A Fast Locality Simulator for GEMM Design-Space Exploration on Multi-Chiplet GPUs

Euijun Chung, Hyesoon Kim
School of Computer Science, Georgia Institute of Technology
ModSim 2026: Workshop on Modeling & Simulation of Systems and Applications

HPArch
@ Georgia Tech

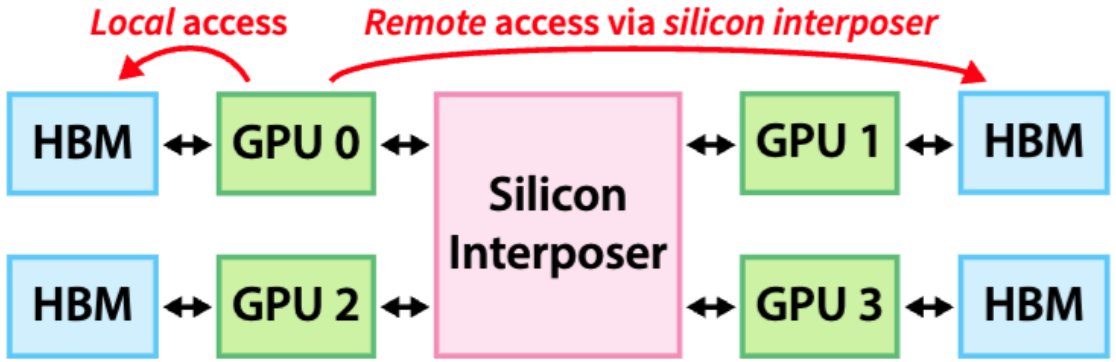
Georgia Tech

Multi-Chiplet GPUs & Chiplet Locality

- **Multi-chiplet GPUs** have multiple NUMA regions (E.g., AMD MI300X).

Remote HBM accesses exhibits:

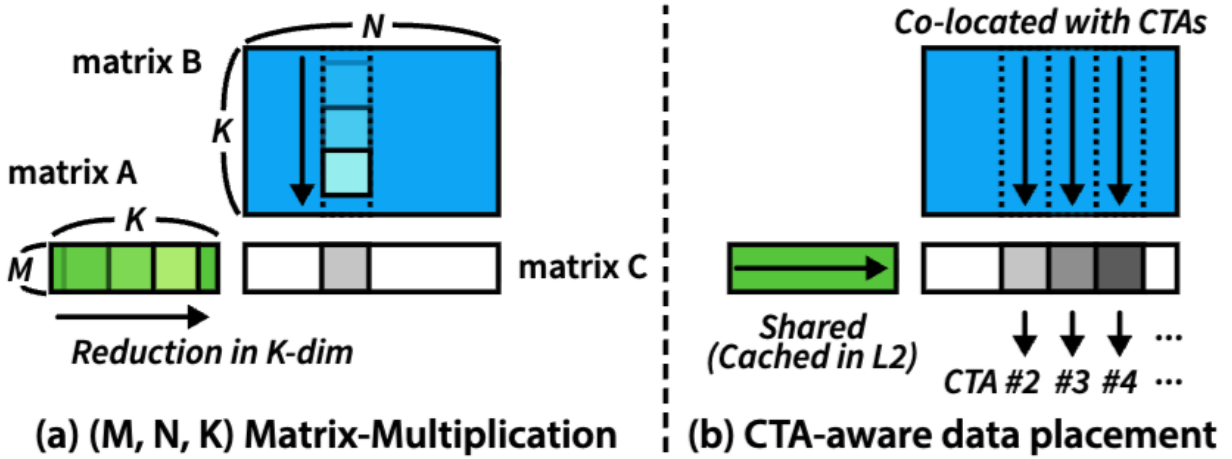
1) Higher energy consumption
2) Longer access latency
3) Inter-chiplet BW consumption



