Power Efficient Photonic Network-on-Chip for a Scalable GPU

Janibul Bashir
Indian Institute of Technology, Delhi
janibbashir@cse.iitd.ac.in

Khushal Sethi
Indian Institute of Technology, Delhi
ee1160556@iitd.ac.in

Smruti R. Sarangi
Indian Institute of Technology, Delhi
srsarangi@cse.iitd.ac.in

ABSTRACT

In this paper, we propose an energy efficient and scalable optical interconnect for GPUs. We intelligently divide the components in a GPU into different types of clusters and enable these clusters to communicate optically with each other. In order to reduce the network delay, we use separate networks for coherence and non-coherence traffic. Moreover, to reduce the static power consumption in optical interconnects, we modulate the off-chip light source by proposing a novel GPU specific prediction scheme for on-chip network traffic. Using our design, we were able to increase the performance by 17% and achieve a 65% reduction in ED^2 as compared to a state-of-the-art optical topology.

CCS CONCEPTS

• **Networks → Photonic Network on chip**.

KEYWORDS

Nanophotonics, Graphics Processing Unit (GPU)

ACM Reference Format:
Janibul Bashir, Khushal Sethi, and Smruti R. Sarangi. 2019. Power Efficient Photonic Network-on-Chip for a Scalable GPU. In *International Symposium on Networks-on-Chip (NOCS '19), October 17–18, 2019, New York, NY, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3313231.3352370

1 INTRODUCTION

[1] GPUs are now the default platforms for high performance machine learning and large scale computing. In this paper we present the design of an optical NoC for a GPU, which is power-efficient and scalable.

By analyzing the *RODINIA* [3] benchmarks, we observed that last level cache units are the major contributors to the on-chip traffic, and hence should be given priority while accessing the on-chip network. Moreover, we observed that by delaying some messages originating from the symmetric multiprocessors (SMs), the effect on the overall performance of the system is minimal – this effect can be used for proposing some optimizations.

Based on these observations, we build a GPU specific optical interconnect. In addition, we propose to modulate the off-chip light source based on a novel traffic prediction scheme, thereby reducing the static power consumption to a large extent. We evaluated the design in Section 3 and finally conclude the paper in Section 4.

[1]This work has been sponsored in part by the Semiconductor Research Corporation (SRC).

NOCS '19, October 17–18, 2019, New York, NY, USA

ACM ISBN 978-1-4503-6700-4/19/10…$15.00
https://doi.org/10.1145/3313231.3352370

2 DESIGN OF GPUOPT

In this paper, we propose a GPU specific topology aiming at ❶ enhancing performance, and ❷ decreasing the static power consumption. To achieve these goals, we propose ❶ an efficient optical topology, called *GPUOPT*, and ❷ a novel power scaling technique. We incorporated our power scaling technique on *GPUOPT* and developed an efficient NoC for GPUs, called *PS_GPUOPT*.

Topology: Figure 1 shows our proposed 3D optical NoC. The chip has two layers: logical layer – containing SMs, L2 banks, and memory controllers (MC), and a photonic layer – containing optical components (powered by an off-chip laser array). The logical layer is divided into 16 clusters: 8 *SM_Clusters* and 8 *LLC_Clusters*. Each *SM_Cluster* has 8 SMs, and each *LLC_Cluster* has an L2 bank tied to a MC. The intra-cluster communication is done electrically, whereas for inter-cluster communication, we incorporate a separate silicon photonics layer underneath the logical layer. The optical layer has optical stations: one for each *SM_Cluster* (called *SM_station*), and *LLC_Cluster* (called *LLC_station*). These stations are connected together using two different optical crossbars (with a separate serpentine layout): *C_network* and *NC_network*.

The *C_network* is used to carry the coherence messages between the SMs, whereas the *NC_network* is used to carry the non-coherence messages (L1-to-L2 and L2-to-L1). We assume a single-writer-multiple-reader (SWMR) topology for the *C_network* and a multiple-writer-single-reader (MWSR) topology for the *NC_network* that requires token based arbitration [2]. Moreover, to further improve the performance, we allow any optical station to source power from any power waveguide (arbitration is required).

