Rechnerarchitekturen für Deep-Learning Anwendungen (RADL)

Dustin Heither, Maximilian Achenbach and Robert Kagan

Application

Dustin Heither, Maximilian Achenbach and Robert Kagan

- Application purpose: Classification
- Dataset: MNIST
- Kind of application: General purpose laptops and desktops

Targeted architecture

Dustin Heither, Maximilian Achenbach and Robert Kagan

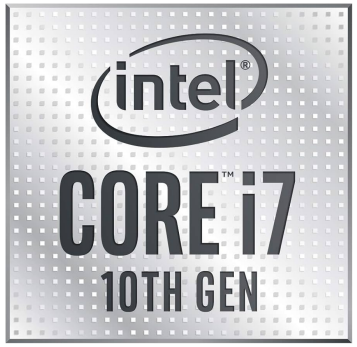
- Hardware: CPU vs. GPU
- Device:
  - Apple M3 Pro (ARMv8.6-A)
  - Intel Core i7 1065G7 (x86-64-v4)
  - Nvidia GeForce RTX 2080

Developer:

Dustin Heither

Robert Kagan

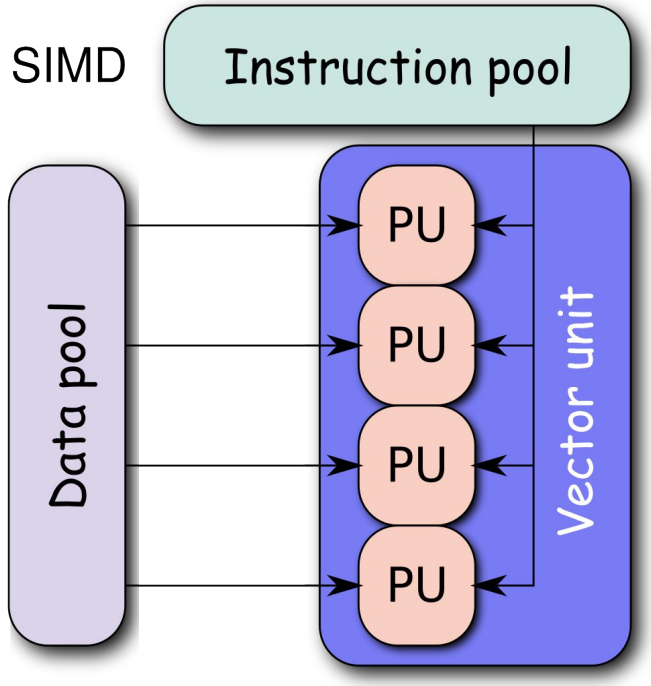
Maximilian Achenbach



Approach and responsibilities

Dustin Heither, Maximilian Achenbach and Robert Kagan

- The final result:
  - Deep Learning Framework with CPU-GPU switch
  - Combination of Multithreading and SIMD
  - Evaluation of:
    - Performance gains
    - Resource consumption
    - Performance per watt
    - PCIe latency



Approach and responsibilities

Dustin Heither, Maximilian Achenbach and Robert Kagan

Kind of optimization:	Developer:
– (Apple M3 Pro NPU)	Dustin Heither
– Multithreading	Dustin Heither, Robert Kagan
– SIMD	
– Arm Neon	Dustin Heither
– Quantization	Dustin Heither, Robert Kagan
– SSE vs. AVX2 vs. AVX-512	Robert Kagan
– ICC vs. GCC	Robert Kagan
– CUDA tuning	Maximilian Achenbach