NeuPIMs: NPU-PIM Heterogeneous Acceleration for Batched LLM Inferencing

Guseul Heo
Sangyeop Lee
Jaehong Cho
Hyunmin Choi
Sanghyeon Lee
Hyungkyu Ham[†]
Gwangsun Kim[†]
Divya Mahajan[§]
Jongse Park

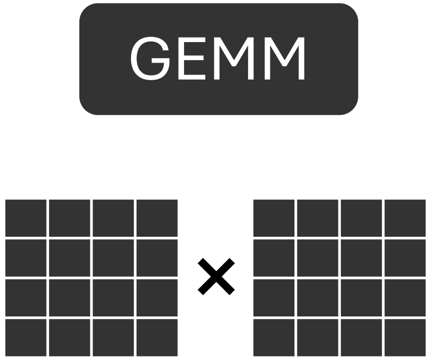
KAIST

POSTECH[†]

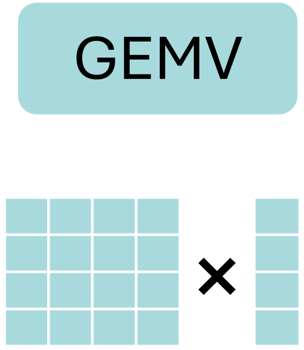
Georgia Institute of Technology[§]

ASPLOS 2024

LLM batched inference comprises **GEMM** and **GEMV**


GEMM

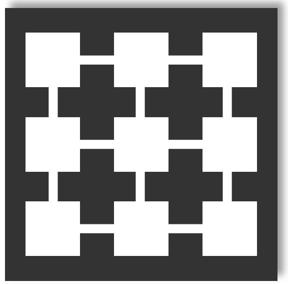
matrix-matrix multiplication


GEMV

matrix-vector multiplication

GEMM

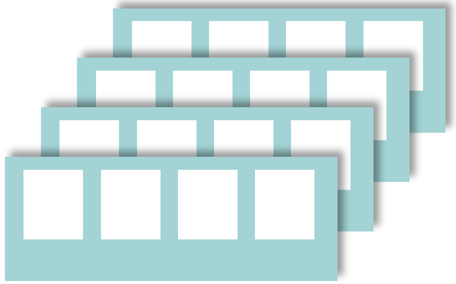
- compute-intensive
- well-suited to NPU



