Chapter 10
OPMDC: Optical Pyramid Data Center Network

Maria Yuang and Po-Lung Tien

10.1 Introduction

Data center networks (DCNs) [1–3] have been designed and deployed to provide a reliable and efficient infrastructure for supporting a wide variety of emerging cloud and enterprise applications and services. Evidence shows that these applications and services not only involve much client-server (north-south) traffic flowing in and out of DCNs but also spawn a massive amount of east-west server-to-server traffic within DCNs. These applications and services are data rich by nature and demand high-bandwidth and low-latency transport of data. Besides, recent studies have further shown an ever-growing trend toward the variety and complexity of new cloud and enterprise applications and services. Such a trend places a higher demand for large-scale DCNs [4–6] that can deliver substantially high bandwidth, low latency, and reduced power consumption. These facts altogether bring about an urgent need for the design and implementation of next-generation DCN architectures and technology that can meet the demand.

There has been an increasing tendency toward a *modular* [3] and *incremental* [4] design for large-scale DCNs. A modular data center is constructed from purpose-engineered modules (e.g., pods, containers) that are flexibly expanded to the original data center infrastructure in an architecture compliant manner. The incremental design allows small rollouts and seamless expansion, resulting in agile and economical deployment and delivering resources on fully as-needed basis.

M. Yuang (✉)
Department of Computer Science and Information Engineering, National Chiao Tung University, 1001 University Road, Hsinchu 30050, Taiwan
e-mail: mariayuang@gmail.com; mcyuang@cs.nctu.edu.tw

P.-L. Tien
Department of Electrical and Computer Engineering, National Chiao Tung University, 1001 University Road, Hsinchu 30050, Taiwan
e-mail: polungtien@gmail.com

F. Testa, L. Pavesi (eds.), *Optical Switching in Next Generation Data Centers*,
DOI 10.1007/978-3-319-61052-8_10

Current state-of-the-art DCNs [2, 3] embrace optical transmissions but electrical switching of packets via electrical switches, such as top of rack (ToR), aggregation, and core switches. The electrical switches are interconnected based on two architecture designs: scale-up and scale-out. The scale-up approach uses a hierarchical tree structure in which the switches toward higher level of the hierarchy demand higher capacity and port count. On the other hand, the scale-out approach, aka the leaf-spine architecture, uses a large number of identical low-cost tier-1 ToR and tier-2 aggregation switches to deliver full bisectional bandwidth with extensive path diversity between servers. Both approaches have different pros and cons, but result in high power consumption [7] due to using power-hungry electrical-to-optical (E/O) and optical-to-electrical (O/E) transceivers. By and large, the electrical switching-based approaches have been deemed to be incapable of meeting the aforementioned DCN demands. This fact, coupled with recent advances in semiconductors and silicon photonics, becomes key driving forces for developing new optical architectures and technologies for next-generation DCNs.

Thanks to advances in silicon photonics and wavelength division multiplexing (WDM) technologies, optical WDM switching networks and systems have been proposed and widely deployed in long-haul and metro networks. Examples are optical wavelength cross-connects (OXCs/WXCs) [8] and reconfigurable optical add-drop multiplexers (ROADMs) [9, 10]. Optical WDM switching possesses some attributes, such as high bandwidth, low latency, and low power consumption, which are proved advantageous to future DCNs. A number of optical WDM DCN architectures that have been proposed [11–13] are based on various types of optical switching devices. Of these devices, the wavelength selective switch (WSS) has been considered the most promising candidate for building next-generation DCNs due to its flexible per-wavelength switching capability, besides being technologically mature and commercially available.

Being a key enabler for ROADMs, WSS is tailored to flexible per-wavelength provisioning. It is typical a $1 \times N$ optical switch that flexibly routes each wavelength from the input port to any of the N multiwavelength output ports, independent of how other wavelength channels are routed. WSS features [14] simple electronic control, low cost, high reliability (low FIT rate), and low power consumption (e.g., <2 W for a typical 1x9 WSS), but at the expense of a reconfiguration delay of a few milliseconds. Such a delay poses a challenge of supporting dynamic packet-based transport that is of crucial importance for future DCNs. The major contribution of our work lies in the design of a unique DCN architecture that operates in conjunction with SDN-based resource management, with the result that, despite the high reconfiguration delay limitation, the DCN efficiently achieves ultra-low-latency packet-based communications.

10.2 OPMDC Architecture

The architecture of a full-scale optical pyramid modular data center network (OPMDC) [4] is shown in Fig. 10.1. It consists of three types of WSS-based optical switching nodes in three tiers: (tier-1) ROADM, tier-2 WXC, and tier-3 WXC.

