The Hi-Ring DCN Architecture

Galili, Michael; Kamchevska, Valerija; Ding, Yunhong; Berger, Michael Stübert; Oxenløwe, Leif Katsuo; Dittmann, Lars

Published in:
Proceedings of Asia Communications and Photonics Conference 2016

Link to article, DOI:
10.1364/ACPC.2016.AS1E.1

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Citation (APA):
Galili, M., Kamchevska, V., Ding, Y., Berger, M. S., Oxenløwe, L. K., & Dittmann, L. (2016). The Hi-Ring DCN Architecture. In *Proceedings of Asia Communications and Photonics Conference 2016*
https://doi.org/10.1364/ACPC.2016.AS1E.1

The Hi-Ring DCN Architecture

Michael Galili, Valerija Kamchevska, Yunhong Ding, Michael Berger, Leif K. Oxenløwe, Lars Dittmann
DTU Fotonik, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark
e-mail: mgal@fotonik.dtu.dk

Abstract: We will review recent work on the proposed hierarchical ring-based architecture (Hi-Ring) proposed for data center networks. We will discuss the architecture and initial demonstrations of optical switching performance and time-domain synchronization.

OCIS codes: (060.4262) Networks, ring; (060.6718) Switching, circuit

1. Introduction

Current datacentre network (DCN) architectures are severely challenged by the continuous growth of large scale datacentres. Commonly deployed architectures like fat-tree [1] scale to large networks by either using switches with higher port counts (radix) or by adding more layers (tiers) to the network. High-radix Ethernet switches are being developed intensively, however, at any given time limited switch radix will determine the size to which a DCN can grow before additional tiers must be introduced. Many-tiered networks are generally associated with higher cost and energy consumption making this undesirable. Therefore, significant research effort is going into developing alternative DCN concepts [2-4].

In this work we focus on a ring-based topology with hierarchically structured optical nodes. Each node-layer is capable of optical switching in a separate physical dimension. We use this to demonstrate a ring network based on multicore fibre (MCF) carrying 1 Tbit/s/core. The network nodes perform switching at the core-, wavelength-, time-slot level. To enable stable time-slotted connections using time domain multiplexing (TDM) we have developed a scheme for global synchronisation at the optical level between nodes in a ring network.

2. The Hi-Ring Architecture

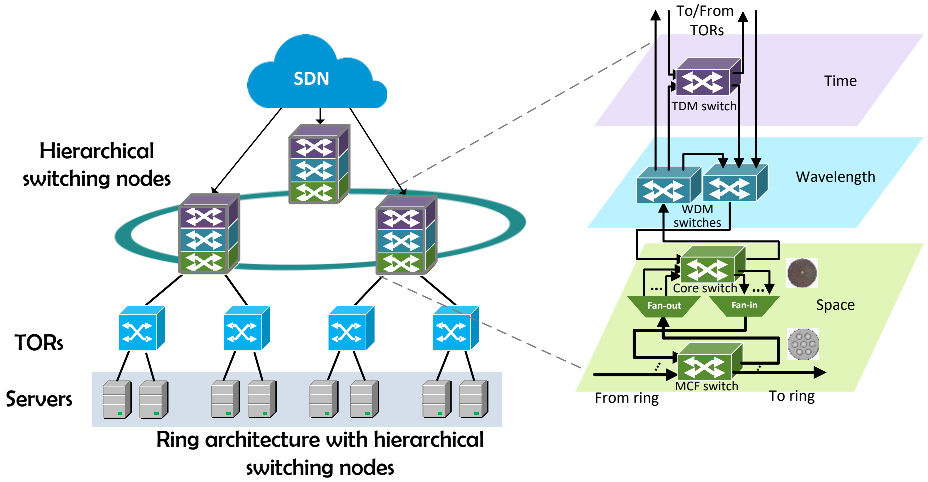


Figure 1 – left: the overall Hi-Ring Architecture. Right – the structure of the proposed optical switching nodes. Adapted from [5]

Figure 1 (left) shows the proposed Hi-Ring architecture. The multidimensional switching nodes are all-optical hierarchically layered nodes capable of switching in three different dimensions (space, wavelength and time). The hierarchical layout enables connections to be bypassed at a low level corresponding to a high degree of data aggregation whenever they are destined to other nodes or the connections can be passed to a higher level in case switching with finer granularity is required. This structure is well suited for a ring topology which inherently has a large amount of bypass traffic at each node. Figure 1 (right) illustrates the structure of one of the nodes.

