A Framework for Memory Oversubscription Management in Graphics Processing Units

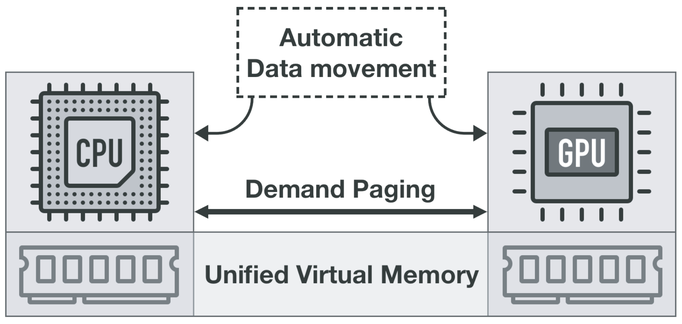
Chen Li, Rachata Ausavarungnirun, Christopher J. Rossbach, Youtao Zhang, Onur Mutlu, Yang Guo, Jun Yang

National University of Defense Technology · University of Pittsburgh · Carnegie Mellon · TEXAS The University of Texas at Austin · vmware · King Mongkut's University of Technology North Bangkok · ETH zürich

Problem & Motivation

Problem

Limited memory capacity becomes a first-order design and performance bottleneck.



Memory oversubscription causes GPU **performance degradation** or, in several cases, **crash**.

Motivation

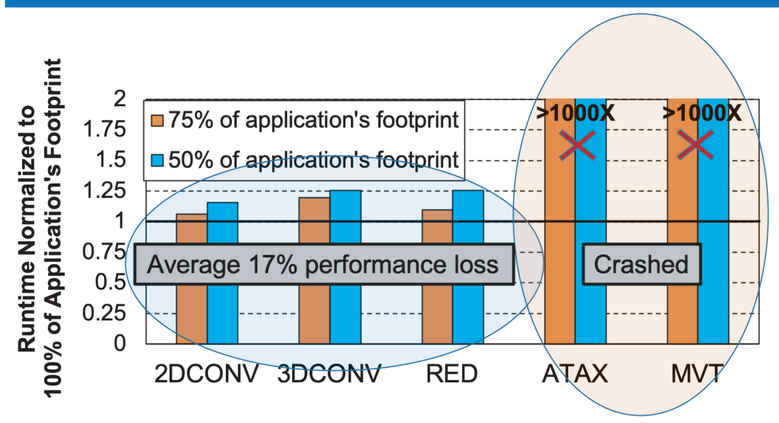
Prior **Hand-tuning** Techniques:

- Overlap prefetch with eviction requests
- Duplicate read-only data

× Requires programmers to manage data movement **manually**
× **No visibility** into other VMs in cloud environment

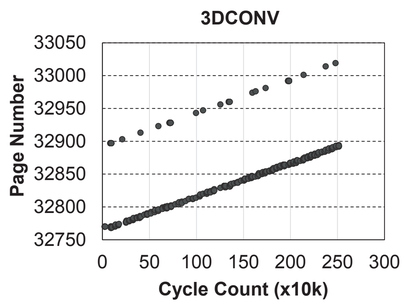
Application-transparent mechanisms are urgently needed.

Observations



5 applications are executed on a real NVIDIA GTX1060 GPU.

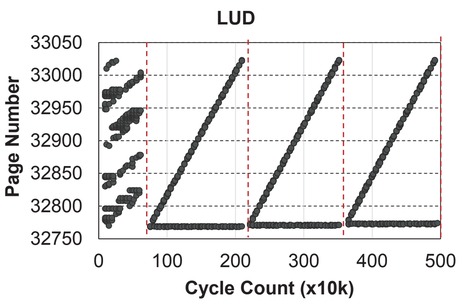
Different techniques are needed to mitigate **different sources** of overhead.



Streaming access
Small working set

Waiting for Eviction

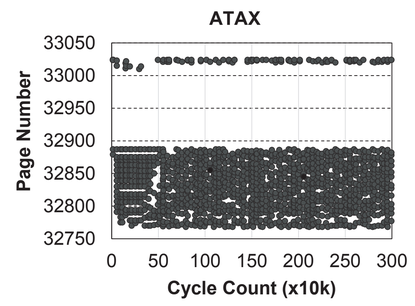
Regular applications with no data sharing