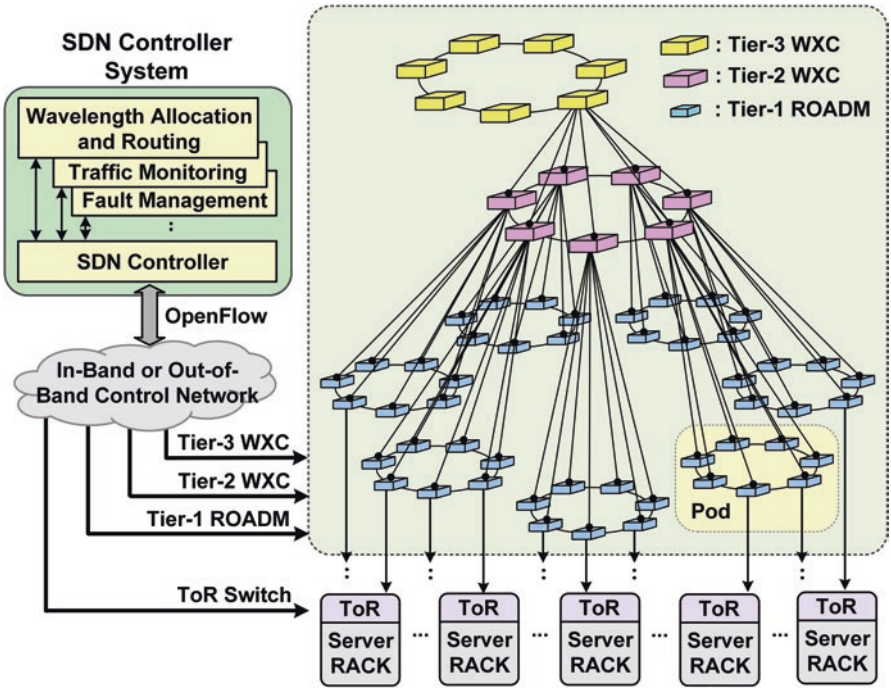

Fig. 10.1 The OPMDC architecture ($B = 7$)

While each ROADM node is directly connected to a ToR switch in tier 1, WXC nodes perform high-bandwidth optical per-wavelength switching in tiers 2 and 3. Further, OPMDC is controlled and managed by a software-defined networking (SDN) controller system in a centralized manner. The system consists of wavelength allocation and traffic engineering modules as well as an SDN controller. The controller governs the operation/configuration of OPMDC switching nodes (optical nodes and ToR switches) based on the OpenFlow protocol via an in-band or out-of-band control network. More details about the implementation of the SDN controller system are given in Sect. 10.4.

OPMDC is recursively built based on a pyramid construct that contains a polygonal base with an odd number (B) of nodes that are mesh connected (not a ring). In the example shown in Fig. 10.1, $B = 7$. The mesh connection is made via ribbon fiber cables, as will be described in detail later. Accordingly, two types of building blocks in OPMDC can be constructed incrementally: *pod* and *macro-pod*. A pod is the basic building block that spans tiers 1 and 2. It consists of B ROADM nodes at the base of its pyramid, each of which is down connected to a ToR switch and up connected to the apex (a tier-2 WXC node) of its pyramid.

A macro-pod is the larger building block that spans three tiers. It consists of B tier-2 WXC nodes (that are mesh connected), each of which is down connected to a pod and up connected to the apex of its pyramid in tier 3. For example, the OPMDC shown in Fig. 10.1 delineates a complete macro-pod that contains B (=7) pods, or B^2

(=49) ROADM nodes. Further, a full-scale OPMDC contains B macro-pods that are connected through B tier-3 WXC nodes that are also mesh connected. As described, these building blocks can be deployed on an incremental basis. For example, to interconnect only a total of $3B$ server racks, OPMDC will contain three pyramids each of which has B ROADMs at the base (tier 1) and one tier-2 WXC at the apex, while the three tier-2 WXCs are mesh connected.

OPMDC boasts four unique features that are deemed crucially significant to support emerging cloud applications. *First*, due to the pyramid topology and horizontal mesh interconnection, OPMDC offers powerful broadcast capability without overloading the network. *Second*, OPMDC allows extensive wavelength reuse. For example, the same set of wavelengths can be reused for transporting traffic within different pods. Such a feature enables OPMDC to employ highly efficient static preallocation of wavelengths, thereby accomplishing ultra-low-latency packet-based transport under a substantial portion of traffic patterns. This will further be described in Sect. 10.3. *Third*, OPMDC is highly fault tolerant [15]. Due to short distances within data centers, failures in optical links are generally disregarded. In addition, as was mentioned earlier, the key device in ROADM and WXC is WSS, which possesses an exceedingly low failure-in-time (FIT) rate [14]. Unlike E-switch-based nodes, each optical ROADM or WXC node contains individual active and passive devices that collaboratively support a number of parallel light paths. Any failure in a node occurs only on the basis of an individual device (also with low probability) rather than the entire node. Additionally, with the rich horizontal mesh connectivity, the occurrence of a few failures results in only minor throughput degradation instead of node disconnections from the rest of the network.

Finally, the pyramid topology allows OPMDC to adopt fairly simple routing under both normal and fault conditions. Under the normal condition, traffic from one ROADM to another ROADM within the same pod is routed through the mesh connection of the pod. For the traffic within a macro-pod but without a pod, packets are passed from the source ROADM to its tier-2 WXC, then through the mesh connection to the destination pod's tier-2 WXC, and finally down to the destination ROADM. By the same token, inter-macro-pod traffic is routed through two tier-3 WXC nodes. Under the condition of a fault in a pyramid, alternative routes can simply be taken through other available horizontal mesh connections or the apex of the pyramid.

10.2.1 Internal Design of ROADM and WXC Nodes

The tier-1 ROADM and tier-2/3 WXC nodes have been designed in such a way that they can be implemented using commercially available components that have already been widely deployed. Their key component is the $N \times 1$ WSS module. Its distinctive features, including low cost, high port count, low power consumption, and high reliability, are ideally suited for data center switching.

10.2.1.1 Tier-1 ROADM Node

The tier-1 ROADM node was originally developed by CoAdna Photonics [16] and then revised to tailor for the OPMDC project. As shown in Fig. 10.2a for $B = 7$, each ROADM node contains an optical multiplexer (MUX) and demultiplexer (DEMUX), a $B \times 1$ WSS, an erbium-doped fiber amplifier (EDFA), ribbon cables, and a series of passive splitters (a 3-way splitter and a number of tap couplers). Each ROADM is horizontally connected to $(B-1)/2$ peer ROADM nodes in the east via $(B-1)/2$ pairs of fibers, and likewise for the west.

