ODTÜ
METU

TÜBİTAK

EURO 4SEE

Optimizing Deep Learning Systems for Hardware

Assoc. Prof. Erdem AKAGÜNDÜZ, METU

✉ ncc@ulakbim.gov.tr

Part I : Fundamentals

- PI.a : Why hardware matters in deep learning?
- PI.b : Performance metrics
- PI.c : Case Study: Edge Devices vs Datacenter vs Supercomputers

Case Study

- Let’s see how the types of computations we discussed manifest in different deployment scenarios.
  - Edge Devices vs
  - Data Centers / Cloud vs (many users simultaneously)
  - Supercomputers / HPC Clusters (to solve difficult problems)

Case Study: Edge Devices

- 1. Edge Devices (e.g., mobile, IoT, embedded AI)
  - Use case:
    - real-time inference,
    - low-latency tasks,
    - small models
  - Performance needs:
    - Low latency → fast single-sample processing
    - Low energy / power consumption
    - Moderate throughput (small batch sizes)

Case Study: Edge Devices

- 1. Edge Devices (e.g., mobile, IoT, embedded AI)
  - Dominant computation types:
    - Matrix multiplications for inference (small/medium matrices)
    - Memory read/write optimized for limited cache
    - Minimal branching
    - complex control flow
      - *Recursion or dynamic graph execution in neural networks etc.*

Case Study: Edge Devices

- 1. Edge Devices (e.g., mobile, IoT, embedded AI)
    - Limitations:
        - Cannot scale to very large models or datasets
        - Limited memory / cache → restricts model size
        - Lower raw FLOPs → slower for heavy computations

Case Study: Data Centers / Cloud

- 2. Data Centers / Cloud
  - Capabilities:
    - Large-scale model training
    - Batch inference for many users
    - High throughput, balanced latency
  - Dominant Computation Types:
    - Large matrix/tensor multiplications
    - Reductions for gradient accumulation
    - Memory read/write; inter-device communication
    - Some branching in dynamic models

Case Study: Data Centers / Cloud

- 2. Data Centers / Cloud
    - Limitations:
        - Latency per single sample is higher.
        - Energy cost can be significant at scale
        - Deployment complexity: managing multi-GPU or multi-node setups

Case Study: Supercomputers / HPC Clusters

- 3. Supercomputers / HPC Clusters
  - Capabilities:
    - Extreme throughput for massive datasets
    - Efficient distributed training across many nodes
    - High precision and mixed-precision arithmetic
  - Dominant Computation Types:
    - Massive parallel matrix/tensor multiplications
    - Reductions and synchronizations across nodes
    - Memory-intensive operations
    - Mixed-precision arithmetic to maximize FLOPs

Case Study: Supercomputers / HPC Clusters

- 3. Supercomputers / HPC Clusters
  - Limitations:
    - High energy consumption
    - Very high cost
    - Complex software stack and maintenance
    - Latency for single sample may be high → not suitable for real-time inference

Next: Part II

- Part I : Fundamentals
- **Part II : Hardware Types & Memory Hierarchy**
- Part III : Model-Level Optimizations
- Part IV : System-Level Optimizations
- Part V : Introduction to Scaling Deep Learning in HPC

Thanks!

Co-funded by
the European Union

EuroHPC
Joint Undertaking

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101191697. The JU receives support from the Digital Europe Programme and Germany, Türkiye, Republic of North Macedonia, Montenegro, Serbia, Bosnia and Herzegovina.