Benchmarking Neural Networks on Heterogeneous Hardware Resources

Christopher Hesse, Holger Eichelberger

eichelberger@sse.uni-hildesheim.de,
christopher.hesse@aptiv.com

Software Systems Engineering
University of Hildesheim

www.sse.uni-hildesheim.de

Motivation

- Artificial Intelligence (AI) is "everywhere"
- HAISEM-Lab (http://haisem-lab.de/)
  - BMBF founded AI lab
  - **H**ardware-optimized **A**rtificial **I**ntelligence Applications using modern **S**oftware **E**ngineering **M**ethods
  - Qualification and training for industry personnel
  - AI/Hardware/SE research
- Partners
  - University of Hannover (L3S, IMS)
  - University of Hildesheim (SSE)

Problem

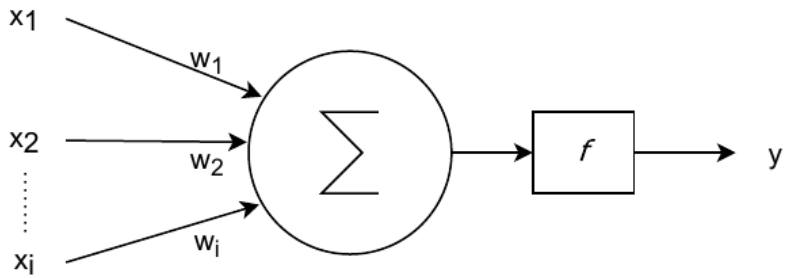
- Hardware for AI
  - GPU servers (NVIDIA Tesla)
  - GPU developer boards, GPU laptops
  - FPGA servers (Maxeler, Intel Arria)
  - Unified memory (Apple M1)

*How to compare NN performance (speed, energy) across **all** heterogeneous hardware resources?*

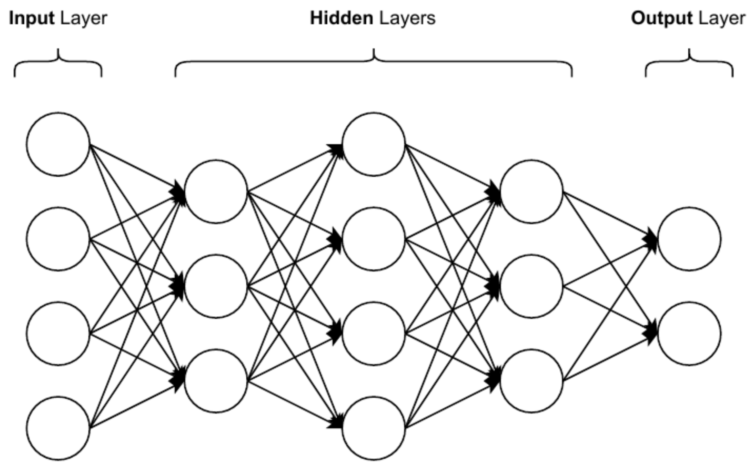
- Existing approaches: At least one hardware type missing