For $B = 7$ as shown in Fig. 10.2, among the 3 pairs of fibers for either direction, 2 fiber pairs are for pass-through traffic and 1 fiber pair for add and drop of traffic to the local node. While travelling along the fibers, packets are tapped into the WSS of each of the three ROADM nodes through tap couplers for broadcasting.

Each ROADM node is directly connected to a ToR switch. There are W uplink ports in the ToR switch that are populated with W wavelength-specific DWDM transceivers, respectively, where W is the total number of wavelength channels. For transmissions, the W channels of optical signals are combined via MUX and passed to EDFA that boosts the peak signal power to ensure sufficient power budget. With a 1×3 splitter, the multiplexed traffic is 3-way broadcast to east and west ROADM nodes and the tier-2 WXC node. For receiving, a $B \times 1$ WSS is used to select W signals from the B input ports of WSS (i.e., $(B-1)/2$ ports from the east, $(B-1)/2$ ports from the west, and 1 from the tier-2 WXC). After DEMUX, W channel signals are passed to the corresponding ports of the ToR switch.

Figure 10.2b depicts how the B (=7) nodes of a pod are horizontally interconnected. Notice that, for any traffic within a pod, packets from the same source node to different destination nodes share the same fiber link. Thus, it requires distinct wavelengths to carry traffic to different destination nodes. On the other hand,

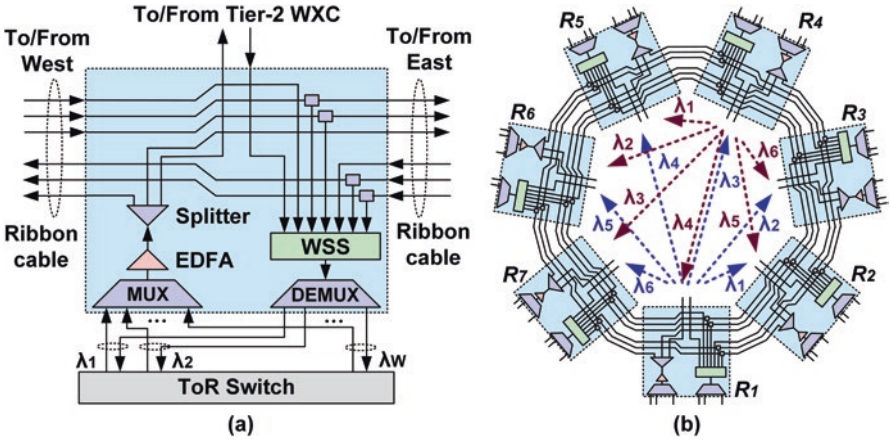

Fig. 10.2 (**a**) Design block diagram of tier-1 ROADM node ($B = 7$), (**b**) Horizontal interconnection of ROADM nodes in a pod

packets from different source nodes are carried by different fiber links, thus causing no wavelength contention. For example, node R_1 can send packets to node R_3 via a wavelength, while R_2 can send packets to node R_4 via the same wavelength without contention. As a result, as shown in Fig. 10.2b, both nodes R_1 and R_4 can send packets independently to six other nodes via λ_1 to λ_6. Thus, it takes a total of six (i.e., $B-1$) wavelengths to facilitate all-to-all independent communications within a pod. Importantly, such wavelength reuse can be applied to all other pods. Namely, the same six wavelengths can be fully and independently reused within the B^2 pods to provide parallel intra-pod transport.

10.2.1.2 Tier-2 WXC Node

The tier-2 switching is performed via a 4-way WXC node, as shown in Fig. 10.3, for the case of $B = 7$. As shown in Fig. 10.3a, the WXC node is south connected to seven ROADM nodes of its pod and north connected to a tier-3 WXC node. The node is also connected to 3 peer east/west WXC nodes via 3 pairs of fibers, in which 2 pairs are for pass-through traffic and 1 pair for traffic being switched to other ports. The key switching element of the WXC node is the 17 × 13 WSS module that can be implemented via commercially available WSS devices, ranging from size 10 × 1 to 17 × 1 (see Fig. 10.3b).

Notice that there are four pairs of parallel fibers for the northbound transport. This is because each edge between the tier-2 and tier-3 WXC node requires a capacity of nearly 4 W to assure the DCN of being congestion free (described in Sect. 10.2.2). Accordingly, for $B = 7$, a WXC node is equipped with a WSS that has 17 input ports (3 + 3 + 7 + 4) and 13 output ports (two east and two west ports are for pass-through traffic only). For the ease of illustration, we delineate in the figure

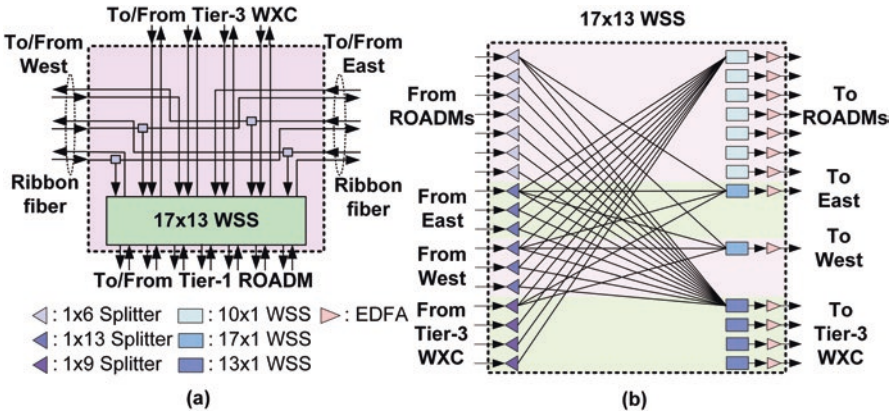

Fig. 10.3 (**a**) Design block diagram of tier-2 WXC node, (**b**) Design block diagram of WSS in tier-2 WXC

the exact number of ports of each splitter and WSS while omitting several direct connections between some splitters to WSSs. For example, it requires a 17 × 1 WSS for switching traffic "to east" from any of the 17 input ports.

