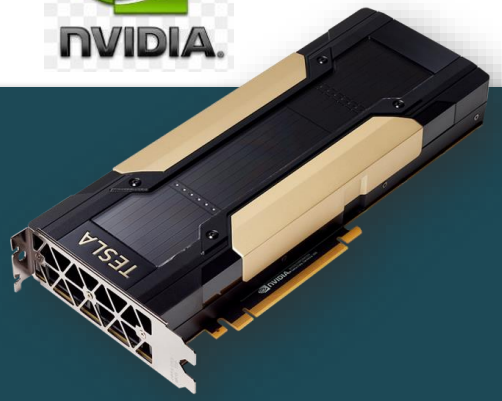
FORTH
INSTITUTE OF COMPUTER SCIENCE



XILINX

Accelerators

intel AGILEX inside

Manos Pavlidakis[1,2]
manospavl@ics.forth.gr

Antonis Chazapis[1]
chazapis@ics.forth.gr

[1] Institute of Computer Science, Foundation for Research and Technology - Hellas, Greece
[2] Computer Science Department, University of Crete, Greece

Agenda

- What is an accelerator?
- What is the difference between accelerators and CPUs?
- How to select the optimal accelerator?
- How to use accelerators?
- What is the performance improvement?

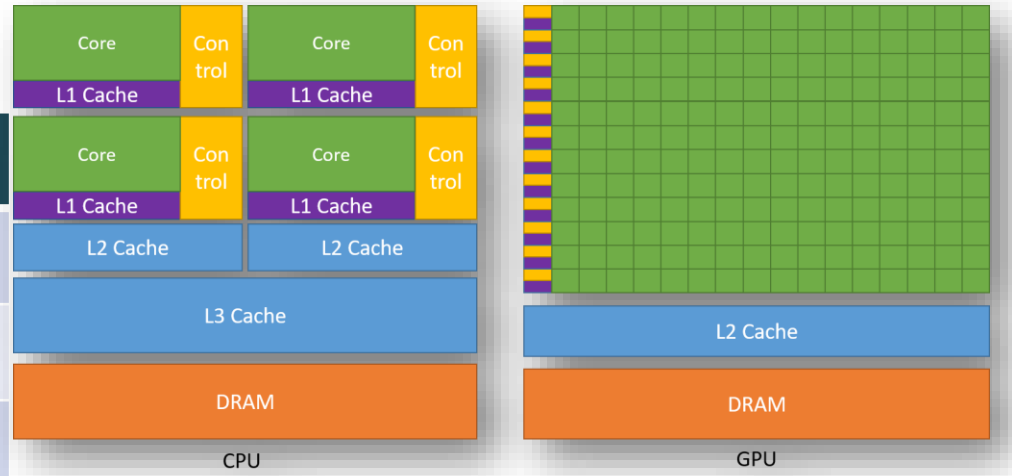
What is an accelerator?

- A device that performs some functions more efficiently than general-purpose CPU
- CPUs have to be good at all functions
  - Run a Browser, Perform mathematical operations etc
- While GPUs are perfect for compute intensive functions
  - Perform Matrix Multiplications

Why accelerators are better than CPUs?

✓Due to massive parallelism

CPU	GPU
Central Processing Unit	Graphics Processing Unit
Several Cores	Thousand Cores
Complex/Larger cores	Simpler/Smaller cores
Low latency	High throughput
Good for serial processing	Good for parallel processing
Good for almost all operations	Perfect for some operations



