each mini-slot, the above procedure of finding vertical pairs of zeros is performed in parallel. For instance, in the first mini-slot, each A_i is matching with each B_i ($1 \leq i \leq k$). In the second mini-slot, each A_i is matching with $B_{((i+1) \bmod k)+1}$. And the procedure is repeated k times. Note that there are at most n matches needed between an IM and OMs in each mini-slot as there are at most n frames departing from the IM.

However, because the FEM scheme maintains the existing matched pairs between the inputs and output ports, the number of new matched IM-OM pairs is fewer than k, especially at high load as shown in Figure 7. For a switch size of 64 (256), there are 8 (16) IMs and 8 (16) OMs. As a result, the required number of mini-slots in the parallel matching is reduced to k', where $k'<k$. Since the sequence of matching the IMs and OMs is no longer predetermined as the k mini-slots case, there will be contention during when matching them. A scheme similar to the iSLIP is used to resolve the contention at the beginning of each mini-slot. Considering that there are a total of 80 IMs and 80 OMs for the 6400 x 6400 switch fabric, the hardware complexity and computation time for contention resolution is considerably low.



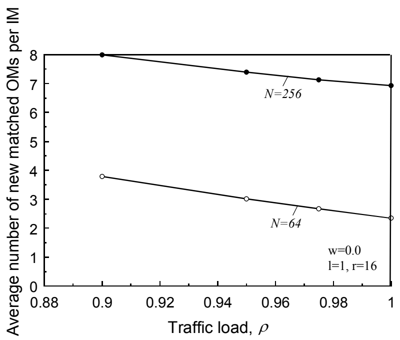


Figure 7. Paralell matching scheme

Figure 8. The average number of new matched OMs per IM in one time slot in uniform Bernoulli Traffic

At the beginning of each mini-slot, a scheme similar to the iSLIP is used to find the matching between the IMs and the OMs that have newly matched input-output outputs (determined by the FEM scheme). As shown in Figure 7(b), there are k IM arbiters and k OM arbiters. After the completion of the FEM in phase 1, each OM arbiter has the information about which IMs has inputs ports destined for its output ports. This thus avoids the request phase in the iSLIP. A register C_j associated with OM_j's arbiter contains exactly k elements, denoting the status of each IM ("0" means that this IM has no new match with OM_j; "1" means it does).

D. Packet Scheduler Implementation Complexity

Here we estimate the implementation complexity of the centralized packet scheduler. For the proposed petabit photonic switch fabric with N=6400, g =16, r=16, the amount of information that needs to be transmitted from each IPC to the packet scheduler is 26 byte (16x$\log_2$(6400) bit) per time slot, or about 4 Gbits/s. In return, the packet scheduler sends back the matching results to each IPC that becomes the cell headers of the frame, and to each IGM that becomes the in-band header of the frame. The former has a total of 93 bits (13+80) and the latter has a total of 30 bits (7+7+7+1+8). Thus, the bandwidth is about 1.8 GHz for the IPC and 600 Mbit/s for the IGM.

As shown in Figure 8, there are 6400 input port arbiters and 6400 output port arbiters. They form a mesh interconnection network and require a total of $(6400)^2$ or about 40M connections. It may require multiple chips and cards to implement the packet scheduler. The FEM achieves satisfactory performance with only one iteration. Let us assume it takes about ¼ of slot time (51.2 ns) to complete the phase 1 task. The remaining ¾ of slot time is used for the parallel matching. Simulation results also show that the system throughput asymptotically saturates at $\log_2 N$ mini-slots, or 13 in our petabit switch case. As a result, each mini-slot has about 3ns to complete the iSLIP-like contention resolution and parallel matching. If this time becomes too constraint, one may consider increasing the slot time from 51.2 ns to 102.4 ns, i.e., increasing the frame size from 16 to 32 cells. The penalty is the slight performance degradation. To implement the change, we just need to read 2 cells from each frame memory as shown in Figure 4 in each time slot. Note that the memory remains unchanged.

E. Performance Study

E.1. Performance of FEM in a Single-Stage Non-Blocking Switch

The performance of the FEM scheme was evaluated using computer simulation for uniform and non-uniform traffic models. We first studied the performance of the FEM in a single-stage non-blocking switch[1]. The delay is measured from when a cell enters the IPC to when it arrives at the OPC.

