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

Performance Metrics

- 1. FLOPs (Floating Point Operations per Second)
  - Measures the raw computational capability of hardware.
  - Indicates how many arithmetic operations hardware can perform per second.
  - Useful for comparing different hardware or estimating theoretical performance.

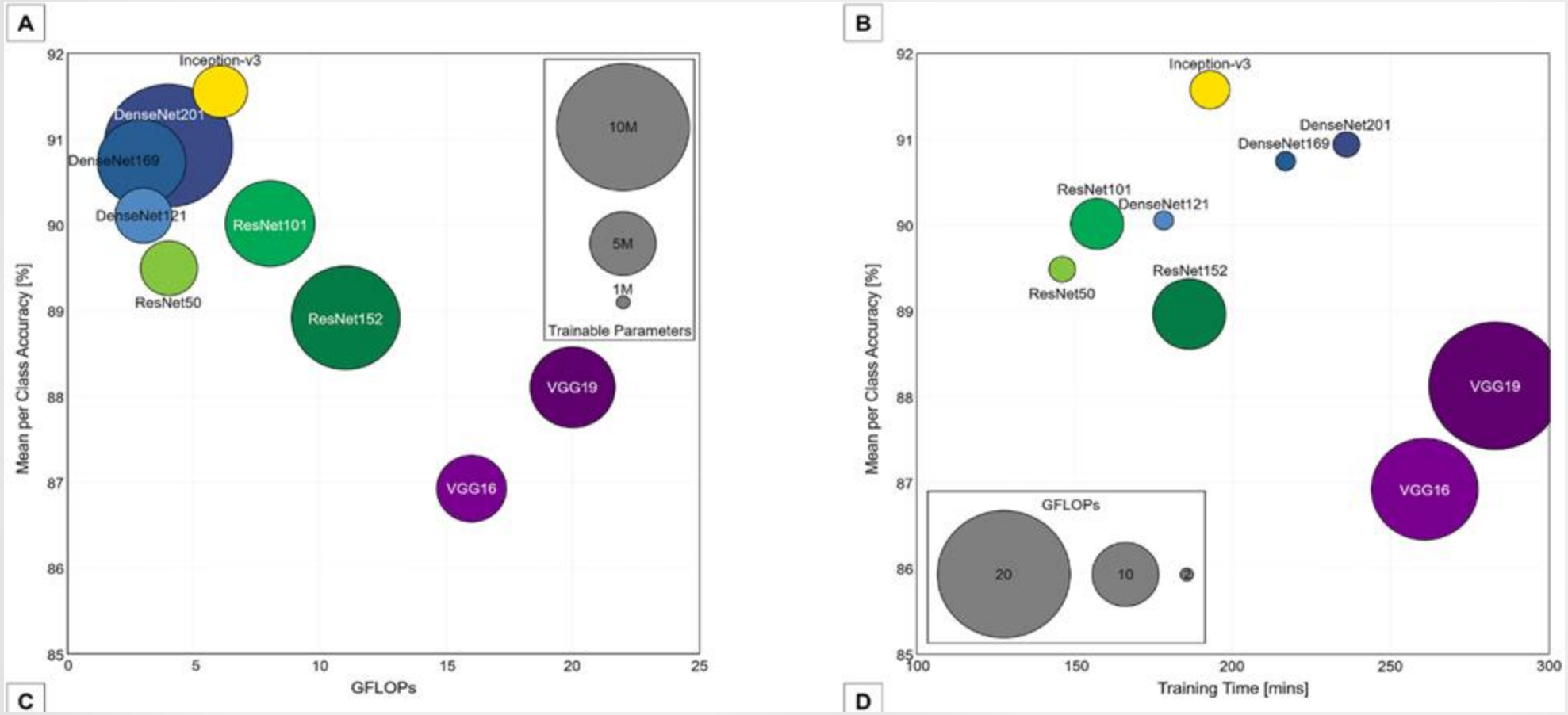
Performance Metrics

- 1. FLOPs (Floating Point Operations per Second)
  - Measures the raw computational capability of hardware.
  - Indicates how many arithmetic operations hardware can perform per second.
  - Useful for comparing different hardware or estimating theoretical performance.