Typical accelerators

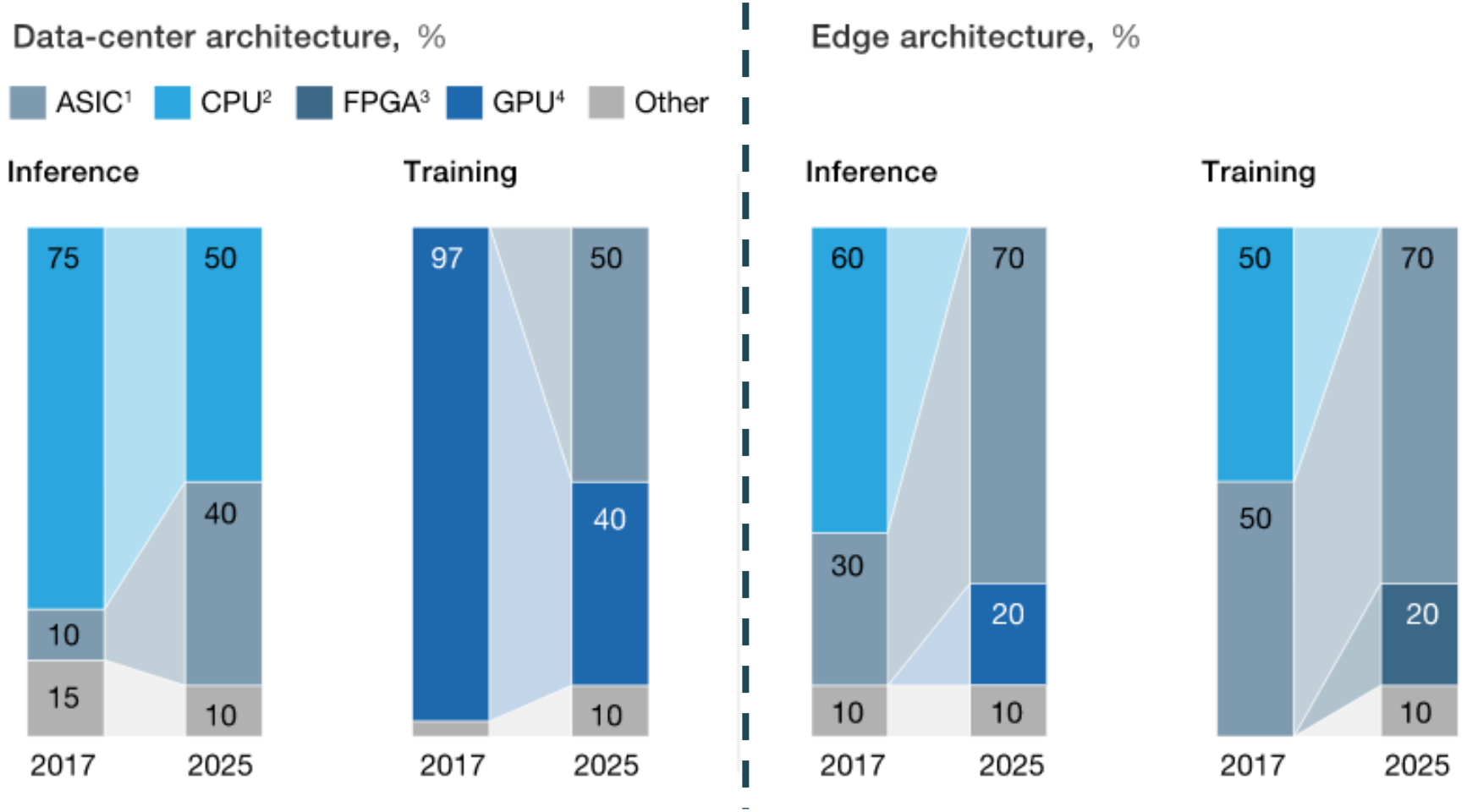
- GPGPUs: General Purpose Graphic Processing Unit (NVIDIA, AMD)
- FPGA: Field-Programmable Gate Array (Xilinx, Intel Altera)
- ASIC: Application-Specific Integrated Circuit
  - TPU: Tensor Processing Unit (Google)
- Accelerators fit perfectly to accelerate compute intensive applications as:
  - Financial
  - Face detection
  - Autonomous driving
  - Language translation
  - Genomics

How to select the optimal accelerator?

Application type	Training	Inference
Speech processing	GPU, ASIC	CPU, ASIC
Face detection	GPU, FPGA	CPU, ASIC
Financial risk stratification	GPU, FPGA	CPU
Route planning	GPU	CPU
Dynamic pricing	GPU	CPU, ASIC
Autonomous driving	ASIC	ASIC, FPGA

https://www.mckinsey.com/industries/semiconductors/our-insights/artificial-intelligence-hardware-new-opportunities-for-semiconductor-companies#

Preferred architectures are shifting!

