Fault-Tolerance Enhanced Design and Analyses for Optical Pyramid Data Center Network (OPMDC)

Maria C. Yuang[1*], Jing-Chun Yang[1], Hsing-Yu Chen[2], Po-Lung Tien[3]

1. Department of Computer Science, National Chiao-Tung University, Hsinchu 300, Taiwan
2. Department of Photonics, National Chiao-Tung University, Hsinchu 300, Taiwan
3. Department of Electrical Engineering, National Chiao-Tung University, Hsinchu 300, Taiwan.
** mcyuang@cs.nctu.edu.tw*

Abstract: This paper presents a fault-tolerance enhanced design and analyses for OPMDC[1]. Capacity-based analysis shows OPMDC accommodates 0.91-capacity under the worst-case single WSS-failure. Emulation analysis justifies that a 26dB-OSNR is retained under the worst-case three-WSS-failure scenario.

OCIS codes: (060.1155) All-optical networks; (060.4257) Network survivability; (060.4510) Optical communications.

1. Introduction

Data center networks (DCNs) [1,2] have been designed to provide an efficient and fault-tolerant infrastructure for supporting a wide variety of cloud and enterprise applications and services. Thanks to advances in silicon photonics and wavelength division multiplexing (WDM) technologies, optical DCNs [2] have been considered the most promising candidate for future DCNs, due to high bandwidth, high reliability, and low power consumption to name but a few. We have earlier proposed an optical pyramid data center network (OPMDC) [1] that has been built on three types of commercially available wavelength-selective-switch (WSS)-based WXC nodes in three tiers. The key component, i.e., the N×1 WSS, is based on the CoAdna's LighFlow™ digital LC platform [3], boasting crucial features such as low cost and high reliability.

Traditional network fault tolerance is a measure of the number of node/link failures the network can sustain before a disconnection occurs [4]. Such a disconnection-based measure, however, is infeasible to OPMDC and optical-switched DCNs owing to three major reasons. First, due to short distances within datacenters, failures in optical links are generally disregarded. Second, unlike E-switch-based nodes, each optical WXC node contains several active and passive devices that collaboratively support a number of parallel light paths. Failures in such a network node should be considered on each individual device rather than the entire node basis. Third, OPMDC possesses a pyramid-based topology that facilitates rich horizontal mesh connections. With the addition of the WSS that possess an exceedingly low failure-in-time (FIT) rate [3], the occurrence of disconnection is near unlikely. Therefore, instead of using the disconnection-based measure, in this work we adopt a new capacity-based measure for evaluating the fault tolerance of OPMDC under the occurrence of a handful of WSS failures, albeit with low probability. In Section 2, we propose a fault-tolerance enhanced design of the WXC nodes in OPMDC. In Section 3, we present the fault-tolerance analysis based on the capacity-based measure. Finally, we give an emulation analysis of OSNR under faults in Section 4.

2. Fault-Tolerant Enhanced Design of OPMDC

Since the key active component of WXC nodes is WSS, we consider the failures of WSS's in the fault-tolerant design and analyses. A full-scale OPMDC (see Fig.1) is built on three types of WSS-based optical switching nodes: (tier-1) ROADM, tier-2 WXC, and tier-3 WXC, which are recursively interconnected according to a pyramid

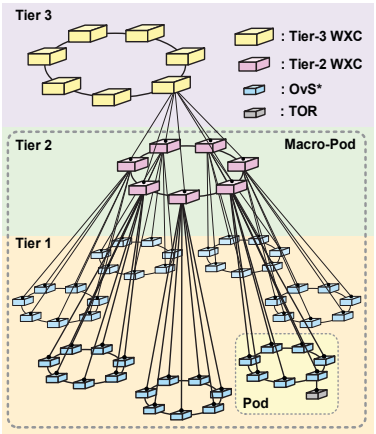

Fig. 1. OPMDC architecture (B=7).

