Machine Learning Hardware

Dominik Scherer
06.11.2017

Seminar „Neuste Trends in Big Data Analytics“

Betreuer: Dr. Julian Kunkel

Universität Hamburg

Motivation

- Maschinelles Lernen in vielen Bereichen angewendet, z.B.
  - Spracherkennung
  - Bilderkennung
  - Robotik
  - Suchmaschinen
  - Konsumenten Analyse
  - Übersetzung

- Neuronale Netze aufgrund ihrer Flexibilität weit verbreitet, insbesondere zur Sprach-, Bild- und Mustererkennung

Motivation

- Umfrage: Welche ML Funktionalitäten werden genutzt

Quelle: Machine Learning Umfrage, Crisp Research AG Kassel, [1]

Motivation

- Training und Anwendung sehr rechenintensiv aufgrund
  - großer Datenmengen (Big Data)
  - komplexer Netze (Deep Learning)

- “Klassische” Hardware (CPUs) nicht optimal für ML
  - besser für sequentielle Probleme
  - müssen breites Spektrum an Aufgaben bewältigen
  - deutlich schlechtere Performance als spezialisierte Hardware
  - High-End CPU bis 1 TFLOPS (Intel Core i9 XE, 18 Kerne)

- Spezielle Beschleuniger zur Verminderung der Rechenzeit

Agenda

- Motivation
- Definition Machine Learning
- Grundlagen
- Machine Learning Hardware
  - Tensor Processing Unit
  - Tensor Cores
  - Nervana
  - Neuromorphic Chip Loihi
- Zusammenfassung