To evaluate the performance under unbalanced traffic distributions, we introduce a parameter called the unbalanced probability, w. The traffic load from input port s to output port d, $\rho_{s,d}$, is defined as,

$$\rho_{s,d} = \begin{cases} \rho(w + \frac{1-w}{N}) & if \quad s = d \\ \rho\frac{1-w}{N} & otherwise \end{cases} \tag{1}$$

where ρ is the offered load for each input port and N is the switch size. When w=0, the traffic is uniformly distributed to all output ports. On the other hand, when w=1, it is circuit switching, meaning that the traffic from each input port i is all destined for output port i. We assume that the burst length (l) is exponentially distributed as the bursty traffic.

[1] It can also be a 3-stage Clos-network switch with $m \geq 2n$-1. The performance and impact of m in a 3-stage bufferless Clos network is further discussed in Section III.E.2

Figure 9 shows the throughput of the FEM and iSLIP under Bernoulli traffic with different unbalanced probabilities (w) and frame sizes (r) for a switch size (*N*) of 64. We assume the number of iterations of finding matches in each time slot is 1. Both the FEM and iSLIP achieve very high throughput under uniform traffic distributions. The latter can even achieve 100% due to the desynchronization of pointers in the input and output arbitrators. However, the FEM provides much higher throughput (close to 100%) than the iSLIP under unbalanced distributions. This is because the FEM significantly reduces the number of unmatched inputs and outputs by maintaining the existing matches, resulting in much less output port contention.

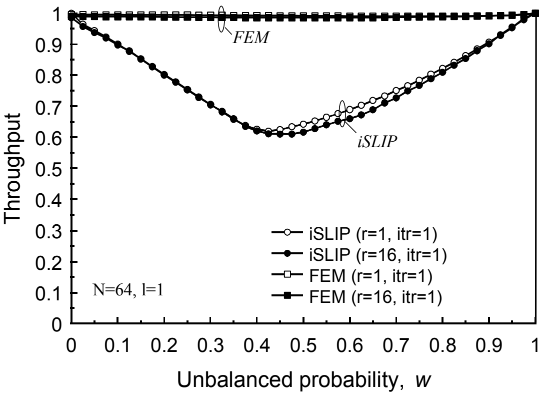


Figure 9. Throughput of the FEM and iSLIP under Bernoulli traffic with different unbalanced probabilities (*N*=64)

The FEM still maintains high throughput even with a large frame size *r* of 16. This is because filled frames have higher priority than those unfilled to be served in order to have higher frame utilization. As a result, currently matched input and output pairs may not be able to maintain matching if their corresponding queue length is less than r.

Let us define "*switch over time*" as the time taken for one input to switch from one VOQ after service completion to another VOQ and start to be serviced. It reflects the idle period of the switch. Based on the random polling system model in the EDRRM scheme **[19]**, we derive a formula that gives the mean switch over time *E(S)* for the FEM scheme in the following equation.

$$\lim_{N\to\infty} E(S) = \lim_{N\to\infty}\left\{ \frac{1-\dfrac{1}{\dfrac{\rho r}{N}(1-\rho^{N-1})(N-1)}\left[1-(1-\dfrac{\rho r}{N})(1-\dfrac{\rho r}{N}(1-\rho))^{N-1}+\dfrac{(1-\dfrac{\rho r}{N}(1-\rho))^{N}-1}{N(1-\rho)}\right]}{\dfrac{1}{\rho r}\left[1-(1-\dfrac{\rho r}{N})(1-\dfrac{\rho r}{N}(1-\rho))^{N-1}\right]} + \frac{\rho^{2(N-1)}}{1-\rho^{N-1}}\right\} = \frac{\rho r}{1-e^{-(1-\rho)\rho r}}-1 \qquad (2)$$

where *N* is the switch size, ρ is the traffic load, and *r* is the frame size. The analysis results for *E(S)* under different offered loads is shown in Figure 10. With the increase of the switch size *N*, the convergence of the mean switch over time *E(S)* to a limit indicates that the throughput of the FEM scheme scales very well with switch size *N*. In other words, the throughput can still approach 100% with large switch sizes.

