PACKET SWITCHED OPTICAL NETWORK IN DATA CENTER

Tanjila Ahmed

Proposed Architecture

Mostly passive optical components used

Control remains in the electrical domain

Synchronous/a synchronous network

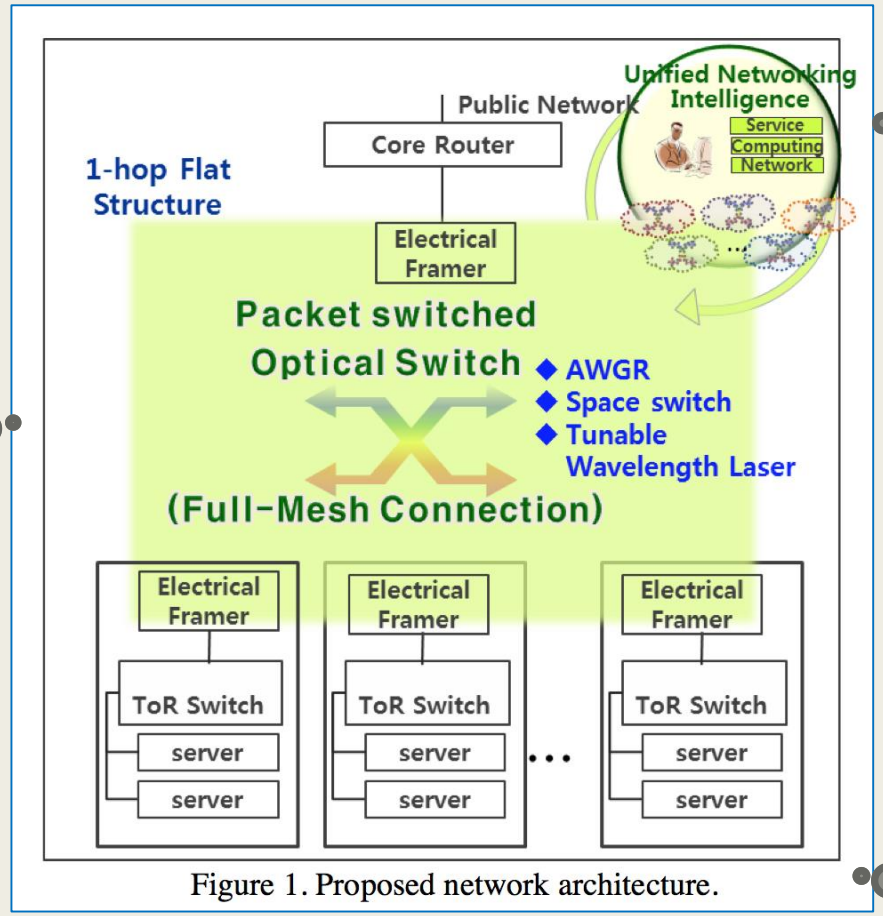

Figure 1. Proposed network architecture.

Electrical framer/deframer at ingress/egress

All optical switching

No electrical buffer and fiber delay loop

Goal

- Design a photonic frame format
- Design a protocol used to synchronize clocks in control plane
- Suggest control message signaling scheme (out of band preferred)

Possible Design Solution 1

40 Gb/s Pure Photonic Packet Switch for Datacenters

Xiaoling Yang, Hamid Mehrvar, Huixiao Ma, Yan Wang, Lulu Liu, H.Y. Fu, Dongyu Geng, Dominic Goodwill, Eric Bernier
OFC 2015.

Goals and characteristics of this design:

1. Demonstrate a 40Gb/s pure photonic packet switch testbed for a data center network
2. Photonic wrap/unwrap packet framing at aggregation layer
3. Having separate control and data plane
4. Proposing synchronization scheme to align packets at photonic switch
5. Wavelength agnostic photonic packet switches at core layer
6. Out of band signaling
7. Handling both unicast and broadcast packets

Network Architecture

1. Three tier hierarchy
2. N=48 servers, M= 32 TOR, P=32 aggregation switch
3. Core Switch: P x P buffer less silicon photonic packet switch
4. Controller: connection management, scheduling, system synchronization
5. Time slot based synchronous network
6. Wrapped photonic frame length = time slot