10.2.1.3 Tier-3 WXC Node

The tier-3 switching for traffic that crosses macro-pods is performed via a 3-way (east, west, and south) WXC node. Since its overall structure is similar to that of a tier-2 WXC node, the designed block diagram is omitted here, and interested readers can refer to [4]. It is worth mentioning that there is a new feature that has been designed at tier 3 for achieving better scalability and fault tolerance purposes. Specifically, recall that each of 7 tier-2 WXCs is connected to a tier-3 WXC via 4 pairs of fiber links. Rather than feeding all 7 × 4 = 28 links from the tier-2 pyramid base into one tier-3 WXC node, the tier-3 WXC node is functionally divided into four smaller-size independent WXC nodes, to which each pair of fiber links is connected. As a result of the division, there are four switching planes at tier 3 (each of which consists of 7 smaller-size WXC nodes) that operate in parallel, thereby offering higher fault tolerance. Further, each main WSS module at each WXC node is reduced in port size to 25 × 13, achieving better scalability.

10.2.2 Edge Capacity and Structure

In this subsection, we are to answer the next design question: how many fiber links are required between any two adjacent optical switching nodes? First, let the *edge* between two adjacent nodes be defined as the inclusion of all parallel fiber links connecting the two nodes. Let W denote the total number of wavelength channels on each fiber link. The *edge capacity* of an edge between two adjacent nodes is defined as the total number of required wavelengths, satisfying an oversubscription ratio of one (i.e., the total output link rate of the first switching node is equal to its total input link rate). Here, we first derive the edge capacities, followed by determining the edge structure, i.e., the number of parallel fiber links on each edge.

It is clear that the determination of all edge capacities depends on the traffic distribution within and outside of pods and macro-pods. Let P_{TL} denote the *traffic locality probability*, which defines the traffic distribution within and outside of a module- pod and macro-pod alike. Specifically, given a source (ROADM) node of a flow that belongs to a macro-pod, P_{TL} is defined as the probability that its destination node falls within the same macro-pod, and $1 - P_{TL}$ the probability outside of the macro-pod. Further, conditional to a given macro-pod, given a source node in a pod within the macro-pod, P_{TL} is defined as the probability that the destination node falls within the same pod, and $1 - P_{TL}$ the probability outside of the pod but within the same macro-pod. Accordingly, a flow is destined to a node within the same pod with

probability, $(P_{TL})^2$; is destined to a node in a different pod but within the same macro-pod with probability, $(1 - P_{TL})P_{TL}$; and is destined to a node in a different macro-pod with probability, $1 - P_{TL}$. Once the locality is determined, the destinations are assumed uniformly distributed. Notice that it is highly expected that the normalized traffic destined to any node within the pod/macro-pod is greater than any node outside of it. For the OPMDC prototyping system, we use $B = 7$, and its edge capacity and structure design is based on a modest locality probability, $P_{TL} = 0.5$.

Let $C_H(B, T)$ denote the edge capacity between two adjacent horizontal nodes in tier T, where B is the number of base nodes in a pyramid. Let $C_V(B, T)$ denote the edge capacity between two adjacent vertical nodes at tiers T and $T + 1$, respectively. First, we are to compute $C_H(B, T)$, where $T = 1$, 2, and 3, and $C_V(B, T)$, where $T = 1$ and 2.

Assume that each traffic flow requires a bandwidth of one wavelength channel, and the total number of flows emitting from any ToR switch (or ROADM) is W. Let $F(B, k)$ denote the mean total number of flows from source ROADM node s to destination node d, where $k = 1$, 2, 3 correspond to three cases for the locality of nodes s and d, as stated in Eq. (10.1). For any given source node (s) in a pod, for case I ($k = 1$), there are $B - 1$ possible destination nodes (d) in the same pod; for case II ($k = 2$), there are $B(B - 1)$ destination nodes in different pods of the same macro-pod; and for case III ($k = 3$), there are $B^2(B - 1)$ nodes in different macro-pods. So, $F(B, k)$ can be given as

$$F(B,k) = \begin{cases} \dfrac{W \cdot (P_{TL})^2}{B-1}, & k = 1,\ if\ s, d \in same\,\text{pod}; \\ \dfrac{W \cdot P_{TL} \cdot (1-P_{TL})}{B(B-1)}, & k = 2,\ if\ s, d \notin same\,\text{pod, and } s, \\ & \quad d \in same\,macro-\text{pod}; \\ \dfrac{W \cdot (1-P_{TL})}{B^2(B-1)}, & k = 3,\ if\ s, d \notin same\,macro-\text{pod}. \end{cases} \tag{10.1}$$

Recall that all traffic emitted from a ROADM node is broadcast via a 1×3 splitter (see Fig. 10.2a) to horizontal ROADM nodes and tier-2 WXC node. Thus, we can directly get $C_H(B, 1) = W$ and $C_V(B, 1) = W$. Next, the capacity $C_H(B, 2)$ needs to accommodate all the traffic from one whole pod to its adjacent pod. Moreover, the capacity is shared for transporting traffic from a tier-2 WXC to $(B-1)/2$ peer WXC nodes. Accordingly, we have

$$C_H(B,2) = F(B,2) \cdot B^2 \cdot \frac{B-1}{2} = \frac{B \cdot W \cdot P_{TL} \cdot (1-P_{TL})}{2} \tag{10.2}$$

