Hierarchical AWGR-based Computing Node Architecture: Performance under Realistic Benchmark Workload

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

Abstract: We propose a scalable multi-board HPC blade (computing node) architecture with hierarchical interconnection based on wavelength routing with Arrayed Waveguide Grating Routers (AWGRs). We validate performance through realistic benchmarking simulations and shows reductions up to 2× in energy consumption and 3× in execution time.

1. Introduction
HPC architectures adopt Multi-Socket Boards (MSBs) to increase the amount of resources (e.g., RAM and cores) as well as the computation density that applications can access with limited overhead due to physical proximity. The latency and bandwidth for accessing the memory hierarchy and, in general, for data movement, are typically the keys to modern HPC applications performance. Optical interconnects, integrated down to the board and chip levels, offer superior performance in terms of energy efficiency and scalable communication bandwidth, compared to their electrical counterpart. In particular, optical interconnects based on AWGR [1] allow the deployment of a contention-free, all-to-all communication network with low-latency and high-bandwidth, capable to offer the required networking support to irregular communication patterns, typical of HPC applications.

2. Hierarchical Multi-Socket Architecture

![Fig. 1. Hierarchical optical interconnected architecture for inter-socket communication. (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): the Multi-Board Blade (MBB) computing node with four MSBs.](image)

Fig. 1 shows the proposed Multi-Board Blade (MBB) node with 16 sockets, four MSBs. Fig. 1(left) shows the socket with four cores electronically interconnected through a Hub switch. The Hub has p wavelength-specific transceivers (TRXs) for all-to-all communication through AWGR within the same MSB, as shown in Fig. 1(center). The Hub also has μ wavelength-specific TRXs for inter-socket communication between different MSBs. This communication involves a relay operation in one Hub in the destination MSB. Note that, in principle, it could be possible to keep the whole architecture flat and interconnect the 16 sockets in an all-to-all fashion like in the MSB, avoiding the relay operation. However, this solution would not scale due to the high number of wavelengths and transceivers per Hub. The hierarchical approach allows to significantly reduce the number of transceivers at the expenses of lower performance due to the relay operations for inter-board traffic [see red line in Fig. 1(right)].

3. Simulation Methodology

Table 1. ONoC inter-cluster parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photodetector sensitivity</td>
<td>-22 dBm</td>
</tr>
<tr>
<td>Worst-case total loss w/ 2 AWGRs</td>
<td>15.8 dB</td>
</tr>
<tr>
<td>Laser efficiency</td>
<td>10%</td>
</tr>
<tr>
<td>Laser power per lambda [mW]</td>
<td>0.240</td>
</tr>
<tr>
<td>P_CDR per bit [mW]</td>
<td>5</td>
</tr>
<tr>
<td>Total-E-Dynamic [pJ/bit]</td>
<td>0.53</td>
</tr>
<tr>
<td>Total-P-static CDR (4-8-bit) [mW]</td>
<td>2957 / 5915</td>
</tr>
<tr>
<td>Total-P-static DVFS (4-8-bit) [mW]</td>
<td>1210 / 2420</td>
</tr>
</tbody>
</table>

Performance evaluations are obtained using GEM5 simulator [2] to model Chip Multi Processors (CMPs) with 64 cores distributed in 16 sockets, booting the Linux 2.6.27 operating system, and running the PARSEC 2.1
benchmarks suite [3]. We considered the execution time of the parallel region of each benchmark and the Energy Delay Product (EDP). Table 2 summarizes the architectural features of the architecture. Table 1 shows the optical energy consumption of all the modules (lasers, microring resonators, and CDR) and the worst-case power loss.

4. Results and Discussion

Fig. 2 shows the results both in term of execution time (left) and EDP (right). We evaluated the performance of the proposed two-level AWGR-based hierarchical architecture using different optimizations and transmission parallelism. Specifically, we analyzed the case of optical links with standard Clock and Data Recovery (CDR) and with source-synchronous techniques [4] and Dynamic Voltage and Frequency Scaling (DVFS) [5] optimizations.

Fig. 2. Execution time (left) and EDP (right) normalized to the electronic baseline in comparison with an optical hierarchical solution with CDR and DVFS and 4 and 8-bit parallelism (low values in the plots indicate superior results).

We compared the results with a state-of-the-art multi-socket board electronic baseline, based on HyperTransport protocol for intra-socket and intra-board communications. The system configuration with CDR always transmits at the maximum speed, but it pays for CDR circuitry consumption and wastes power to send synchronization bits to keep the receivers locked. In terms of execution time, the performance of the optical hierarchical architecture with CDR is always the best, and increasing the communication parallelism from 4- to 8-bit helps to further improve the achieved execution time, as shown in Fig. 2(left), in the first two bars. With DVFS, instead, the system avoids the transmission of bits when no packets are in the Hub buffers (clock lambda is always transmitted though). Indeed, the system dynamically sets the transmitter frequency and voltage supply to a maximum and minimum value depending on the incoming messages (traffic load) in each Hub, reducing the power consumption consequently. We applied the voltage and frequency scaling only to the transmission side to avoid complex signaling protocol in the architecture. DVFS introduces some latencies in the transmissions due to the burstiness of the considered benchmarking traffic. The main benefit of the DVFS is in terms of energy consumption. Fig. 2(right) shows that this approach, when using only 4-bit parallelism, gives significant energy reduction compared to CDR system. This is mainly due to the avoidance of the clock recovery circuitry. We notice that increasing the parallelism for the DVFS solution comports an average higher EDP value [fourth bars in Fig. 2(right)], due the increased number of lambdas needed in this setup and to the slower execution time in comparison to the CDR case, which affects the static energy consumption.

4. Conclusion

Results show that tradeoffs between execution time and energy consumption can be achieved depending on the architecture, communication parallelism, and applied transmission techniques. Up to a 3× time improvement can be obtained exploiting the CDR solution and up to a 2× energy improvement exploiting the DVFS setup technique.

References


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