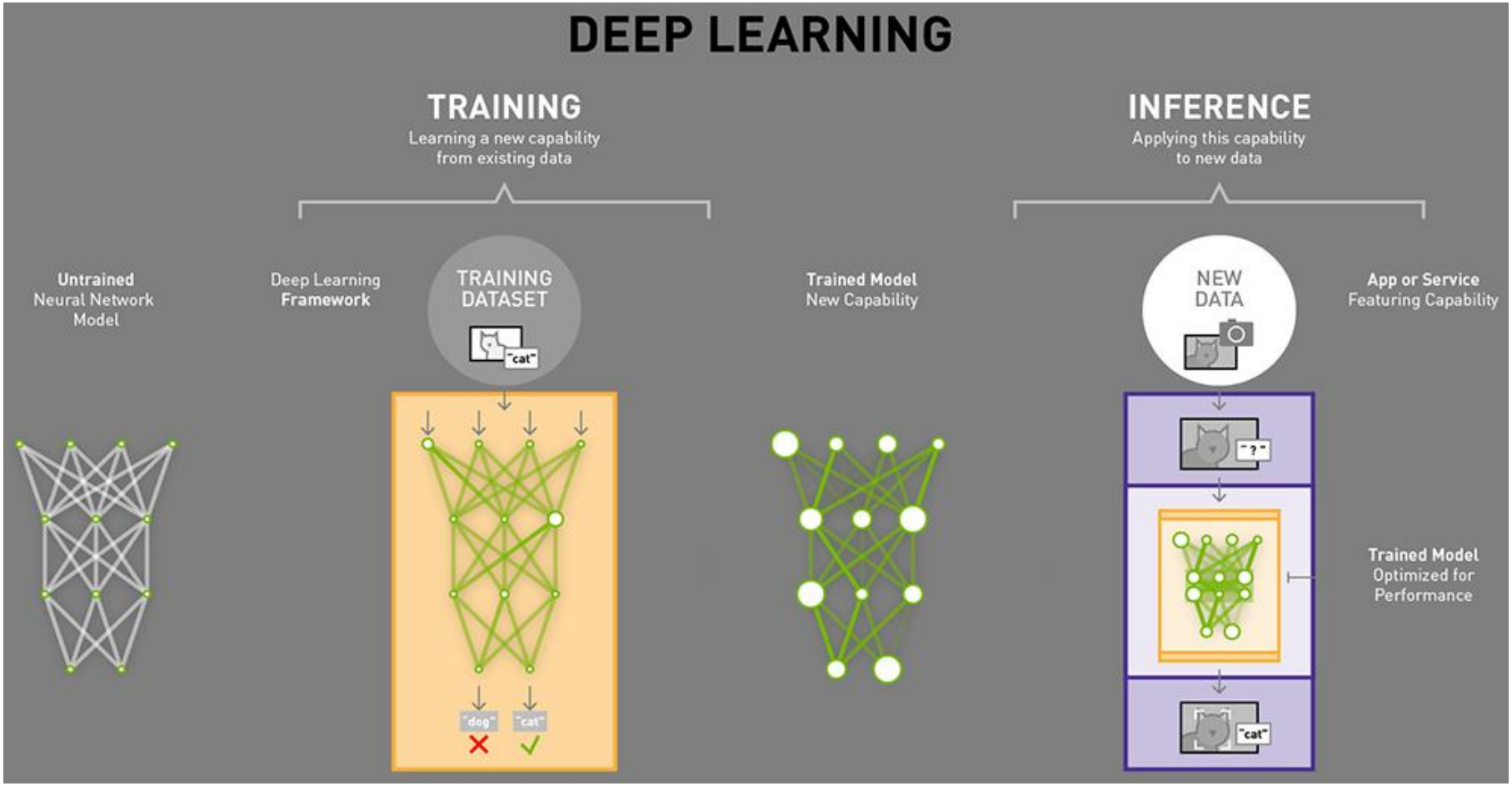
Definition Machine Learning

- “The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.”
  [Machine Learning, Tom Mitchell, McGraw Hill, 1997]

- Unterkategorien
  - Künstliche neuronale Netze
  - Entscheidungsbäume
  - Clustering
  - Genetische Algorithmen
  - Reinforcement Learning

Grundlagen

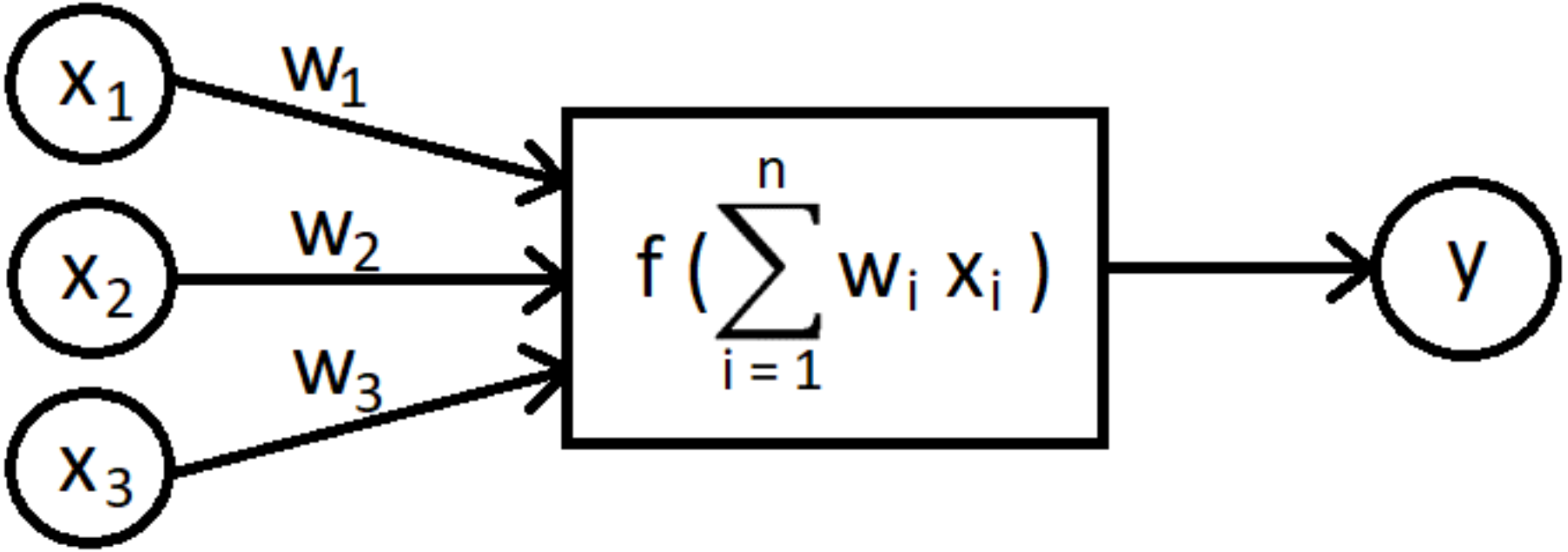
- Neuronales Netz: Training und Inferenz



Quelle: Training and Inference of NNs, Nvidia Corporation, [2]

Grundlagen

- Input x, Output y, Gewicht w, Aktivierungsfunktion f



- Implementation als Matrixmultiplikation

Erinnerung: Matrix Multiplikation

- Skalarprodukte aus Zeilen und Spalten

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} a*e+b*g & a*f+b*h \\ c*e+d*g & c*f+d*h \end{pmatrix}$$