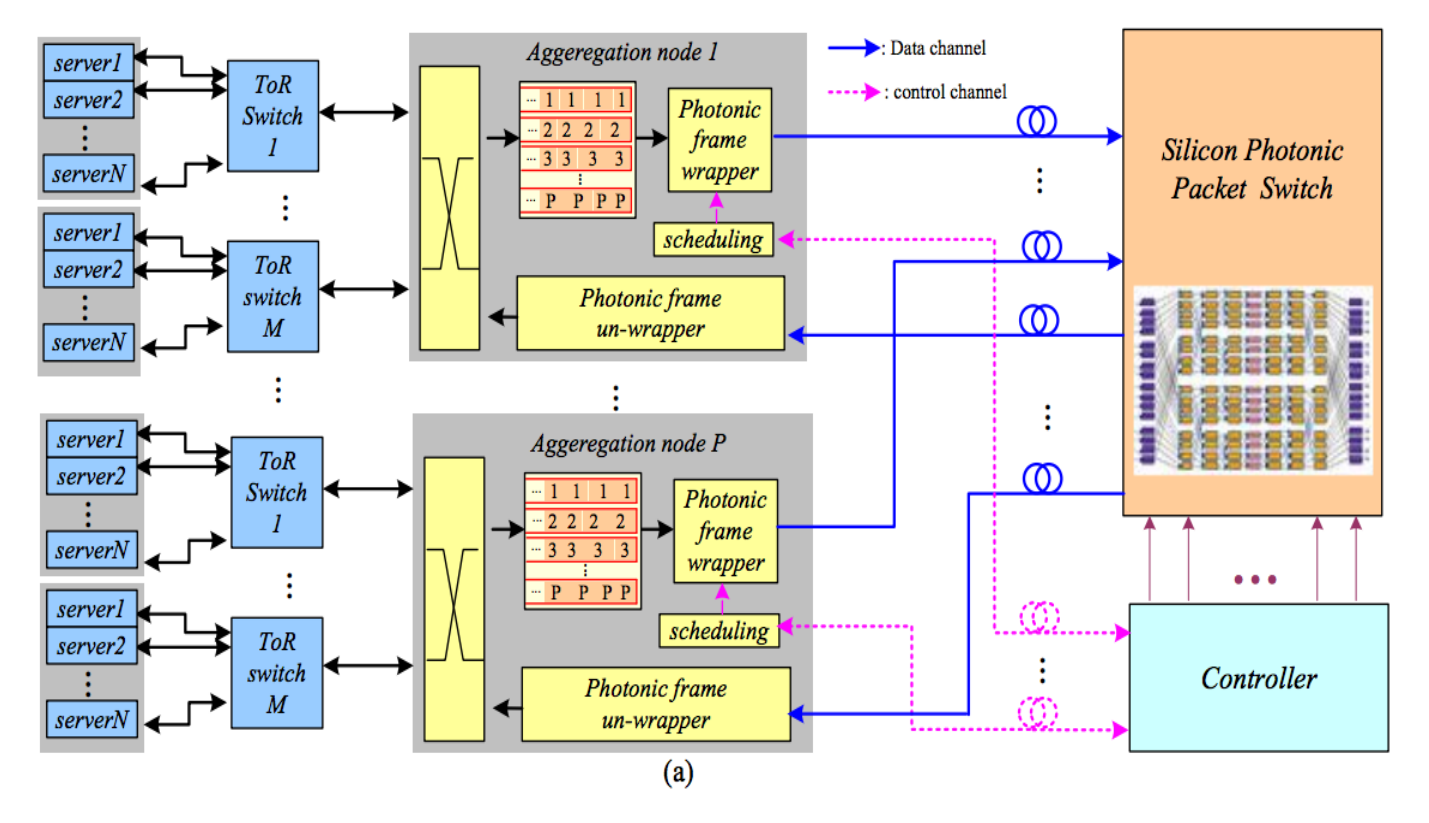

Figure 2.Network structure

Framing Technique

Ingress Card,

1. Ethernet/IP packets are placed into one of the P virtual queues based on their destination MAC
2. For broadcast frame, the packets are copied into all the queues
3. Controller and scheduler determines the packet to be wrapped in next time slot
4. Photonic frame wrapper packs multiple packets from a given virtual queue into a large photonic frame
5. Labels, preamble, start delimiter are also inserted
6. Label transmits into control channel at 1310 nm wavelength
7. Wrapped payload is transmitted through data channel at 1550 nm

