Bandwidth Reconfigurable Optical Switching Architecture for CPU-GPU Computing Systems with Shared Memory

Arastu Sharma[1,*], Qixiang Cheng[1], Nikolaos Bamiedakis[1], Madeleine Glick[2], Fotini Karinou[3], Keren Bergman[2], Richard Penty[1]

[1]Engineering Department, University of Cambridge, 9 JJ Thomson Av. CB3 0FA, UK
[2]Department of Electrical Engineering, Columbia University, New York, NY 10027, USA
[3]Microsoft Research Cambridge, 21 Station Road, CB1 2FB, Cambridge, UK
Author email*: as2708@cam.ac.uk

Abstract: We propose a reconfigurable optical switching architecture for shared-memory CPU-GPU systems. System-level simulations show improved execution time and energy efficiency up to 34% and 25% respectively compared to a static point-to-point architecture for specific application sets.

1. Introduction

Data-centre board-level system architectures are trending towards heterogenous systems where different compute and memory nodes like CPUs, GPUs and accelerators are employed to run multiple applications with a common system memory. Due to the low yield of large silicon chips, smaller chiplet-based systems have received much attention for scaling-up processing power in computing systems as they can offer high energy efficiency and low-cost manufacturing [1]. Cloud computing traffic varies considerably from compute-intensive general-purpose applications to memory-intensive deep learning workloads. Multi-chiplet systems with disaggregated CPU, GPU and memory nodes can create a flexible, reconfigurable board-level computing platform offering higher resource utilization and energy efficiency. However, the practical implementation of large such systems has been limited by the large inter-chiplet latency, the distance-related energy overhead, and the limited IO bandwidth imposed by current copper-based interconnection and packaging technologies. Optical interconnects have been suggested as a promising solution to enable system scalability as they can offer high IO bandwidth, distance-independent energy consumption, and low-latency interconnections [1]. In addition, the use of optical interconnection fabric can allow significant re-design of the system architecture and resource allocation which, otherwise, is not possible with conventional electrical solutions [2]. Here we propose a novel system architecture for heterogenous computing systems that relies on wavelength and space optical switching. The proposed architecture combines an "unconventional" shared memory design, and dynamic bandwidth allocation enabled by a wavelength division multiplexing (WDM) N×N optical fabric and reconfigurable wavelength and space optical switching. In particular, we exploit the properties of a reconfigurable micro-ring resonator (MRR) switching fabric in a heterogeneous CPU/GPU system employing 12-chiplets (8 CPU chiplets with 16 cores and 4 GPU chiplets). We carry out system simulations in gem5-gpu simulation environment and we assess the performance of the system by calculating the execution time and energy efficiency under different heterogenous workloads. The system performance is compared against two static architectures based on WDM-based optical point-to-point links implemented with array waveguide grating routers (AWGRs): (i) a conventional point-to-point non-uniform memory access (NUMA) design and (ii) a uniform memory access (UMA) multi-chiplet architecture. The simulation results show that the proposed architecture can provide significant benefits in both execution time and energy efficiency for a variety of workloads.

2. Shared memory, dynamic bandwidth and switching architecture

The use of CPU and GPUs into a single memory system has a huge advantage of allowing processors to avoid memory copy overheads [5]. Shared memory in parallel computing multi-core systems can be allocated with either NUMA or UMA designs [1]. In NUMA, system chips can access data stored on the local memory of another chip given they issue a specific memory request to the host chip. This generates cache memory coherence issues and results in additional data that needs to be exchanged between the chiplets. On the other hand, UMA-based systems are less sensitive to data coherence issues and offer reduced energy consumption for performing the same task owing to the reduced workload of the inter-chip communication. However, their performance is limited by the latency of the (physically distant) off-chip memory interconnection (typically ~100-150 ns) and restricted bandwidth due to the use of a single memory controller. As a result, conventional computing systems have been based on NUMA architectures. In recent work [2], a static point-to-point optical network based on UMA system employing AWGRs was proposed as it can alleviate the main drawbacks of the NUMA architecture and offer significant performance improvements. This is mainly achieved by reducing inter-chiplet latency. However, such point-to-point networks have limited scalability due to finite size of the AWGRs, pose bandwidth constraints due to fixed channel sizes, and exhibit non-optimum energy efficiency as they typically result in low link utilization (i.e. wasting static laser energy). Space and

