Volume 7, Issue 2
April 2018

ITEE Journal
Information Technology & Electrical Engineering

ISSN: - 2306-708X

Future Data Centers Core Switches Design Challenges

[1] **Arunendra Singh,** [2]**Amod Kumar Tiwari**
[1] Dr. A.P.J. Abdul Kalam Technical University, Lucknow,
Uttar Pradesh, INDIA

[2] Rajkiya Engineering College, Churk, Sonbhadra,
Uttar Pradesh, INDIA
E-mail: [1] arun.sachan@gmail.com , [2]amodtiwari@gmail.com

ABSTRACT

A data center is a collection of servers which are connected in ToR configuration. Currently data centers run on electronic packet switching technology. Due to explosive growth of data in past few years these electronic data centers are facing bottleneck, and therefore it is expected that both optical circuit switching along with optical packet switching will be used in near future. This paper discusses the evaluation of data centers over the years. Basic concept of the data centers are discussed, and challenges that have to meet are also discussed. Notable switch designs are also discussed along with their pros and cons.

Keywords: *Data Centers, OPS, OCS*

1. INTRODUCTION

In the ever increasing demand for higher bandwidth currently operated servers and data centers start to feel bottleneck. This is because of processing speed limitations of electronic devices. The use of optical fiber in back-bone networks has enhanced speed of data propagation, and enormous bandwidth of fiber supports large number of channels moving in parallel using wavelength division multiplexing (WDM) technology [1]. However, it will take much longer time before all-optical data centers can be deployed as even in presence of recent progresses still optical technology not mature enough. In contrast to this electronic technology is matured enough, and recent progresses are also promising. Therefore, in next few years hybrid technology which uses both electronics and optical for efficient usage of data centers will be used [2-3].

Since late 90's an ever rising growth has been seen in internet traffic, and it doubles in each 18 months in core networking, and in servers input and outputs doubles in each 2 years, therefore to cope up with such rising demands very fast technological innovations are desirable at brisk rate. The rising trends for 25 years are shown in Figure 1.

For instance, a data center with 50,000+ servers, each equipped with 40 Gb/s of transfer speed, would need an internal network with 2 Petabits/sec of total transmission capacity to help full- bandwidth communication among each of the servers. While apparently extraordinary, the innovation, both on the software [4-6] and equipment [7-11] side, is accessible at present. By considering the present datacenter switching and interconnect innovation makes it troublesome and expensive to acknowledge such scale and execution.

2. DATA CENTERS REQUIREMENTS

We start by investigating a portion of the communication and network necessities in developing large-scale data centers. The principal question is the objective scale. Even though economies of scale propose that data centers should to be as big as it can be, normally estimated by the measure of energy accessible for the site; data centers ought to likewise be circulated over the planet for adaptation to fault tolerance and latency area. The other thing is the whole computation and communication limit needed by an objective application. Let us consider person to person communication as an example. Their sites should basically store and replicate each client created content over a bunch worth of machines.

