[Position] TensorPRAM: Designing a Scalable Heterogeneous Deep Learning Accelerator with Byte-addressable PRAMs

Sangwon Lee, Gyuyoung Park, Myoungsoo Jung

Computer Architecture and Memory systems Laboratory

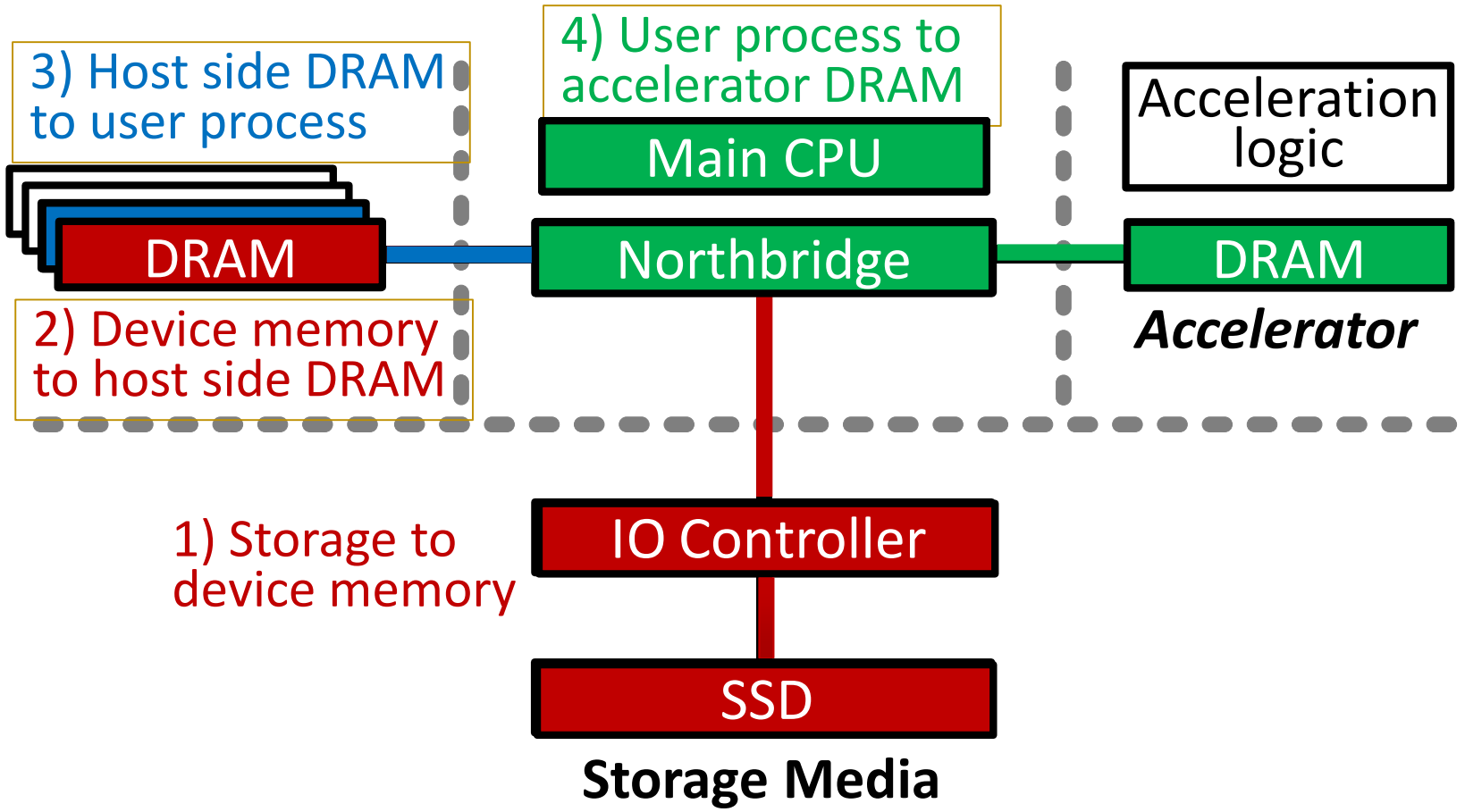
KAIST EE

CAMELab

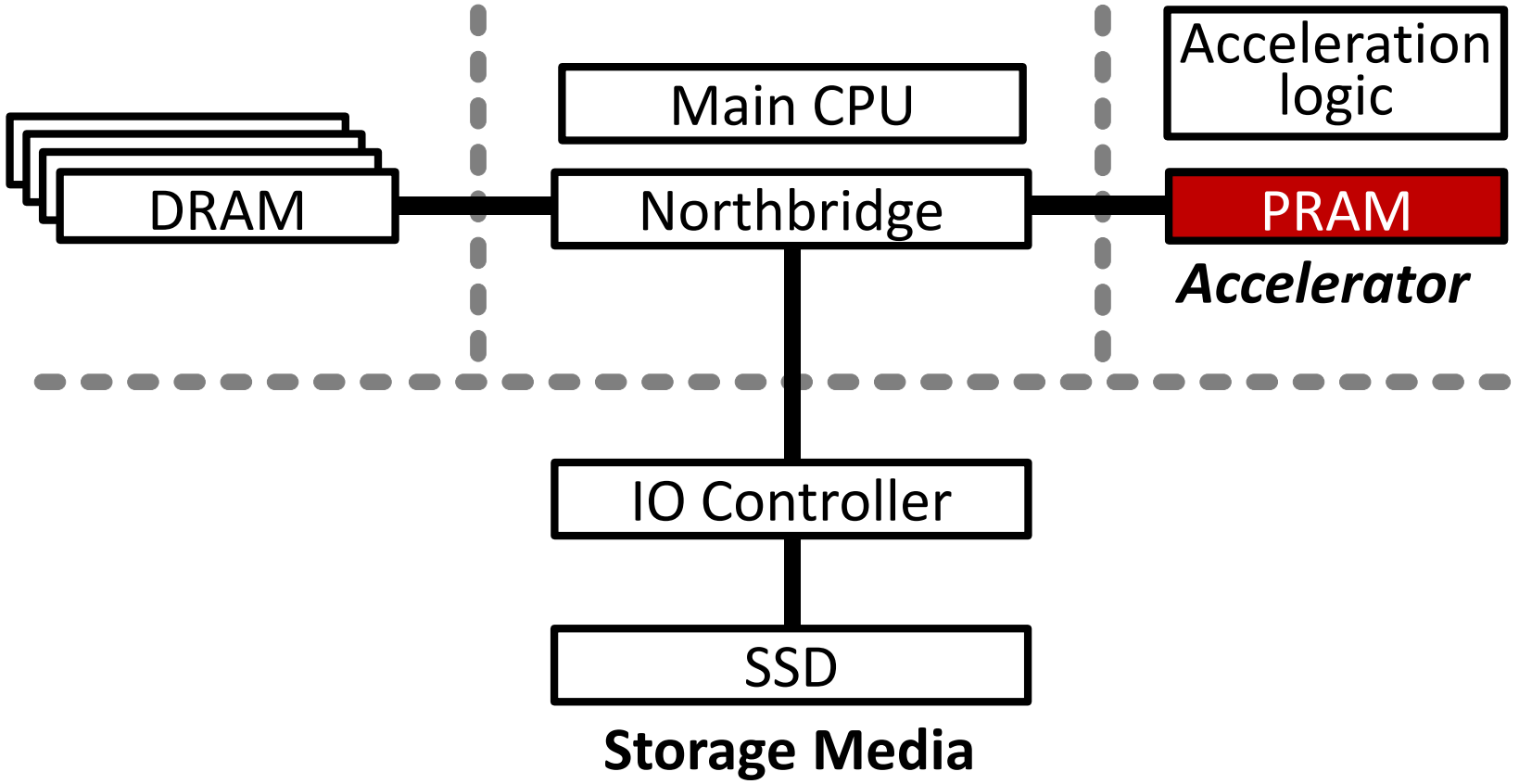
Introduction

- Background
  - Deep neural networks (DNNs) have gained significant attention for a wide spectrum of domains:
    - Visual understanding
    - Speech perception
    - Automated reasoning
  - To accelerate the computation of DNNs, various kinds of approaches are suggested.
  - FPGA-based hardware accelerations are one of promising solutions.