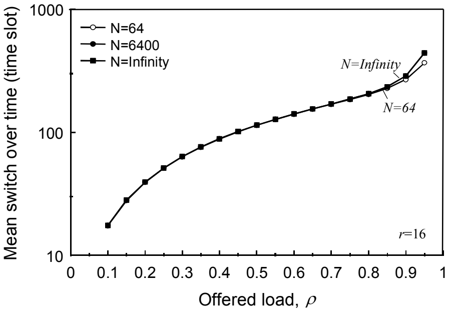


Figure 10. Mean switch over time of FEM in uniform Bernoulli i.i.d.traffic with different switch size *N* (*r=16*)

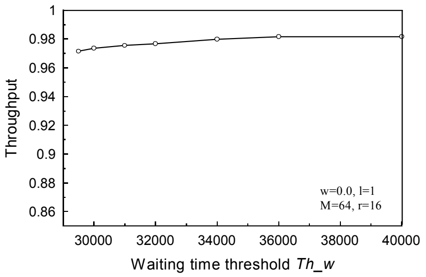


Figure 11. Throughput of the FEM in uniform Bernoulli traffic with different waiting time threshold *Th_w* (*N*=64, *r*=16)

To avoid some pathological traffic conditions that may lead to starvation, we boost the head-of-line (HOL) frame's priority level to the highest when its waiting time exceeds a certain threshold, say *Th_w*. This can be easily implemented by using a counter to keep track of the HOL frame's waiting time. When the HOL frame's waiting time exceeds the *Th_w*, the input port arbiter only sends these highest-priority requests to output port arbiters, which then first grant the highest-priority requests. Figure 11 shows the throughput of the FEM with different waiting time thresholds *Th_w*. We assume that *w*=0.0, *l*=1, *N*=64, and *r*=16 in the simulations. The FEM achieves high throughput, e.g., 0.982 when *Th_w* is set as 36000 (cell time). This is because large waiting time thresholds won't break already established matches that often, and thus starvation is prevented and high throughput is maintained.

E.2. Performance of FEM in 3-Stage Clos-Network Switch

Through computer simulations, we have found that for an *N*×*N* switch it takes about log_2N mini-slots for the parallel matching to converge as shown in Figure 12. This is similar to the results obtained for the PIM **[20]** and iSLIP **[16]**, in which the authors prove that $E(I) \leq log_2N + 4/3$, where *I* is the number of iterations that the PIM takes to converge. For all the stationary arrival process, it is also shown that $E(I) \leq log_2N$ holds for the iSLIP.

show the throughput of the FEM in a 3-stage bufferless Clos-network switch when considering different CM numbers (*m*) and speedup factors (*s*) under unbalanced Bernoulli traffic distributions. We assume that *n*=*k*=8, *w*=0.5, and *r*=16 and the number of mini-slots is set to log_2N. The increase of CM number (*m*) improves the throughput as expected. A small speedup is introduced to compensate for the inefficiency that may occur during the input-output pair matching and the parallel matching for finding a routing path among the CMs. shows that when there is no expansion (*m*=*n*), speedup of 1.5 can achieve high throughput even under unbalanced traffic distributions.

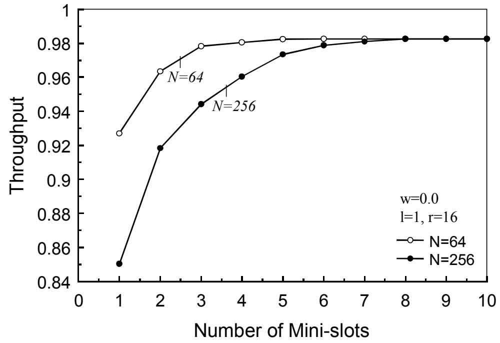


Figure 12 Throughput of the FEM combined with parallel matching with different number of mini-slots.

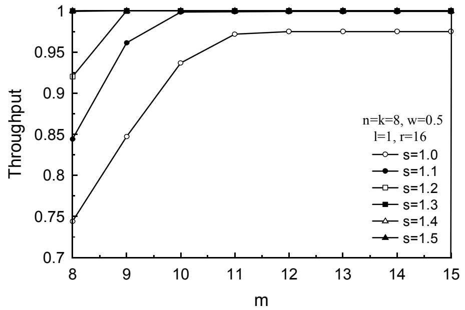


Figure 13. Throughput of the FEM in a 3-stage bufferless Clos-network switch with different CM numbers and speedup factors under unbalanced traffic (*n*=*k*=8, *w*=0.5, *l*=1, *r*=16)

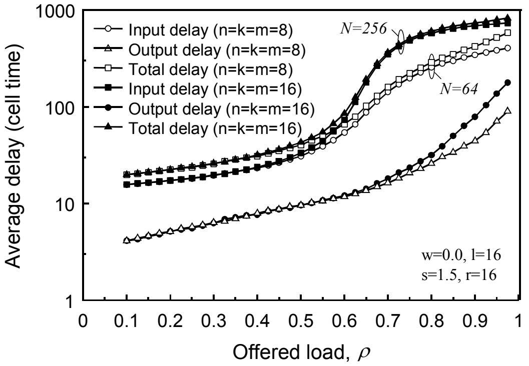


Figure 14. Delay performance of FEM scheme without expansion and with speedup factor of 1.5 in 3-stage Clos-network switch.