By the same token, $C_H(B, 3)$ at tier 3 needs to accommodate all the traffic from one whole macro-pod to its adjacent macro-pod, and the capacity is also shared for transmitting $(B-1)/2$ sets of traffic. We thus get

$$C_H(B,3) = F(B,3) \cdot B^2 \cdot B^2 \cdot \frac{B-1}{2} = \frac{B^2 \cdot W \cdot (1-P_{TL})}{2} \tag{10.3}$$

Finally, $C_V(B, 2)$ for a tier-2 WXC is to interconnect its entire pod to any node outside of the macro-pod it belongs to, with a probability of $1 - P_{TL}$. This implies

$$C_V(B,2) = B \cdot W \cdot (1-P_{TL}) \tag{10.4}$$

Given $P_{TL} = 0.5$ and $B = 7$, we get from Eqs. (10.2, 10.3, and 10.4) that $C_H(7, 2) = 0.875W$, $C_V(7, 2) = 3.5W$, and $C_H(7, 3) = 12.25W$ in OPMDC. This explains the requirements of 4 pairs of parallel fiber links to connect a tier-2 WXC to its tier-3 WXC and 12 pairs of parallel fiber connections to connect two adjacent tier-3 WXC nodes before the division into four switching planes is applied.

10.3 Wavelength Allocation Strategies

OPMDC strives for flexible optical packet-based and circuit-based transport based on three innovative wavelength allocation strategies. They are (1) static pre-allocation, (2) relay-based allocation, and (3) dynamic allocation. While strategies 1 and 2 require no further WSS reconfiguration and thus are best suited for supporting packet-based transport, strategy 3 provides efficient dynamic establishment of new optical paths for circuit-based transport. They are described in the following.

10.3.1 Strategy 1: Static Wavelength Pre-allocation

The first strategy caters to all intra-pod and intra-macro-pod transport, based on static wavelength pre-allocation. Significantly, as will be shown, the static pre-allocation allows all intra-pod and intra-macro-pod transport to be facilitated fully in parallel using only a total of $2(B-1)$ wavelengths. Due to the avoidance of WSS reconfiguration, this strategy is capable of meeting the demand of ultra-low latency for supporting packet-based transport.

In the case of intra-pod communications, as was illustrated in Fig. 10.2b, a small fixed number ($= B-1$) of wavelengths can be pre-allocated and reused to simultaneously support all intra-pod transport for all B^2 pods in OPMDC. Due to incurring no WSS reconfiguration and any other delays, this class of the packet-based transport receives near-zero latency.

For intra-macro-pod communications, since the horizontal mesh connection at tier 2 is the same as that at tier 1, the same wavelength reuse principle (illustrated in Fig. 10.2b) can be applied to tier 2. That is, for OPMDC with $B = 7$, it takes only a total of six (i.e., $B-1$) wavelengths to facilitate all pod-to-pod communications

independently within any macro-pod in OPMDC. For example, assume a wavelength is designated for the communications from s-pod to d-pod. Since there are seven ROADM nodes in any pod, the designated wavelength can be used to connect one random ROADM node (say R_1) in s-pod to another random ROADM node (say R_3) in d-pod. Then, all other communications between two different pairs of ROADM nodes from s-pod to d-pod can be established by means of packet relays at nodes R_1 and R_3 (and their ToR switches). Such a relay operation is employed in Strategy 2, which is described in detail in the next subsection.

Further, such wavelength reuse can be applied to all other macro-pods. Therefore, it takes only a total of $B-1$ wavelengths to establish all intra-macro-pod packet-based transport fully in parallel for all B macro-pods in OPMDC. Without any WSS reconfiguration, this class of packet-based transport experiences an ultra-low delay resulting from two (or less) relays at the ToR E-switches (described next).

10.3.2 Strategy 2: Relay-Based Wavelength Allocation

The relay-based wavelength allocation aims to facilitate intra-macro-pod and inter-macro-pod transport using existing optical paths between the source and destination pods located at different pods and macro-pods, respectively. Notice that such existing optical paths for intra-macro-pod transport are to be statically pre-allocated based on Strategy 1. Basically, this relay-based strategy employs a combined *relay* and *aggregation* operation at the source and/or destination pods, referred to as SDRA. The SDRA mechanism allows new flows to be transited using existing optical paths by means of flow relay and aggregation through the horizontal mesh connections in tier 1 at the source and/or destination pods.

The SDRA mechanism can be best explained via an example illustrated in Fig. 10.4. Suppose there is a new traffic flow from node S in s-pod to node D in