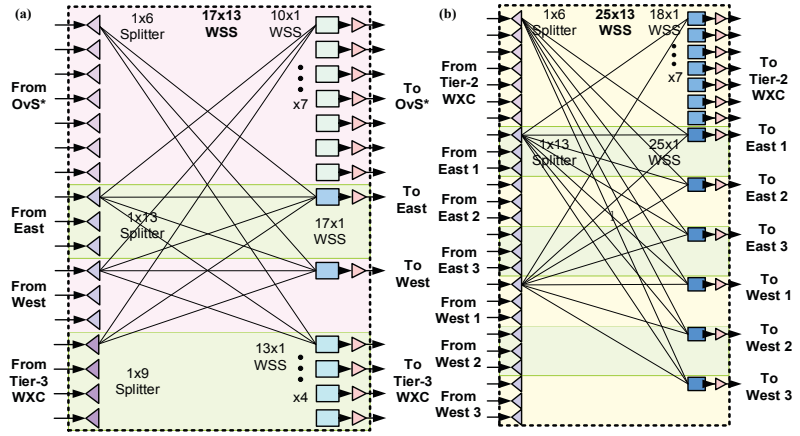

Fig. 2. Enhanced design: (a) 17×13 WSS in tier-2 WXC; and (b) 25×13 WSS in tier-3 WXC.

978-1-943580-07-1/16/$31.00 ©2016 Optical Society of America

structure parameterized by the number of nodes (B) that are mesh connected via ribbon fiber cables. By taking advantage of the pyramid-based topology, OPMDC employs a fairly simple and fixed routing under the normal operation. This can be illustrated via the following example. Traffic from a tier-1 ROADM to a tier-3 WXC is first routed to the immediate vertical tier-2 WXC through the northbound port of the ROADM, and then through the northbound WSS port of tier-2 WXC to the tier-3 WXC. Namely, such traffic will not travel along the mesh infrastructure in tier 2. However, when a failure, say the northbound WSS of tier-2 WXC fails, in the same example, traffic is re-routed to a neighboring tier-2 WXC node (through the tier-2 mesh connections), from which the northbound WSS port is taken instead.

Accordingly, fault-tolerant OPMDC is required to offer alternative routing. To achieve it, we simply enhance the internal design of the 17×13 WSS in tier-2 WXC, and 25×13 WSS in tier-3 WXC, as shown in Fig. 2. Notice that the changes of design are made only on the port sizes of splitters and WSS's, while the general interconnection structure of the WXC nodes remains the same [1].

3. Capacity-based Fault-Tolerant Analysis

In the analysis, we compute the maximum load, referred to as capacity, under which all traffic flows can be accommodated when WSS failures occur. For simplicity, we only consider a single WSS failure, whereas the analysis for multiple WSS failures can be similarly derived. There are seven possible locations of the occurrence of a WSS failure: tier-2 southbound (SB), northbound (NB), eastbound (EB), and westbound (WB), and tier-3 SB, EB, and WB. Due to traffic symmetry, the seven cases can be boiled down to four cases, as shown in Fig. 3. In the following, we only present the most complex one, namely a tier-3 EB/WB WSS failure. In principle, when a WSS fails, the affected traffic is distributed to all alternative routes subject to maximizing the throughput and minimizing the number of WSS's traversed. Accordingly in the analysis, we first determine the *candidate* routes, each of which includes the same minimal number of WSS's. For each candidate routes, we then compute its minimum residual capacity (the bottleneck edge). The amount of traffic is distributed to all candidate routes depending on the minimum residual capacity of the routes, subject to minimizing potential bottleneck and maximizing the throughput.

Suppose a tier-3 EB-WSS fails, affected traffic flows (i.e., to be routed through any of three EB tier-3 WXC nodes) are transported through two sets of candidate routes. Recall that there are four planes in tier-3. The first set of routes is from the northbound WSS of the source tier-2 WXC to any of the three functioning tier-3 WXC nodes. The second candidate route set transports traffic through other intact tier-3 WXC nodes using the horizontal mesh connection in tier 3. First, the total number of affected flows can be derived as $1/24\cdot\left[B^2LW\left(1-P\right)\right]$, where W is the number of wavelength channels, P the traffic locality probability, and L the normalized load. Now, consider the first candidate route set, the residual capacity of these candidate routes can be given as

$$r_{2n}=W\left[1-\frac{1}{4}BL\left(1-P\right)\right],\quad \text{and}\quad r_{3e}=r_{3w}=W\left[1-\frac{1}{24}B^2L\left(1-P\right)\right], \qquad (1)$$

