Wavelength Reuse for Scalable Multicasting: A Cross-Layer Perspective

Wavelength Reuse for Scalable Multicasting: A Cross-Layer Perspective

Houman Rastegarfar[1], Kamran Keykhosravi[2], Erik Agrell[2], and Nasser Peyghambarian[1]

[1] *College of Optical Sciences, University of Arizona, 1630 E. University Blvd., Tucson, AZ 85721, U.S.A.*
[2] *Dept. of Electrical Engineering, Chalmers University of Technology, 412 96 Gothenburg, Sweden*
houman@optics.arizona.edu

Abstract: We examine the feasibility of ultrahigh-scale datacenter multicasting by simultaneously taking into account the choice of architecture, modulation, and coding. Our Monte Carlo simulations indicate the dominant impact of in-band crosstalk on the throughput performance.

OCIS codes: (060.4255) Networks, multicast; (060.4265) Networks, wavelength routing.

1. Introduction

The ever-increasing cloud traffic demands and the overwhelming challenges of current datacenter deployments are calling for disruptive solutions simultaneously optimizing physical-layer signal transmission, network-layer architecture and scheduling, and application-layer control perspectives. A major class of jobs in datacenters, including the MapReduce type of applications, depend on the simultaneous dissemination of the same information copy to a large number of processing nodes, i.e, multicasting. Multicasting is an efficient mechanism for group communications that results in a reduced network load while improving the throughput and delay performance of applications. The lack of proper multicasting mechanisms in electronic datacenter designs has motivated the design of optical solutions [1]. However, due to the limited spectral and spatial resources (i.e., wavelength and coupler port counts), provisioning efficient and scalable multicasting directly in the optical domain is not straightforward.

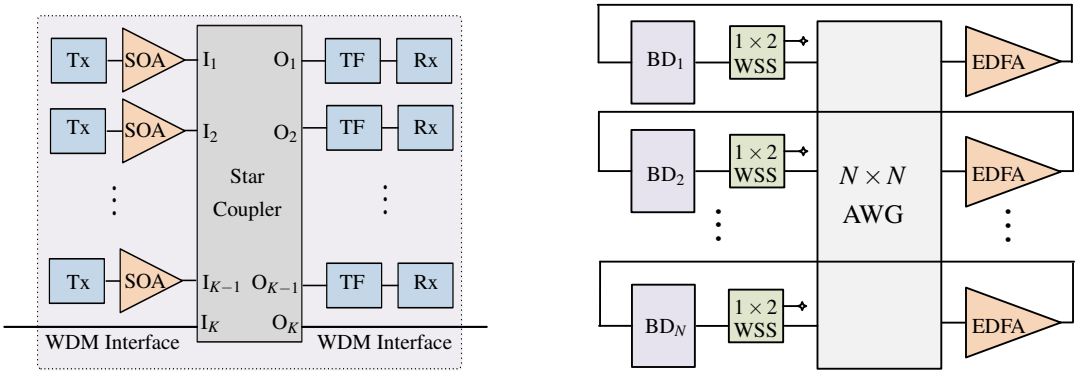
To cope with the burgeoning datacenter demands, cross-layer designs that seek to simultaneously maximize the capacity per connection as well as the number of concurrent connections in the network become significant. From an architectural viewpoint, wavelength parallelism is an effective technique for reusing the limited spectral resources in multiple broadcast domains. From the physical layer point of view, the all-optical switching of M-level pulse amplitude modulation (M-PAM) signals, based on intensity modulation and direct detection (IM/DD), is an energy- and cost-efficient candidate for spectral efficiency and scalability improvements in datacenters [2,3]. In addition, to compensate for the physical layer impairments, PAM can be combined with short block-length error-correcting codes with rate adaptation to help minimize the redundancy overheads and processing latencies [2,4].

In this paper, we consider the scalability that can be achieved by a multicast-enabled switch, considering different design factors from both the network and physical layers. These include wavelength routing, pulse amplitude modulation, and code rate adaptation. We examine a wavelength-reuse architecture based on an arrayed waveguide grating (AWG) core. The cyclic routing pattern of the AWG leads to a crosstalk-rich environment. We show how tweaking crosstalk parameters impacts the capacity, provided that programmable transceivers adapt their code rate to counteract the highly random crosstalk impairments. We especially point to the sensitivity of PAM throughput to crosstalk.

2. Wavelength-Reuse Optical Multicast Architecture

Fig. 1(a) depicts a flexible datacenter building block for the switching of optical signals. A star coupler combined with tunable transceivers supports unicast, multicast, or broadcast traffic delivery in an energy-efficient and bit-rate transparent fashion [1, 2]. Semiconductor optical amplifiers (SOAs) are used to compensate for the coupler loss by amplifying single-wavelength signals, and tunable filters (TFs) tune over the range of all available wavelengths. Due to the limited coupler port count and the tuning range of tunable components, the baseline switch is not scalable. Hence, wavelength-division-multiplexed (WDM) ports are required for interfacing with other broadcast modules in a scalable architecture.