- Einschränkungen:
  - nur Multiplikation von m x n mit n x o Matrizen möglich
  - nicht kommutativ

Grundlagen

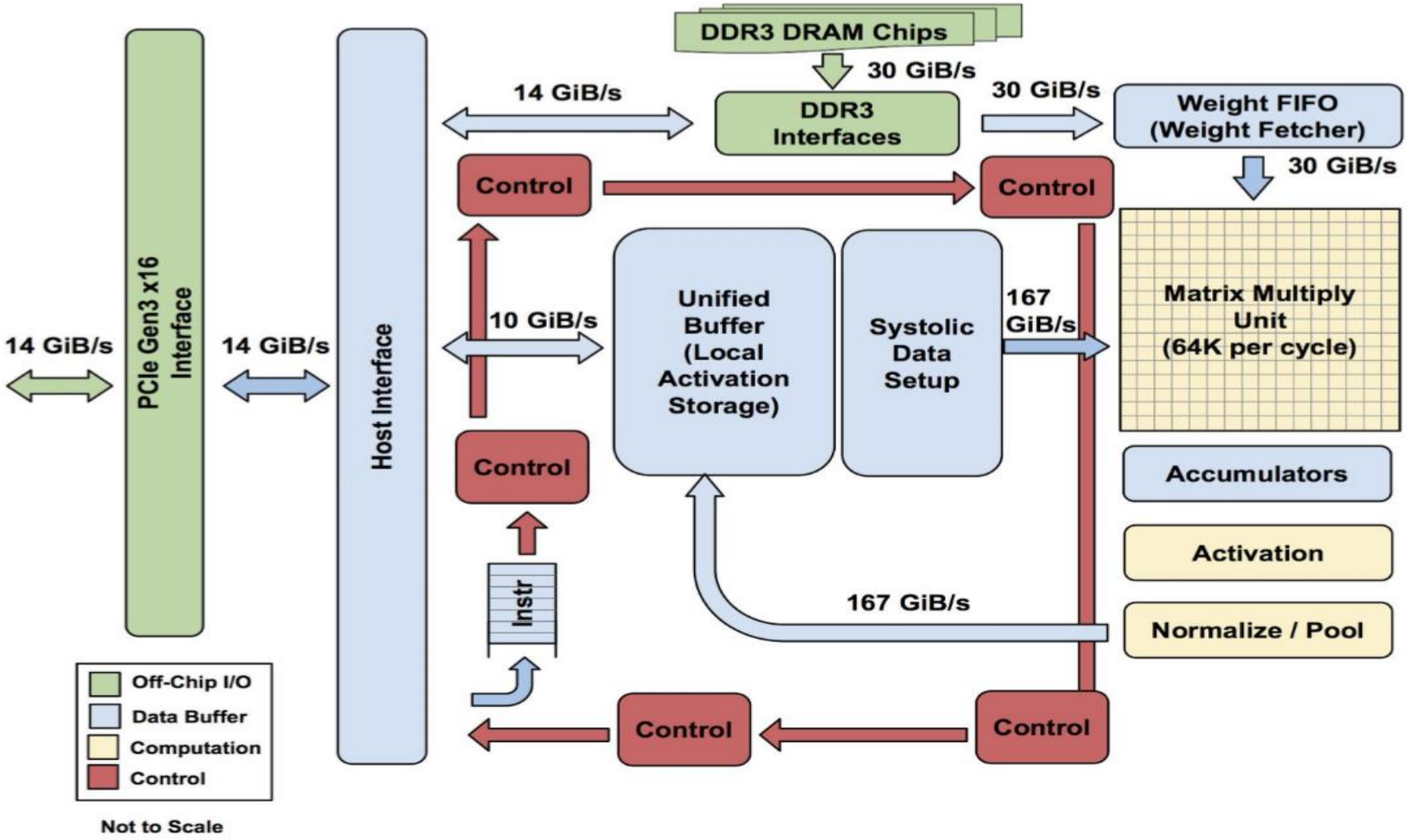
- Anforderungen an die Hardware:
  - Parallelität (Matrix Rechnungen inhärent parallel)
  - Skalierbarkeit (viele Einheiten miteinander verbunden)
  - hoher Durchsatz (Berechnungszeit möglichst gering)
  - Effizienz (sehr viel Hardware zeitgleich im Betrieb)
  - low-precision Arithmetik (ausreichend für NNs)
  - teilweise niedrige Latenz (wenige ms)

- Viele CPU Features ungenutzt, Verwendung spez. Hardware
  - GPGPU
  - Beschleunigerkarten

