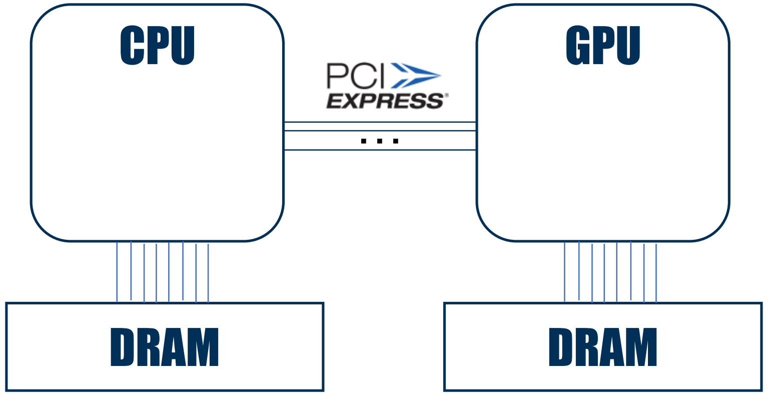
Trace Generation of Machine Learning Workloads with GTReplay for Intel Integrated-GPU Modeling

Jaewon Lee, (Georgia Tech)
Konstantin Levit-Gurevich, (Intel)
Sunpyo Hong, (Intel)
Hyesoon Kim (Georgia Tech)

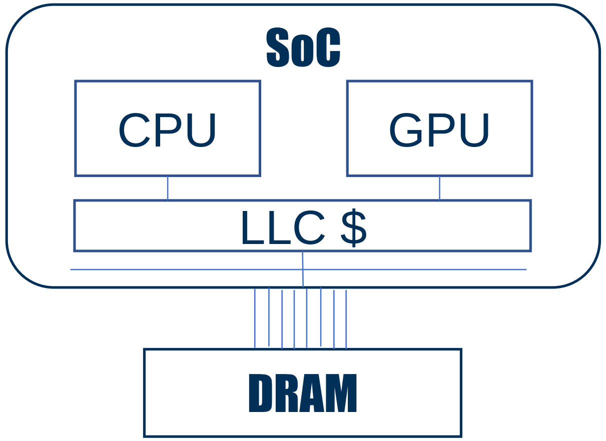
Georgia Tech

Introduction

- Integrated GPUs are prevalent. (2021 GPU market share of Intel iGPU: 68.3%)
  https://www.statista.com/statistics/754557/worldwide-gpu-shipments-market-share-by-vendor/
  - Cheap and small packaging size
  - ML with edge devices
- However, evaluating the performance of iGPUs is still hard
  - Have focused on discrete GPUs
- Goal
  - iGPU simulation environment driven by the traces of actual machine learning workload



