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
Problem

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

- **Limited memory capacity** becomes a first-order design and performance bottleneck
- **Unified virtual memory** and **demand paging** enable **memory oversubscription** support
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** Technique 1:
  - Overlap prefetch with eviction requests
Motivation

- Prior **Hand-tuning** Technique 2:
  - Duplicate read-only data

Reduce the number of evictions

**Duplicate** read-only data instead of migration

No need to evict duplicated data
Motivation

- Prior Hand-tuning Techniques:
  - Overlap prefetch with eviction requests
  - Duplicate read-only data

- Manually managing data movement
- No visibility into other VMs in cloud environment
Motivation

• Prior **Hand-tuning** Techniques:
  - Overlap prefetch with eviction requests
  - Duplicate read-only data

× **Manually** managing data movement
× **No visibility** into other VMs in cloud environment

**Application-transparent** mechanisms are urgently needed
Our Proposal

- Application-transparent Framework

**ETC Framework**
Our Proposal

- Application-transparent Framework

ETC Framework

Proactive Eviction
Our Proposal

- Application-transparent Framework

**ETC Framework**

- Proactive Eviction
- Memory-aware Throttling
Our Proposal

- Application-transparent Framework

ETC Framework

Proactive Eviction

Memory-aware Throttling

Capacity Compression
Our Proposal

Regular Applications With No Data Sharing

ETC Framework
Proactive Eviction

ETC fully mitigates the oversubscription overhead
Our Proposal

ETC improves the performance by **60.4%**
Our Proposal

Regular Applications With No Data Sharing

ETC Framework
Proactive Eviction

Regular Applications With Data Sharing

ETC Framework
Proactive Eviction

ETC Framework
Memory-aware Throttling
Capacity Compression

Irregular Applications

ETC improves the performance by 270%
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