530 Project Proposal

Team Members: Rishabh Jain and Pranitha Malae
Date/Time of meeting: [Wednesday: Between 2-5pm or Friday: Between 1-5pm]. Please let us know if this time doesn't work.

1. **Motivation:** GPU performance is crucial as they are primary hardware used in the execution of many general purpose applications.
2. **Problem Statement:**
   Improving the L1 memory design and interconnection network for a GPGPU arch. Following are the steps:
   a. Workload characterization:
      i. Memory and network analysis specific for metrics: memory latency, network bandwidth at different levels, network latency, and L1 miss rate.
      ii. Looking for applications which are memory latency sensitive, contention sensitive(as shared resources give rise to contention)
      iii. Overall performance impact measured in IPC.
      iv. Coming up with other sensitivity metrics.
   b. This analysis should help us in figuring out the causes of bad performance at cache, network and overall GPU level. Thus, we would be able to make observations like replication of cache lines, underutilization of bandwidth, etc.
   c. Once we have figured out certain issues, we would try to propose a design to solve the problem. Like, Decoupled L1 caches [1] helps in improving the L1 cache performance.
   d. Verifying whether the observations mentioned in [1] holds valid for multi-core CPU architectures. This is because L1 comes in the critical path of processor frequency, and thus scope of shared L1 cache needs to be checked.
3. **Plan of attack:**
   a. Workload characterization
   b. Analysis of bottlenecks
   c. Design proposal
   d. Design implementation and Evaluation of our results
   Along with (a, b), we would explore literature works in the cache and network design for GPUs (high throughput processors): [8], [9], [10], [11]

4. **Resource needs:** gem5, GPGPU (version 3.0 or 4.0?) and DSENT on a lab machine.
5. **Benchmark applications list:** Broad Categories are: Deep Neural Networks and Graphics. Examples: CUDA-SDK [3], Rodinia [4], SHOC (S) [5], PolyBench (P) [6], and Tango (T) [7]).
6. **Open Questions:**
   a. Further discussion on references and fine tuning the first set of tasks.

b. Scope of improvement in private vs shared cache in [1] via Dynamic partitioning of crossbar to have a dynamic sizing of clustered shared caches.

c. Work plan: milestones and their expected time to complete.

d. Suggestions on what application to pick which are L1 hit latency sensitive? Applications which would like private caches or low contention?

e. Having a mesh network instead of crossbar for [1]? How do architect's decide on which network is suitable? With a mesh network, the proposed shared cache design may suffer from high L1 hit latency due to the new network latency component.

f. Learning GPGPU sim: any useful tutorials or references to begin?

7. **References**:

(1) Primary paper: https://adwaitjog.github.io/docs/pdf/decoupledl1-hpca21.pdf

(2) M. A. Ibrahim, O. Kayiran, Y. Eckert, G. H. Loh, and A. Jog, "Analyzing and Leveraging Shared L1 Caches in GPUs," in Proceedings of the International Conference on Parallel Architecture and Compilation Techniques (PACT), 2020 - https://adwaitjog.github.io/docs/pdf/sharedl1-pact20.pdf

(3) NVIDIA, "CUDA C/C++ SDK Code Samples," 2011. [Online]. Available: http://developer.nvidia.com/cuda-cc-sdk-code-samples

(4) S. Che, M. Boyer, J. Meng, D. Tarjan, J. Sheaffer, S.-H. Lee, and K. Skadron, "Rodinia: A Benchmark Suite for Heterogeneous Computing," in Proceedings of the International Symposium on Workload Characterization (IISWC), 2009.

(5) A. Danalis, G. Marin, C. McCurdy, J. S. Meredith, P. C. Roth, K. Spafford, V. Tipparaju, and J. S. Vetter, "The Scalable HeterOgeneous Computing (SHOC) Benchmark Suite," in Proceedings of the Workshop on General Purpose Processing Using GPU (GPGPU), 2010.

(6) L.-N. Pouchet, "Polybench: The Polyhedral Benchmark Suite," 2012. [Online]. Available: http://web.cs.ucla.edu/ ~ pouchet/software/ polybench/

(7) A. Karki, C. P. Keshava, S. M. Shivakumar, J. Skow, G. M. Hegde, and H. Jeon, "Detailed Characterization of Deep Neural Networks on GPUs and FPGAs," in Proceedings of the Workshop on General Purpose Processing Using GPU (GPGPU), 2019.

(8) Z. Jia, M. Maggioni, B. Staiger, and D. P. Scarpazza, "Dissecting the NVIDIA Volta GPU Architecture via Microbenchmarking," arXiv, April 2018.

(9) A. Li, S. L. Song, W. Liu, X. Liu, A. Kumar, and H. Corporaal, "Locality-Aware CTA Clustering for Modern GPUs," in Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS), 2017.

(10) S. Dublish, V. Nagarajan, and N. Topham, "Cooperative Caching for GPUs," ACM Transactions on Architecture and Code Optimization (TACO), 2016.

(11) X. Zhao, S. Ma, Z. Wang, N. E. Jerger, and L. Eeckhout, "CD-Xbar: A Converge-Diverge Crossbar Network for High-Performance GPUs," IEEE Transactions on Computers (TC), 2019.