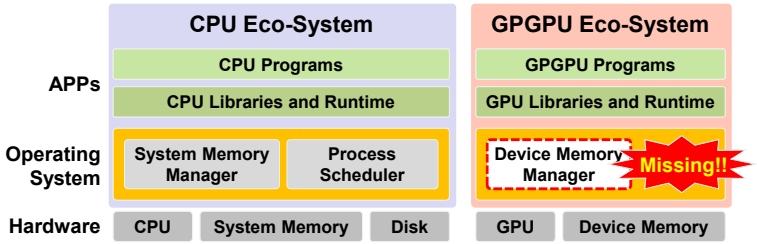
GDM: Device Memory Management for GPGPU Computing

Kaibo Wang, Xiaoning Ding, Rubao Lee, Shinpei Kato+, Xiaodong Zhang*

*The Ohio State University *New Jersey Institute of Technology +Nagoya University*

NSF SI2 Project: OCI-1147522, PI: Xiaodong Zhang, The Ohio State University

Background



- Problems with the lack of a system-level GPU device memory manager
  - Managing GPU device memory directly by individual applications is a heavy burden for programmers
  - Even more difficult when multiple applications or application components with conflicting demands share the same GPU
  - Causing application crashes, device memory underutilization, vulnerability to malicious device memory usage, and low tolerance to device memory leaks

Research Statement

- **GDM**: The first system-level device memory manager to unleash the power of GPUs on general-purpose commodity computing systems
  - A memory management framework to dynamically reclaim under-utilized device memory for better uses
  - A set of optimization techniques to minimize the overhead associated with device memory management and to ensure system performance
  - Application-transparent: no modifications to existing GPGPU programming APIs

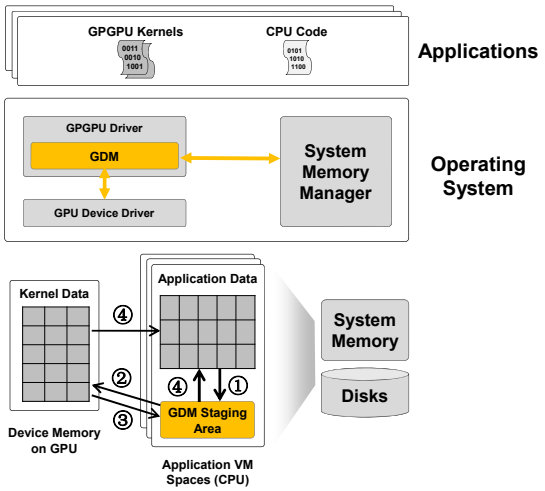
GDM Overview

① Data to be transferred to device memory is first copied to the staging area in each application's VM space.

② Data is transferred to the device memory when the kernel accessing the data is launched.

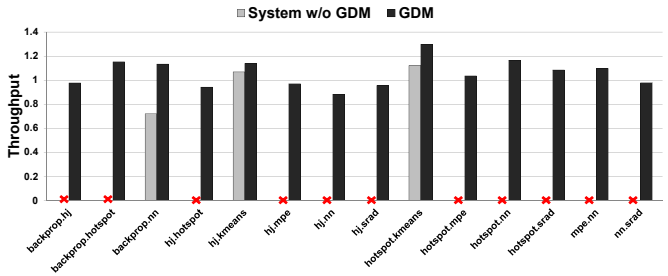
③ When the device memory is short of free space, GDM evicts some data to staging area and reclaims the space.

④ When the application wants to copy some data from device memory, GDM locates the latest version of the data and copies the data to user buffer.



Evaluation

- Experimental Setup
  - Prototype implementation of GDM in open-source GPU driver, Gdev
  - Core i7 860 CPU; 8GB system memory; GTX 480 GPU with 1.5GB device memory; PCIe 2.0 bus
- Multitasking Performance



Conclusion

- We have designed and implemented GDM to manage device memory for GPGPU computing
- A set of optimization techniques and principles are proposed to improvement the performance of GPU device memory management
- Extensive experiments with insights verify the effectiveness and usability of GDM