GAIA

Glob**A**l un**I**fied p**A**ge cache for Heterogeneous systems

Tanya Brokhman

Pavel Lifshits

Mark Silberstein

Technion, Israel Institute of Technology

Tighter integration of GPU memory into the OS page cache and file I/O mechanisms is required!

mmap() for GPUs

Disciplined CPU-GPU file sharing

Improved performance
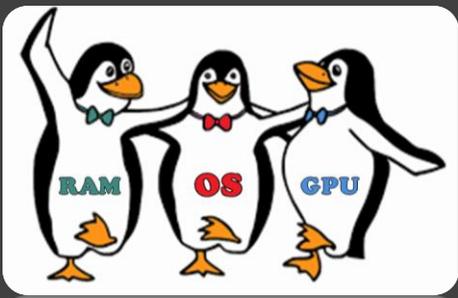
ACSL
Technion Accelerated Computing Systems Lab

THE ANDREW & ERNA VITERBI
FACULTY OF ELECTRICAL ENGINEERING

TECHNION Israel Institute of Technology

1

Consistency model considerations

GAIA

Glob**A**l un**I**fied p**A**ge cache for Heterogeneous systems

Tanya Brokhman

Pavel Lifshits
Mark Silberstein

Technion, Israel Institute of Technology

PCIe
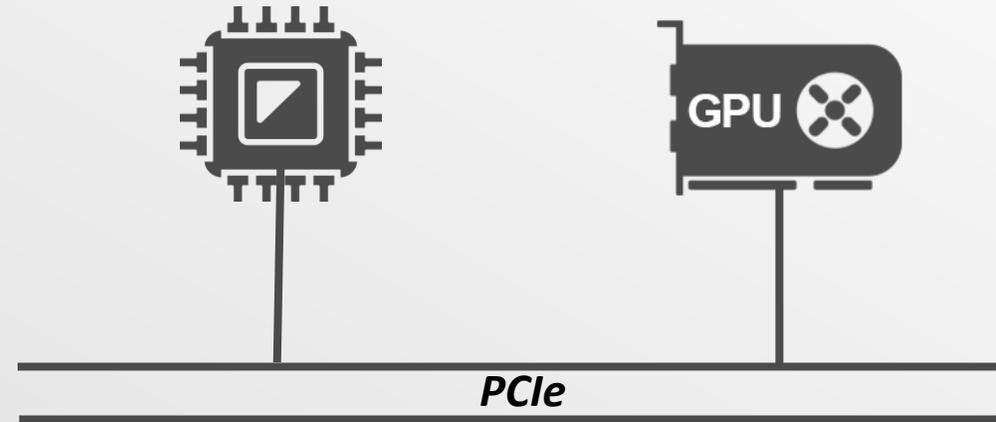
Strong consistency unsuitable!

Up to 27X performance degradation due to false sharing
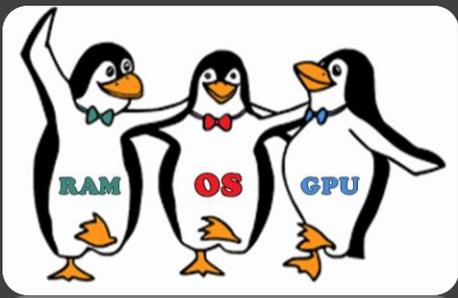
GAIA

GAIA
Global unified pAge cache for Heterogeneous systems

Tanya Brokhman

Pavel Lifshits

Mark Silberstein

Technion, Israel Institute of Technology

GAIA
A distributed, weakly-consistent page cache architecture for heterogeneous multi-GPU systems

✓ Implements Lazy Release Consistency

✓ Extends OS Page cache to GPU memory

✓ New CPU and GPU I/O optimizations
   ✓ Up to 3X speedup in CPU file I/O
   ✓ Up to 8X speedup in unmodified realistic workloads

USENIX 2019, 4:35 PM,

Track 1, on Thursday July 11

ACSL
Technion Accelerated Computing Systems Lab

THE ANDREW & ERNA VITERBI
FACULTY OF ELECTRICAL ENGINEERING

TECHNION
Israel Institute of Technology