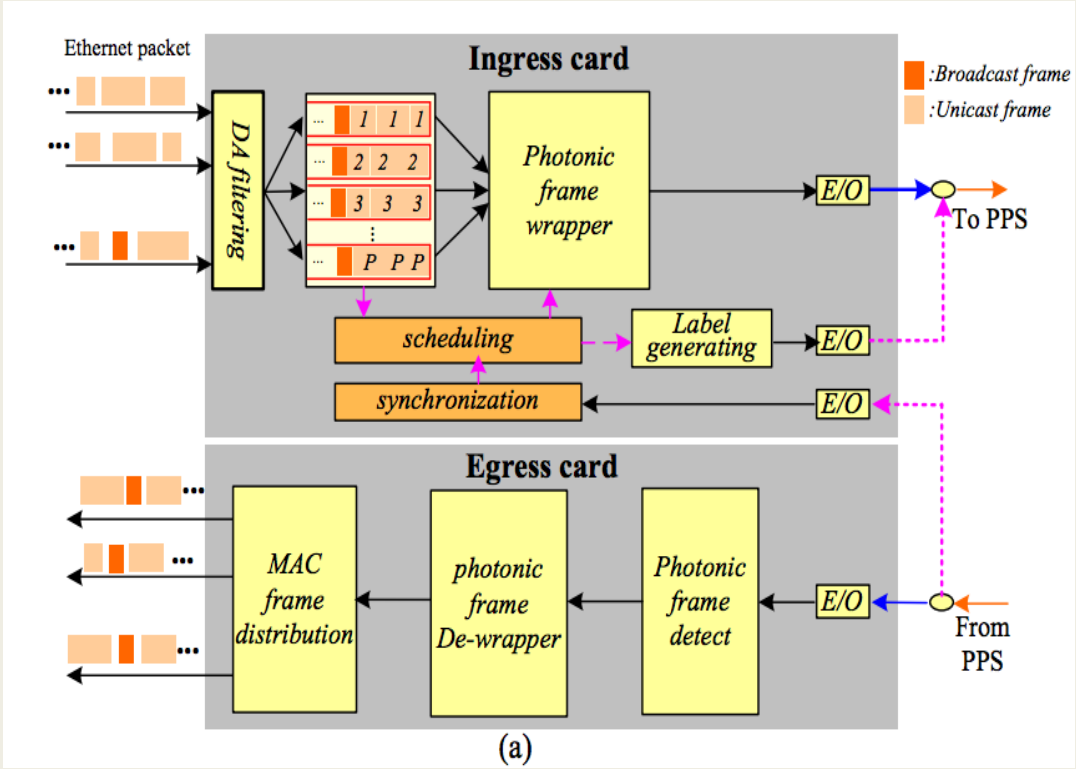

Figure 3.Egress /ingress of aggregation switch

Framing Technique(Contd.)

Egress Card,

1. Received optical packet is disassembled and routed to destination TOR
2. For broadcast packet it is send to each TOR

Frame Format:

1. Packets of same destination are concatenated into a wrapper with no inter-packet-gap
2. Preamble: burst receivers at egress card
3. Delimiter: fixed pattern used for frame alignment and locating label/payload
4. Offset time: label decoding
5. IFG: switch setup
6. ID: identifier for destination aggregation node
7. Full BW utilization:

*packets-per-wrapper * inter-packet-gap > preamble+start delimiter+inter-frame-gap*

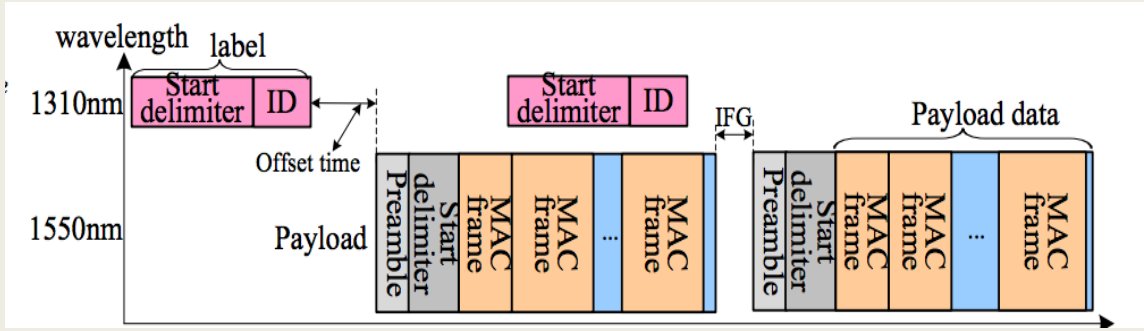

Figure 4.Photonic frame format.