where r_{2n}, r_{3e}, and r_{3w} denote the residual capacity of a tier-2 NB link, tier-3 EB link, and tier-3 WB link respectively. From Equ. (1), with $B=7$ and $P=1/2$, we have: $r_{2n}=3W\left[1-(7/8)\cdot L\right]$, and $r_{3e}=r_{3w}=9W\left[1-(49/48)\cdot L\right]$. Notice that these two functions intersect at $L=96/105$, resulting in different bottleneck locations under different loads (see Fig. 3). Specifically, if $L\le 96/105$, the bottleneck is on the tier-2 NB link, capable of carrying r_{2n} amount of flows; otherwise the bottleneck is on the tier-3 EB/WB link, capable of carrying r_{3e} amount of flows. Next, we examine the second candidate route set. Through simple derivation, the bottleneck is on the WB edge of the failed tier-3 WXC node. The residual capacity for the second route set that can accommodate affected traffic flows is $5W\left[1-(49/48)\cdot L\right]$. Subtracting the residual-capacity sum of both candidate route sets from the total number of affected flows, we attain the total amount of flows that cannot be accommodated due to the WSS failure. The results for all four types of failures are summerized in Fig. 3. From Fig. 3, we arrive at the worst-case capacity being equal

Failure	Total number of traffic flows that fail to be accommodated
Tier-2 SB-WSS	$W\left[(7/8)\cdot L-1\right]$
Tier-2 EB-WSS Tier-2 WB-WSS	$W\left[(21/4)\cdot L-5\right]$
Tier-2 NB-WSS Tier-3 SB-WSS	$W\left[(21/4)\cdot L-5\right]$
Tier-3 EB-WSS Tier-3 WB-WSS	$L\le 96/105$: $W\left[(35/4)\cdot L-8\right]$; $L>96/105$: $W\left[(245/16)\cdot L-14\right]$

Fig. 3. Capacity-based fault tolerance analysis for OPMDC.

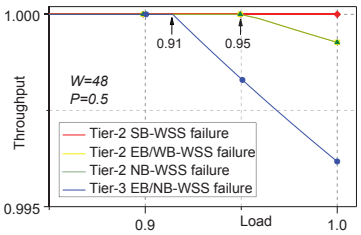

Fig. 4. Analytic results.

to 91% (the L value such that the un-accommodated traffic is zero) due to the worst-case single WSS failure. We further plot in Fig. 4 the throughput based on the results in Fig. 3, and observe that OPMDC retains exceedingly high throughput (0.996) under the worst-case single WSS failure.

4. Emulation Analysis of OSNR

We conduct an emulation analysis to examine the impact of WSS failures on OPMDC signal performance. Specifically, we study optical signal-to-noise ratio (OSNR) after traversing the alternative longest path while focusing on the accumulated noise effect due to cascaded EDFA's. First, the optical signal-to-noise ratio (OSNR) after passing an EDFA can be formulated [1] as:

$$OSNR_{out} = \left(\frac{1}{1/OSNR_{in} + \left[NF \cdot h \cdot \nu \cdot \Delta f \cdot (G-1) \right] / (P_{in} \cdot G)} \right), \qquad (2)$$

where $OSNR_{in}$ and P_{in} are the OSNR and optical power of the signal at the input the EDFA, NF, h, ν, and Δf are noise figure (NF), Plank constant, frequency of light, and the bandwidth that measures the NF (0.1nm), respectively, and G is the gain of the EDFA.

For the OPMDC system, the gain and NF of any EDFA in WXC are 21.5dB and 5dB, respectively. The insertion loss of MUX/DEMUX is 3dB. The insertion losses of the 7×1 WSS in tier-1 OvS, 10×1, 13×1, and 17×1 WSS's in tier-2 WXC, and 18×1, and 25×1 WSS's in tier-3 WXC are 6dB, 6dB, 6dB, 7dB, 7dB, and 8dB, respectively. The insertion losses of the 1×4 splitter in tier-1 OvS, the 1×8 and 1×16 splitters in tier-2 and tier-3 WXC's are 6dB, 9dB, and 12dB, respectively. The loss of each connector is 0.5dB. At tiers 2 and 3, the signal broadcasts horizontally, yielding the insertion losses at the first node (30% tapped), second node (70%×45% tapped) and third adjacent node (70%×55% tapped) being equal to 5.2dB, 5dB and 4.1dB, respectively. Assume that the optical power of the transmitter at OvS-1 is 0dBm, and the noise can be neglected at the input of OPMDC. According to Equ. (2), for any given path, the OSNR at the output of each node of the path can be calculated.