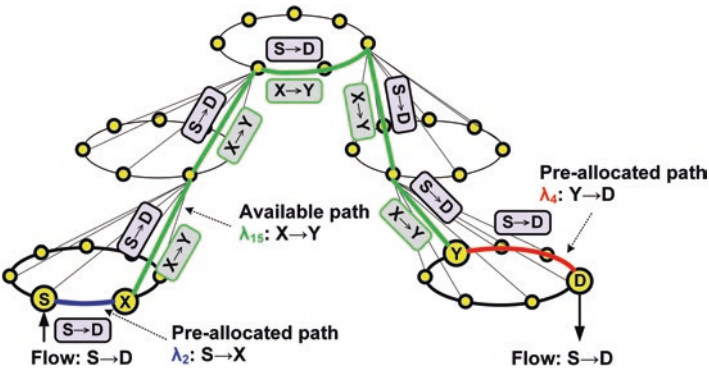

Fig. 10.4 Relay-based wavelength allocation for low-latency packet-based transport

d-pod, and there exists an available optical path, λ_{15}, for the traffic flow between nodes X and Y in s-pod and d-pod, respectively. S-to-D packets are first sent to X through the horizontal pre-allocated path (λ_2) and then relayed at X's ToR switch. These packets are optical-to-electrical (OE) converted, aggregated with X-to-Y packets, and together delivered through available optical path λ_{15}. Upon having arrived at node Y in d-pod, S-to-D packets are OE-converted again and transported together with Y-to-D packets through horizontal pre-allocated path λ_4. The S-to-D packets are finally dropped after reaching node D. Notice that if node S = X, the relay/aggregation operation is not invoked in the source pod. Likewise, if node D = Y, the operation is avoided in the destination pod.

The SDRA mechanism results in high utilization due to packet aggregation. In addition, the mechanism offers low-latency transport owing to the avoidance of WSS reconfiguration. The price paid is only no more than two additional hops of E-switch processing delay. Simulation results [4] show that employing the SDRA mechanism yields a substantial throughput improvement from 42.5 to 87.9%, due to taking advantage of available optical paths.

10.3.3 Strategy 3: Dynamic Wavelength Allocation

The third strategy aims to establish new optical paths for inter-macro-pod transport, based on dynamic wavelength allocation. Due to the need for WSS reconfiguration that causes a few millisecond delay, this strategy is best suited for supporting circuit-based transport.

The dynamic wavelength allocation problem can be formally defined as: given a set of available wavelengths and circuit flows to be served, the problem is to assign the wavelengths to a maximum number of circuit flows, subject to being free from wavelength contention at any fiber link. To maximize the throughput, the wavelength allocation problem boils down to the proper determination of the order by which the circuit flows are assigned wavelengths. Specifically, the flow that contends with a higher number of flows should be served first in order to reduce the contention probability. Accordingly, we have proposed a heuristic algorithm, called the most contentious first (MCF) [4]. The algorithm first ranks each flow according to the total number of all other flows in its *most congested* link of the path the flow travels through. The MCF algorithm then assigns wavelengths to the flows sequentially in descending order of the flow ranks.

The performance of the MCF algorithm was evaluated via experimental testbed results (see Sect. 10.4) as well as simulation results [4]. In particular, the simulation results show that OPMDC achieves 95.8% throughput under $P_{TL} = 0.5$ and 80% throughput even under poor traffic locality, $P_{TL} = 0.3$.

10.4 Prototype and Performance Assessment

In this section, we present our OPMDC prototype system and give an assessment of its performance with respect to scalability, power consumption, wiring complexity, fault tolerance, and mean packet latency.

10.4.1 OPMDC Prototyping System

We have built a prototyping system of OPMDC with $B = 7$ using seven ROADM nodes, in which the WSS module is based on the CoAdna's LightFlow™ digital LC platform [14]. Each ROADM node is directly connected to a Pica8 P-3295 ToR switch that is compliant with the OpenFlow software-defined networking (SDN) [17] standard interface. Each ToR switch provides 48 10-Gb/s ports. Among them, 16 ports are populated with 16 10-Gb/s DWDM transceivers associated with 16 wavelength channels, respectively. We have implemented an OpenFlow-1.3 [18] SDN controller based on a Ryu 3.9 open source system. The SDN controller system runs the wavelength allocation algorithms and, in turn, performs the reconfiguration and control of ROADM nodes and ToR SDN switches.

Each ROADM node is equipped with a Raspberry Pi [19] embedded firmware system that operates with a ROADM controller, developed under a Debian 7.2 Linux kernel-based operating system. The controller is primarily responsible for the real-time reconfiguring and monitoring and periodic reporting of the health of the hardware devices, e.g., WSS and EDFA. Specifically, the controller performs the configuration of its optical WSS whenever it receives a control message from the SDN controller after making new wavelength allocation decisions. The control message includes information such as wavelength channel IDs and WSS ports. For the time being, the communication between the SDN controller and ROADM controller is facilitated through a TCP socket interface.

10.4.2 Performance Assessment

An overall performance assessment of OPMDC is made in the following:

Scalability: OPMDC supports up to B^3 ROADM nodes. Each ROADM provides dynamic switching of up to 96 (W) wavelengths under the 50-GHz channel spacing. Each wavelength channel can currently accommodate a capacity (C) of 100 Gb/s. Thus, OPMDC can support a maximal capacity of B^3WC, which is 3.3 Pb/s with $B = 7$, or 7 Pb/s with $B = 9$. As such, OPMDC provides high and scalable bandwidth for both cloud and enterprise DCNs.

Power Consumption: Power consumption is often estimated on a per port basis. For E-switch-based DCNs, ARISTA has claimed to provide industry leading power

efficiency, achieving a typical power consumption of 5 watts per 40-GbE port in its 7050X Series of products. For the OPMDC that supports 48 wavelengths under the 100-GHz channel spacing, the typical power consumptions of a ROADM and a tier-2/tier-3 WXC node, dominated by the EDFA module, are around 15 watts and 12 watts, respectively. With $B = 7$, there are $7^3 = 343$ ROADM nodes, 49 tier-2 WXC, and $7 \times 4 = 28$ tier-3 WXC nodes; and each WXC node is populated with 13 pairs of WSS/EDFA modules. The total power consumption becomes $(343 \times 15w) + (49 \times 13 \times 12w) + (28 \times 13 \times 12w) = 17,157$ watts. For OPMDC with 343 48-port ROADM nodes, i.e., 16,464 ports altogether, the typical power consumption is around 1 watt per port. Thus, OPMDC currently achieves five times as power efficient as E-switch DCNs. In our future work, we aim at designing an optical DCN architecture using fewer EDFAs, for achieving more power saving.