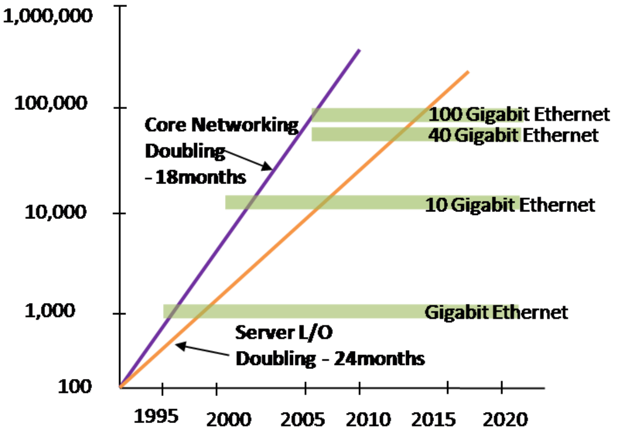
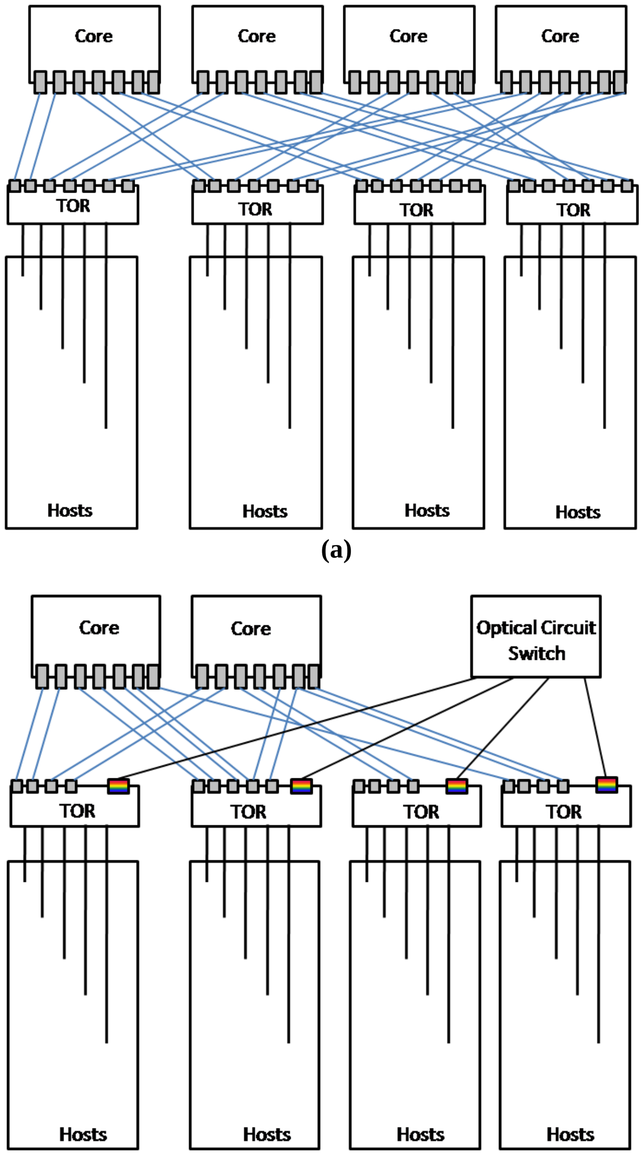


Figure 1: Background: Data Center Network Architecture

The network prerequisites giving support to these kinds of administrations are likewise noteworthy. For every outside demand, a huge number of servers must be reached in parallel to fulfill such demand. The last inquiry is the measure that particular servers are multiplexed crosswise over applications and properties. For example, any portal, Yahoo! may have several individual client confronting administrations alongside a comparable number of inward applications to help mass information handling, index production, advertisements arrangement, and general business bolster.

Although we do not have any definite information regarding these questions, on adjust we place a pattern to expanding register densities in data centers positively at the level of a huge number of servers. It is obviously conceivable to segment singular applications to keep running on committed machines with a devoted interconnect, bringing about smaller-scale networks.



(a)

(b)

Figure 2: (a) Traditional Data centers design (b) Emerging Data centers design

The incremental expense of scaling the network will in a perfect world be unobtrusive [9-14] and the adaptability advantages of both moving calculation powerfully and supporting ever-bigger applications are extensive. Thus, we take interconnects that must generally scale with the quantity of servers in the data center. Figure 2(a) demonstrates the structure of a normal data center networks. Individual racks house contains several servers, which associate with a Top of-Rack (ToR) switch through copper cables. These ToR switches connects to core switching layer through optical transceivers [12].

To develop the bigger scale networks, every ToR switch would associate with all accessible core switches. If a ToR has *m* uplinks, then it connects to *m* core switches. If each core switch has *n* number of ports, then it would supports *n* ToR connections. In the event that every ToR utilizes *u* downlinks hosts then the total network scales to *nxu* ports.

