FORTH
INSTITUTE OF COMPUTER SCIENCE

Why accelerators?

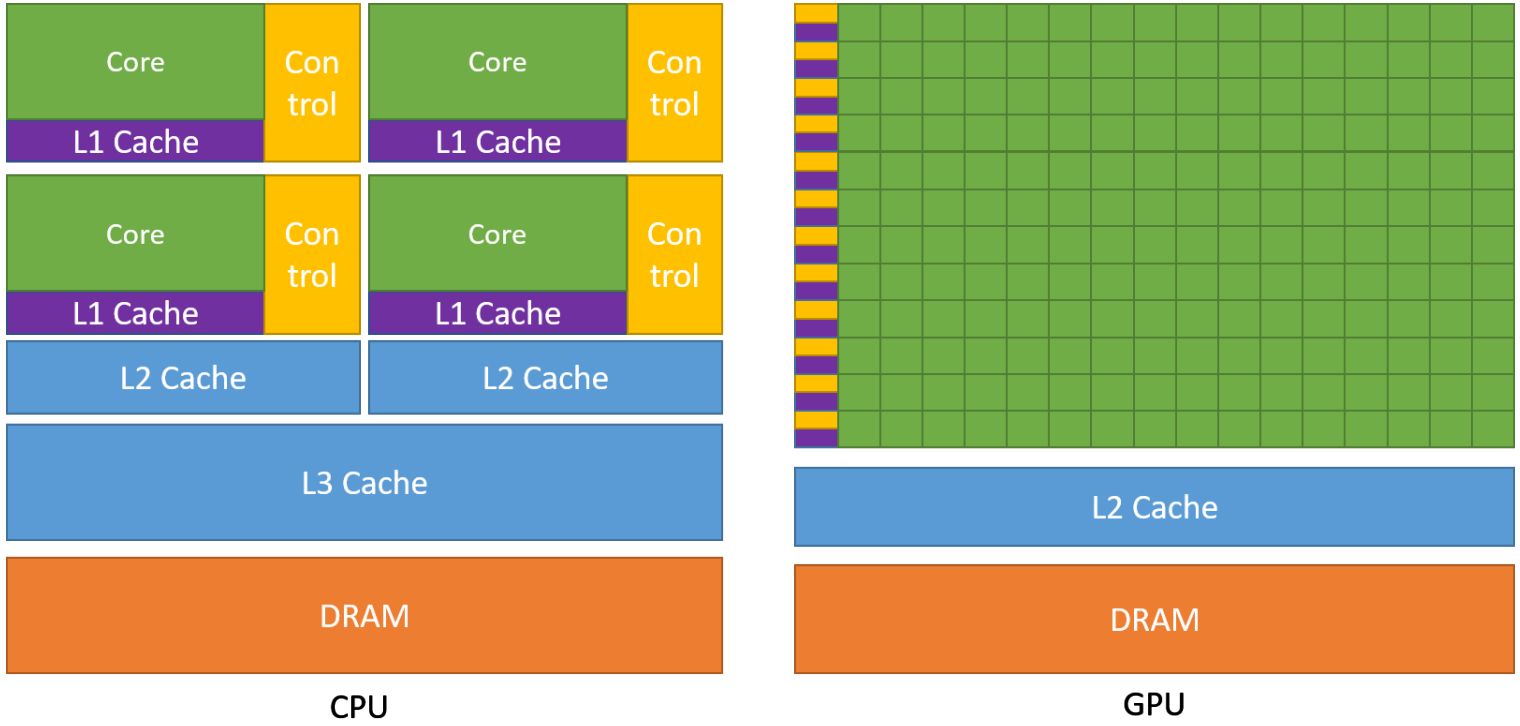
Manos Pavlidakis[1,2]
manospavl@ics.forth.gr

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

What is an accelerator?

- A device that performs some functions more efficiently than general-purpose CPU
- CPUs have to be good at all functions
- E.G. GPUs are perfect for M.M.