Power Scaling : To decrease the laser power consumption in optical NoCs, we: ❶ divide the execution time into fixed size durations, called *epochs*, ❷ predict the laser power requirement for the next epoch , and then ❸ modulate the off-chip laser (reconfiguration) [1].

Every *SM_station* uses a function, Ψ, to predict the laser power requirement in the next epoch. It takes three inputs: waiting time of a station ($\mathcal{W}$), messages received ($\mathcal{M}_\mathcal{R}$), and messages sent in the current epoch ($\mathcal{M}_\mathcal{S}$), and produces a 1-bit output (station active/inactive in the next epoch) to be sent to the *Laser Controller (L_Cntrlr)* at the end of every epoch. The rules are:

$$\Psi(\mathcal{M}_\mathcal{R}, \mathcal{M}_\mathcal{S}, \mathcal{W}) = \begin{cases} 1 & (\mathcal{M}_\mathcal{R} \geq R_T \wedge \mathcal{M}_\mathcal{S} \leq \alpha * R_T) \vee (\mathcal{W} \geq W_T) \\ & \vee (\mathcal{M}_\mathcal{R} \leq R_T \wedge \mathcal{M}_\mathcal{S} < \alpha * M_R) \\ 0 & default \end{cases}$$

Here, R_T, W_T, α, $and \beta$ are hyper-parameters (generated empirically). Similarly, *LLC_stations* also send a 1-bit prediction to the *L_Cntrlr* based on the rules given below.

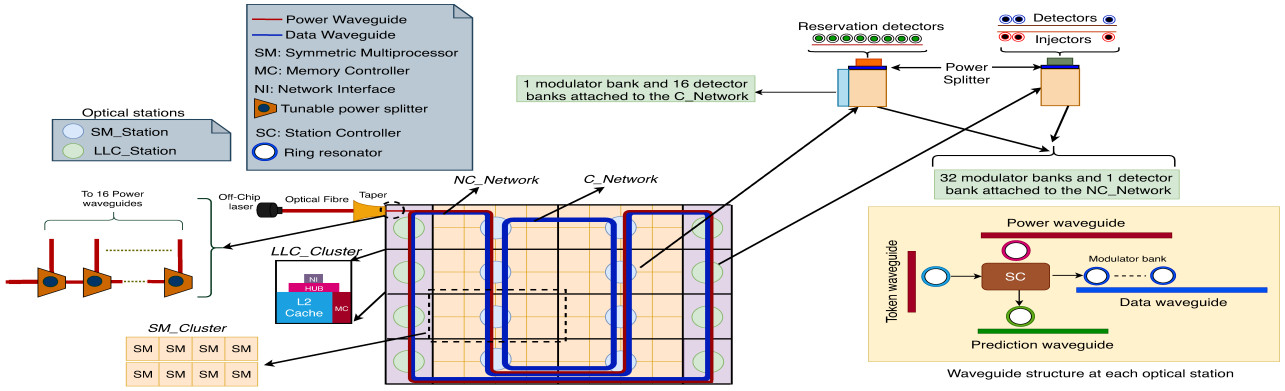

Figure 1: Topology of the optical NoC

$$\delta(\mathcal{M}_{\mathcal{R}}, \mathcal{P}_{\mathcal{E}}, \mathcal{W}) = \begin{cases} 1 & (\mathcal{M}_{\mathcal{R}} \geq \alpha * R_T) \vee (\mathcal{W} \geq \alpha * W_T) \\ & \vee (\mathcal{P}_{\mathcal{E}} \geq \beta * \mathcal{M}_{\mathcal{R}}) \\ 0 & default \end{cases}$$

Here, $\mathcal{P}_{\mathcal{E}}$ is the number of pending events at the *LLC_station*.

In the reconfiguration phase, the *L_Cntrlr* collects 16 1-bit predictions sent by the 16 optical stations. Based on these recommendations, the *L_Cntrlr* calculates the amount of laser power required in the next epoch and accordingly modulates the off-chip light source.