Figure 2(b) demonstrates a data center design for future generation that utilizes optical circuit switching (OCS) as a major technology. We make the replacement of certain part of the center electrical switches with optical circuit switches. Numerous 10G SFP+ (enhanced small form-factor pluggable) transceivers are supplanted with integrated CWDM (coarse wavelength division multiplexing) transceivers (e.g., 4x10G QSFP-LR4) to total electrical channels with a typical goal. While OCS can't carry out per- packet switching, it can switch all the long-lived flows between aggregation points. The expense of per-port of an OCS is aggressive with, if not inherently less expensive than, the comparable EPS. In any case, it has greater limit through wavelength division multiplexing and less consumption of power. WDM decreases cabling complexity, a critical test in the data center. At long last, OCS dispenses with some part of the optical transceivers and EPS ports by wiping out a subset of the needed OEO variation.

Optics assumes a basic part in conveying on the capability of the data center network and tending to the above difficulties. Notwithstanding, completely understanding its potential in the data center network will require a reconsidering of the optical innovation segments generally utilized for telecom and will require enhancements focusing on the particular data center network arrangement conditions. In this research article, we introduce an outline of present data center network arrange deployments, the pretended by optics in this condition, and open doors for creating variations of existing advancements particularly focusing on substantial scale sending in the data center. Specifically, we take WDM innovation streamlined for data center deployments alongside the advantages of consolidating OCS alongside EPS in the data center [13-14].

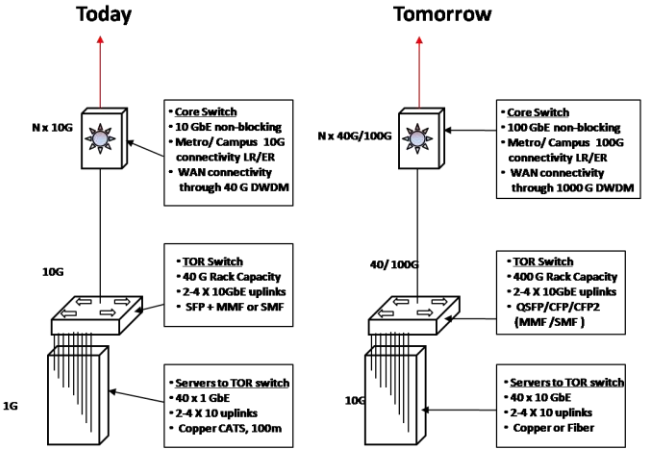


Figure 3: Technology requirements: Today and Tomorrow

Today networks are basically runs on electronics, with 1 Gigabyte Ethernet (GbE) (Fig. 3). Considering 40 links, therefore ToR switches connects to server with maximum capacity of 40 G. these ToR switches connects to core switches. These core switches provide 10 GbE non-blocking connectivity, while in LAN, a connection speed of 40 Gbps can be achieved using DWDM., while in near future it is desirable that the servers speed should be at-least 10 times, while hierarchy above this should be 4-10 times now.

3. CHALLENGES IN DATA CENTERS

In data center systems the cost and performance of communication depends on the distance along with the hierarchy, in terms of layer 2 domains, the hierarchal architecture forms clusters of servers. Further, as spreading a service outside a single layer 2 domain leads to the overhead of re-configuring IP addresses and VLAN trunks, spare capacity throughout the data center is over provisioned per individual service so that each service can scale out to nearby servers to respond rapidly to demand spikes or to failures. Preventing resource sharing, this approach could suffer significant disturbance in case of increasing resource needs [15].

The challenges to meet out are as follows:

Scalability

Due to the rapid growth of network traffic and the fixed growth in the processing power of multi core servers, future data centers should enable the support for millions of microprocessor cores. We expect NGDCs (Next Generation Data Centers) of more than 10 million server cores, which is one order of magnitude larger than the numbers supported by available designs and proposals.

Flexibility

NGDCs are expected to support the requirements of various applications by providing nonblocking connectivity among clusters of computing racks. This is a key requirement of cloud computing. A uniform high capacity, non-blocking network design provides flexibility in slicing the data center, enabling the efficient mapping of application requirements to virtualized resources.

Space Management

Providing the non-blocking data center connectivity would require a massive amount of Ethernet cables when the data center should accommodate millions of server cores, resulting in severe implementation, management, and maintenance problems. So an alternative technology is required that reduces the cabling requirements in future data center systems.