Wiring Complexity: The wiring complexity and overall cost of DCNs are directly proportional to the total number of distinctive optical fibers interconnecting all switching nodes. We now draw a comparison of the total number of fibers required between OPMDC and a typical DCN, populated with the same port number. We consider $B = 7$ in this comparison. Recall that OPMDC supports $7^3 = 343$ ROADM nodes, and each ROADM connects to the 6 40-Gb/s ports of a ToR switch. To support the same number of 40-Gb/s ports, i.e., $343 \times 6 = 2058$ ports, a typical leaf-and-spine DCN demands 2058 fibers altogether. Due to the use of ribbon fibers for all the horizontal connections, for example, each tier-1 pyramid pod needs 7 horizontal and 7 vertical fibers. Therefore, OPMDC requires a total of 959 fibers, including $(7 + 7) \times 49$ fibers for 49 tier-1 pods, $[7 + (4 \times 7)] \times 7$ fibers for 7 tier-2 macro-pods, and 7×4 fibers for 4 planes of tier-3 nodes. Compared to a typical spine-and-leaf DCN, OPMDC achieves a reduction of more than half in wiring complexity.

Fault Tolerance: Existing DCNs achieve fault tolerance and high availability by means of switch redundancy and path diversity. However, the price paid is high wiring complexity and poor resource efficiencies. For designing large-scale but manageable cost DCNs, we incorporate additional resiliency means. OPMDC achieves fault tolerance especially for two lower tiers solely by using highly reliable WSS-based switching modules and periodic monitoring and fast recovery of hardware devices. For tier 3, as explained in Sect. 10.2.1, OPMDC additionally employs the segregation of the tier-3 WXC backbone into four identical but scale-down switching planes. The division results in path diversity, thus offering an additional level of reliability.

Cost: Due to not yet being prevalent on the market, the costs of the major components (purchased in small quantity) are currently high. For example, a 1×4 WSS costs around US$3000, and a DWDM 10G transceiver with a channel spacing of 100 GHz costs around a range of US$300–$400. Nevertheless, the costs can be greatly reduced if the optical components are produced in large quantity. Furthermore, with the advances in emerging photonic integrated circuit (PIC) technology, these optical switching nodes can be PIC-based designed and implemented, resulting in significant cost reduction.

10.4.3 Packet Latency Performance

We take the direct measurements of packet latency from experiments on the prototype system. In the experiments, we send a large video file via python-based TCP socket, from the source server (S) (socket client) to the destination server (D) (socket server). The buffer size of the socket is set to be 60 Kbytes. To measure the latency for both circuit- and packet-based transport, we adopt two types of flows: c-flow (2000 S-to-D transmissions of 60-Kbyte packets) and p-flow (50 S-to-D transmissions of 60-Kbyte packets). In particular, the c-flow requires the setup of a new optical path based on the MCF algorithm (allocation Strategy 3). The p-flow is transported via existing optical paths based on the SDRA algorithm (Strategy 2). Moreover, to measure the one-way S-to-D packet delay, we implement a simple ACK program at the D-node, sending an acknowledgment packet to the S-node as soon as a 60-Kbyte packet has been received from its TCP socket. The round-trip time is measured at the S-node upon having received the ACK packet. Since the ACK packet undertakes the same overhead as that of its video packet, the S-to-D packet latency is calculated as half of the round-trip time. The measurement of mean latency performance is summarized in Table 10.1.

For c-flow, the ROADM reconfiguration delay consists of two parts: firmware response time and WSS response time. The firmware response time, including Ethernet delay, is 10 ~ 12 ms. The single-channel (sc) and multichannel (mc) WSS response time, including WSS switching time and universal asynchronous receiver/transmitter (UART) delay, are 3 ms and 35 ms, respectively. Therefore, the total reconfiguration delay is 13 ~ 15 ms for a single wavelength and 45 ~ 47 ms for

Table 10.1 Mean packet latency performance of the OPMDC prototyping system

Task	Description	Time
c-flow: Circuit-based transport (requiring new optical paths using Strategy 3)		
A. ROADM reconfiguration	Firmware response time (12 ms)	15 ms(sc); 47 ms(mc)
	WSS response time (single-channel/multichannel) (3 ms/35 ms)	
B. Run time	MCF algorithm (48 × 7 flows)	0.184 ms
C. ToR switch configuration	Packet-in/packet-out (max. two times)	6.47 ms
	Flow entry setup in 2 switches	
D. Packet transmission	Packet size = 60 Kbytes	2410 ms
	Flow duration = 2000 packets	
Mean S-to-D packet latency = (A + B + C + D)/2000		1.232 ms
p-flow: Packet-based transport (using existing optical paths based on Strategy 2)		
E. Run time	SDRA algorithm (48 × 7 flows)	0.07 ms
F. ToR switch configuration	Packet-in/packet-out (max. two times)	6.54 ms
	Flow entry setup in 4 switches	
G. Packet transmission	Packet size = 60 Kbytes	79.2 ms
	Flow duration = 50 packets	
Mean S-to-D packet latency = (E + F + G)/50		1.716 ms

multiple wavelengths. For p-flow, the configuration of four SDN-enabled ToR switches takes place in parallel, with the result that it takes no more than two times of packet-in/packet-out as that for transporting c-flow. With the SDRA mechanism employed, the bottleneck of latency lies in populating the flow tables in the ToR switches.