We consider a modular architecture to interconnect a large number of broadcast domains by reusing the same set of available wavelengths. Fig. 1(b) depicts the proposed design, comprising an $N \times N$ AWG and N broadcast domains (BDs). With $K \times K$ star couplers, this architecture interconnects $N \times (K-1)$ computing nodes. In Fig. 1(b), 1×2 wavelength selective switches (WSSs) are employed to only allow connections with source and destination nodes in



(a) (b)

Fig. 1: (a) Basic structure for an optical broadcast domain (BD). The WDM interfaces enable connectivity to the rest of the network. (b) Interconnecting broadcast domains with wavelength routing.

different BDs to pass through the AWG. As only one interface is set up between a BD and the AWG input, one WSS tributary port is employed and the other is terminated. Erbium-doped fiber amplifiers (EDFAs) compensate for the losses of WDM signals due to a star coupler and a TF [2]. For connections that traverse the AWG, two amplification stages are required, one realized by an SOA in the input port of the source BD and the other by an EDFA.

The multicast-enabled, wavelength-routing switch supports datacenter traffic locality and avoids long physical paths as much as possible. This gives rise to two connection types. A connection whose source and destination belong to the same BD is an intradomain connection. Otherwise, it is considered as an interdomain connection. In terms of scheduling, we consider a distributed, greedy, and offline algorithm by adopting a request–grant approach. In servicing connection requests within each domain, the scheduler first addresses the interdomain connections due to the connectivity constraints imposed by the AWG cyclic routing pattern, and then resolves the contentions among intradomain connection requests and allocates them wavelengths based on the first-fit wavelength assignment strategy.

In Fig. 1(b), a signal with wavelength index $i = 1, 2, ...$ on input fiber port $j = 1, 2, ..., N$ of the AWG is routed to its output port $1 + \text{mod}\,(i - j, N)$. This routing pattern allows for multiple signals with the same wavelength to coexist within the AWG, resulting in in-band crosstalk. On the other hand, the nonideal TF shape results in out-of-band crosstalk. These two impairments along with several other impairments, including amplified spontaneous emission and laser intensity noise, determine the bit error rate (BER) of PAM signals [2]. They differ from other noise terms in that their variance depends on the number of connections within the AWG or BDs. Hence, Monte Carlo analysis is required to evaluate their impact based on a scheduling algorithm that determines the loading of the switch.

3. Cross-Layer Performance Analysis

We developed a cross-layer Monte Carlo simulator to examine the PAM performance in a crosstalk-rich environment. Code rate adaptation was employed to assign as little redundancy as possible based on short Reed-Solomon codes with a block length of 255 bytes, assuming a maximum pre-FEC BER of 3×10^{-2} and a target post-FEC BER=10^{-12}. As the code rate drastically drops past the mentioned pre-FEC BER [2], we assume connections with a pre-FEC BER greater than 3×10^{-2} to be irretrievable. In our simulations, both AWG and coupler port counts are set to $N = K = 64$. The symbol rate is 28 GBaud. The average transmit power is 3 dBm. The tunable filter has a Gaussian amplitude response and a crosstalk ratio of $R_{\text{F}} = -30$ dB. We consider a low-crosstalk AWG [5] with adjacent (R_{AX}) and nonadjacent (R_{NX}) crosstalk ratios of -31 dB and -43 dB, respectively. For the physical layer model details, please refer to [2].

In a simulation run, each node generates a connection request with a nonzero probability (i.e., load). With probability $1/4$, the generated request is interdomain. Although the majority of datacenter traffic is localized, interdomain connections have to traverse a longer chain of components including the AWG. Hence, we only report results for the interdomain portion of the traffic so as to study the impact of cascaded impairments in isolation. Each reported value has been averaged over 100 runs. Fig. 2 depicts the total interdomain goodput (i.e., the sum of the *net* rate of all established interdomain connections) versus load, considering both imperfect and ideal physical layers (PHYs). While the PHY impact is negligible for 4-PAM, 8-PAM experiences a 44.9% goodput degradation at full load. Interestingly as the load exceeds 0.64, 4-PAM outperforms 8-PAM due to PHY impairments.

Due to the high sensitivity of 8-PAM to PHY impairments, we examine its performance in different crosstalk scenarios. Three cases are considered in Fig. 3, assuming 1) default crosstalk parameters, 2) an ideal AWG and an