Energy Efficiency

In the information and communication technology the data centers are the major power consumers. The key areas that have become the major concern for the research community are the cost associated with fuelling data centers as well as their greenhouse gas emissions. In heavy data center systems the electronic links and switches consumes a large amount of power and it is difficult to perform power savings by turning of underutilized machines. An energy efficient interconnect is required for significant energy saving in future data center [16-17].

4. NOTABLE SWITCH DESIGNS

A. Datacenter Optical Switch (DOS)

Datacenter Optical Switch (DOS) is packet-based optical architecture presented by X.Ye et al. [18]. The key component in the switching system is Arrayed Waveguide Grating Router(AWGR), which permits contention resolution only in the wavelength domain. AWGR is capable of multiplexes a large number of wavelength into a single optical fibre at the transmission end and demultiplexes to retrieve individual channels at the receiving end. Apart from the AWGR, the switching fabric consists also an array of Tuneable Wavelength Converters (TWCs), Label Extractors (LEs), a loopback shared Synchronous Dynamic Random Access Memory(SDRAM) buffer and a control plane. Figure 4, depicts the high level overview diagram of the DOS architectures. The AWGR can convey optical signals through from any input port to any output port. The wavelength channel that carries the signal would decide the routing path inside the AWGR. Having the TWC set up before the AWGR, each for one node, it is possible to configure an appropriate transmitting wavelength at each input of AWGR separately with distinct wavelengths, so that a non-blocking desired routing path with different optical signal is established.

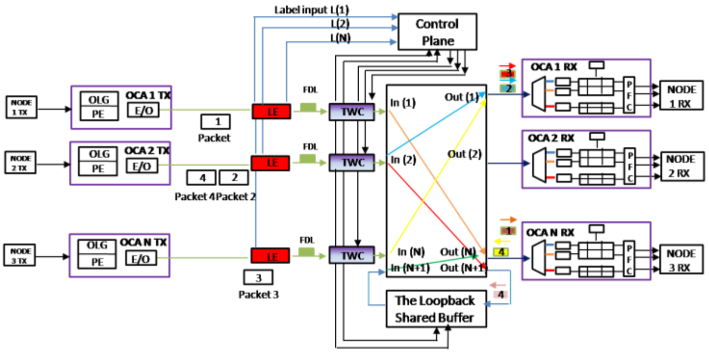


Figure 4: Schematic of Data center optical switch

After the LEs receives a packet from ToR switches, the optical labels are detached from the optical payloads and sent to the DOS control plane, shown in Figure 4. The label has information of the packet length and destination address. Inside the control plane, the optical signal is converted to electrical signal by an optical-to-electrical (O/E) module and then forwarded to the label processor, which sends a request to the arbitration unit for content resolution. The control plane configures control signal to TWCs after arbitration, and sending proper wavelength to the inputs of AWGR. For the outputs of TWCs with no assignment, the control plane sends them wavelengths to carry packets to the AWGR outputs which connect with the shared buffer.

A shared buffer is need for contention resolution when the number of nodes is more than the number of output receivers. It is used to store temporarily for the transmitted packets, which cannot reach the desire outputs, so that they can try it later. Figure 5 shows the loopback shared SDRAM with electrical-to-optical (E/O) converters, optical DEMUX and MUX. The wavelengths which failed to receive a grant in arbitration are routed to the buffer system. Out from the same output of AWGR, the wavelength are split by the optical DEMUX then converted to electrical signal through the optical to electrical converts. Then the packets stay in SDRAM which connects to a shared buffer controller. This controller generates requests to the control plane according to the queue status in the buffer and waits for a grant. The packet is retrieved from the buffer, when the grant arrives.

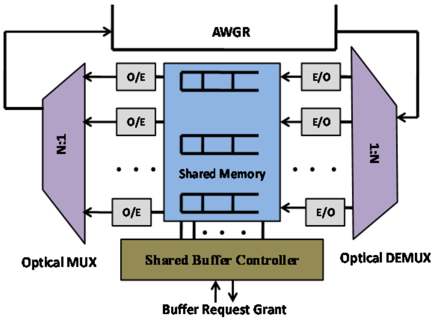


Figure 5: Schematic of SDRAM

Advantages:

1. The DOS architecture has quite low latency which also stays independent of the number of inputs. Because the ToR packets only travel through optical switches, no delay from buffer of electrical switches.