In a node with all switching layers present, switching in the space dimension is performed using two different types of switches, one operating at a multicore fibre (MCF) granularity and another one operating at the granularity of a single fibre core. If the traffic carried in one multicore fibre is not intended for the node through which it is passing, then bypass at the lowest level minimises the use of switch resources for that connection. Switching of individual cores is used to rearrange the content of cores dropped by the lowest switching layer. It also enables passing the content of individual cores to the next higher level in the node.

The WDM switch is a reconfigurable wavelength switch which allows for dynamic wavelength multiplexing and demultiplexing. WDM provides support for connections requiring significant bandwidth. In the envisioned structure direct connections exist between the WDM layer and the communicating hosts e.g. Top-of-Rack (TOR) switches. WDM connections can thus be established utilizing switches in the space and wavelength dimensions only.

The topmost layer of the node is TDM switching of time slots. This allows for connections to be established with sub-wavelength granularity. Strict synchronisation of the nodes is required for TDM transmission to succeed.

The proposed architecture has several advantages. It provides high connectivity in a scalable way for interconnecting server racks. It allows for simultaneous connections with different granularity. The hierarchical node structure results in highly aggregated traffic on the actual physical links between the nodes. This means that a relatively simple physical topology with a reasonable number of nodes and physical links can support a vast amount of traffic and a large number of hosts.

Unlike electrical switching, the multidimensional switching node is bit rate independent. Changes in channel bandwidth and flex grid operation can be supported by simple reconfiguration of the WDM switch and edge transceivers. As all switching within the node is performed optically the latency experienced by the signal is only that associated with propagation delays. For most applications this will have negligible impact given the small dimension of the node. Finally, by simultaneous optical switching of highly aggregated data i.e. a large number of data connections, the consumed energy to manage each connection is expected to be very low.

3. Switching in the optical nodes

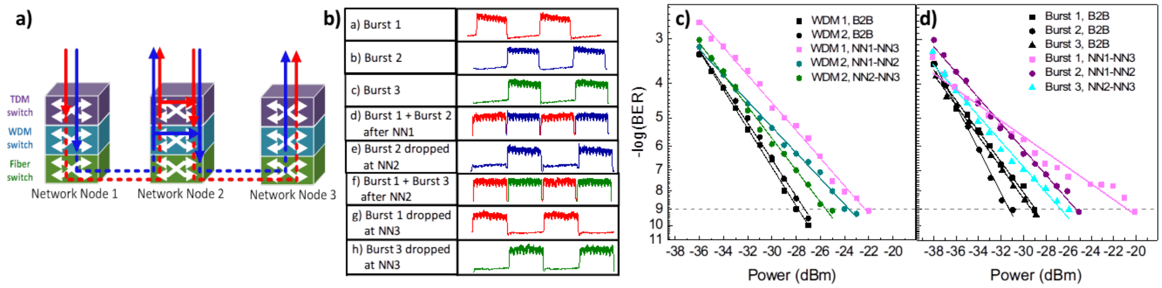
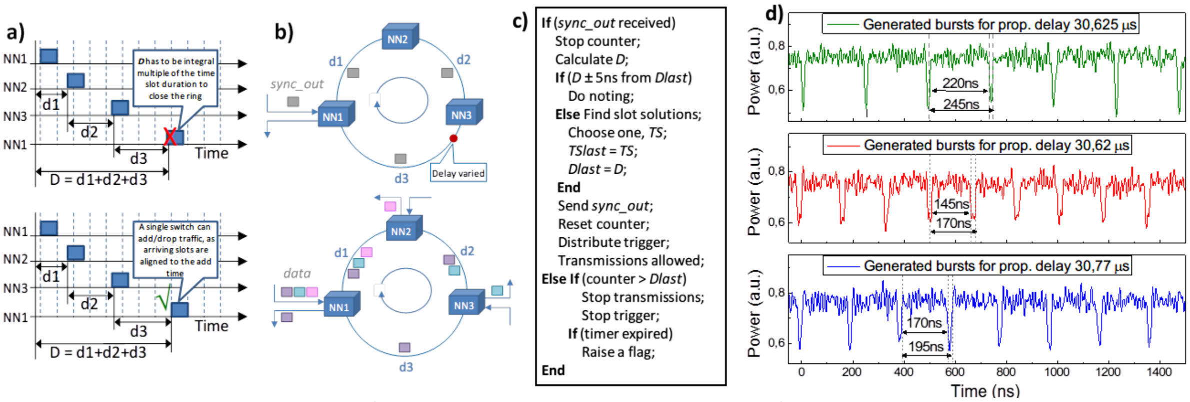