Tensor Processing Unit

- TPUs:
  - anwendungsspezifische Chips für ML von Google
  - insbesondere zum Verarbeiten neuronaler Netze
  - speziell für TensorFlow entworfen
- Spezifikationen (1. Generation):
  - Accelerator connected via PCIe 3.0 Bus
  - CISC Instruction Set
  - 700 MHz Clock Rate
  - 65.536 8-bit Integer Multiply-and-Add Units
  - 28 MB On-Chip Activation Memory
  - 4MB On-Chip Accumulator Memory
  - 8GB Off-Chip DRAM mit 34 GB/s

TPU: High Level Chip Architektur



Quelle: TPU Block Diagram, Google, [3]

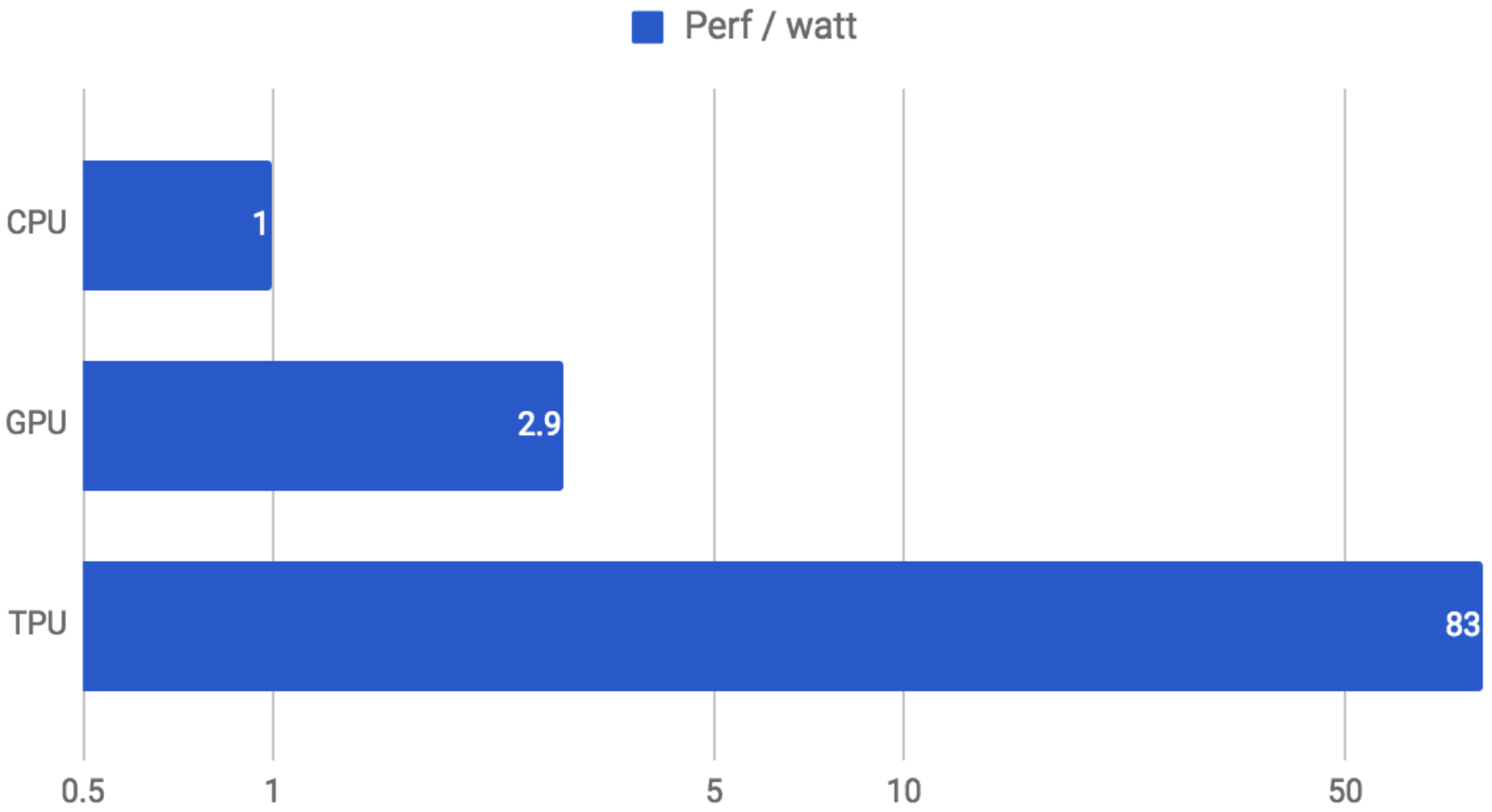
TPU: Instructions

• Read_Host_Memory	Lese Daten/Input vom Host Speicher
• Read_Weights	Lese Gewichte ein
• MatrixMultiply	Führe Matrix Multiplikationen der Gewichte mit den Daten aus, akkumuliere die Ergebnisse
• Activate	Berechne Aktivierungsfunktion
• Write_Host_Memory	Schreibe das Ergebnis in den Host Speicher