2. The TWC has rapid reconfiguring time of a few nanoseconds, which is useful to meet the demand of bursty traffic fluctuation.

Dis-Advantages:

1. In terms of congestion resolution, the electrical buffer together with O/E, O/E converters draw power consumption and increase packet latency.

2. The cost of TWCs is quite high compared with other commodity optical devices.

B. IRIS (Integrated Router Interconnected Spectrally)

The IRIS project is one of the research results from the program Data in the Optical Domain Networking, which is proposed for exploring photonics packet routers technologies [19]. IRIS is a three-stage architecture using Wavelength Division Multiplexing (WDM) and Arrayed Waveguide Grating Routers (AWGR) with all optical wavelength converters. Though the two space switches are partially blocking, IRIS is still a dynamically non-blocking system.

The architecture of IRIS is illustrated in Figure 6, In the first stage, a ToR switch on each node is linked to a port of the first space switch via N WDM wavelengths channels. After a NxN AWGR, the packets are distributed consistently to the second stage in a random schedule or through a simple round-robin way [20]. The second stage is a time switch that contains N optical time buffers to hold the packets until next stage. Inside the time buffer there are an array of WC and two AWGRs which are connected with multiple shared optical delay lines, each of them carries with different delays. The optical signal is converted by the WC to a specific wavelength, and then it is routed to the AWGR with the needed time delay. After a second AWGR, the delayed signals are multiplexed and sent to the third stage, another round-robin space switch, where the signal is converted back to the required wavelength and sent to the destination port via multiples of the packet-slot duration , the optical time buffer can delay N simultaneous packets.

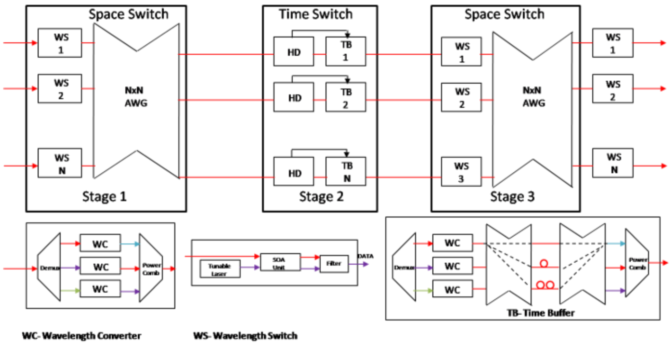


Figure 6: Schematic of IRIS

In case that the buffer overflows, the packets can be dropped too. Through configuring the AWGRs which connected with delay lines, the packets can enter the time buffer and reach the corresponding output port with the independent delay path. The third space switch in the architecture is a periodic operation and the scheduling is deterministic and local to each optical time buffer, so that it significantly reduces the complexity of the control centre and complete without optical random access memory.

Advantages:

1. It is easy to scale the architecture. A 40Gb/s wavelength converters and 80x80 AWGRs allows the system to scale to 256Tb/s.

Dis-Advantages:

1. Speed is a limitation in the design due to the use of space switch.

2. Cost is high.

C. WDM-Passive Optical Network (PON)

In [21], a novel hybrid architecture which introduces passive optical components such as Arrayed Wave Guide Routers (AWGR) is proposed by Christoforos Kachris and Ioannis Tomkos. It contains of both commodity Ethernet electronic switches and WDM PON devices. The performance of the simulation is reported a 10 % power reduction using different traffic ratios for both inter and intra rack flows. The design of the system as shown in Figure 7, in each rack, there are a ToR switches and an optical WDM PON. The ToR is used for the intra-rack communication while WDM PON participates in offloading inter-rack traffic to eliminate additional processing in ToR switch. Hence, the power consumption waste between ToR Switches for inter-rack is reduced and high throughputs are achieved with low latency.