Primer on Neural Networks

- Neuron: “fancy weighted sum”
  - “perceptron”: f is a threshold

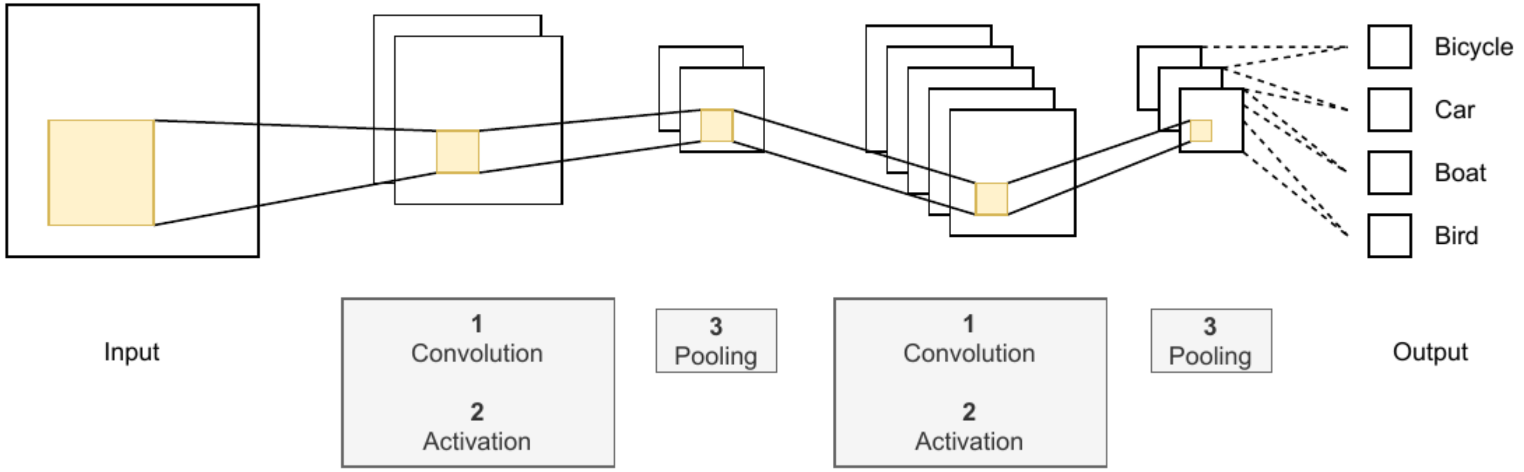


- Deep Neural Network
  - at least two hidden layers



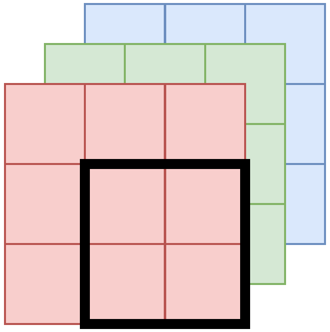
Approach

- Focus: Convolutional Neural Networks
  - "What's in this image?“