Programming Guide :: CUDA Toolkit Documentation (nvidia.com)

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

Why accelerators are better than CPUs?

- Accelerators can process data several orders of magnitude faster than CPUs
  - Due to massive parallelism

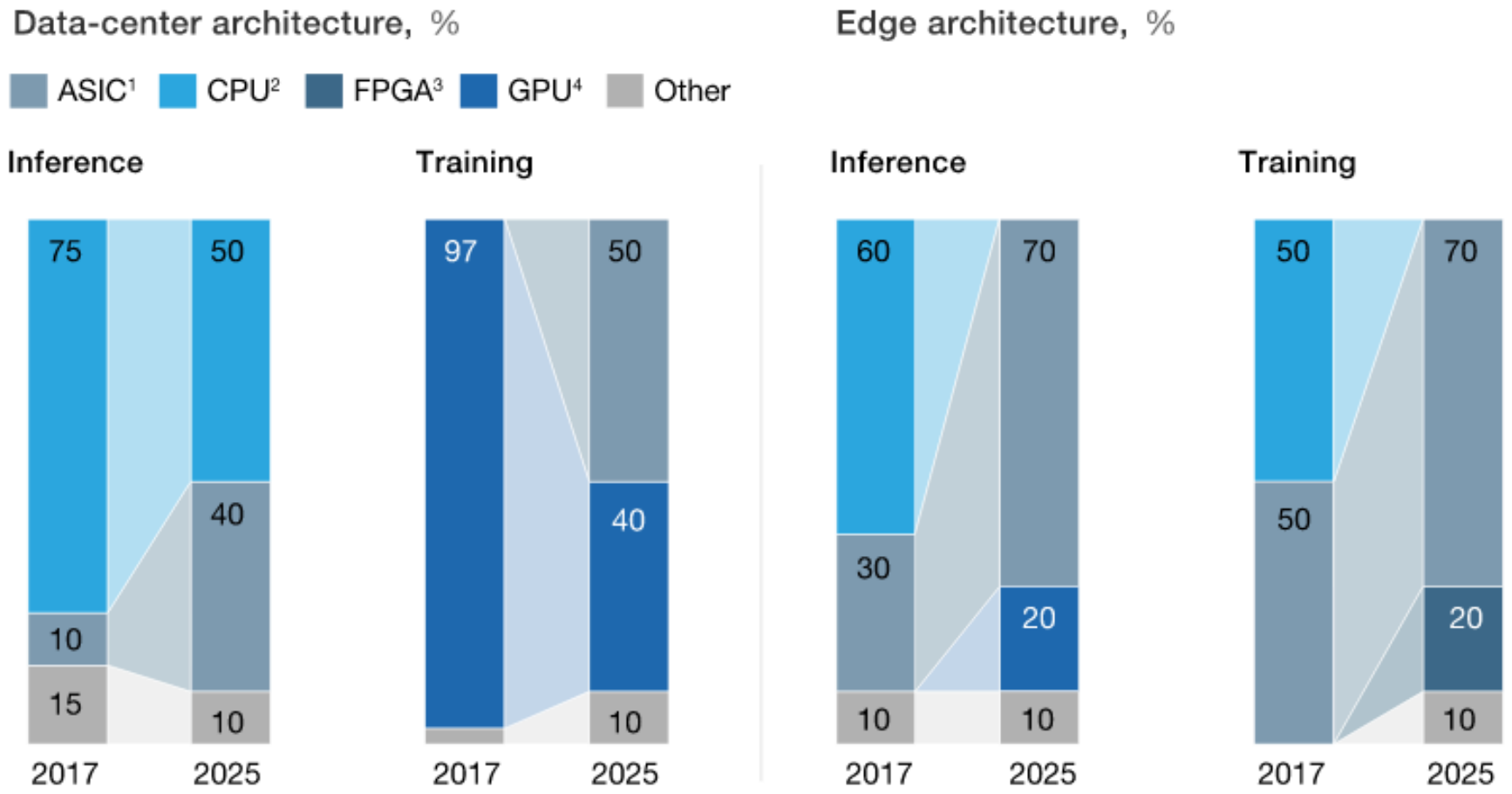
CPU	GPU
Central Processing Unit	Graphics Processing Unit
Several Cores	Many Cores
Complex/Larger cores	Simpler/smaller cores
Low latency	High throughput
Good for serial processing	Good for parallel processing
Good for almost all operations	Perfect for some operations