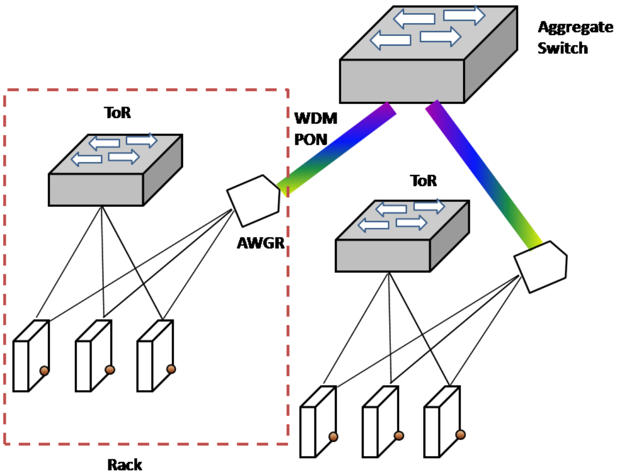


Figure 7: Passive WDM PON

Compared with the telecommunication system, in this architecture, the ToR switches are used as an optical network units and the Aggregate switch is used as optical link terminator. In each server, a commodity Ethernet transceiver is set for intra-rack communication and an WDM transceiver is for inter-rack communication.

Normally in the reference system the power consumption consists of the power from the ToR, aggregate switches and the Ethernet transceivers, which including the edge links and the aggregate links work. But in WDM PON network, the power consumption is mainly the power dissipation in ToRs, aggregate switches, the Ethernet transceivers and the WDM SFP transceivers.

Advantages:

1. The WDM-PON system can provide 10 % reduction of power consumption with no side effect on the packet latency.
2. The architecture can be further developed to the core layer of the data centre, saving more power consumption in operational cost.

Dis-Advantages:

1. Due to the lack of flexibility, a pure WDM PON architecture tends to waste bandwidth.

4. CONCLUSIONS

This paper discusses the needs and requirements of near future data centers. We begin our paper with investigating rise in data requirements over the period of time. The basic layout design of data centers is discussed. The challenges faced in data centers designs are also discussed. The recent data centers designs are also discussed with their advantages and dis-advantages. From above, it is conclusive that AWGR will be an integral part of the optical switch and data centers design in future. The commercialization of TWC will be break-through in the field of optical data centers design.

REFERENCES

[1] Astfalk, G., 2009. Why optical data communications and why now?. *Applied Physics A*, *95*(4), pp.933-940.

[2] Davis, A., 2010, August. Photonics and future datacenter networks. In *Hot Chips 22 Symposium (HCS), 2010 IEEE* (pp. 1-38). IEEE.

[3] Kachris, C. and Tomkos, I., 2012. A survey on optical interconnects for data centers. *IEEE Communications Surveys & Tutorials*, *14*(4), pp.1021-1036.

[4] Cvijetic, N., Tanaka, A., Ji, P.N., Sethuraman, K., Murakami, S. and Wang, T., 2014. SDN and OpenFlow for dynamic flex-grid optical access and aggregation networks. *Journal of Lightwave Technology*, *32*(4), pp.864-870.

[5] Cui, L., Yu, F.R. and Yan, Q., 2016. When big data meets software-defined networking: SDN for big data and big data for SDN. *IEEE network*, *30*(1), pp.58-65.

[6] Channegowda, M., Nejabati, R. and Simeonidou, D., 2013. Software-defined optical networks technology and infrastructure: Enabling software-defined optical network operations. *Journal of Optical Communications and Networking*, *5*(10), pp.A274-A282.

[7] Yeow, T.W., Law, K.E. and Goldenberg, A., 2001. MEMS optical switches. *IEEE Communications magazine*, *39*(11), pp.158-163.

[8] Richardson, D.J., Fini, J.M. and Nelson, L.E., 2013. Space-division multiplexing in optical fibres. *Nature Photonics*, *7*(5), p.354.

[9] Priolo, F., Gregorkiewicz, T., Galli, M. and Krauss, T.F., 2014. Silicon nanostructures for photonics and photovoltaics. *Nature nanotechnology*, *9*(1), p.19.

[10] Fitsios, D., Alexoudi, T., Kanellos, G.T., Vyrsokinos, K., Pleros, N., Tekin, T., Cherchi, M., Ylinen, S., Harjanne, M., Kapulainen, M. and Aalto, T., 2014. Dual SOA-MZI Wavelength Converters Based on III-V Hybrid Integration on a µm-Scale Si Platform. *IEEE Photonics Technology Letters*, *26*(6), pp.560-563.