- **Locality-aware data & CTA placement**:
  - Co-location of data and compute to minimize remote HBM traffic
  - Supported by changing memory allocation and CTA scheduling
  - In GEMMs, best locality-aware strategy can reduce remote traffic by up to 58x

Problem: Optimal locality for GEMM is non-trivial

- **GEMM** is a dominant operation in **LLM inference & training**.
  - E.g., Attention weight (QKV) projections & Feed-forward network (FFN)



- However, numerous configs make locality-optimal placement hard.
  - (M, N, K) **dimensions** and (T_M, T_N, T_K) **tile sizes**
  - **Data layout** (e.g., Row-major or Column-major) for each matrix
  - **Data access pattern** from CTA traversal order ⇔ L2 cache behavior

Georgia Tech

Our Solution: A Fast Functional Locality Simulator

- **CTA-granularity** GPU modeling:
  - Per-chiplet L2 cache slices
  - Wave*-synchronous execution model
  - Performance & power with roofline analysis
  - **Data placement**: 4KB interleaving / coarse-grained** / fine-grained**

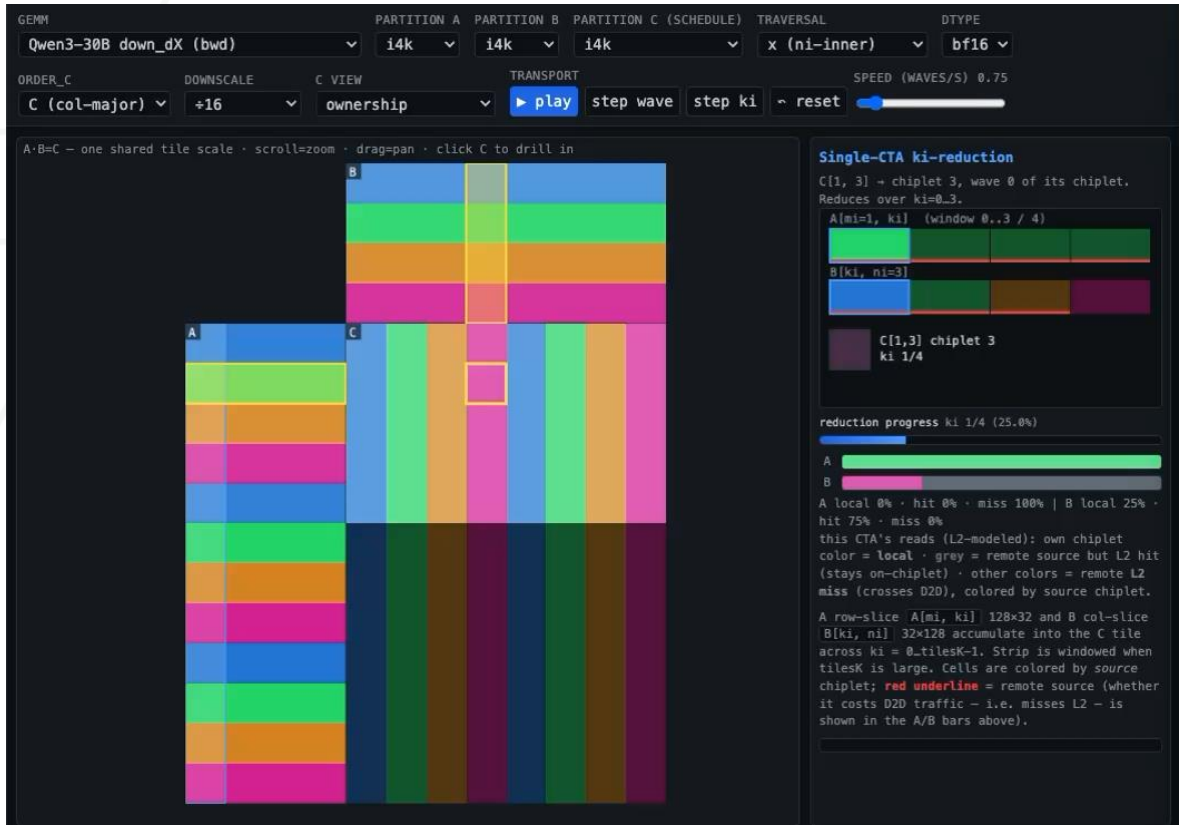
- Input & Output parameters:
  - **GEMM dimensions** (M, N, K) and **tile sizes** (T_M, T_N, T_K)
  - **Data layout**: Row- / Column-major layout
  - **CTA traversal order**: X / Y / 2D block-swizzle
  - **Outputs**: L2 hit rate, local & remote HBM traffic (GB), performance & power

