Flexible-bandwidth Optical Interconnects for Datacom Networks

Roberto Proietti, Paolo Grani, Zheng Cao, and S.J. Ben Yoo
Department of Electrical and Computer Engineering, University of California, Davis, CA, 95616, USA.
*proietti@ucdavis.edu

Abstract: This paper presents flexible bandwidth solutions for Datacom optical networks exploiting AWGR technology, DVFS and channel bonding techniques. Benchmarking simulations and experiments demonstrate up to 2x reduction in energy consumption and 1.77x throughput increase.

Keywords: Flexible Bandwidth; Data Centers; Arrayed Waveguide Grating Routers.

I. INTRODUCTION

The exponential increase of data in today’s data centers brings significant challenges in terms of power consumption. While optical interconnects enable a scalable bandwidth interconnection (independent of the communication distance) with low energy requirements below 1 pJ/bit, the power consumption due to communications alone still represents a significant portion of the overall power consumed in data centers today. A significant part of this communication power is actually wasted because the conventional communication systems cannot adapt the communication to the traffic patterns, which are bursty [1] with high peak-to-average ratios. As a result, a lot of energy is consumed even when no meaningful bits are transmitted [2]. However, by employing flexible-bandwidth optical communication techniques, it is possible to significantly improve the energy efficiency of the communication systems. Within the rack and at the board level, architectures usually rely on shared memory and cache-coherency for the communication and synchronization between the working threads. Exploiting flexible bandwidth optical communications with this kind of traffic is challenging because of its burstiness and because of the very limited size of most of the packets (control packets). Between racks and clusters, where aggregation takes place, the link utilization and average packets or flow sizes can be higher, making it easier to implement flexible-bandwidth optical transmission and allocation schemes. Typical systems are designed to support either highest-peak-demands (energy and resource inefficient) or average-demands. In particular, several applications create hot-spots that can change dynamically over time, overloading the electronic switches at the higher hierarchy, resulting in limited system-wide performance. The next two sections will present two use cases for board-level (intra-rack) and inter-rack optical interconnection networks and demonstrate by simulation and experiments the benefits of using different flexible bandwidth assignment techniques [3, 4].

II. ALL-TO-ALL AWGR INTERCONNECTION WITH DYNAMIC VOLTAGE FREQUENCY SCALING

For board-level communication between multiple sockets, we propose to use all-to-all communication based on wavelength routing in Arrayed Waveguide Grating Routers (AWGRs), and Dynamic Voltage and Frequency Scaling (DVFS [5]) to dynamically adjust the transceivers bandwidth according to the link utilization. Figure 1 shows the proposed architecture.

![Figure 1. Hierarchical optical interconnected architecture for inter-sOCKET communication within a board and between multiple boards. (Left): the Socket (S) topology with the Hub switch connecting the four computing cores with private and shared cache memory; (Center): the Multi-Socket Board (MSB) with four sockets based on passive AWGR all-to-all interconnection; (Right): a Multi-Board Blade (MBB) computing node with four MSBs. [3]](image)

It is well known that the dynamic power of CMOS transistors scales as \( \propto V_{dd}^2 \cdot f \), where \( V_{dd} \) is the driving voltage and \( f \) is the clock speed. If \( V_{dd} \) can be lowered for circuits with low \( f \), it is then possible to obtain significant improvement in energy efficiency by lowering the clock speed in combination with the driving voltage (nearly 2x improvements in power efficiency for 20% underclocking). We studied the benefits of using DVFS when applying this power scaling technique.

This work was supported in part under DoD Agreement Number: W911NF-13-1-0090.
technique to the $p$ and $\mu$ TRXs of each node (Hub switch) in the architecture of Figure 1. We modeled the overall system by using the GEM5 simulator with 64 cores distributed in 16 sockets. GEM5 boots the Linux 2.6.27 operating system and runs the PARSEC 2.1 benchmarks suite [6]. Figure 2 shows the benchmark results in terms of execution time (left) and Energy Delay Power (EDP, right). We evaluated the performance of the proposed AWGR-based architecture with DVFS using different transmission parallelism (4 and 8 bits). Specifically, we compared the DVFS approach with source-synchronous technique [7] against a conventional approach using optical links with Clock and Data Recovery (CDR). We also compared the results with a state-of-the-art Hyper Transport electronic system.