In our proposed switch architecture, cell delay can occur in the IPC, OPC, and egress line cards (for packet re-assembly and buffering). Here we only consider the delay at the IPC (input delay) and at the OPC (output delay). The delay at the egress line cards will be studied in the future. Assume the service rate of the OPC is equal to the port speed. Figure 14 shows the delay performance, input delay, output delay, and the total delay (sum of the input and output delays) for different switch sizes without internal bandwidth expansion (i.e., *m*=*n*) but with a small speedup factor of 1.5. We evaluate the average delay under non-uniform bursty traffic. The increase of the switch size to 256 doesn't affect the delay performance much, even under non-uniform traffic distributions. It demonstrates that the FEM scheme has good scalability to the switch size.

IV. PHOTONIC SWITCH FABRIC

A. *Photonic Switching Fabric (PSF)*

Figure 15 shows the structure of the PSF consisting of input modules (IMs), central modules (CMs), and output modules (OMs) in a 3-stage Clos network. The switch dimensions of the IM and OM are chosen to be *n*x*m* while CM is *k*x*k*. The total number of modules in the network to achieve $N=n^2$ interconnections between input and output ports scales linearly with *n*. IM at the first stage is a simple array waveguide grating (AWG) device. CMs and OMs share the same design that consists of the following elements: sub-carrier unit (SCU), wavelength conversion unit (WCU), and an *n*x*n* AWG. Using a total of *k* modules for IMs and OMs and *m* modules as CMs, the center stage of the Clos network can be expanded with *m*≥ *n*. With more components and interconnections, the expanded Clos network provides extra paths between the input and output ports and thus improves the throughput. It has been proven that a strictly non-blocking circuit switch can be achieved if *m*≥2*n*-1. As each OTDM frame traverses the PSF, its route path has been pre-determined by the packet scheduler before entering the PSF.

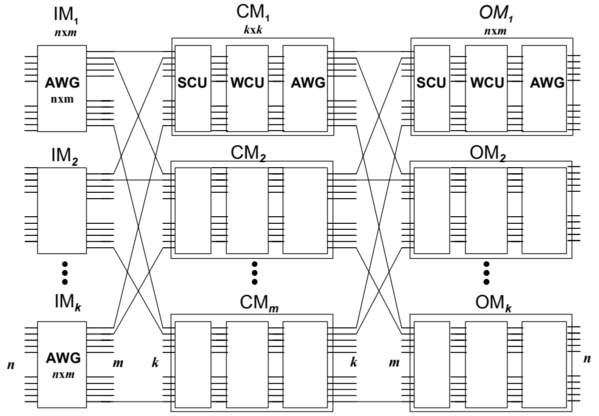


Figure 15. The photonic switching fabric (PSF) architecture.

A.1. *First Stage: IM*

Based on the cyclic routing property of an *n*x*n* AWG router, full connectivity between input and outputs of the IM can be established by arranging input wavelength **[8]**. In general, the wavelength λ_k from input i (i = 1,2,..,n) to output j (j = 1,2,..,n) can be calculated according to the following formula: $k = (i + j - 1)$ *modulo n*. The fast wavelength tuning capability at the inputs of the AWG enables the passive device to turn into a space switch. The re-configuration of this space switch is solely determined by the active wavelength tuning of the input tunable laser. The wavelength switching can be reduced to a couple of nanoseconds by rapidly changing the control currents for multiple sessions in tunable semiconductor lasers **[7]**. The stringent time constrain on the hardware switch time (wavelength tuning + TDM) can be relaxed by extending the frame size (r), which requires a slight higher speedup.