(a) Discrete GPU System



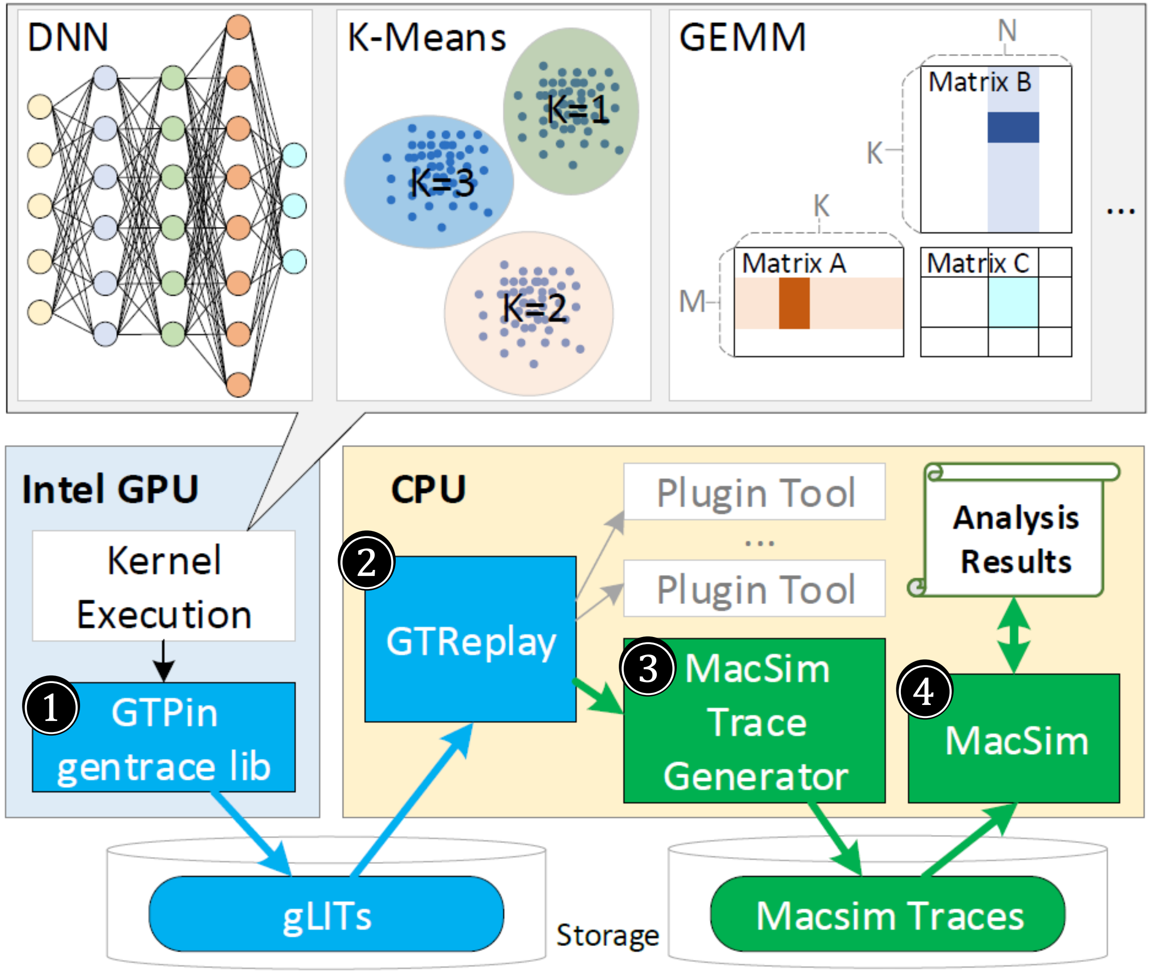
(a) SoC-style GPU System

< Benefit >

- less IOs
- DRAM sharing
- Direct comm. btw CPU and GPU

GTPin-Macsim Simulation Flow

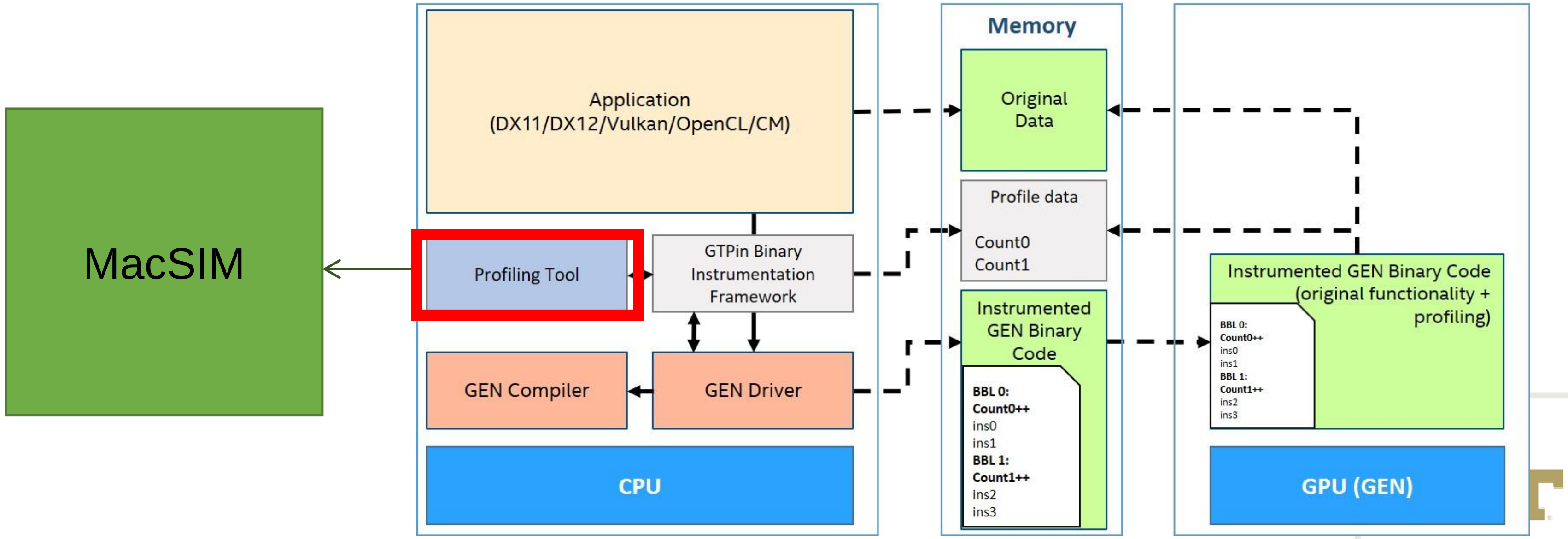
1. Generate a trace with GTPin in gLITs format while running ML workloads on target iGPU
2. Analyze the gLITs with GTReplay
   - Interpret memory instructions
3. Generate MacSim trace With MacSim trace generator plug-in for GTReplay
4. Evaluate performance on MacSim for generated MacSim traces



gLITs: Long Instruction Trace for GEN

GTPin and GTReplay

- GTPin
  - dynamic binary instrumentation framework for GEN (Intel graphics) Architecture