Control and Bandwidth utilization

Controller:

1. The controller send synchronization request
2. The ingress cards returns ACK to the controller
3. From the ACK arrival time the controller determines the ‘start indicator 'should be sent
4. The controller extracts the labels and determines the switching

Bandwidth Utilization vs wrapper size:

1. For a given rate bandwidth utilization increases for lesser preamble length/burst recovery time
2. For given burst recovery time lower rate achieves higher bandwidth utilization

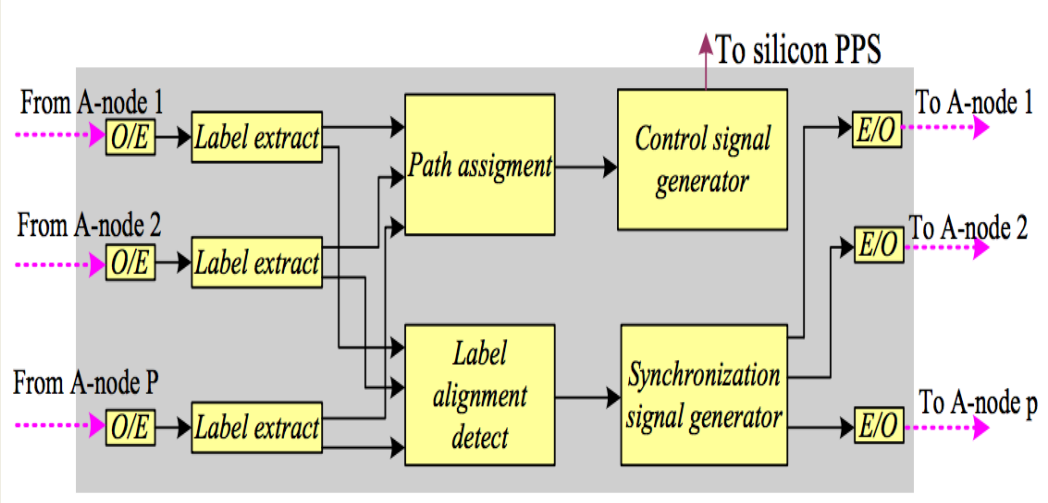

Figure 5.Photonic packet switch controller

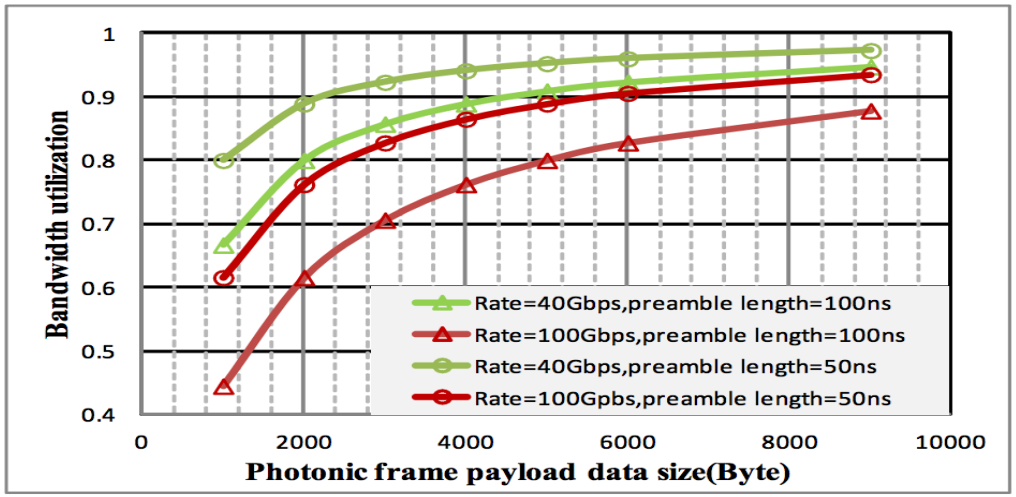

Figure 6.Bandwidth utilization ratio vs photonic frame wrapper size

Possible Design Solution 2

Scalable Photonic Packet Switch Test-bed for Datacenters

Hamid Mehrvar , Yan Wang , Xiaoling Yang , Mohammad Kiaei , Huixiao Ma, Jianchao Cao, Dongyu Geng , Dominic Goodwill , Eric Bernier
OFC 2016

Goals:

1. Using stacked silicon photonic space switches that are time slot synchronized
2. Using efficient low latency algorithm to assign connections to the aggregation nodes via the photonic fabric