A.2. Second and Third Stages: CM and OM

To cascade AWGs for multi-stage switching, CMs and OMs have to add wavelength conversion capability, where the incoming wavelengths from the previous stage are converted to new wavelengths. An all-optical technique is deployed to provide the necessary wavelength conversion without O/E conversion. The detailed design of an nxn non-blocking switch module (CM and OM) is illustrated in Figure 16). Three key elements used to implement the switch module are the sub-carrier unit (SCU) for header processing and recognition, the wavelength conversion unit (WCU) for performing all-optical wavelength conversion, and the AWG as space switch.

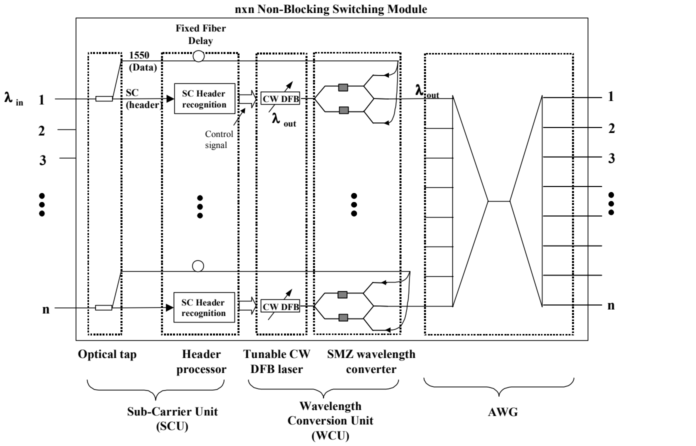


Figure 16. Switch module (CM and OM) architecture. All-optical wavelength conversion is added to cascade the modules.

A.3. Sub-Carrier Unit (SCU)

The main function of the SCU is to process the SC header information for setting up the switch path. The SC header information, which consists of 3 bytes, is readily available at each stage of the PSF as these bytes are carried out-of-band along with each OTDM photonic frame. Upon arriving at each module, a portion of the power from the OTDM photonic frame is stripped by an optical tap and fed into the SCU for sub-carrier demodulation. At the front end of the SCU, a low-bandwidth photo-detector and low-pass filter is able to recover the header information from the OTDM photonic frame. The SC header information is converted to the control signals to set the wavelength of the continuous-wave (CW) tunable laser in the WSU. The fields in the SC header are extracted and processed by IMs, CMs, and OMs, respectively, to determine the designated output at each stage. On the data path, a fixed fiber delay is added to allow the SCU to have sufficient time to perform header recognition and processing.

A.4. Wavelength Conversion Unit (WCU)

Recently, wavelength conversion at the OTDM rate up to 168 Gbits/s has been demonstrated by using a Symmetric Mach-Zehnder (SMZ)-type all-optical switch **[21]**. The strong refractive index change from the carrier-induced resonance nonlinearity in the semiconductor optical amplifiers (SOAs), coupled with the differential interferometric effect, provides an excellent platform for high-speed signal processing. In this paper, we propose to utilize an array of such devices to accomplish the all-optical wavelength conversion at an ultra-high bit rate.

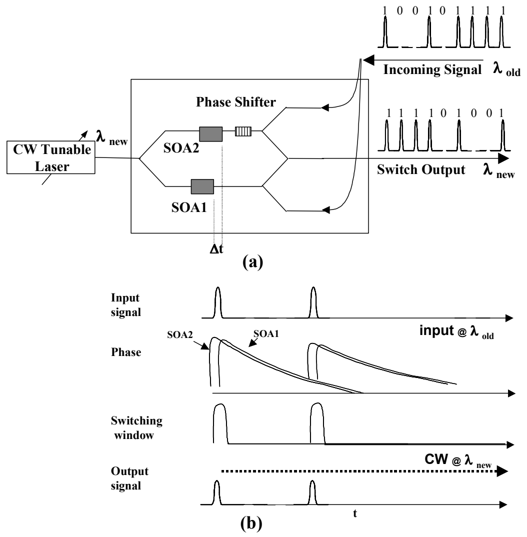


Figure 17. a) Wavelength conversion unit based on an SOA Mach-Zehnder interferometer. b) The timing diagram of the ultrafast wavelength conversion process.