NPU

GEMV

- bandwidth-intensive
- well-suited to PIM



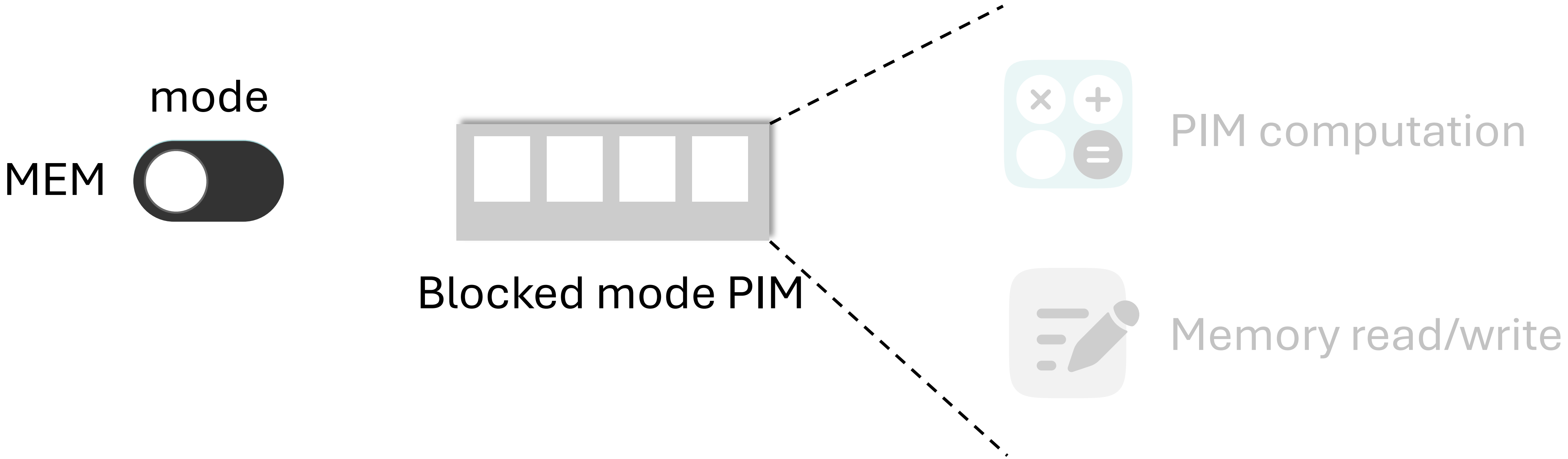
PIM

However, with naïve NPU+PIM integration,

system suffers from resource underutilization

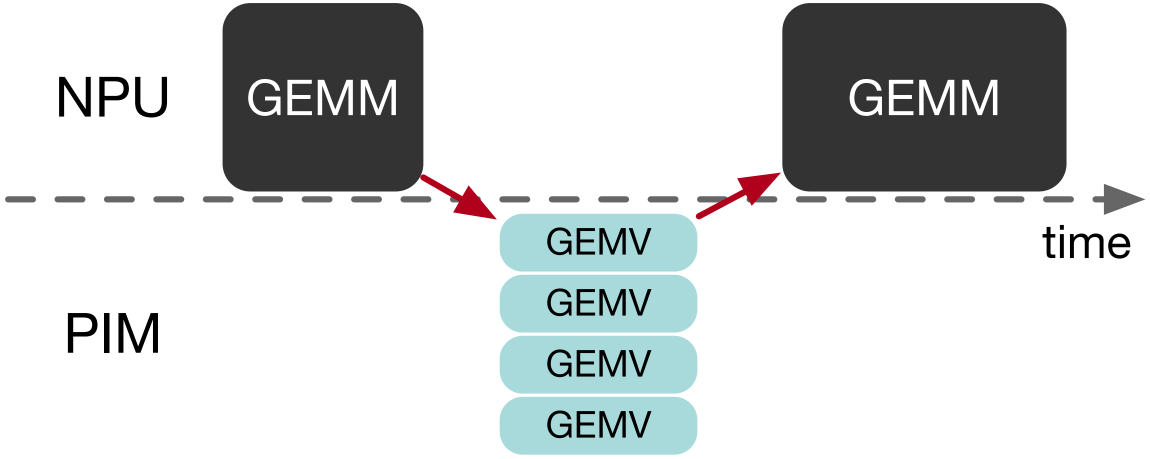
Challenge #1

Existing PIM operates in “blocked” mode

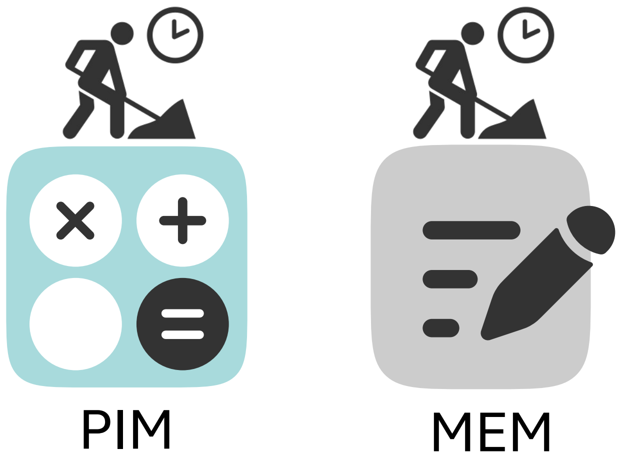


Challenge #2

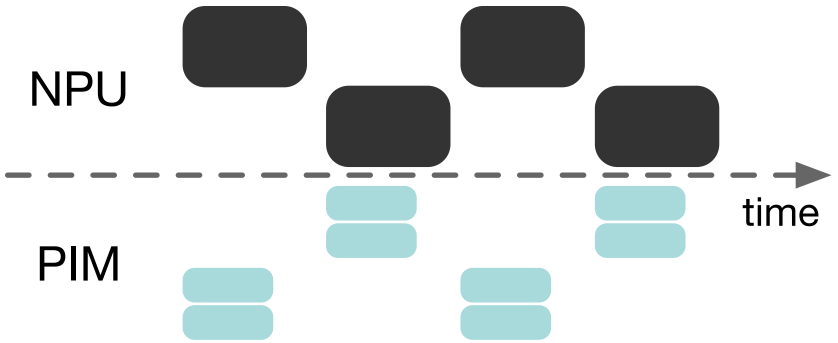
GEMM and GEMV have algorithmic dependency in LLM



We devise **NeuPIMs**, NPU-PIM heterogeneous acceleration solution for batched LLM inferencing



PIM with dual row buffers



Sub-batch interleaving

NeuPIMs: NPU-PIM Heterogeneous Acceleration for Batched LLM Inferencing

Session 6B

04/30 (Tue)

14:30

ASPLOS 2024

2.4×

throughput improvement over NPU

1.6×

throughput improvement over naïve NPU+PIM

Our simulator code is available

https://github.com/casys-kaist/NeuPIMs