TPU: Matrix Multiplier Unit

- Systolisches Array mit 256x256 ALUs
- 65.536 Multiply-and-Adds in jedem Zyklus
- $65.536 \times 700{,}000{,}000 = 46 \times 10^{12}$ Multiply-and-Add Ops/s
  bzw. 92 Teraops/s
- Mit *MatrixMultiply* bis zu 128.000 Operationen in einem Zyklus ohne Speicher Zugriffe für Zwischenergebnisse
  => signifikante Verbesserung der Effizienz

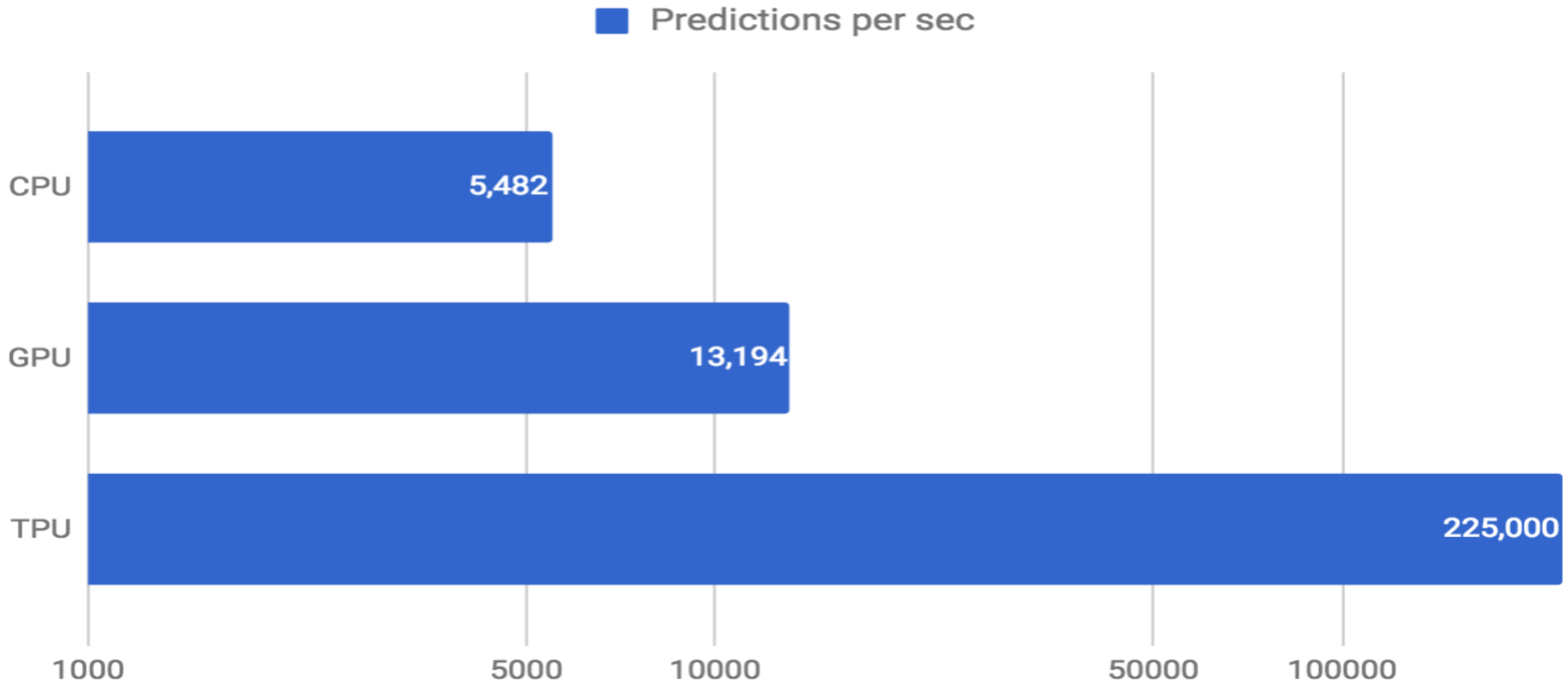
TPU: Leistung pro Watt



Quelle: Performance/watt, relative to contemporary CPUs and GPUs, Google, [3]