How to select the optimal accelerator?

Application type	Processing speed	Processing/ Watt	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	GPU, ASIC, FPGA

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

Preferred architectures are shifting!

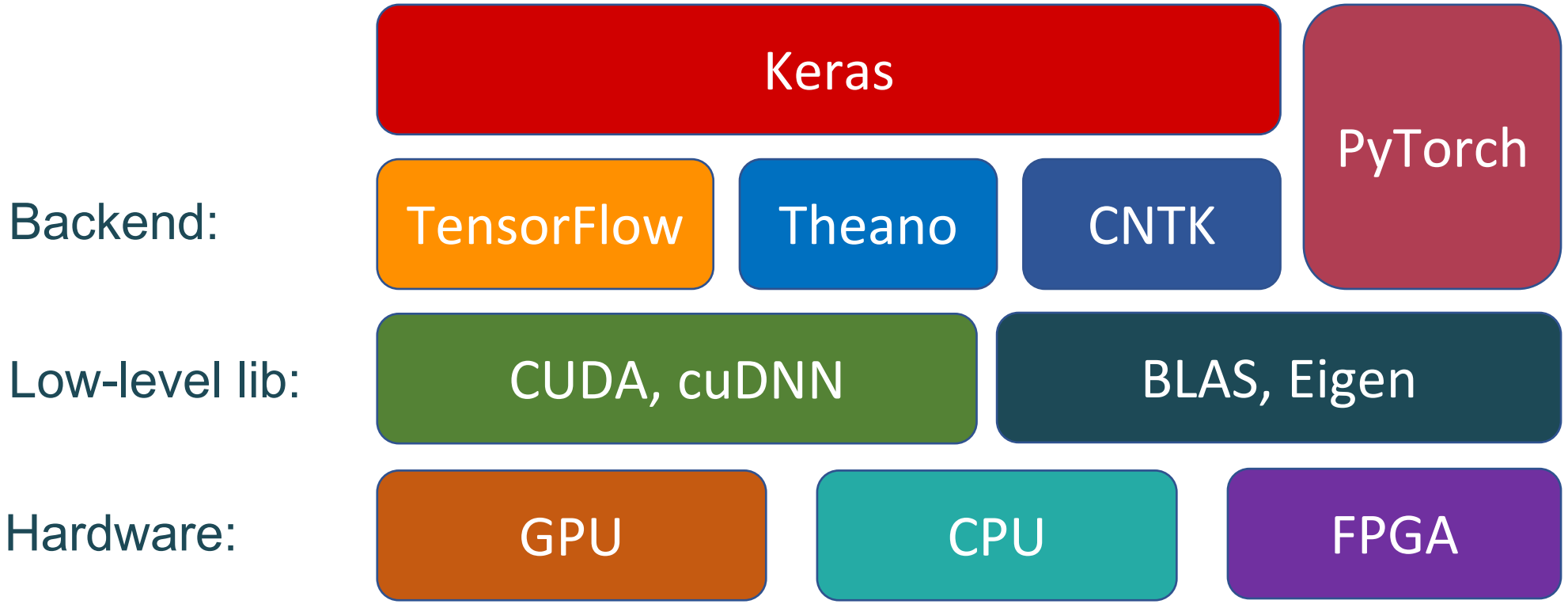


https://www.mckinsey.com/industries/semiconductors/our-insights/artificial-intelligence-hardware-new-opportunities-for-semiconductor-companies#

How to use an accelerator?