The basic structure, based on a Mach-Zehnder (MZ) interferometer with in-line SOAs at each arm, is shown in Figure 17. The incoming signal with wavelength (λ_{in}) is split and injected to the signal inputs, entering the MZ from the opposite side of the switch. Figure 17(b) shows the operation of the wavelength conversion. A switching window at time domain can be set up (rising edge) by the femto-second ultrafast response induced by the signal pulses through carrier resonance effect of SOAs. The fast response of the SOA resonance is in the femto-second regime, considerably shorter than the desired rise time of the switching window. Although the resonance effect of each individual SOA suffers from a slow tailing response (~100 picoseconds), the delayed differential phase in the MZ interferometer is able to cancel the slow-trailing effects, resulting in a fast response on the trailing edge of the switching window. By controlling the differential time between the two SOAs accurately, the falling edge of the switching window can be set at the picosecond time scale. The timing offset between two SOAs located at each arm of the MZ interferometer controls the width of the switching window. To be able to precisely control the differential timing between two arms, a phase shifter is also integrated in the interferometer. The wavelength conversion occurs when a continuous wave (or CW) light at a new wavelength (λ_{new}) enters the input of the MZ interferometer. An ultrafast data stream whose pattern is the exact copy of the signal pulses at λ_{old} is created with the new wavelength at the output of the MZ interferometer , completing the wavelength conversion from λ_{old}. to λ_{new}.

B. OTDM Input Grooming Module (IGM)

Optical Time Division Multiplexing is capable of operating at ultrafast bit rates beyond the current electronics limit around 40 Gbits/s. By interleaving short optical pulses at the time domain, aggregated frames can be formed to carry data at bit rates of hundreds of gigabits per second. Utilizing the OTDM technique, there can be at least one order of magnitude in bandwidth increase comparing from the existing electronics approach. IGM interfaces with *r* (grooming factor) parallel electronic inputs from the IPC. shows the structure of the IGM based on the OTDM technology. It consists of a short-pulse generation unit, modulator array, and a passive *r*x*r* fiber coupler with proper time delays for time-interleaved multiplexing.

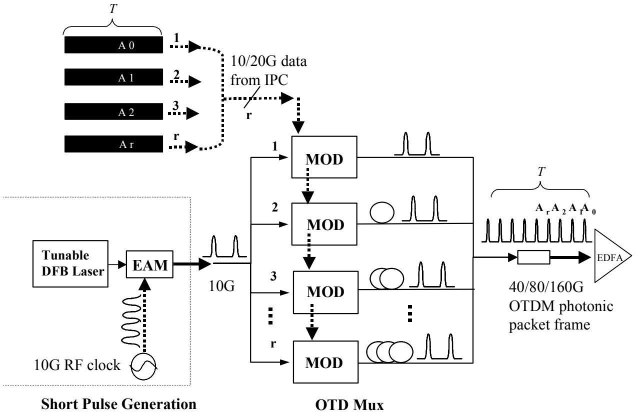


Figure 18. The structure of the IGM based on optical time division multiplexing. EAM = electro-absorption modulator, EDFA = erbium-doped fiber amplifier.

B.1. Short Pulse Generation Unit

Optical pulses with widths of several picoseconds can be generated using electro-absorption modulators (EAMs) over-driven by a 10 GHz sinusoidal clock signal. Using a tunable CW distributed feedback (DFB) laser as the light source, the wavelength of the output ultra-short pulses can also be tunable. The pulse width will be around 7 to 10 picoseconds generated by cascaded EAMs, which is suitable for data rates up to 100 Gbits/s. To generate pulses suitable for higher bit rates (>100 Gbits/s), nonlinear compression with self-phase modulation (SPM) is utilized. The pulses, generated from the EAMs, are injected into a nonlinear medium (a dispersion shifted or photonic bandgap fiber) followed by a compression fiber (dispersion compensation fiber) to further compress the pulse width to about 1 picosecond. Such short pulses allow multiplexing in the time domain, resulting in an aggregated OTDM photonic frame of bandwidth equivalent to hundreds of gigabits.

B.2. OTD Multiplexing Unit (MUX)

The parallel *r* input lines from the IPC modulate electronically the modulator array to encode the bit stream onto the optical pulse train. Precise time delays on each branch of the fiber coupler ensure the time-leaved multiplexing of *r* inputs. Through the parallel-to-serial conversion in the Mux, *r* cells at 10 Gbits/s from the IPC, a total bit rate of *r*x10 Gbits/s in parallel, are now effectively compressed in the time domain as the RZ-type photonic frame that operates at *r*x10 Gbits/s in serial. The fiber coupler and time delays can both be integrated using planer waveguide structures. To support the internal speedup factor of two for system performance improvement, the electronics, including drivers and modulators must support up to 20 Gbits/s of bandwidth inside the IGM.