TPU: Durchsatz pro Sekunde

- Inferenz Durchsatz bei einem MLP mit 5 Layern und 20M Gewichten mit <= 7ms Latenz [4]



Quelle: Throughput under 7 ms latency limit, Google, [3]

Tensor Cores

- Teil der GPU-Architektur Volta von Nvidia
- Bisher nur auf dem Tesla V100 Chip vorhanden
- Speziell auf Deep Learning ausgerichtet
- Konsumenten Chips voraussichtlich Anfang 2018

Tesla V100: Daten

- 21 Milliarden Transistoren auf 815mm²
  - größter GPU-Chip der Welt
- 80 Streaming Multiprozessoren (SM)
  - 64 FP32 Cores / SM, 5120 pro GPU
  - 32 FP64 Cores / SM, 2560 pro GPU
  - 8 Tensor Cores / SM, 640 pro GPU
- 1530 MHz Taktrate
- 16GB HBM2 mit 900GB/s

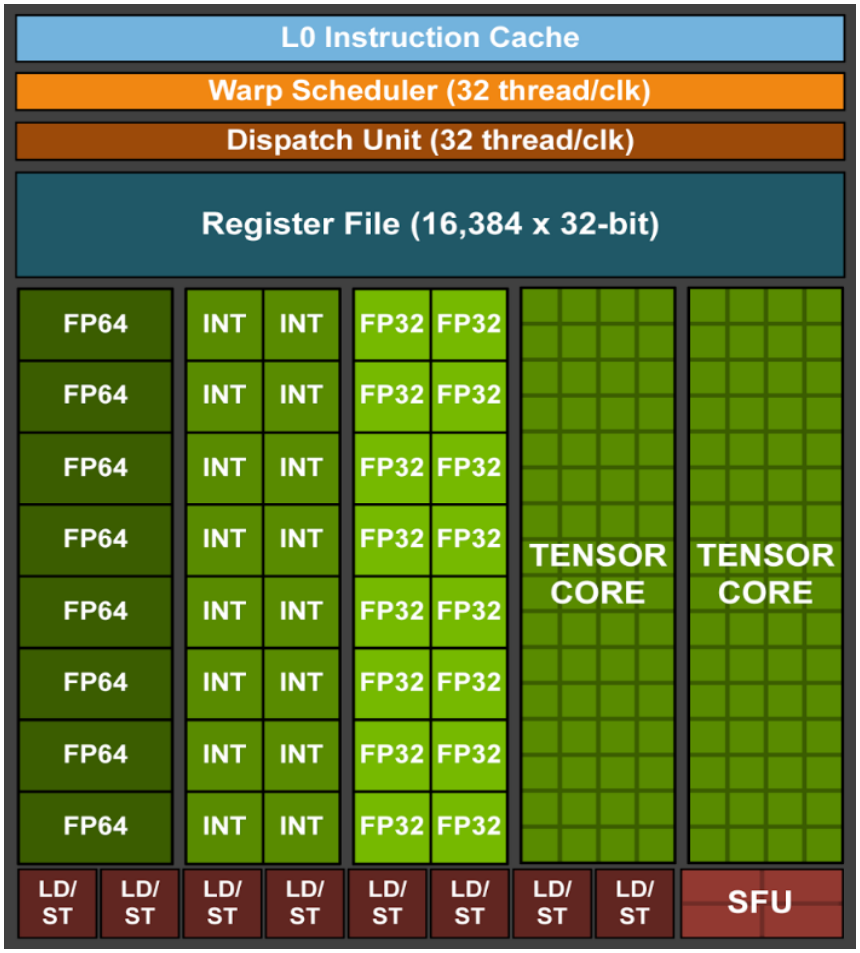
Tesla V100: Leistung

- 7.5 TFLOPS Double Precision Performance
- 15 TFLOPS Single Precision Performance
- 120 TFLOPS Tensor Performance
- Stromverbrauch: 300 Watt

Volta vs. Pascal

- GPUs noch Standard im Machine Learning Bereich
- Leistungsfähigste Grafikkarten vor Volta basierten auf der Pascal Architektur von Nvidia
- GPUs aber nicht spezialisiert auf Machine Learning
- Volta erste Architektur mit Deep Learning Hardware
- Volta 9x schneller in Deep Learning Anwendungen