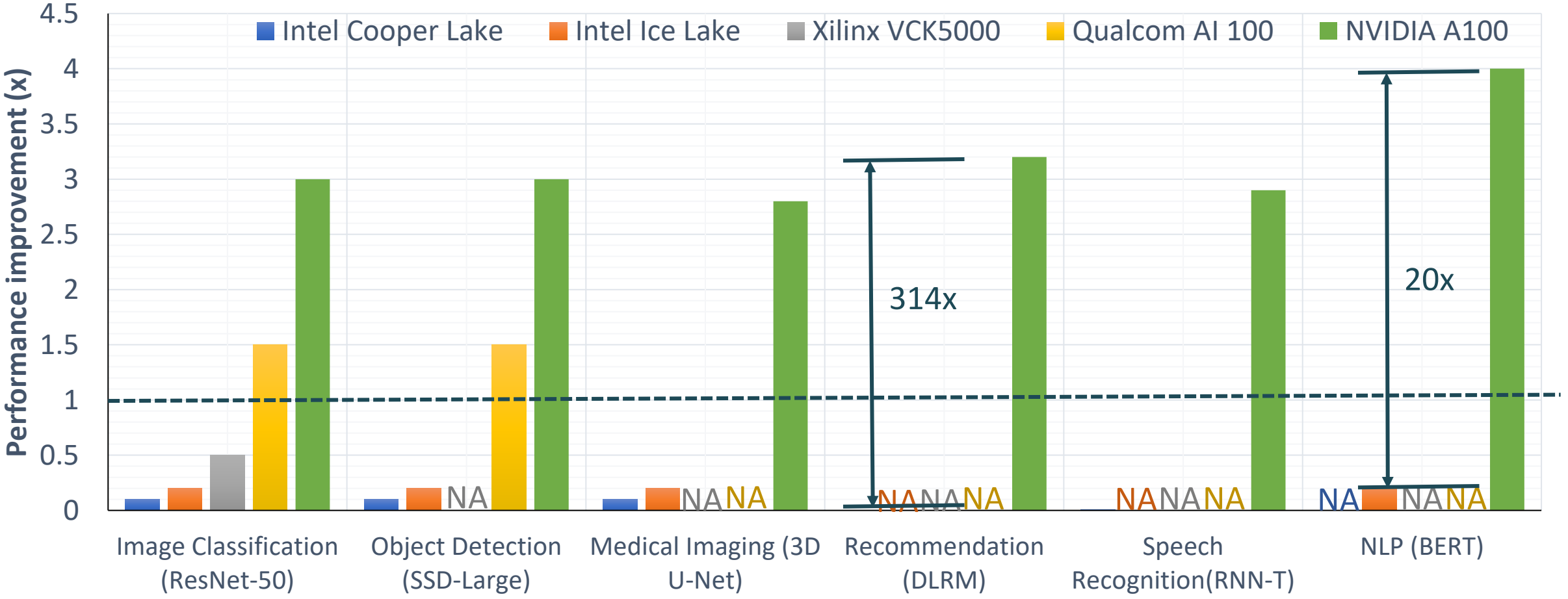
- Use the accelerator programming language and libraries
  - CUDA → NVIDIA GPUs, OpenCL → Intel Altera FPGAs
  - cuDNN, cuBLAS → NVIDIA GPUs, clBLAS → Intel Altera FPGAs
- Generic Programming languages
  - OneAPI, OpenCL
- High level languages
  - Python, Java
  - For instance CUDA offers plugins for 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://inaccel.com/cpu-gpu-or-fpga-performance-evaluation-of-cloud-computing-platforms-for-machine-learning-training/

Thank you

Questions?

Manos Pavlidakis
manospavl@ics.forth.gr