*Wave: the set of CTAs concurrently resident across all SM/CUs, bounded by per-SM/CU occupancy
**[MICRO '20] Locality-Centric Data and Threadblock Management for Massive GPUs

Georgia Tech

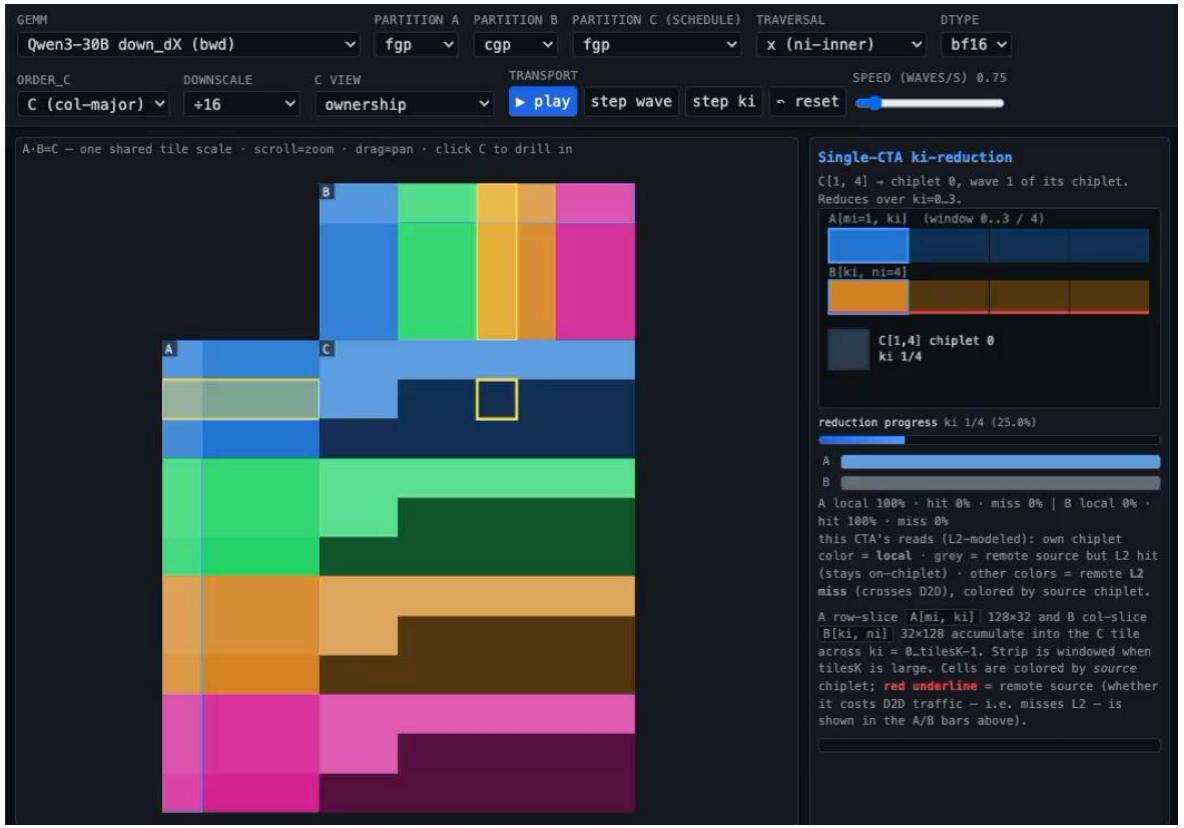
Visualization of the Locality Simulator

4KB-granularity data interleaving



A: 100% remote, but cached to L2
B: 25% remote

Locality-aware data & CTA placement

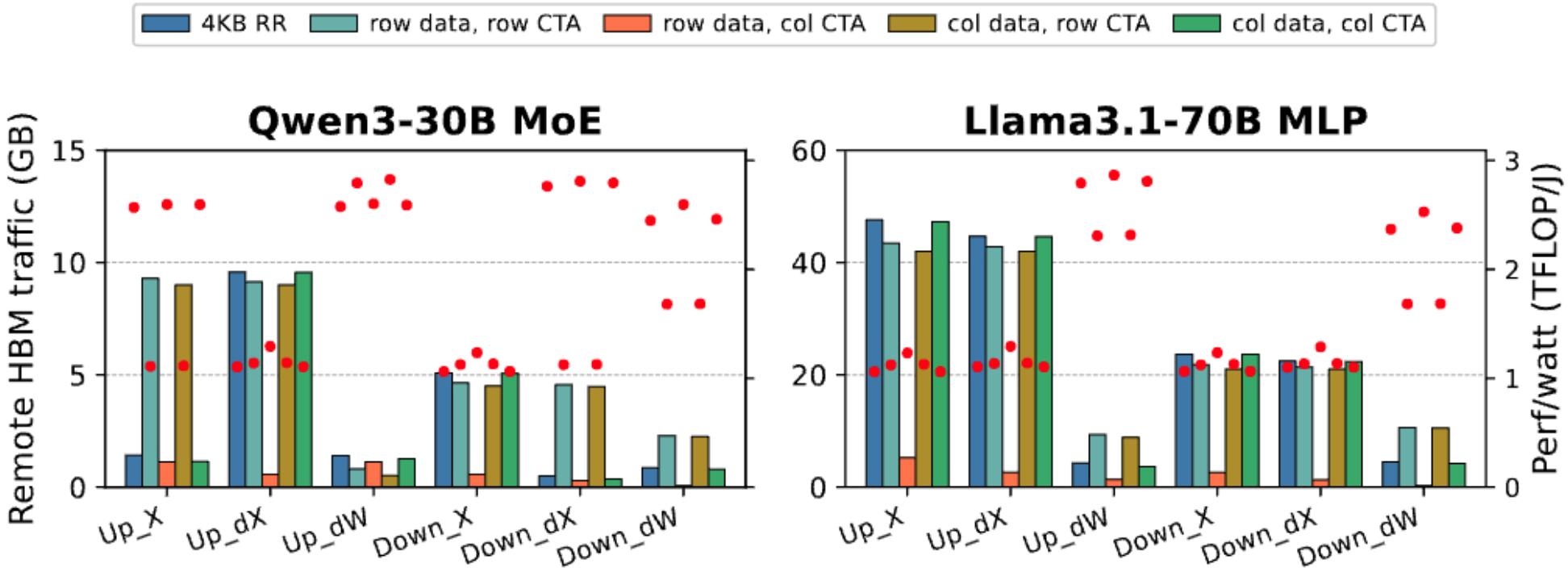


A: 100% local, cached to L2
B: 100% remote, but cached to L2

chiplet 0 chiplet 1 chiplet 2 chiplet 3 waiting computing done

Georgia Tech

Results: Locality & Performance per Power



- Locality-aware placement reduces **remote traffic by up to 18.3x**, and improves **perf-per-power by up to 1.5x** over 4KB round-robin interleaving.
- A **mis-aligned partition**, however, inflates remote traffic to as much as 9.3× that of 4KB RR, which **our simulator can rapidly identify**.

Georgia Tech