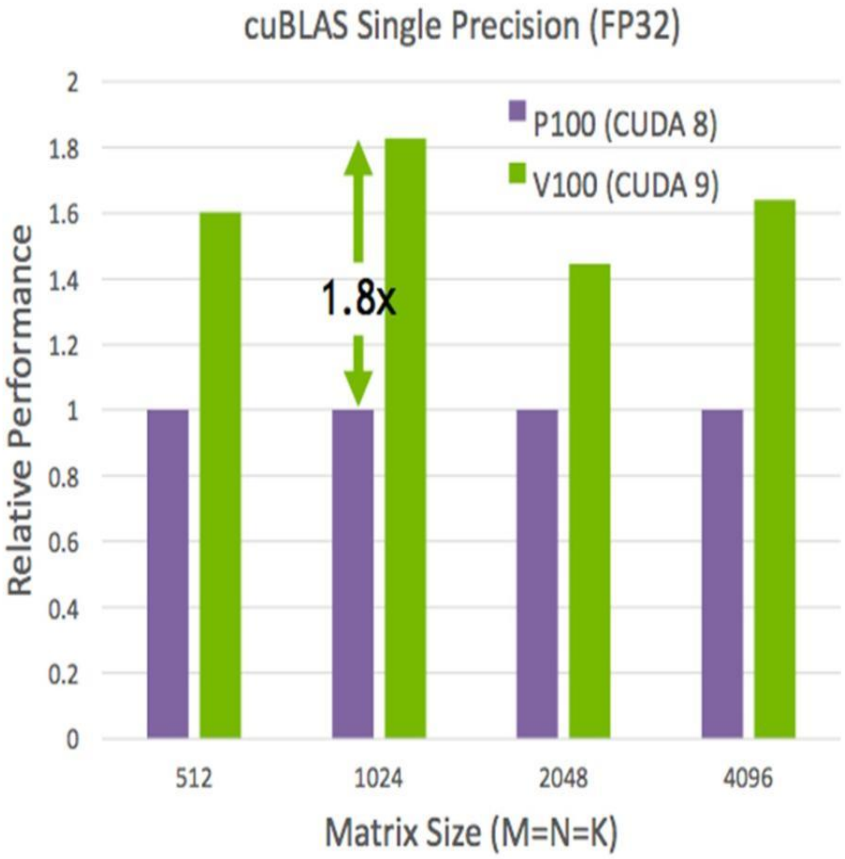
Volta vs. Pascal

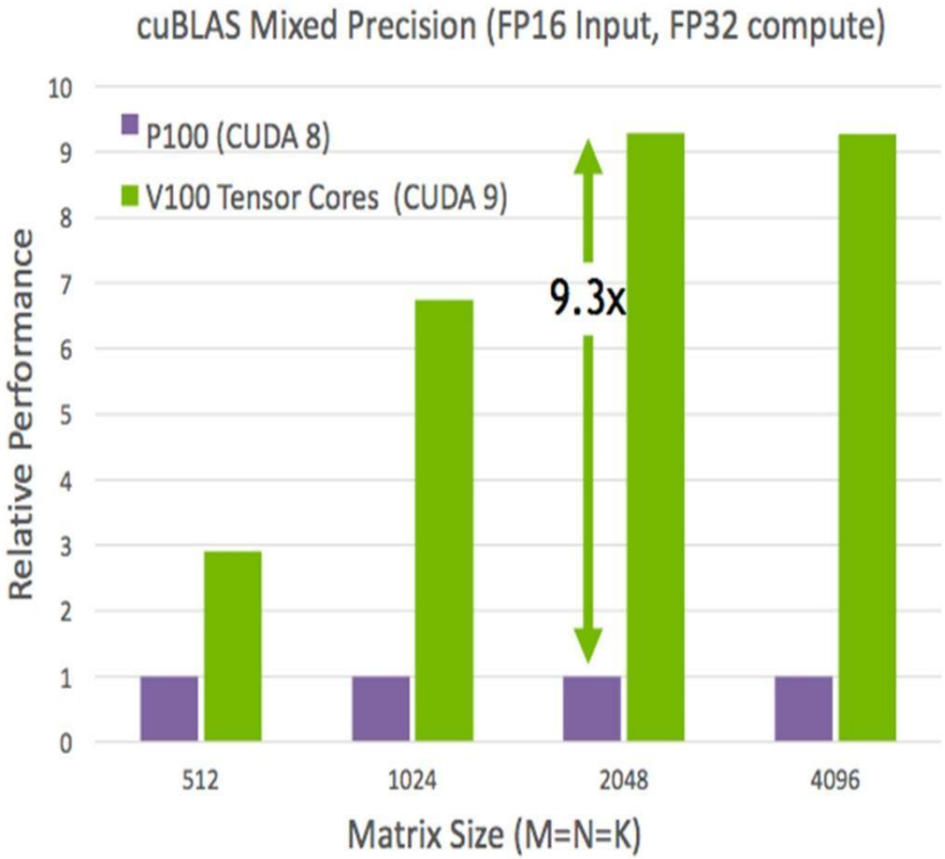


Quelle: Tesla V100 PB, Nvidia, [4]