Data reuse by kernels
Small working set

Moving data back and forth for several times

Regular applications with data sharing

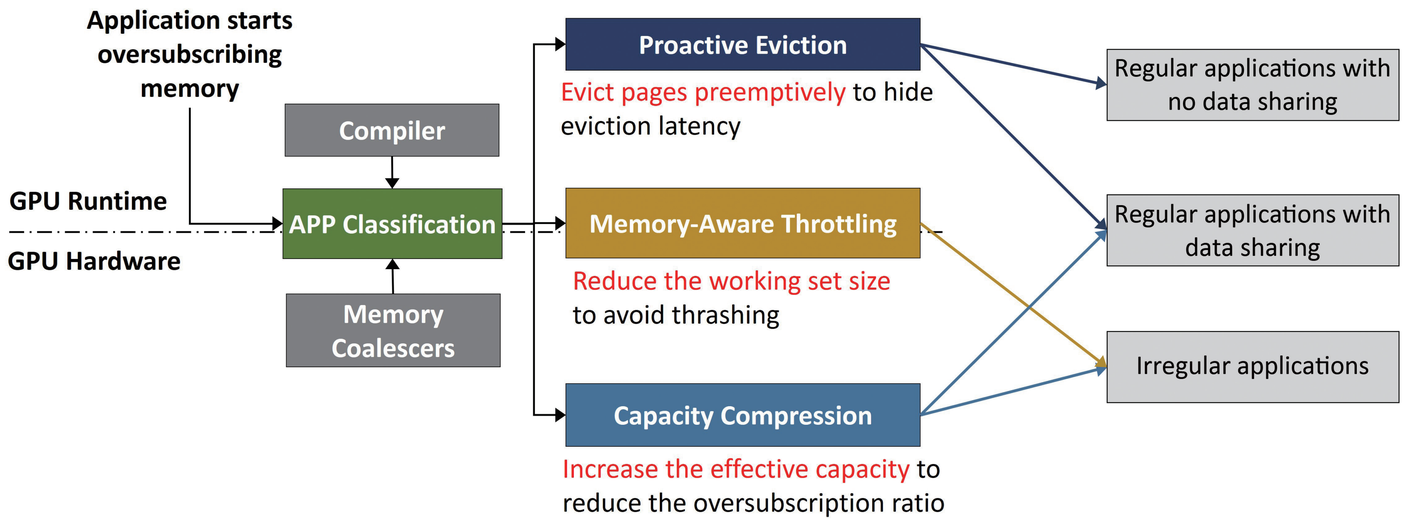


Random access
Large working set

Thrashing

Irregular applications

ETC: an application-transparent framework



No **single technique** can work for **all applications**. **ETC** dynamically selects the **most effective combination** of techniques to mitigate the **memory oversubscription overhead** in GPU.

Methodology

Simulator

- Based on Mosaic simulation platform *[MICRO'17]*, enhanced GPGPU-Sim with address translation and page table walk
- Models demand paging and memory oversubscription support

Real GPU evaluation

- NVIDIA GTX1060 GPU with 3GB memory

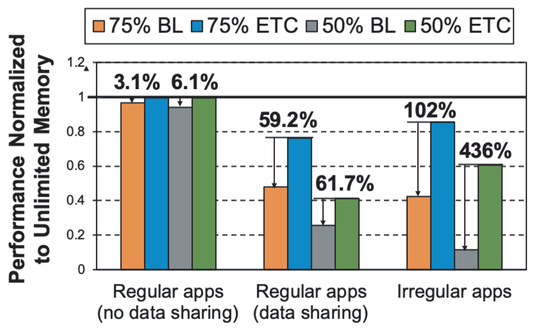
Workloads

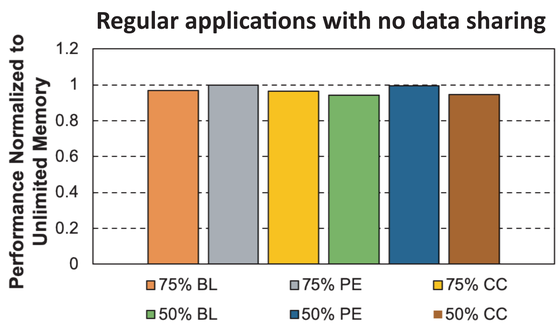
- CUDA SDK, Rodinia, Parboil, and Polybench benchmarks

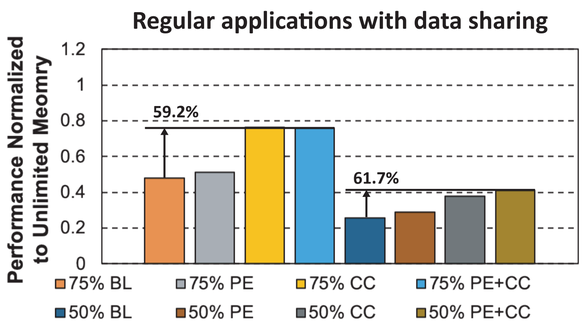
Compared baseline

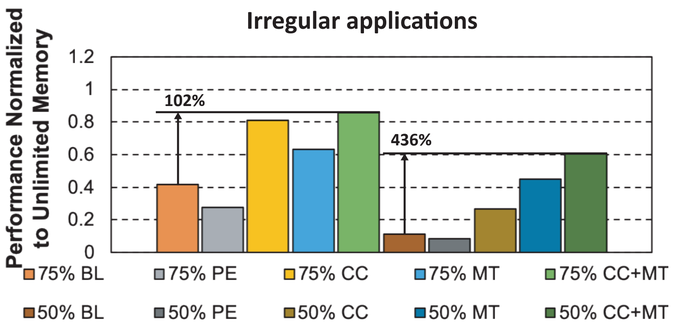
- BL: the state-of-the-art baseline with prefetching *[Zheng et al., HPCA'16]*
- An ideal baseline with unlimited memory capacity

Experimental Results









- **Fully mitigates** the oversubscription overhead of **regular applications with no data sharing**.
- Improves the performance of **regular applications with data sharing** by **60.4%**.
- Improves the performance of **irregular applications** by **270%**.

	Regular applications with no data sharing	Regular applications with data sharing	Irregular applications
Proactive Eviction (PE)	✓	✓	×
Memory-aware Throttling (MT)	×	×	✓
Capacity Compression (CC)	×	✓	✓