We consider one to three WSS failures occurring at different locations, therefore resulting in six different paths, as given in the legend of Fig. 5(b). Among six paths, we show in Fig. 5(a) the worst-case scenario, due to three WSS failures, in which the signal traverses the longest path that contains the highest number (=8) of WSS's/EDFAs. Notice that when there are more than three WSS failures (but less than numerous WSS failures such that any node along the path is completely disconnected from the system), the number of traversed WSS's/EDFAs remains the same. As shown in Figure 5(a), when one WSS failure occurs at 1* (the northbound path), the signal is first passed westbound to the adjacent tier-2 WXC, from which the northbound path is taken. For the six paths, the OSNR results at the output of traversed nodes are plotted in Fig. 5(b) and summarized in Fig. 6. The worst-case scenario under three WSS failures (after passing through 8 EDFA's) results in an OSNR of 25.61dB, and the lowest output power of around -10dBm at the destined OvS. With -10dBm received power and 26dB OSNR, we also show in Fig. 6 the experimental results of the eye diagrams for a 10-Gbps OOK signal under three different extinction ratios (ER) of a 10-GHz Mach-Zehnder Modulator.

5. References

[1] Maria Yuang, *et al.*, "OPMDC: Architecture Design and Implementation of a New Optical Pyramid Data Center Network," *Journal of Lightwave Technology*, vol. 33, no. 10, May 2015, pp. 2019-2031.

[2] C. Kachris, K. Kanonakis, and I. Tomkos, "Optical Interconnection Networks in Data Centers: Recent Trends and Future Challenges," *IEEE Communications Magazine*, vol. 51, no. 9, Sep. 2013, pp. 39-45.

[3] CoAdna, "50GHz Wavelength Selective Switch- High performance with integrated functionalities in a small footprint," http://www.coadna.com/2/ products.html#_top.

[4] W. Najjar, and J. Gaudiot, "Network Resilience: A Measure of Network Fault Tolerance," *IEEE Transaction on Computers*, vol. 39, no. 2, Feb. 1990, pp. 174-181.

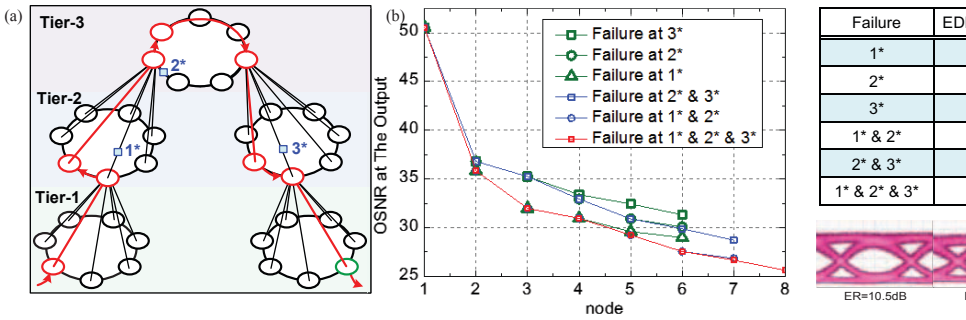

Fig. 5. (a) The worst-case longest path when three WSS failures occur; and (b) the corresponding OSNR results at each node.

Failure	EDFA number	OSNR (dB)
1*	6	28.98
2*	6	30.08
3*	6	31.32
1* & 2*	7	26.83
2* & 3*	7	28.72
1* & 2* & 3*	8	25.61

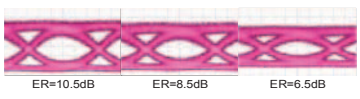

Fig. 6. OSNR and eye diagrams (26dB).