  - Generates traces by using gentrace()
- GTReplay
  - a GEN emulator allowing replaying special trace generated by GTPin
  - User can develop flexible analysis tools on top of GTReplay
- Both are open to public



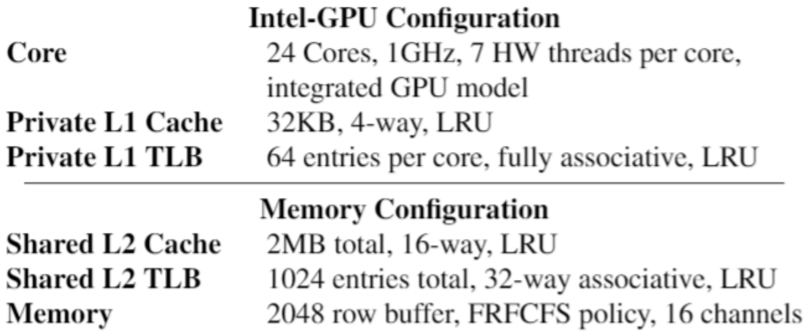
https://software.intel.com/sites/landingpage/gtpin/index.html

MacSim Simulator

- A cycle-level, heterogeneous architecture simulator for x86, ARM, NVIDIA PTX, and Intel GPU instructions
- Can be configured as either a trace driven or execution-drive cycle level simulator
- Support performance evaluation and architecture exploration with various statistical results