C. OTDM Output Demultiplexing Module (ODM)

At the receiving end of the system, the output demultiplexing module (ODM) demultiplexes OTDM photonic frames from the output of the PSF into *r*-parallel electronic signals at 10 Gbits/s. As shown in Figure 19, the ODM consists of a quarter-phase detector and quarter-phase shifter, an array of OTDM demultiplexers (DEMUX) based on EAMs, and the photo-detector (PD) array for O/E conversions.

C.1. OTDM Demux

We have previously demonstrated ultrafast demultiplexing at 40, 80, 100, and 160 Gbits/s using cascaded EAMs as the gating device. As shown in the inset of Figure 19, the OTDM demultiplexer consists of two cascaded EAMs based on multiple quantum well devices **[12,13].** A SOA section is also integrated with the EAM to provide optical amplification at each stage. The optical transmission of the EAM, controlled by the driving electronic signal, responds highly nonlinearly and produces an ultra-short gating window in the time domain. Cascading the EAMs can further shorten the gating window compared to a single EAM. The incoming optical signal is split by an 1x*r* optical coupler into *r* modulators located in the array structure. Each EAM is over-driven by a 10-GHz sinusoidal RF clock to create the gating window for performing demultiplexing. The RF driving signals supplied to adjacent modulators in the array structure are shifted by a time τ, where τ is the bit interval inside the OTDM photonic frame. As a result, r modulators are able to perform demultiplexing from rx10 Gbits/s down to 10 Gbits/s on consecutive time slots of the OTDM photonic frame.

C.2. Quarter Phase Detector and Shifter

The incoming frames may inherit timing jitters induced by either slow thermal effects (fiber, device, and component thermal lengthening) or system timing errors. The result is a slow walk-off from the initial timing (phase). Since the frames are operating on a burst mode, traditional phase lock loop cannot be applied here. To track on the slow varying jitters on the burst frames, we devised a new quarter-phase locking scheme using phase detection and a shifter. The quarter-phase shifter has been demonstrated using a digital RF switched delay lattice **[13].** The semiconductor switch is used to set the state at each stage. Depending on the total delay through the lattice, the output phase can be shifted by changing the state at each lattice. The resulting clock is synchronized with the incoming packet with a timing error less than $\pm 1/8\ \tau$.

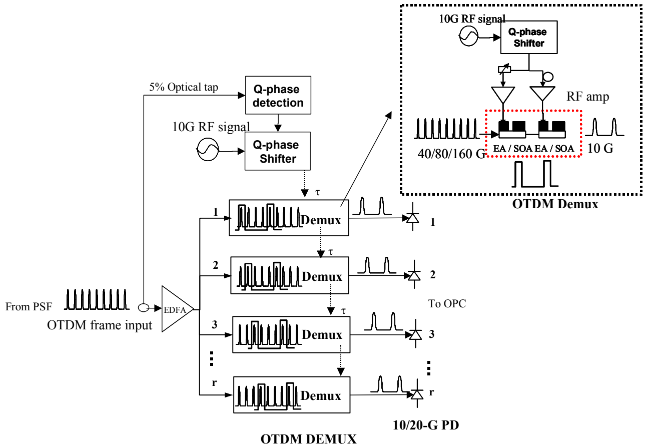


Figure 19. The structure of the ODM. The OTDM demultiplexer, based on cascaded EAMs, is shown in the inset.

V. CONCLUSIONS

A 6400x6400 switch based on a three-stage Clos network can be realized using 80 wavelengths at each switch module while the port speed can be increased to 160Gbit/s by the OTDM technology, resulting in a total capacity of 1.024petabits/s. The fiber connections are dramatically reduced to about 12k compared to 800k in an equivalent electronic approach. By increasing the port speed by 1.5 times for better system performance, it results in a total of 680 terabit/s usable bandwidth. Without expansion in the photonic switching fabric, the throughput close to 100% can be readily achieved with the internal speedup. We believe that P^{3}S has potential to be the feasible platform in building next-generation switches with capacity from terabit to perabits per second while offering excellent system performances.