Quelle: Tesla P100 PB, Nvidia, [5]

Volta vs. Pascal





Quelle: V100 vs. P100 Performance, Nvidia, [4]

Tensor Core Operation

- Ein 4x4x4 Matrix Processing Array pro Tensor Core
- Parallele Verarbeitung von 4x4 Matrix Multiplikation
  - 64 Produkte pro Instruktion mit Akkumulation
  - vgl. Pascal: 4 Produkte pro Instruktion mit Akkumulation
- Multiplikation mit 16FP Input, Addition mit 16FP oder 32FP
- 64 FMA Op. pro Takt pro Tensor Core
  - 8 x 64 x 2 = 1024 Op. pro Takt pro SM
  - 1024 x 80 = 81920 Op. pro Takt pro GPU

Komplettsysteme

- Nvidia DGX-1, “400 servers in a box“
  - 8x Tesla V100
  - 960 Tensor TFLOPS
  - 3200W
  - Preis: 149.000$
- Nvidia DGX Station
  - 4x Tesla V100
  - 480 Tensor TFLOPS
  - 1500W
  - Preis 69.000$

Intel Nervana

- Neue Neural Network Prozessor Familie von Intel
  - ehemals Lake Crest Deep Learning Architecture
- Nervana Systems 2016 von Intel gekauft
  - auf künstliche Intelligenz spezialisierte Software Firma
- In Zusammenarbeit mit Facebook entstanden
- Erster Release für Ende 2017 angekündigt

Intel Nervana: Technische Details

- Eigens entwickelte Bi-Directional High Bandwidth Links
  - 6 pro Prozessor
  - ermöglichen nahezu linearen Speedup
- Keine Standard Cache Hierarchie
  - vorhandener on-chip Memory per Software verwaltet
  - 32GB HBM2
- Neues numerisches Format: Flexpoint
  - beschrieben als Mischung aus floating point und fixed point
  - gleicher Exponent für einen Datenblock

Loihi Chip

- Self-Learning Neuromorphic Chip von Intel
  - erste Ausgabe Anfang 2018 an ausgewählte Einrichtungen
- Imitiert Funktionsweise des Gehirns
- On-Chip Learning, Training und Inferenz
- 130.000 Neuronen,130.000.000 Synapsen
- 1 Million mal schneller als typische NNs, gemessen an der Anzahl von Operationen bei MNIST Ziffer Erkennung

Zusammenfassung

- Machine Learning Branche stark am Wachsen
  - enorm große Datenvolumen zu verarbeiten
  - hohe Nachfrage nach Beschleunigern
- Tensor Leistung der Architekturen im Vergleich:
  - TPU: 92 TOPS 8bit @ 75W
  - TESLA V100: 120 TFLOPS @ 300W
  - TESLA P100: 13 TFLOPS @ 300W
- ML Hardware noch sehr junges Forschungsgebiet
  - aktuelle Lösungen noch nicht optimal

Quellen:

- [1] https://www.computerwoche.de/g/die-it-welt-in-zahlen,116626,64#galleryHeadline
- [2] http://www.analyticsearches.com/wps-differentiating-beween-ai-machine-learning-and-deep-learning/
- [3] https://cloud.google.com/blog/big-data/2017/05/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu
- [4] https://devblogs.nvidia.com/parallelforall/inside-volta/
- [5] https://devblogs.nvidia.com/parallelforall/inside-pascal/
- https://www.servethehome.com/case-study-google-tpu-gddr5-hot-chips-29/
- https://blogs.nvidia.com/blog/2016/08/22/difference-deep-learning-training-inference-ai/
- https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/dgx-1/NVIDIA-DGX-1-Volta-AI-Supercomputer-Datasheet.pdf
- https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/dgx-station/dgx-station-data-science-supercomputer-datasheet-10232017.pdf
- https://www.intelnervana.com/nervana-engine-delivers-deep-learning-at-ludicrous-speed/
- https://www.anandtech.com/show/11942/intel-shipping-nervana-neural-network-processor-first-silicon-before-year-end
- https://newsroom.intel.com/editorials/intels-new-self-learning-chip-promises-accelerate-artificial-intelligence/