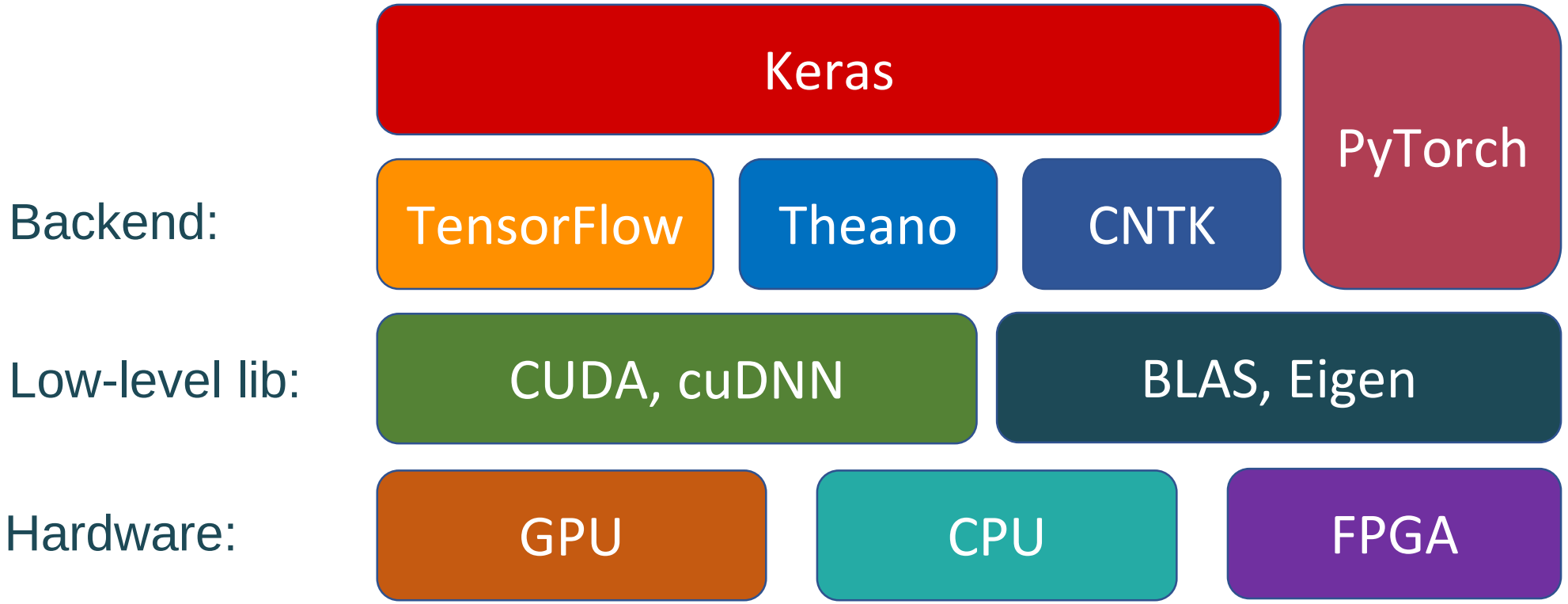


https://www.mckinsey.com/industries/semiconductors/our-insights/artificial-intelligence-hardware-new-opportunities-for-semiconductor-companies#

How to use an accelerator?

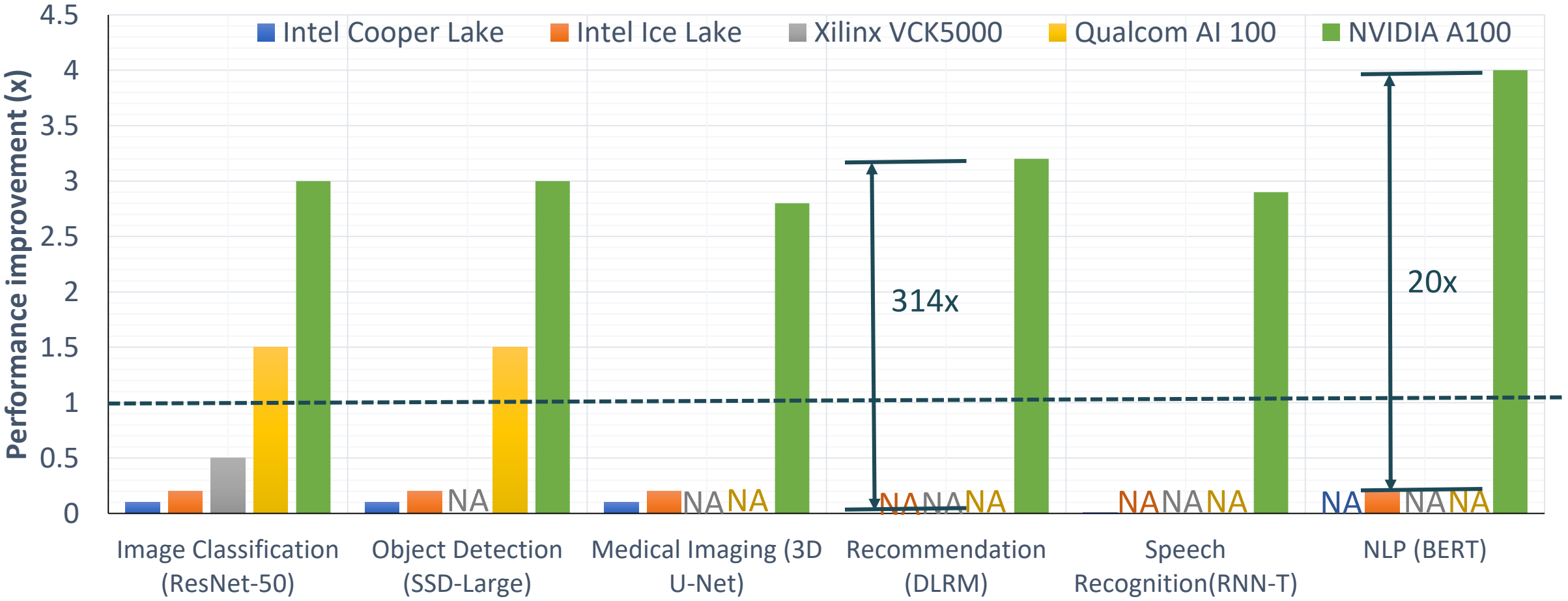
- Use the accelerator programming language and libraries
  - CUDA → NVIDIA GPUs, OpenCL → Intel Altera FPGAs
  - cuDNN, cuBLAS → NVIDIA GPUs, clBLAS → Intel Altera FPGAs
- Generic Programming languages
  - OneAPI, OpenCL
- High level languages
  - Python, Java
  - For instance CUDA offers plugins for high-level languages (PyCUDA, JCUDA)
- Frameworks
  - TensorFlow, PyTorch, MatLab, Caffe, Wolfram Language, mxnet etc.
  - Have implementations for different accelerator types
  - Have simple and flexible APIs that simplify their use (e.g. Keras)