VI. REFERENCES:

[1] J. E. Smith and F. W. Weingarten, Eds., "Research Challenges for the Next Generation Internet," *Computing Res. Assoc*., May 12-14, 1997

[2] http://www.velio.com/product/product_crosspoint.html

[3] H. J. Chao, et. al, "A photonic front-end processor in a WDM ATM multicast switch," *IEEE Journal of Lightwave Technology*, Vol. 18, No. 3, pp. 273-285, March 2000

[4] S. Verma, H. Chaskar, R. Ravikanth, Optical Burst Switching - A viable solution for Terabit IP backbone," *IEEE NETWORK 2000*, Vol 14, Iss 6, pp 48-53

[5] R. Ryf, et. Al., "1296-port MEMS transparent optical cross-connect with 2.07 petabit/s switch capacity," *Optical Fiber Communications Conference 2001, OFC 2001*, PD22-1

[6] P. Bernasconi, C.Doerr, C.Dragone, M. Cappuzzo, E. Laskowski, A. Paunescu, "Large N x N waveguide grating routers, " *J. of Lightwave Technology*, vol. 18, issu. 7. pp. 985 –991, 2000

[7] Chun-Kit Chan, K.L. Sherman, M. Zirngibl, "A fast 100-channel wavelength-tunable transmitter for optical packet switching," *IEEE Photonics Technology Letters* , vol. 13, issu. 7, July 2001 , pp: 729 -731

[8] R. Ramamurthy, B. Mukherjee, **"**Fixed-alternate routing and wavelength conversion in wavelength-routed optical networks," IEEE/ACM Transactions on Networking, vol., 10, Issue:3 , Jun 2002, pp.351 -367

[9] F.S. Choa and H. J. Chao, "All-optical packet routing - architecture and implementation," *Journal of Photonic Network Communications*, Vol. 1, No. 4, pp. 303-311, 1999.

[10] H. J. Chao and T. S. Wang, ``An optical interconnection network for Terabit IP routers," *IEEE Journal of Lightwave Technology*, vol. 18, no. 12, pp. 2095-2112, Dec. 2000

[11] K. -L. Deng, R. J. Runser, P. Toliver, I. Glesk, and P. R. Prucnal, "A highly scalable, rapidly-reconfigurable, multicasting-capable, 100-Gbit/s photonic switched interconnect based upon OTDM technology," *Journal of Lightwave Technology* 18 (12) 1892-1904 (2000).

[12] D.T.K. Tong, K.-L Deng, B. Mikkelsen, G. Ranbon, K.F. Dreyer, J. Johnson, "160 Gbit/s clock recovery using electroabsorption modulator-based phase locked loop", *Electron. Lett*., vol. 36, pp.1951-1952, (2000)

[13] K.-L Deng, D.T.K. Tong, C.-K Chan, K.F. Dreyer, J.E. Johnson, "Rapidly re-configurable optical channel selector using RF digital phase shifter for ultra-fast OTDM networks", *Electron. Lett*., vol.36, pp.1724-1725, (2000)

[14] E. Oki, Z. Jing, R. Rojas-Cessa, and H. J. Chao, "Concurrent Round-Robin Dispatching scheme for a Clos-Network Switches," *Prof. IEEE ICC 2001*, Helsinki, Finland, June 2001 and to appear in *IEEE/ACM Trans. on Networking*

[15] Y. Li, S. S. Panwar, and H. J. Chao, "The Dual Round Robin Matching Switch with Exhaustive Service", *in Proc. IEEE HPSR* 2002.

[16] N. Mckeown, "The iSLIP Scheduling Algorithm for Input-Queues Switches," *IEEE/ACM Trans. On Networking*, vol. 7, no. 2, pp. 188-200, April, 1999.

[17] H. J. Chao, "Saturn, A Terabit Packet Switch Using Dual Round-Robin," *Proc.IEEE Communication Magazine*," vol. 38, no. 12, pp. 78-84, Dec., 2000

[18] M. Karol, and C-L. I, "Performance Analysis of a Growable Architecture for Broadband Packet (ATM) Switching", Globecom'89.

[19] Y. Li, S. S. Panwar, and H. J. Chao, "Performance Analysis of a Dual Round Robin Matching Switch with Exhaustive Service", in Proc. IEEE High-Speed Networking Workshop 2002. June 23, 2002 , New York.

[20] 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, Nov.1993.

[21] S. Nakamura, Y. Ueno, K. Tajima, **"**168-Gb/s all-optical wavelength conversion with a symmetric-Mach-Zehnder-type switch**,**" *IEEE Photonics Technology Letters*, vol.13, isuu. 10, pp.1091-1093, (2001)