![Figure 2](image.png)

Note that, the CDR system always transmits at the nominal speed (i.e., 10 Gb/s) while paying for the CDR circuitry consumption and power to send synchronization bits to keep the receivers locked. In terms of execution time, the performance of with the system with CDR is always the best. By increasing the communication parallelism from 4- to 8-bit helps to further improve the execution time, as shown in the first two bars of Figure 2(left). When using DVFS, the system avoids the transmission of bits when the Hub buffers are empty. The system sets the transmitter frequency and voltage supply to a maximum and minimum values according to the traffic load in each Hub. DVFS introduces some latency [see Figure 2(left)] due to the burstiness of the considered benchmarking traffic. However especially when using only 4-bit parallelism, DVFS gives significant energy reduction compared to CDR system [see Figure 2(right)].

Note that, increasing the bit parallelism for the DVFS solution comports an average higher EDP value [fourth bars in Figure 2(right)] due to the higher number of transceivers as well as the higher execution time in comparison to the CDR case.

### III. Flexible Bandwidth Inter-rack All-to-all Optical Network With Channel Bonding

Figure 3(top) shows an example of rack-to-rack optical interconnect architecture exploiting wavelength routing in AWGR to perform "channel bonding" (see Figure 3(bottom)) to dynamically, rapidly, and flexibly assign additional bandwidth upon demand between hot spots. To support both high scalability and connectivity, $\mu$ AWGRs interconnect with each other in an all-to-all pattern and each AWGR connects with $p$ clusters. The architecture scales to $p \times p \times \mu$ clusters and the radix of AWGR is $p \times (p + \mu - 2) + \mu - 1$. Therefore, the full system can reach up to 103,680 servers using six 65-port AWGRs when $p = 6$, $\mu = 6$, 40 servers/racks, 72 racks/clusters. By using two intra-region transceivers (the grey TRXs in Figure 3), three clusters connected with the same AWGR can achieve contention-free all-to-all communication. Each cluster communicates with other AWGRs with two inter-region TRXs (the green ones). To achieve flexible-bandwidth reconfiguration between the hot spots, the TRXs make use of fast Tunable Lasers (TLs) which can achieve fast wavelength tuning in ~10 nanoseconds [8]. When hot spots arise, the control plane can the TRXs’ wavelengths to increase the number of connections between the hot clusters. For instance, originally, the four clusters in Figure 3(bottom) connected with each other in an all-to-all fashion. When the bandwidth between C0 and C3 exceeds the peak bandwidth of a single link, TRX for C0−C2 with $\lambda_1$ (blue link) tunes to $\lambda_2$ (red link) for C0−C3. Eventually, the bandwidth between C0 and C3 is doubled by bonding the signals transmitted by the two red TRXs. The TRXs that require tuning are selected based on the following rules: 1) all the Cluster are still reachable after tuning; 2) the tuning introduces limited additional forwarding to non-hot Clusters; 3) the TRXs to be tuned have light traffic load. We experimentally demonstrated the channel bonding on an 8-node demo using 8 FPGA boards forming two regions with four FPGAs per region. In particular, we demonstrated the following four scenarios: 40Gbps hot-spot traffic between FPGA 1 and 4 (intra-region) with and without flexible bandwidth adjustment by channel bonding; 40Gbps hot-spot traffic between FPGA 1 and 5 (inter-region) with and without flexible bandwidth adjustment.

This work was supported in part under DoD Agreement Number: W911NF-13-1-0090.
Specifically, we performed benchmarking simulations and experimental measurements to prove that adapting the architecture capabilities to the traffic requirements is crucial to achieve good performances. This can be explained considering that the congestion caused by the hot spot traffic is released even though the link reconfiguration causes additional forwarding between other clusters.

Figure 4 shows the statistics (measured experimentally) for the four scenarios described above. Figure 4(left) shows that network with channel bonding achieves up to 1.77x improvement in accepted hot-spot traffic. Figure 4(right) shows how the links reconfiguration dedicated to certain clusters does not necessarily reduce but can actually increase the accepted background bandwidth. This is fundamental to meet the requirements of next generation data center systems.

IV. CONCLUSION

In this paper we demonstrated the benefits of applying flexible bandwidth techniques in optically-interconnected architectures for intra and inter-rack communication. Specifically, we performed benchmarking simulations and experimental measurements to prove that adapting the architecture capabilities to the traffic requirements is crucial to achieve good performances. This is fundamental to meet the requirements of next generation data center systems.

V. REFERENCES


This work was supported in part under DoD Agreement Number: W911NF-13-1-0090.