Figure 2 – a) Schematic showing the different tested connections (blue arrows indicate WDM, red arrows indicate TDM). b) Measured traces of the TDM data bursts in different configurations. c) BER performance for the WDM channels. d) BER performance for the TDM bursts. Adapted from [5]

Figure 2 a) shows the test case used to demonstrate joint switching of fibre cores, wavelengths and TDM channels. A number of WDM channels exist in the network. Most of them are dedicated to full wavelength channels while one wavelength contains TDM timeslots forming data connections at sub-wavelength granularity. The blue arrows in the figure illustrate the wavelength connections which are switched to the WDM-level in the hierarchical structure of node 2 and subsequently de-multiplexed. One channel is dropped while one channel is forwarded to node 3. The resource vacated by the dropped channel is used to support a new channel connecting node 2 to node 3. The wavelength carrying TDM-slotted channels is passed to the TDM switch in node 2 where one TDM channel is dropped and a new channel is added in its place. The two TDM channels are then forwarded to node 3. Figure 2 b) shows the TDM slots; bursts 1 and 2 are transmitted from node 1, burst 2 is dropped at node 2 where burst 3 is added in its place. Burst 1 and 3 are then transmitted to node 3 where they both terminate. Figure 2 c) shows the bit error rate (BER) performance of the WDM channels. A reasonably small variation in penalty is achieved for the different connections in the network. Figure 2 d) shows the BER performance of the TDM connections in the network. The ‘single hop’ connections from node 1 to node 2 and from node 2 to node 3 are very similar. However, the ‘multihop connection’ from node 1, through node 2 to node 3 has somewhat reduced performance. Bypassing a node should thus be done at the lowest possible level in the switch hierarchy. Appropriate traffic grooming to minimise the number of TDM bypasses combined with TDM switches with high extinction ratio is expected to allow scaling this architecture to much larger networks.

4. Synchronization of TDM connections


a)
NN1
NN2
NN3
NN1
d1
d2
d3
Time
D = d1+d2+d3
D has to be integral multiple of the time slot duration to close the ring
NN1
NN2
NN3
NN1
d1
d2
d3
Time
D = d1+d2+d3
A single switch can add/drop traffic, as arriving slots are aligned to the add time
b)
NN2
d1
d2
sync_out
NN1
NN3
d3
Delay varied
NN2
d1
d2
data
NN1
NN3
d3
c)
d)
Generated bursts for prop. delay 30,625 µs
220ns
245ns
Generated bursts for prop. delay 30,62 µs
145ns
170ns
Generated bursts for prop. delay 30,77 µs
170ns
195ns
Power (a.u.)
0,8
0,6
0
200
400
600
800
1000
1200
1400
Time (ns)

```
If (sync_out received)
   Stop counter;
   Calculate D;
   If (D ± 5ns from Dlast)
      Do noting;
   Else Find slot solutions;
      Choose one, TS;
      TSlast = TS;
      Dlast = D;
   End
   Send sync_out;
   Reset counter;
   Distribute trigger;
   Transmissions allowed;
Else If (counter > Dlast)
         Stop transmissions;
         Stop trigger;
      If (timer expired)
         Raise a flag;
End
```