10.5 Conclusions and Research Directions

In this chapter, we have presented a novel optical pyramid DCN architecture and its prototype, OPMDC, including the design and implementation of its three types of WSS-based optical switching nodes. After introducing a traffic locality parameter, P_{TL}, we derived the edge capacities and determined the edge structure that satisfies the need of OPMDC for being bottleneck free. Owing to the pyramid architecture, OPMDC enables powerful wavelength reuse and broadcast capability. Specifically, we proposed three wavelength allocation strategies, boasting flexible ultra-low-latency optical packet-based transport and high-throughput circuit-based transport. We demonstrated experimental testbed results to justify that OPMDC achieves high and scalable bandwidth, low latency, high fault tolerance, and reduced power consumption and wiring complexity.

We are currently undertaking a number of research work related to OPMDC. First, recall that one of the distinguishing features of OPMDC is its modular and incremental design of the architecture. As a result, such a design makes OPMDC highly flexible to serve the needs for providing different-scale data centers at different locations of networks. Thus, one of our current research tasks is to design micro/mini-data centers that are constructed based on a few pods or macro-pods of OPMDC. These smaller-scale data centers are targeted at facilitating near-zero-latency mobile computing at the edge of 5G mobile networks. The second research task is to design broadcast-based traffic control mechanisms that cater for supporting parallel-processing cloud applications, such as Big Data computing. Last but not least, we have continually been refining the OPMDC architecture and the internal designs of optical switching nodes in an effort to reduce the number of power-hungry EDFA devices for achieving greater power efficiency.

References

1. N. Bitar, S. Gringeri, and T. Xia, "Technologies and Protocols for Data Center and Cloud Networking," *IEEE Commun. Mag.*, vol. 51, no. 9, Sep. 2013, pp. 24–31
2. Y. Liu, J. Muppala, M. Veeraraghavan, and D. Lin, *Data Center Networks: Topologies* (Springer, Architectures and Fault-Tolerance Characteristics, 2013)
3. C. Kachris, K. Bergman, and I. Tomkos, *Optical Interconnects for Future Data Center Networks* (Springer, 2013)

4. M. Yuang, P. Tien, H. Chen, W. Ruan, S. Zhong, J. Zhu, Y. Chen, and J. Chen, "OPMDC: Architecture Design and Implementation of a New Optical Pyramid Data Center Network," *IEEE/OSA Journal of Lightwave Technology*, vol. 33, no. 10, May 2015, pp. 2019–2031
5. Z. Li, Z. Guo, and Y. Yang, "BCCC: An Expandable Network for Data Centers," *IEEE/ACM Trans. Networking*, vol. 24, no. 6, Dec. 2016, pp. 3740–3755
6. N. Han, Y. Chung, and M. Jo, "Green Data Centers for Cloud-Assisted Mobile Ad Hoc Networks in 5G," *IEEE Network*, vol. 29, no. 2, March/April 2015, pp. 70–76
7. C. Kachris, K. Kanonakis, and I. Tomkos, "Optical Interconnection Networks in Data Centers: Recent Trends and Future Challenges," *IEEE Commun. Mag.* **51**(9), 39–45 (Sep. 2013)
8. T. Ban, H. Hasegawa, K. Sato, T. Watanabe, and H. Takahashi, "A Novel Large-scale OXC Architecture and an Experimental System that Utilizes Wavelength Path Switching and Fiber Selection," *Opt. Express* **21**(1), 469–477 (Jan. 2013)
9. J. Homa and K. Bala, "ROADM Architectures and Their Enabling WSS Technology," *IEEE Commun. Mag*. **46**(7), 150–154 (June 2008)
10. Y. Li, Li Gao, G. Shen, and L. Peng, "Impact of ROADM Colorless, Directionless, and Contentionless (CDC) Features on Optical Network Performance," *IEEE/OSA J. Optical Communications and Networking*, vol. 4, no. 11, Nov. 2012, pp. 58–67
11. C. Kachris and I. Tomkos, "A Survey on Optical Interconnects for Data Centers," *IEEE Communications Surveys and Tutorials* **14**(4), 1021–1036 (2012)
12. Jordi Perello, et al., "All-Optical Packet/Circuit Switching-Based Data Center Network for Enhanced Scalability, Latency, and Throughput," *IEEE Network*, vol. 27, no. 6, Nov./Dec. 2013, pp. 14–22
13. K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen, and Y. Chen, "OSA: An Optical Switching Architecture for Data Center Networks With Unprecedented Flexibility," *IEEE/ACM Trans. Networking* **22**(2), 498–511 (April 2014)
14. CoAdna, "50GHz Wavelength Selective Switch- High performance with integrated functionalities in a small footprint," http://www.coadna.com/2/products.html#_top
15. M. Yuang, J. Yang, H. Chen, and P. Tien, "Fault-Tolerance Enhanced Design and Analyses for Optical Pyramid Data Center Network (OPMDC)," *Optical Fiber Communication (OFC) Conference*, 2016
16. S. Zhong and Z. Zhu, "Distributed Optical Switching Architecture for Internal Data Center Networking," USA Provisional Patent, OMB 0651–0032, Jan. 2014
17. P. Goransson and C. Black, *Software Defined Networks: A Comprehensive Approach* (Morgan Kaufmann, San Francisco, 2014)
18. OpenFlow Switch Consortium and Others. OpenFlow Switch Specification Version 1.4.0. 2013. https://www.opennetworking.org/images/stories/downloads/sdn-resources/onf-specifications/openflow/openflow-spec-v1.4.0.pdf, Jan. 2014
19. Raspberry Pi. Available online: http://www.raspberrypi.org