Network Architecture

- Stack of M small N×N photonic switch matrix with capacity of (M×N)×(M×N)
- Time slot synchronous controller
- N=32 and M=16, the total switch capacity in 100Tb assuming 100G interconnectivity
- Each aggregation node has M photonic wrapper/un-wrapper interfacing with all M switches
- Aggregation nodes perform buffering and reporting to the main controller the status of the virtual output queues
- Main controller decides on the connectivity map

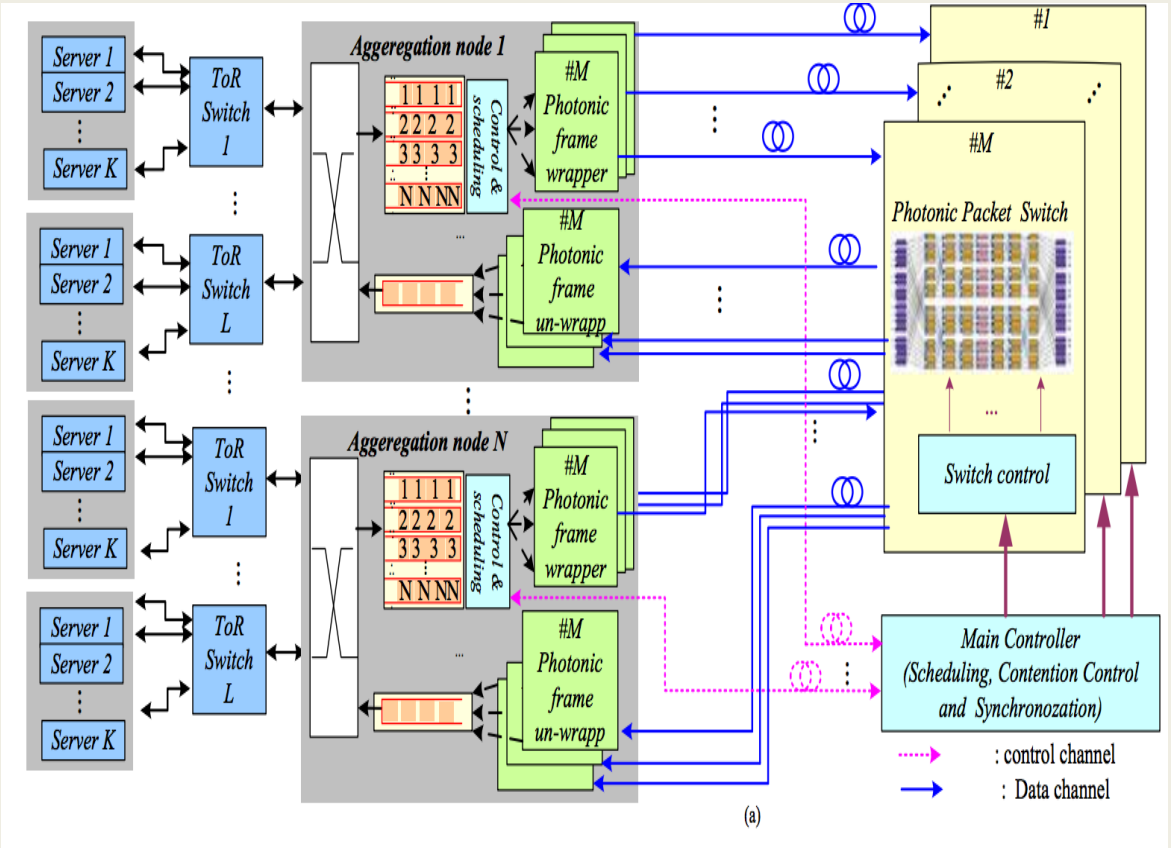

Figure 7.Scaled datacenter using M bufferless photonic switch chips.

Low latency Control Algorithm

- Variable size packets wrapped into frame that fits a fixed timeslot
- Aggregation nodes report status of top R queues ($M \leq R \leq N$)
- Control scheme: number of bytes in the queue is called Longest Queue First (LQF), number of packets is called Largest Number of Packets First (LNPF)
- Controller sorts report in descending order and grants connection if available
- Short Queue faces starvation
- Starvation avoidance:

$$Q = q/Qth + d/Dth$$
$$P = p/Pth + d/Dth$$

p= *number of packets*
q=length of packets
d=oldest packet in a queue
If d>Dth, chances of getting a grant

- Reporting d alone is called oldest packet first (OPF)