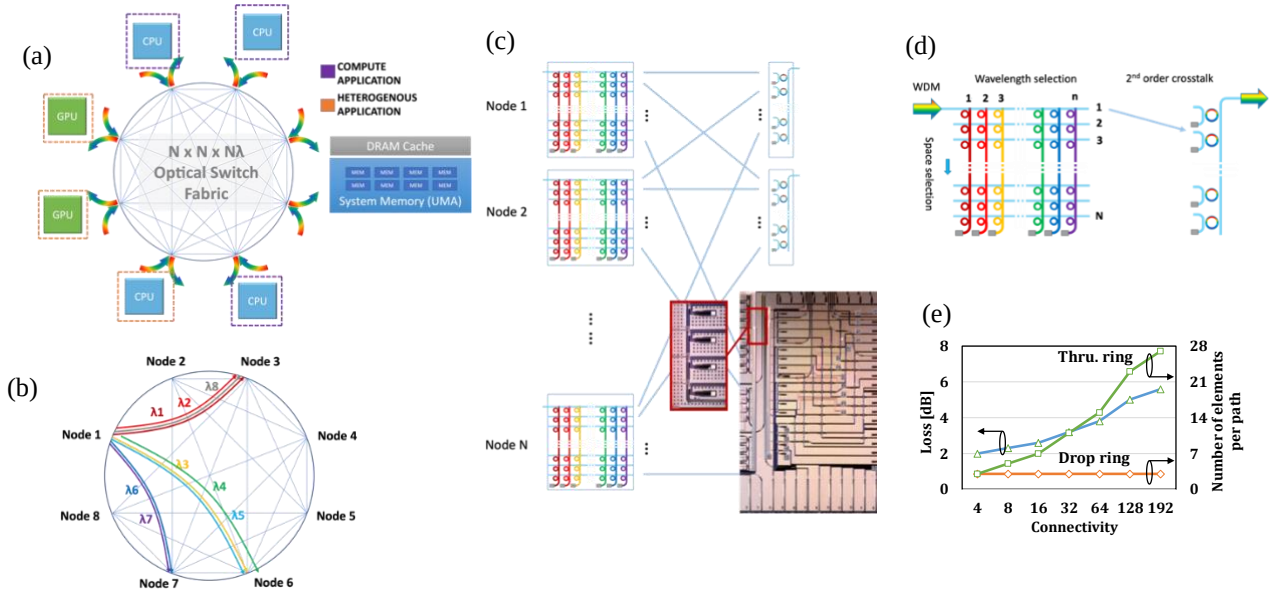

Fig 1 (a) Schematic of the proposed optical switch architecture. The allocation of the two applications to the system is also noted. (b) Illustration of the dynamic bandwidth allocation, (c) planer view of space-and-wavelength switch design based on [3], (d) operating principle and (e) emulated loss function for the N×M λ crosspoint matrix up to 192 connections.

wavelength optical switches can, in comparison, provide reconfigurability and help overcome these constraints by dynamically providing additional connections and bandwidth between nodes when needed. Fig. 1(a) shows a schematic of the UMA architecture used in this work while Fig. 1(b) illustrates the space-and-wavelength switching functionality that allows arbitrary combinations of wavelengths to be selected and routed across the on-board optical network. The work presented here is based on the scalable switch architecture presented in [3]. The space and wavelength switch uses both arrays of MRR-based wavelength selectors and free spectral range (FSR)-matched comb aggregators (Fig. 1c) and can implement an N×M λ crosspoint matrix of up to 192 connectivity (16 ports × 12λ) (Fig. 1d). Here, we assume that the switch supports 160 Gb/s for each channel, with WDM using 8 λs, each operating at 20 Gb/s. The insertion loss of each path is calculated by identifying the number of drop and through rings (Fig 3e) [3].