Challenge: Data movement overhead

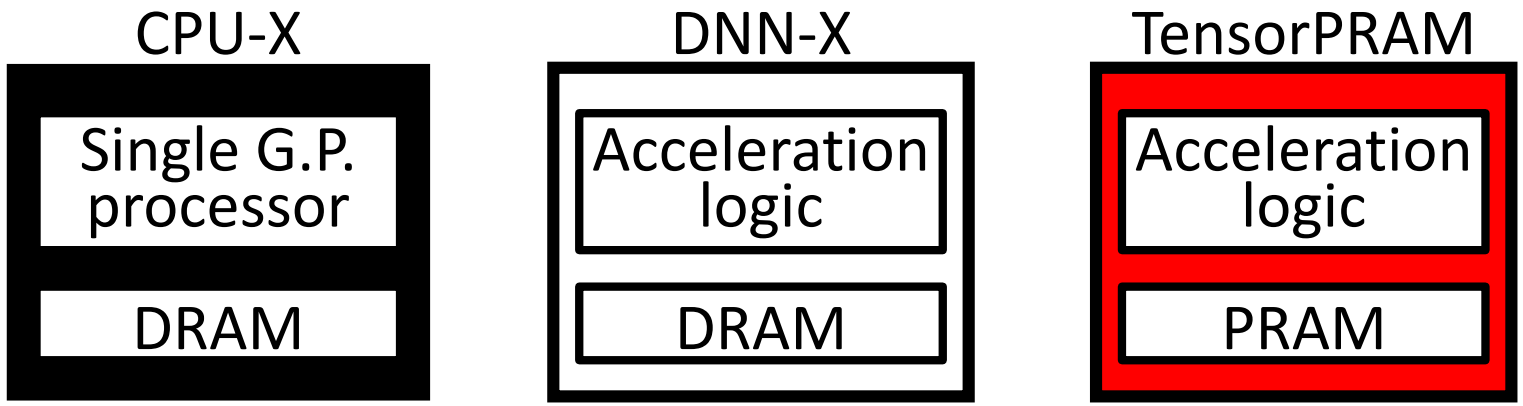


Solution: Reduction of data movement overhead



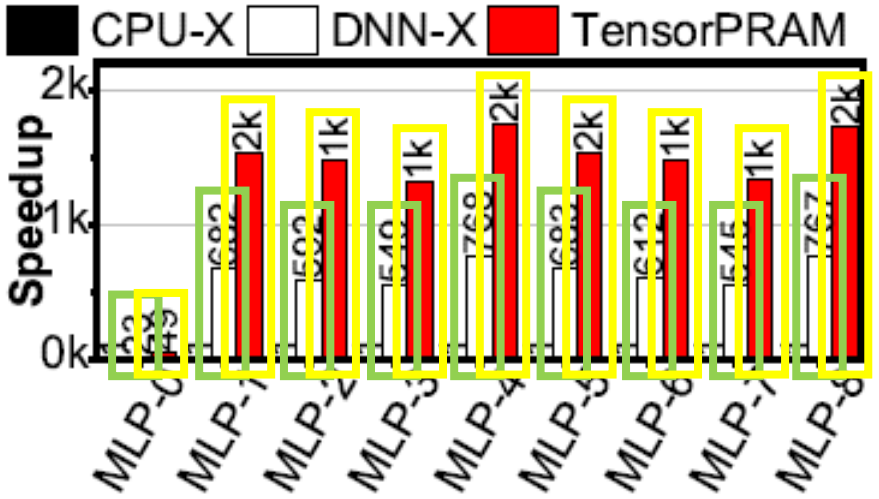
Evaluation

- Compared the TensorPRAM with other two types of accelerators



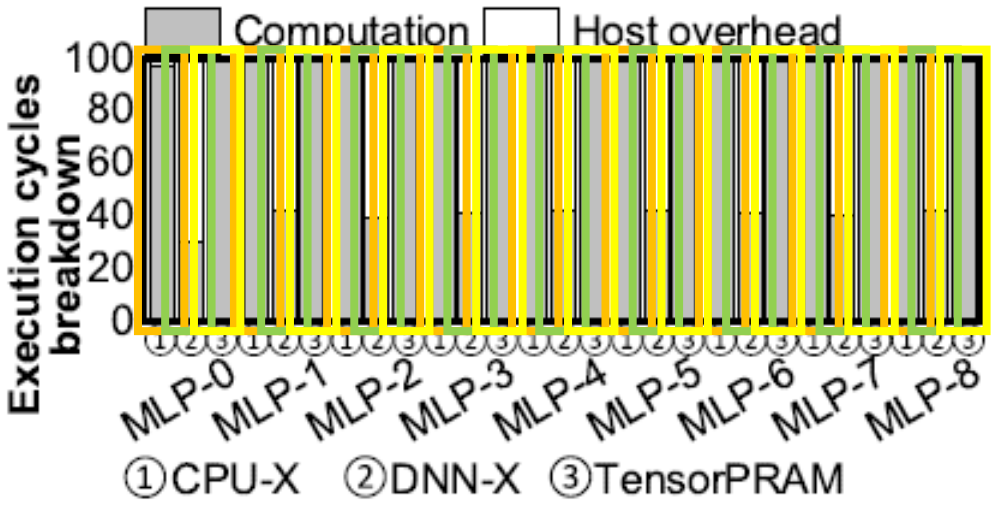
- Workloads – MLP (multi-layer perceptron)

Evaluation: Execution Cycle Comparison



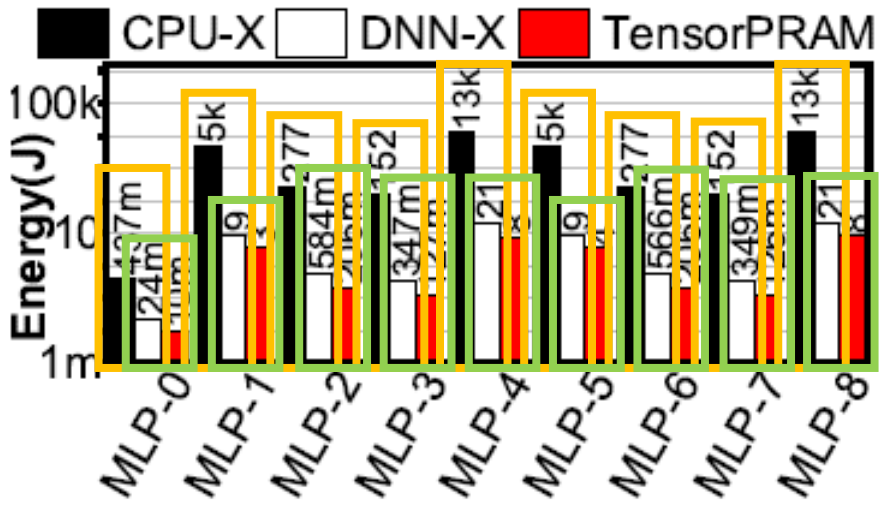
- All results of execution cycles are normalized to those of **CPU-X**
- **DNN-X** improves the performance by **580X**, on average, compared to **CPU-X**.
- **TensorPRAM** further improves the performance by **1350X**, on average, compared to **CPU-X**.

Evaluation: Execution Cycle Breakdown



- **CPU-X** spends most of cycles to perform matrix multiplications.
  - The host overhead looks relatively small.
- The host overhead of **DNN-X** accounts for **60%** of the total execution cycles.
- The **TensorPRAM** greatly reduce the host overheads.

Evaluation: Energy Consumption



- **DNN-X** reduces the average energy consumption by **99%**, compared to **CPU-X**.
- **TensorPRAM** reduces the average energy consumption by **61%**, compared to **DNN-X**.

Conclusion

- We present TensorPRAM, a deep learning accelerator implemented on a FPGA with an integration of PRAMs.
- TensorPRAM outperforms conventional DRAM-based accelerators in the execution cycles and energy efficiency.
- Reduction of the data movement overhead in TensorPRAM contributes to the performance gain and lower energy consumption.

Thank you for listening!