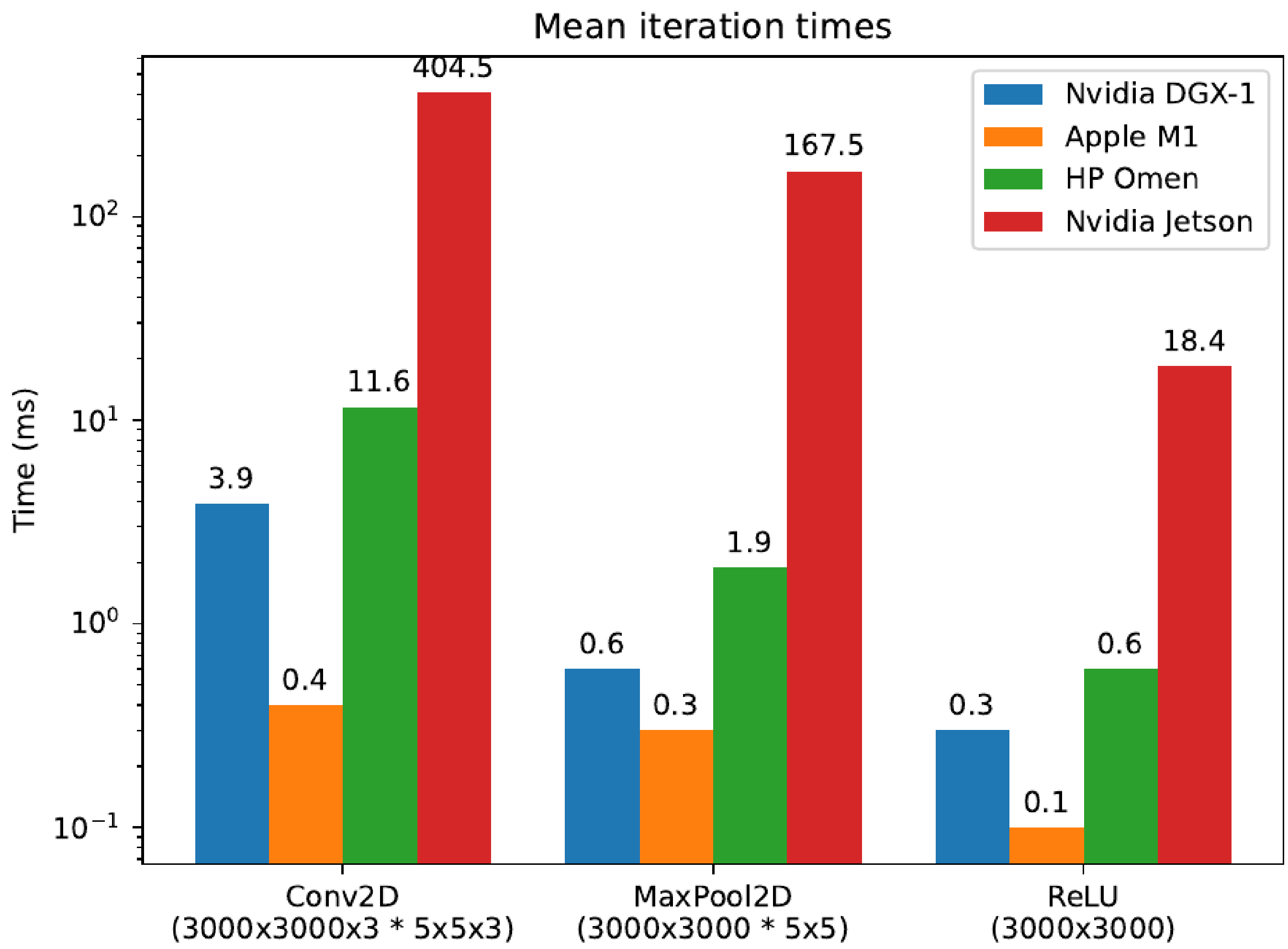
Approach

- Bottom-Up
  - Microbenchmarks
    - Convolution
    - Activation
    - Pooling
  - Macrobenchmarks
    - Training
    - Inference

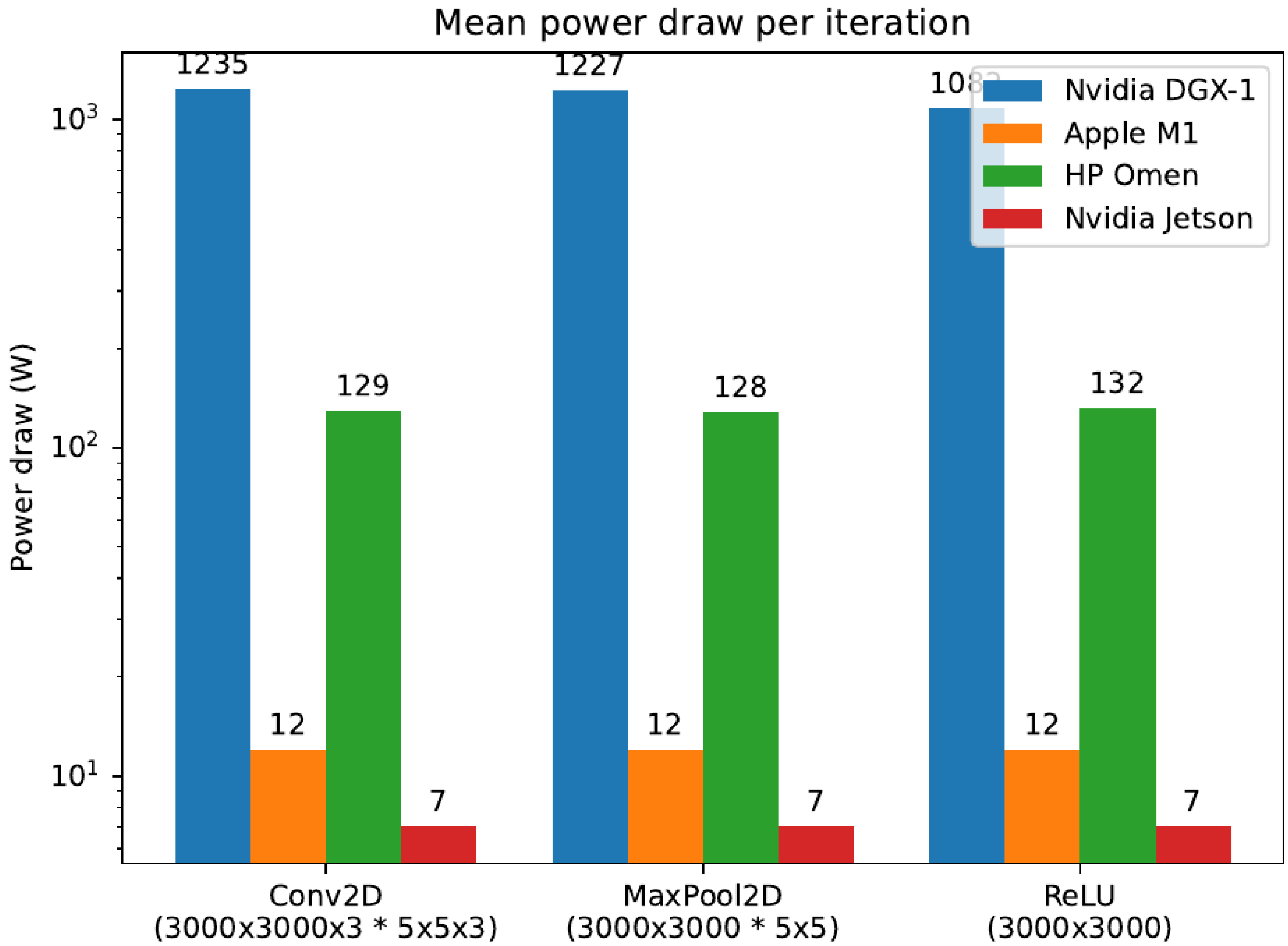


- Methodology
  - Varying parameters, e.g. input size
  - Measure per iteration / benchmark
    - CPU/GPU time
    - Energy

