ODTÜ METU

TÜBİTAK

EURO 4SEE

Optimizing Deep Learning Systems for Hardware

Assoc. Prof. Erdem AKAGÜNDÜZ, METU

ncc@ulakbim.gov.tr

Part V : Final Notes

- PV.a : Introduction to DeepSpeed
- PV.b : Conclusions and Directions

Our aim was to…

- Understand why hardware matters in deep learning efficiency
- Explore fundamentals of compute & memory bottlenecks
- Learn model-level and system-level optimizations
- Connect hardware, algorithms, and software frameworks
- Gain a starting point for scaling deep learning in HPC

What we covered:

- Fundamentals: training vs inference bottlenecks, performance metrics, case studies
- Hardware & Memory: CPU/GPU/TPU/FPGA/ASIC, hierarchy, bandwidth, precision
- Model-Level Optimizations: pruning, quantization, knowledge distillation, efficient architectures
- System-Level Optimizations: parallelism, mixed precision, runtime/system optimizations
- Scaling for DL: Introduction to DeepSpeed and HPC training setups

Next Steps:

- Hands-on with DeepSpeed
  - Start small: run with 2 GPUs to understand scaling trade-offs
  - Explore the config file: optimizer states, zero redundancy (ZeRO), precision settings
- Experiment with Mixed Precision: compare fp32, fp16, and bfloat16
- Profiling: measure FLOPs, latency, throughput on your own models
- Explore model compression: try pruning or quantization on a pretrained model
- Study trade-offs: edge vs datacenter vs HPC environments
- Keep updated: follow hardware and framework developments (NVIDIA, AMD, Google, Microsoft DeepSpeed, PyTorch, etc.)

Final Words:

- Deep learning efficiency is as much about systems as it is about models
- Scaling requires balancing hardware, memory, and algorithmic trade-offs
- The best path: experiment, profile, iterate

  - Think hardware-aware,
  - Learn to optimize system-level,
  - Scale smart.

EURO 4SEE

Thanks!

"I find your lack of precision… fantastic!"

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.