Figure 3 - Schematic showing the requirement for global ring synchronisation. b) Illustration of sync pulse used to measure propagation delay and data bursts propagating in time slots in a properly synchronised ring. c) Pseudocode describing the process on delay monitoring and time slot selection. d) Traces of automatically adapting time slots when propagation delay is varied. Adapted from [6]

Global synchronisation of the network nodes is a strict requirement in the TDM-based network. We have chosen the approach of pushing all signal-buffering to the edges of the network avoiding the challenging task of buffering optical signals. Bursts of data are thus only sent into the network in a TDM time slot when a predetermined and guaranteed path is available through the network. The ring topology simplifies global synchronisation of the network. We have developed a mechanism for real time monitoring of the ring-roundtrip delay simultaneously providing synchronisation signals to the nodes in the ring. Based on this monitoring and knowledge of hardware limitations in the network nodes (rise- and fall time of switches etc.) we calculate and continuously update a table of allowed TDM structures. The TDM structure defines the number of TDM slots, their duration and the duration of the guard bands. The preferred TDM solution will then be selected by the network controller. Figure 3 a) illustrates the synchronisation challenge for a 3-node ring. The round trip delay has to be a multiple of the time slot duration to maintain synchronisation after a full round trip. Figure 3 b) shows the basic principle of the scheme. A synchronisation signal is regularly launched into the ring measuring the delay and synchronising the nodes. This allows for real data bursts to be subsequently launched into stable TDM slots creating reliable connections between all nodes in the ring. Figure 3 c) shows pseudo code for the algorithm running on the FPGA which is tracking the roundtrip delay and updating the TDM slots. Finally, Figure 3 d) shows three solutions for allowed TDM slots which are resolved and implemented in real time as the roundtrip delay was manually changed.

6. Conclusion and Acknowledgements

Based on our recent work we believe we have made significant progress towards offering optical switching technologies as viable means of improving capacity and scaling of high density short range networks, specifically in large scale datacentres. The Hi-Ring architecture exploits key benefits of optical switching as low loss and transparency to data rate or format, while also enabling global TDM synchronisation for sub-wavelength granularity connections and a highly simplified cabling structure compared to many other architectures.

This work is supported by ECFP7 grant no. 619572, COSIGN, by the Danish Council for Independent Research (DFF-1337-00152 and DFF-1335-00771) and finally by the Danish National Research Foundation through SPOC (DNRF123).

5. References

[1] M. Al-Fares et al., “A scalable, commodity data centre network architecture,” Proc. ACM SIGCOMM, (2008).
[2] N. Farrington et al., “Helios: A hybrid electrical/optical switch architecture for modular data centers,” Proc. ACM SIGCOMM, (2010).
[3] G. Wang et al., “c-Through: part-time optics in data centers,” Proc. ACM SIGCOMM, (2010).
[4] Z. Cao et al., “Experimental demonstration of dynamic flexible bandwidth optical data center network with all-to-all interconnectivity,” Proc. ECOC, PD.1.1, (2014)
[5] V. Kamchevska et al., “Experimental Demonstration of Multidimensional Switching Nodes for All-Optical Data Center Networks,” Journal of Lightwave Technology, vol: 34, issue: 8, pages: 1837-1843, 2016
[6] V. Kamchevska et al., “Synchronization Algorithm for SDN-controlled All-Optical TDM Switching in a Random Length Ring Network,” in Proc. OFC 2016, Th3I.2, Anaheim, USA