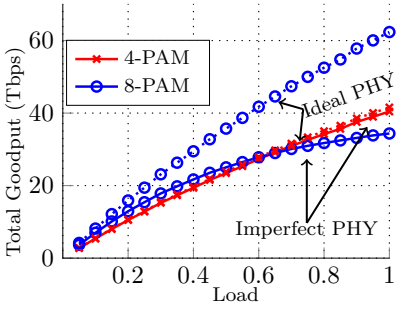
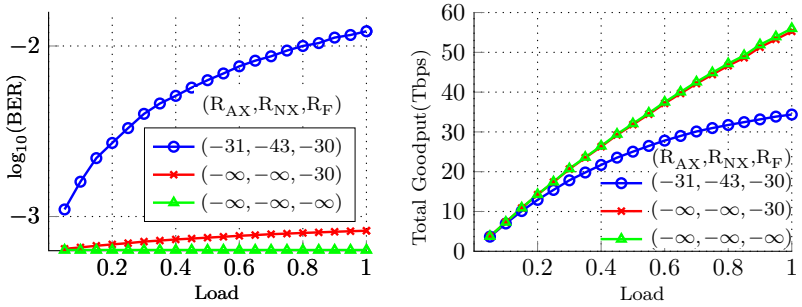
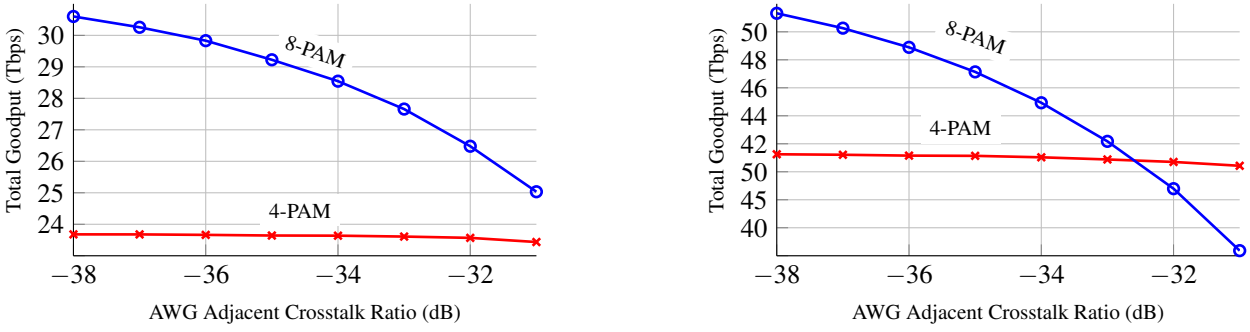


Fig. 2: Interdomain goodput vs. load for 4-PAM and 8-PAM.



(a) (b)

Fig. 3: Impact of crosstalk parameters on 8-PAM performance: (a) average interdomain pre-FEC BER, and (b) overall interdomain goodput vs. load. The crosstalk parameters are expressed in dB.



(a) (b)

Fig. 4: Sensitivity of PAM to AWG crosstalk: total interdomain goodput versus AWG adjacent crosstalk ratio (R_{AX}) for (a) load=0.5 and (b) load=1. R_{NX} is assumed to be 12 dB less than R_{AX}.

imperfect TF, and 3) an ideal AWG and an ideal TF. The choice of crosstalk parameters alone poses a significant impact on BER and goodput. In the absence of any crosstalk terms, the PHY degradation impact on goodput is upper-bounded by 10.2% (comparing the achievable goodput for imperfect and ideal PHYs at full load). Our analysis points to a negligible impact of out-of-band crosstalk compared with in-band AWG crosstalk.

To quantify the sensitivity of 8-PAM to AWG crosstalk, in Fig. 4 we report the overall interdomain goodput for a range of AWG crosstalk ratios (assuming a fixed offset between R_{AX} and R_{NX}). For comparison purposes, values for 4-PAM are also depicted. 8-PAM is remarkably sensitive to parameter variations. An average goodput sensitivity of 0.8 Tbps/dB and 2.4 Tbps/dB could be observed for loads of 0.5 and 1, respectively. By reducing R_{AX} to -38 dB, the PHY penalty could be limited to 14.5% for load=0.5 and 17.7% for load=1 (compared to 44.9% with $R_{\mathrm{AX}} = -31$ dB).

4. Conclusion

We examined the impact of crosstalk mechanisms on PAM performance in a wavelength-routing optical multicast switch. While 4-PAM proved to be robust to impairments, 8-PAM was found highly susceptible. However, improved AWG crosstalk ratios were shown to reduce the physical layer penalty by as much as 60.6%. In order to fully capitalize on the potentials of PAM, besides the high-precision fabrication of AWG devices, it is crucial to seek novel architectural solutions (e.g., through using multiple free spectral ranges of a smaller AWG) and crosstalk-aware scheduling.

5. Acknowledgement

This work was supported by the NSF Center for Integrated Access Networks (CIAN) under grant no. EEC-0812072 and the Swedish Research Council under grant no. 2014-6230.

References

1. P. Samadi *et al.*, "Optical multicast system for data center networks," Opt. Express **23**, 22162–22180 (2015).
2. H. Rastegarfar *et al.*, "PAM performance analysis in multicast-enabled wavelength-routing data centers," J. Lightwave Technol. **35**, 2569–2579 (2017).
3. K. Szczerba *et al.*, "70 Gbps 4-PAM and 56 Gbps 8-PAM using an 850 nm VCSEL," J. Lightwave Technol. **33**, 1395–1401 (2015).
4. L. Yan *et al.*, "Sensitivity comparison of time domain hybrid modulation and rate adaptive coding," in Proc. OFC, W1I.3, Anaheim (2016).
5. S. Kamei *et al.*, "$N \times N$ cyclic-frequency router with improved performance based on arrayed-waveguide grating," J. Lightwave Technol. **27**, 4097–4104 (2009).