Performance Metrics

- 1. FLOPs (Floating Point Operations per Second)
  - When people say “X GFLOPs” for a model or hardware, they usually mean the total number of multiply-add operations (MACs) required for a forward pass, aggregated across all matrix/tensor operations.
  - Logic/branching and memory ops are not included in FLOPs.

Performance Metrics

- 

Harriet L. Dawson, Olivier Dubrule, Cédric M. John, "Impact of dataset size and convolutional neural network architecture on transfer learning for carbonate rock classification", Computers & Geosciences, vol. 171, 2023.

Performance Metrics

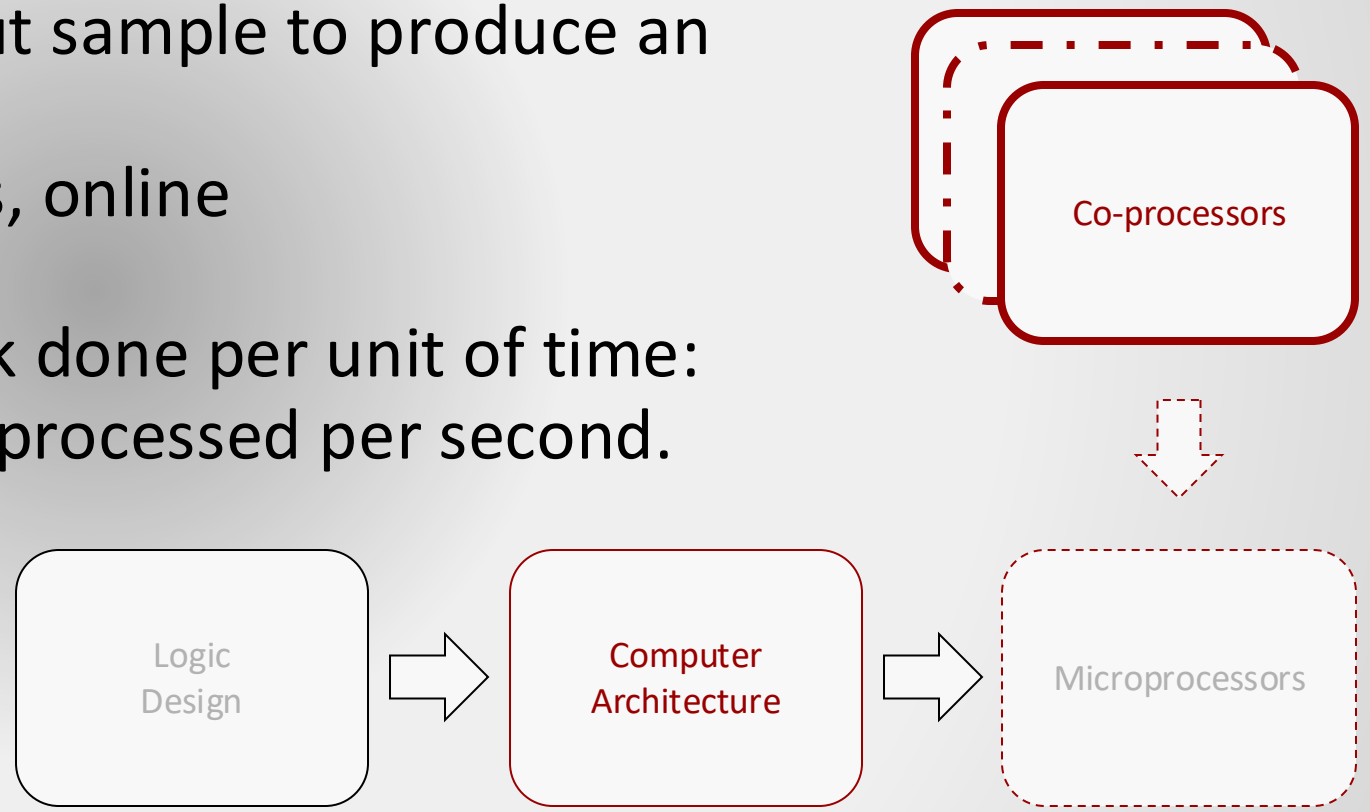
- 2. Latency
  - Time to process a single input sample.
  - Critical for real-time inference.

Performance Metrics

- 3. Throughput
  - Number of samples processed per unit time (e.g., images/sec)
    - in parallel?
  - Important for training efficiency and batch processing.

latency vs throughput? (do not use the word fast!)

