ODTÜ
METU

TÜBİTAK

EURO 4SEE

Optimizing Deep Learning Systems for Hardware

Assoc. Prof. Erdem AKAGÜNDÜZ, METU

ncc@ulakbim.gov.tr

Part II : Hardware & Memory Hierarchy

- PII.a : Existing Solution: CPU, GPU, TPU, FPGA, ASIC basics
- PII.b : Memory hierarchy, bandwidth bottlenecks, movement costs
- PII.c : Precision

Numerical Precision

- Numerical precision refers to the number of bits used to represent a number in computation.
- Common formats in deep learning: FP32 → FP16/BF16 → INT8.
- Precision affects:

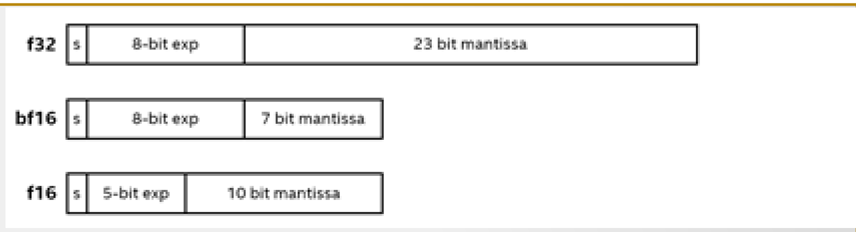
  ↑ Memory footprint

  ↑ Computation cost

  ↑ Energy consumption

  ↓ Accuracy of the model

Numerical Precision



Format	Bits	Use Case	Pros	Cons
FP32	32	Standard training	High accuracy	High memory & energy cost
FP16 / BF16	16	Mixed-precision training	Reduced memory & energy, higher throughput	Slight accuracy degradation
INT8	8	Inference	Minimal memory & energy, very high throughput	Needs quantization, may lose accuracy

Why change precision?

- Scientific computing and HPC often require FP64 (double precision):
  - e.g.: Climate simulations, fluid dynamics, molecular modeling
- Sensitive numerical operations where rounding errors accumulate

- However, Training of DL models usually requires FP32 for gradients,
  - **but** forward pass may allow lower precision.

Quantization

- Lower precision increases throughput and reduces memory pressure.
- But lowes accuracy
- May require **quantization-aware training** to preserve accuracy.
- Hardware-aware design:
    - Use FP32 for sensitive computations (gradients)
    - FP16/INT8 for forward pass and inference
    - Goal: balance accuracy, energy, and speed.

Precision choice

- Precision choice is workload- and hardware-dependent.
- Lower precision improves energy efficiency, throughput, and memory usage.
- High precision is critical for scientific accuracy.
- Mixed strategies are often the optimal compromise.
  - A quantization research.

Next: Part III

- 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.