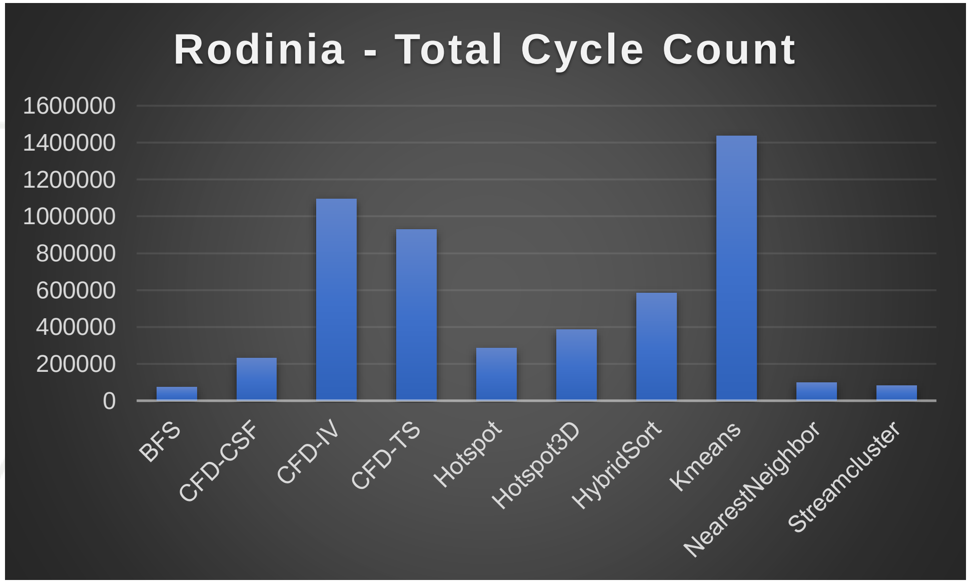
Simulation Results

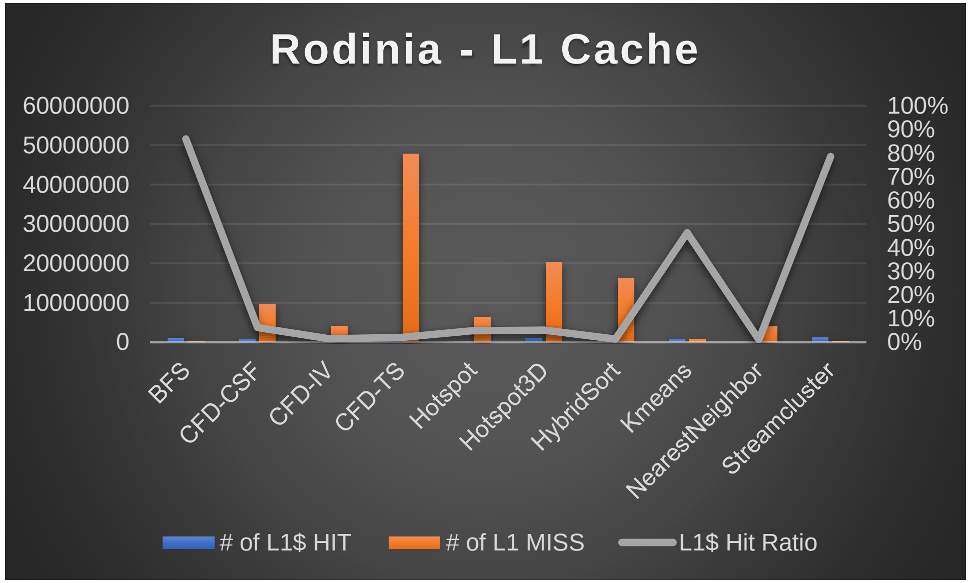
- System Configuration



	Intel-GPU Configuration
Core	24 Cores, 1GHz, 7 HW threads per core, integrated GPU model
Private L1 Cache	32KB, 4-way, LRU
Private L1 TLB	64 entries per core, fully associative, LRU
	Memory Configuration
Shared L2 Cache	2MB total, 16-way, LRU
Shared L2 TLB	1024 entries total, 32-way associative, LRU
Memory	2048 row buffer, FRFCFS policy, 16 channels

- Rodinia Simulation Results


Rodinia - Total Cycle Count
BFS
CFD-CSF
CFD-IV
CFD-TS
Hotspot
Hotspot3D
HybridSort
Kmeans
NearestNeighbor
Streamcluster


Rodinia - L1 Cache
BFS
CFD-CSF
CFD-IV
CFD-TS
Hotspot
Hotspot3D
HybridSort
Kmeans
NearestNeighbor
Streamcluster
# of L1$ HIT
# of L1 MISS
L1$ Hit Ratio