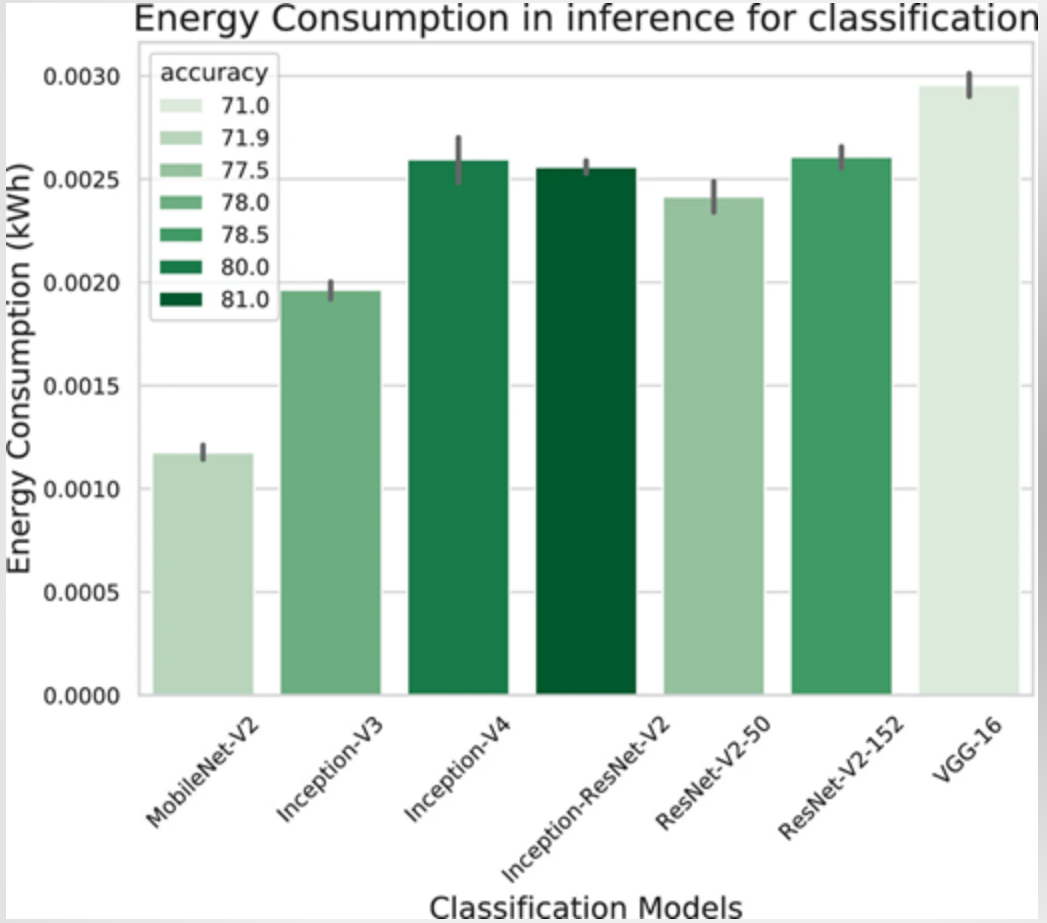
- Latency: The time it takes to complete a single task or operation from start to finish:
  - how long it takes for one input sample to produce an output.
  - (e.g., edge inference, robotics, online recommendations).
- Throughput: The amount of work done per unit of time:
  - how many input samples are processed per second.
  - (e.g. batch processing etc.)



Performance Metrics

- 4. Energy / Power Efficiency
  - Total energy consumed to perform computation (Joules per operation or per sample).
  - Key for edge devices or large-scale HPC where energy cost is significant.

Performance Metrics



Carastan-Santos, D., Pham, T.H.T. (2022). Understanding the Energy Consumption of HPC Scale Artificial Intelligence. In: Navaux, P., Barrios H., C.J., Osthoff, C., Guerrero, G. (eds) ***High Performance Computing****. CARLA 2022.*

Performance Metrics

- 5. Cost / Price-Performance
  - Hardware acquisition cost relative to performance.
  - Guides decisions for budgeted deployments or cloud-based training.

Next: Part I.c Case Study

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

EURO 4SEE

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.