3 EXPERIMENTAL RESULTS

To evaluate our design, we used a cycle-accurate GPU simulator *GPUTejas* [4], and extended it to model optical networks. We used an in-house thermal simulator to calculate the temperature variations on the die and accordingly calculated the amount of tuning power required [2].

We used the workloads from the *RODINIA* benchmark suite for simulations and compared our design with a state-of-the-art photonic network proposed by Ziabari et al. [5] (*Prior_Opt*). This is the baseline design.

Performance Comparison: Figure 2 compares the performance (inverse of simulated execution time) of three different optical NoC configurations. Form the plot, we conclude that *GPUOPT* is the best configuration that performs 17% better than *Prior_Opt*. This is because the *GPUOPT* allows all the optical stations to share the available optical power. In addition, separating the coherence and non-coherence messages, further increases the performance of the overall system. In comparison, the *PS_GPUOPT* scheme does not perform so well (11% better) because it sometime produces less power than what is needed – stations need to wait. In *GPUOPT* the optical power is available all the time since it does not use any laser modulation scheme.

Figure 2: Performance comparison

Laser Power Consumption: As compared to *GPUOPT* and *Prior_Opt*, *PS_GPUOPT* results in 59% and 67% reduction in laser power consumption respectively. The main reason is the modulation of the off-chip laser in *PS_GPUOPT*. The lower laser power consumption in *GPUOPT* as compared to *Prior_Opt* is attributed to its ability to allow on-chip optical stations to share the available optical power.

ED^2 **Comparison:** In Figure 3 we compare the energy-delay-square (ED^2) (energy of entire system) product of the three optical configurations. From the plot we observe that *PS_GPUOPT* is the best configuration. As compared to *Prior_Opt* and *GPUOPT*, it has a 65% and 29% reduction in ED^2 respectively. The lower ED^2 in the case of *GPUOPT* as compared to *Prior_Opt* is attributed to its higher performance, whereas in the case of *PS_GPUOPT* the large reduction is due to a significant reduction in laser power consumption.

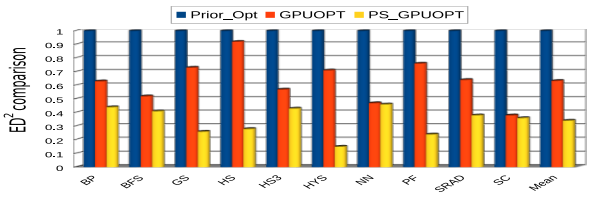

Figure 3: Energy-Delay-Square product comparison

4 CONCLUSION

In this paper, we designed an efficient optical NoC for GPUs where we use a combination of SWMR and MWSR topologies to decrease the contention. In addition, we propose to use a GPU specific laser modulation scheme in order to reduce the static power consumption. Using these set of techniques, we were able to reduce the laser power consumption by 67% with a 65% reduction in ED^2 values as compared to a state-of-the-art optical topology.

REFERENCES

[1] J. Bashir, E. Peter, and S. R. Sarangi. 2019. BigBus: A Scalable Optical Interconnect. *ACM JETC* 15, 8 (2019), 8:1–8:24.
[2] J. Bashir, E. Peter, and S. R. Sarangi. 2019. A Survey of On-Chip Optical Interconnects. *ACM Comput. Surv.* 51, 6 (Jan. 2019), 115:1–115:34.
[3] S. Che, M. Boyer, J. Meng, D. Tarjan, J W Sheaffer, S. Lee, and K. Skadron. 2009. Rodinia: A benchmark suite for heterogeneous computing. In *IISWC*.
[4] G. Malhotra, S. Goel, and S. R. Sarangi. 2014. Gputejas: A parallel simulator for gpu architectures. In *HiPC*.
[5] A. K. Ziabari, J. L. Abellán, R. Ubal, C. Chen, A. Joshi, and D. Kaeli. 2015. Leveraging silicon-photonic noc for designing scalable gpus. In *ICS*.