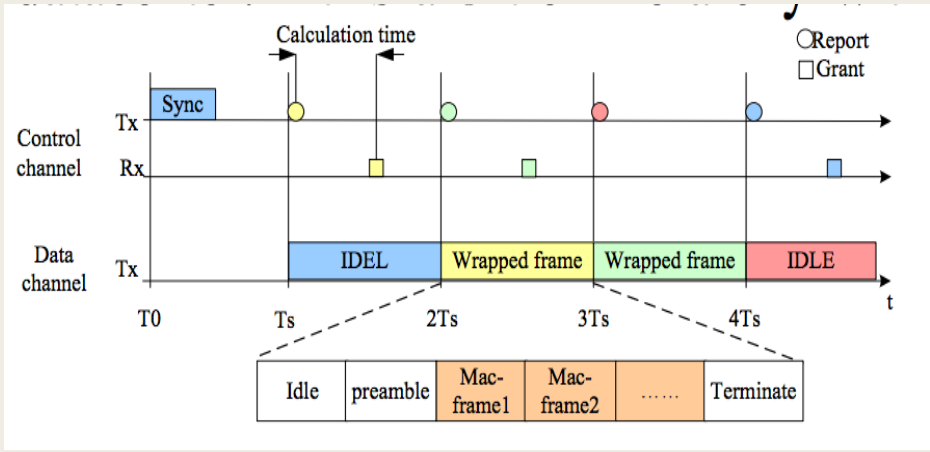

Figure 8. Control and Data path diagram.

Results

1. Scheme 1: no dependency between the interfaces, and assesses the report from each interface separately. If there is an output contention, only one of the requesting interfaces is granted
2. Scheme 2: considers that each aggregation device can send their top M=2 queue traffic using any of the two interfaces upon contention
3. As seen low latency scheme of LQF/SA scheme, outperforms the other two approaches

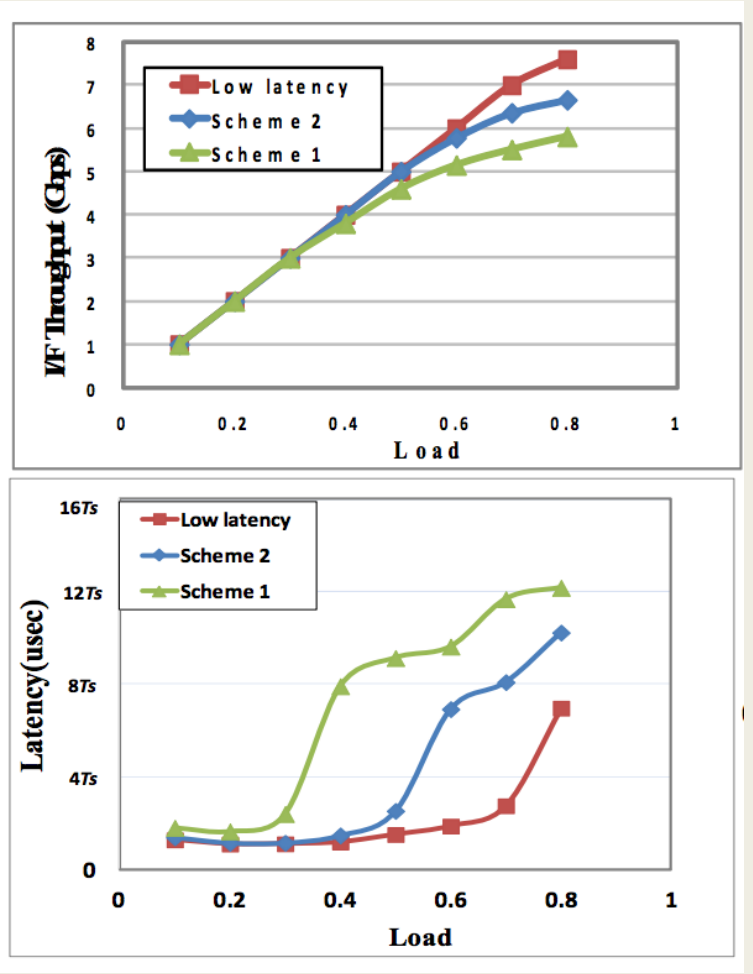

Figure 9.Throughput & Measured average delay graph

Future Work

- Propose architecture also for asynchronous network
- Simulate and compare this design approach with existing designs
- Comparison in performance replacing silicon photonic switch with AWGR and tunable wavelength converters