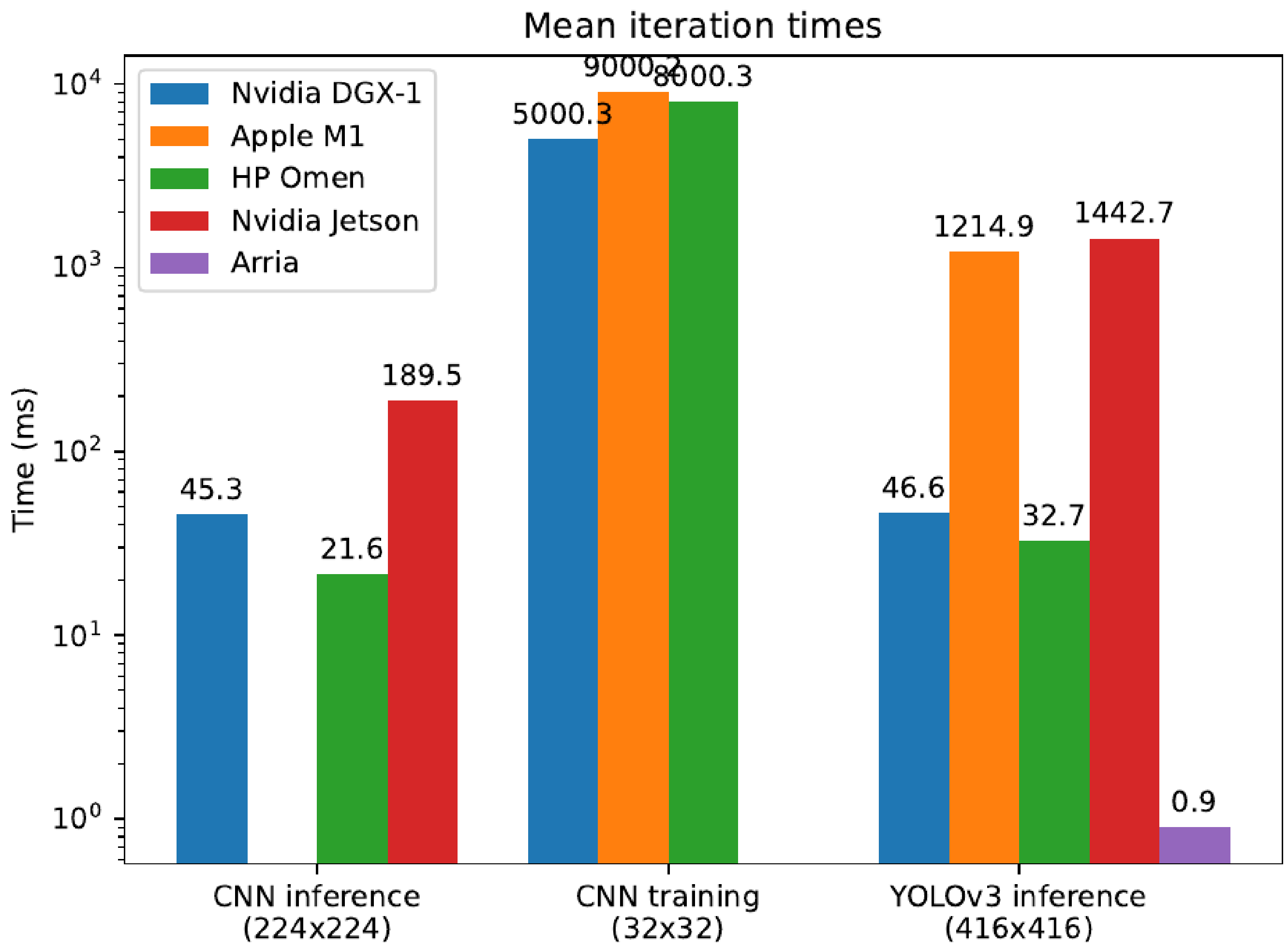
Results - Microbenchmarks



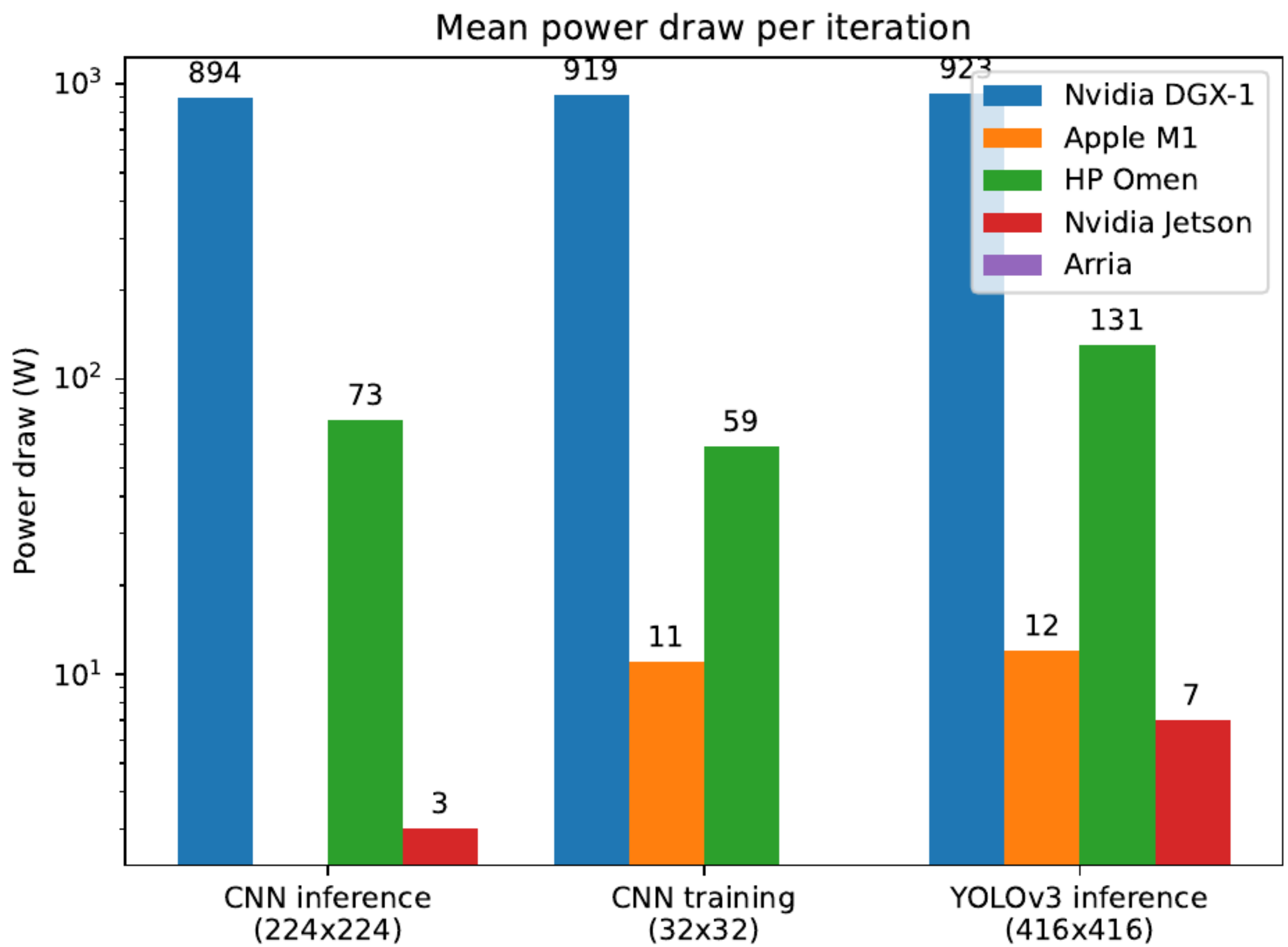
Results - Microbenchmarks



Results - Macrobenchmarks



Results - Macrobenchmarks



Conclusions & Lessons Learned

- Intel FPGA 7x - 45x faster than Pro GPU
  - while using 10x less power

- Apple M1 3x - 5x faster than current x86 CPU/GPU combo
  - while using 5x - 10x less power

- Accelerator programming / DSL is awkward
- TensorFlow for M1 *was* (?) alpha quality
  - Microbench: M1 is most cost & energy effective

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

HAISEM-Lab
http://haisem-lab.de/