[11] Schares, L., Kuchta, D.M. and Benner, A.F., 2010, August. Optics in future data center networks. In *High Performance Interconnects (HOTI), 2010 IEEE 18th Annual Symposium on* (pp. 104-108). IEEE.

[12] Glick, M., 2013. Optical interconnects in next generation data centers: An end to end view. In *Optical Interconnects for Future Data Center Networks* (pp. 31-46). Springer, New York, NY.

[13] Davis, A., Jouppi, N.P., McLaren, M., Muralimanohar, N., Schreiber, R.S., Binkert, N. and Ahn, J.H., 2013. The role of photonics in future datacenter networks. In *Optical Interconnects for Future Data Center Networks* (pp. 67-93). Springer, New York, NY.

[14] Chen, L., Sohdi, A., Bowers, J.E., Theogarajan, L., Roth, J. and Fish, G., 2013. Electronic and photonic integrated circuits for fast data center optical circuit switches. *IEEE Communications Magazine*, *51*(9), pp.53-59.

[15] Xu, L., Zhang, W., Lira, H.L., Lipson, M. and Bergman, K., 2011. A hybrid optical packet and wavelength selective switching platform for high-performance data center networks. *Optics Express*, *19*(24), pp.24258-24267.

[16] Beloglazov, A., Abawajy, J. and Buyya, R., 2012. Energy-aware resource allocation heuristics for efficient management of data centers for cloud computing. *Future generation computer systems*, *28*(5), pp.755-768.

[17] Chen, J., Gong, Y., Fiorani, M. and Aleksic, S., 2015. Optical interconnects at the top of the rack for energy-efficient data centers. *IEEE Communications Magazine*, *53*(8), pp.140-148.

[18] Ye, X., Yin, Y., Yoo, S.B., Mejia, P., Proietti, R. and Akella, V., 2010, October. DOS: A scalable optical switch for datacenters. In *Proceedings of the 6th ACM/IEEE Symposium on Architectures for Networking and Communications Systems* (p. 24). ACM.

[19] Gripp, J., Stiliadis, D., Simsarian, J.E., Bernasconi, P., Le Grange, J.D., Zhang, L., Buhl, L. and Neilson, D.T., 2006. IRIS optical packet router. *Journal of Optical Networking*, *5*(8), pp.589-597.

[20] Gripp, J., Simsarian, J.E., LeGrange, J.D., Bernasconi, P. and Neilson, D.T., 2010, March. Photonic terabit routers: The IRIS project. In *Optical Fiber Communication Conference* (p. OThP3). Optical Society of America.

[21] Kachris, C. and Tomkos, I., 2011, July. Power consumption evaluation of hybrid WDM PON networks for data centers. In *Networks and Optical Communications (NOC), 2011 16th European Conference on* (pp. 118-121). IEEE.

AUTHOR PROFILES

Arunendra Singh is presently working as an Assistant Professor at Pranveer Singh Institute of Technology, Kanpur in Department of Information Technology. He is pursuing his Ph.D. in Computer Science & Engineering from Dr. A.P.J. Abdul Kalam Technical University, Lucknow, Uttar Pradesh. He received M.Tech. from Motilal Nehru National Institute of Technology(MNNIT) ,Allahabad in 2011. He received his B.Tech.(IT) from Harcourt Butler Technological Institute (HBTI), Kanpur(U.P.) in 2005. His area of interest is Optical Communication, Computer Networks, GIS.

Dr. Amod Kumar Tiwari is working as an Associate Professor in Department of Computer Science and Engineering in Rajkiya Engineering College, Sonbhadra, Uttar Pradesh. He has completed his PhD in Computer Science from Dr. A.P.J Abdul Kalam Technical University. He has a vast experience of more than 6 years of guiding many PhD and M.Tech. scholars. His areas of research include Image Processing, Algorithms Design, Geometrical Modelling, Artificial, Intelligence, Computer Networks and Optical Communication.