Machine learning stack



JETSON XAVIER NX

MLPerf Data-center benchmarks



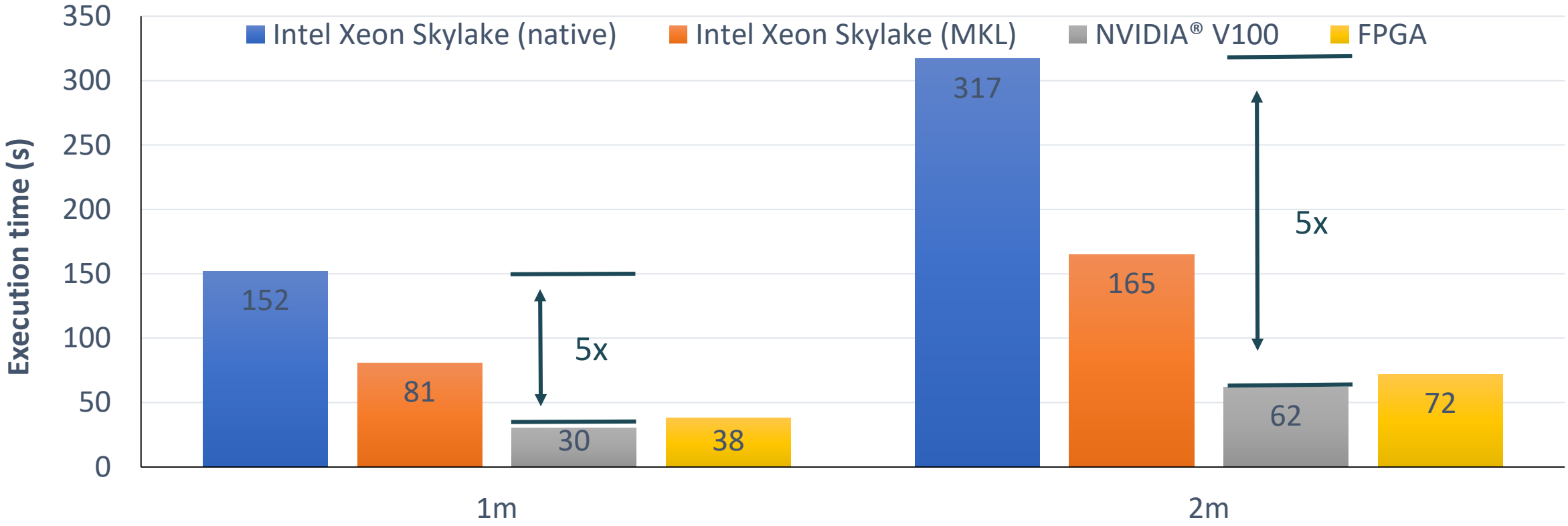
https://www.hpcwire.com/2021/04/21/mlperf-issues-new-inferencing-results-adds-power-metrics-nvidia-wins-again/

Thank you

Questions?

Manos Pavlidakis
manospavl@ics.forth.gr

What is the performance gain with 1xaccel?



https://inaccel.com/cpu-gpu-or-fpga-performance-evaluation-of-cloud-computing-platforms-for-machine-learning-training/