3. Simulation and application workloads

We use gem5-gpu [5], to simulate the heterogeneous CPU-GPU computing system. The computing system comprises 8 CPU chiplets with 2 CPU cores each and 4 GPUs with 16 Fermi GPU cores (Fig. 2a). For the memory system configuration, we integrate the Ramulator, an open-source code for DRAM systems [8], to gem5-gpu and simulate the configured system in a cache mode. Two different system architectures are studied to understand the trade-offs of bandwidth allocation in low and high utilization use-cases with reconfiguration either before (Case A) or during (Case B) runtime. Case A represents a reconfigurable architecture which is however static during runtime: the switch (and therefore the bandwidth allocation) is configured at the start-up time using the traffic load of executing the specific application assuming a baseline static point-to-point UMA case (e.g. for application set #7 shown in Fig 3b). On the contrary, Case B represents the proposed novel dynamic energy proportional architecture, with only 4 λs per channel, in which the bandwidth is reconfigured several times during the run-time to achieve high utilization of optical links. This on-the-fly network reconfiguration adds delay in the application execution time in Case B but reduces laser energy wastage by reducing the time operating at low utilization. The two architectures are evaluated against NUMA and UMA based optical point-to-point system proposed in [2], where each channel has the same non-reconfigurable fixed bandwidth of 160 Gb/s (8λ × 20 Gb/s).

CPU	GPU	Memory
16 cores (8 chiplets) @ 3GHz	4 x 16 Fermi SMs @ 700MHz 32KB	8 Channels
L1_I:64KB/core, assoc:4, private L1_D: 32KB/core, assoc:8, private L2: 512KB/core, assoc:8, private L3: 8MB/(8 cores), assoc:16, shared	4-way (private) 4096KB	DDR4, 2GB

Optical Parameter	Value	Optical Parameter	Value
Optical Fiber	5e-6 dB/cm	Laser efficiency	15%
Modulator Insertion Loss	1 dB	Power Margin	3 dB
Coupler Loss (Package to Chiplet)	0.5 dB	Waveguide loss	0.5 dB/cm
Receiver Sensitivity for 10^{-12} BER	-17 dB		
Photodetector Loss	0.1 dB		
AWGR Loss	1.8 dB		
AWGR crosstalk	-20 dB		

Fig. 2. (a) Primary simulation set-up, and (b) parameters for optical components used in energy efficiency calculations.

Two memory-intensive workloads from the PARSEC [9] and Rodinia [10] suite which are simultaneously running over our computing platform are employed for the simulations. We use only-CPU workloads from PARSEC and CPU-GPU workloads from Rodinia which are distributed among the 8 CPU chiplets in our system as shown in Fig. 3(a): 4 CPU chiplets (C1-C4) for only-CPU workloads and 4 CPU chiplets (C5-C8) for CPU-GPU workloads. The two applications differ in bandwidth needs, thus introducing dynamic bandwidth reconfigurability can extensively achieve better execution time performance and reducing time of laser energy consumption. We use state-of-the-art values for

energy consumption and loss of optical devices (Fig. 2b). The energy consumption of the switch is calculated based on a Si/SiN multi-layer platform and incoming traffic from the simulation.

4. Results and Discussions

The traffic between all pairs of nodes in the network is obtained for each workload for the baseline point-to-point UMA system. These are then used to determine the bandwidth allocation of our switching architecture either before (Case A) or during (Case B) runtime. Fig 3(b) shows an example of the obtained traffic load between all pairs of nodes for the workload #7. It can be observed that the particlefilter application generates considerable amount of memory-intensive traffic (Gx to M and C5-C8 to M) in comparison to freqmine (C1-C4 to M). As a result, providing additional bandwidth over the network to this application provides a considerable improvement in execution time and energy consumption (28% and 9 % respectively for Case A – see #7 in Fig. 3c and 3d). Similar analysis is carried out for the other workloads and the results are shown in Fig.3(c) and 3(d) for the two cases and compared to the NUMA and UMA baseline systems.

