Embracing Heterogeneity in Modern GPUs

Xulong Tang, School of Computing and Information, University of Pittsburgh

Motivation

- GPUs are widely used in modern applications for parallel acceleration.
  - Deep Learning; Personal medicine; VR/AR.
- The scaling of GPUs is behind the ever-growing complexity of application algorithms and ever-increasing data volume of inputs.
  - Memory wall and expensive data movement.

Project Description

- This project targets near-data computing (NDC) aware GPU systems to mitigate the memory wall [1].
- This project aims to enable scalability, high performance, and energy efficiency of applications running on NDC GPUs.
- The project goal is to deliver a tailored GPU runtime system that automatically and opportunistically schedules the application kernels in an NDC-aware fashion.

Context

- Given NDC GPUs, the fundamental question remains to answer:
  - How do the applications realize and take full advantage of the underlying NDC feature?
- Modern GPU runtime management is not aware of underlying heterogeneity.

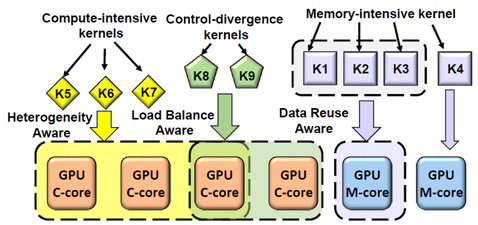

Figure 1: Scheduling different types of kernels on heterogeneous GPU cores.

- **Task I: Application aware Device Kernel Launch.**
  - ***What*** should be launched as GPU kernels to benefit from execution on M-cores and C-cores.
- **Task II: Locality aware Kernel Execution.**
  - ***How*** to execute the dynamic kernels.
- **Task III: Heterogeneity aware Dynamic Load Balancing.**
  - ***Where*** to execution the dynamic kernels.

University of Pittsburgh

Primary project goal is that, with the proposed **Near Data Computing (NDC)** aware GPU runtime framework, NDC-GPUs will become one of the first-class computing platforms for processing modern applications, delivering promising performance, energy-efficiency, and quality of service in the big data era.

Project Deliverables

- We propose a flexible NDC aware GPU runtime that is
  - Application-aware launching.
  - Locality-aware execution.
  - Heterogeneity-aware dynamic load-balancing.
- We will leverage the NDC-GPU[2] simulation framework to design and test our proposed runtime optimizations.

Potential Impact

- Impact on future research
  - The success of this project will broaden and have profound impact on the adoption of heterogeneous NDC-GPUs for large-scale graph processing applications.
  - This research will influence many application domains such as social networking, machine learning, scientific computing, bioinformatics, medical imaging, and virtual reality.
  - It will also inspire future graph library developers to employ the NDC feature.
- Education and Curriculum Development
  - Heterogeneous computing and GPUs parallel computing In graduate course.
  - Introduce parallel computing and GPUs to undergraduate students.

References and/or Acknowledgements

- [1]K. Hsieh, E. Ebrahim, G. Kim, N. Chatterjee, M. O'Connor, N. Vijaykumar, O. Mutlu, andS. W. Keckler. Transparent offloading and mapping (tom): Enabling programmer-transparentnear-data processing in gpu systems. InISCA, 2016.
- [2] A. Pattnaik, X. Tang, A. Jog, O. Kayiran, A. K. Mishra, M. T. Kandemir, O. Mutlu, and C. R.Das. Scheduling Techniques for GPU Architectures with Processing-In-Memory Capabilities. In PACT, 2016.