(a)

Name	Parsec (CPU apps)		Rodinia (Heterogenous apps)	
#1	freqmine	4 CPUs	Needle	4 CPUs + 4 GPUs
#2	bfs	4 CPUs	Needle	4 CPUs + 4 GPUs
#3	Fft, bfs	2 CPUs each	Needle	4 CPUs + 4 GPUs
#4	freqmine	4 CPUs	Backprop	4 CPUs + 4 GPUs
#5	bfs	4 CPUs	Backprop	4 CPUs + 4 GPUs
#6	fft, bfs	2 CPUs each	Backprop	4 CPUs + 4 GPUs
#7	freqmine	4 CPUs	Particlefilter	4 CPUs + 4 GPUs
#8	bfs	4 CPUs	Particlefilter	4 CPUs + 4 GPUs
#9	Fft, bfs	2 CPUs each	Particlefilter	4 CPUs + 4 GPUs

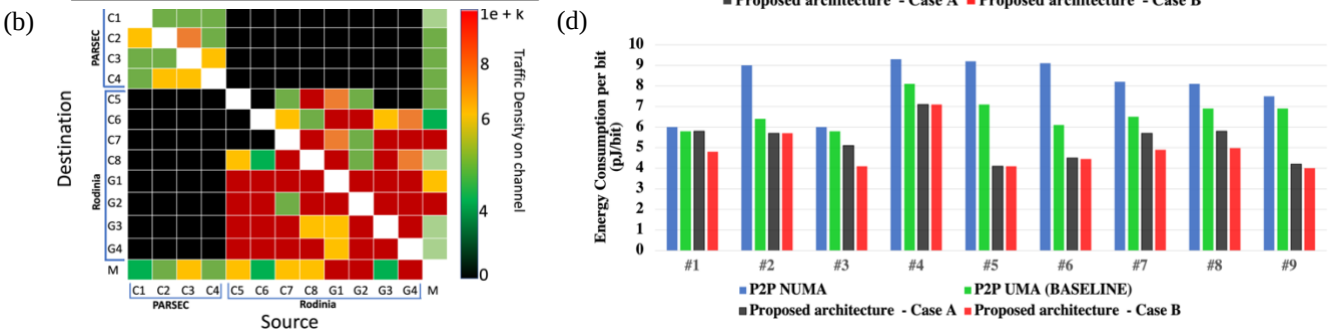

Fig 3 (a) Application sets used in simulations and (b) example of traffic pattern for application set #7 for the baseline UMA system, used to optimise the optical switch reconfiguration. (c) Application speed-up compared to the baseline UMA system and d) average energy per bit (pJ/bit) of the optical links.

The mean speedup and improvement in energy efficiency over all the application sets are calculated for both architectures: Case A: 18% and 16% respectively, and Case B: -5% (slow-down) and 27% respectively. Case B provides a similar average execution time as the baseline point-to-point system due to the on-the-fly reconfiguration overhead but offers much larger gains in energy efficiency. It is essential to note that to maximise the benefits that the proposed architectures bring, some knowledge of the traffic patterns of the running applications in the system is required. This can be inferred by the type of application to be executed (e.g., these are well defined for some machine learning workloads) before runtime (Case A) or can be observed during runtime using a monitoring system (Case B).

5. Conclusion

We propose a new system architecture for shared memory heterogeneous multi-chiplet architectures based on wavelength and space switching. Simulation results show that the proposed system achieves better or at least similar execution times compared to a static point-to-point optical network while providing significantly improved energy efficiency.

6. References

[1] Sharma, Arastu, et al. "Multi-Chiplet System Architecture with Shared Uniform Access Memory Based on Board-Level Optical Interconnects." 2021 Optical Fiber Communications Conference and Exhibition (OFC). IEEE, 2021.

[2] Fotouhi, Pouya, et al. "Enabling scalable chiplet-based uniform memory architectures with silicon photonics." Proceedings of the International Symposium on Memory Systems. 2019.

[3] Cheng, Qixiang, et al. "Scalable space-and-wavelength selective switch architecture using microring resonators." CLEO, IEEE, 2019.

[4] Choi, Sungji, et al. "BODCA: Heterogeneous CPU-GPU computing system with Bandwidth-Optimized DRAM cache design." 2020 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia). IEEE, 2020.

[5] Power, Jason, et al. "gem5-gpu: A heterogeneous cpu-gpu simulator." IEEE Computer Architecture Letters 14.1 (2014): 34-36.