Data Science Survival Skills

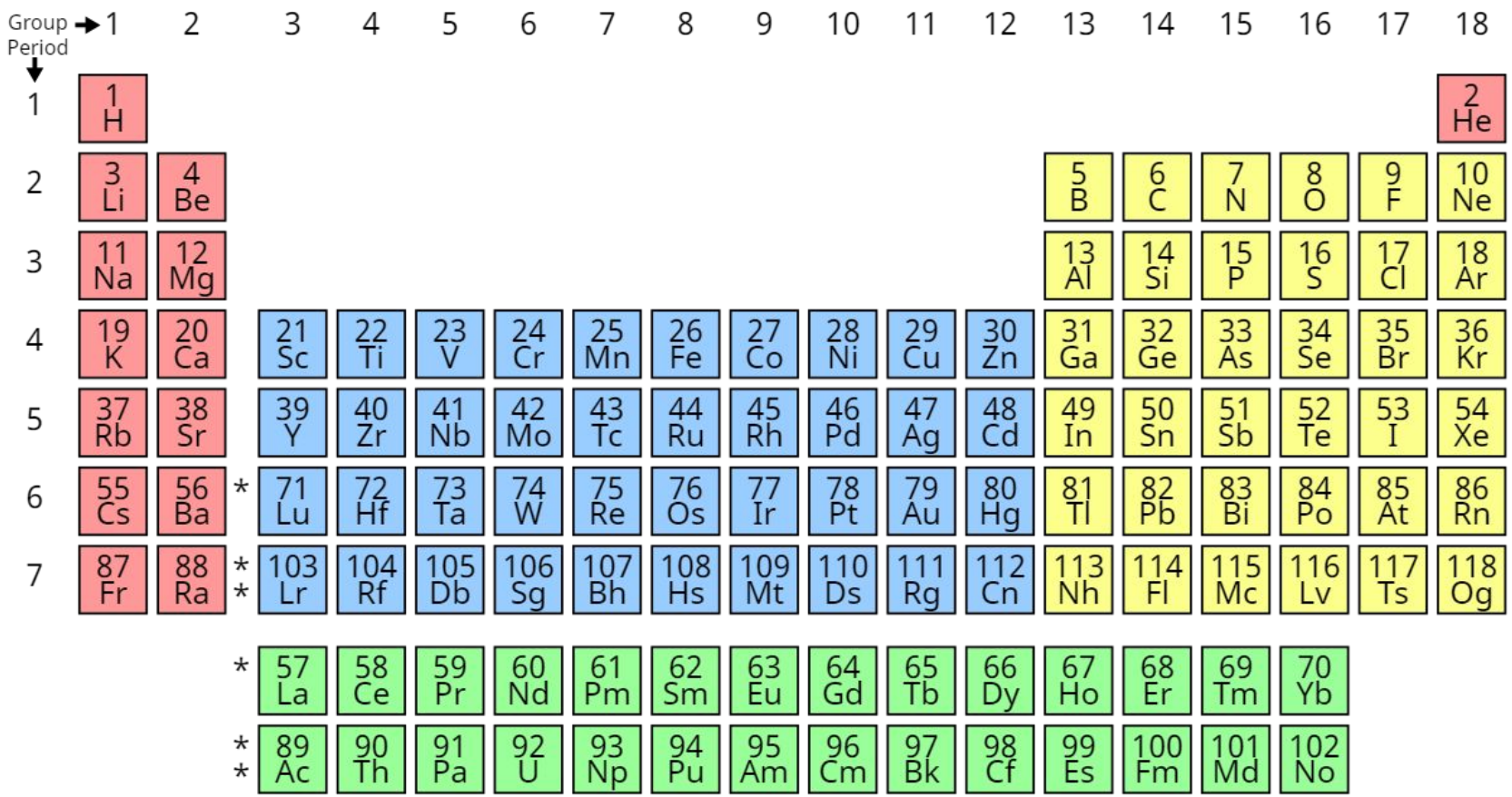
Hardware relevance in data science

A computer from the inside

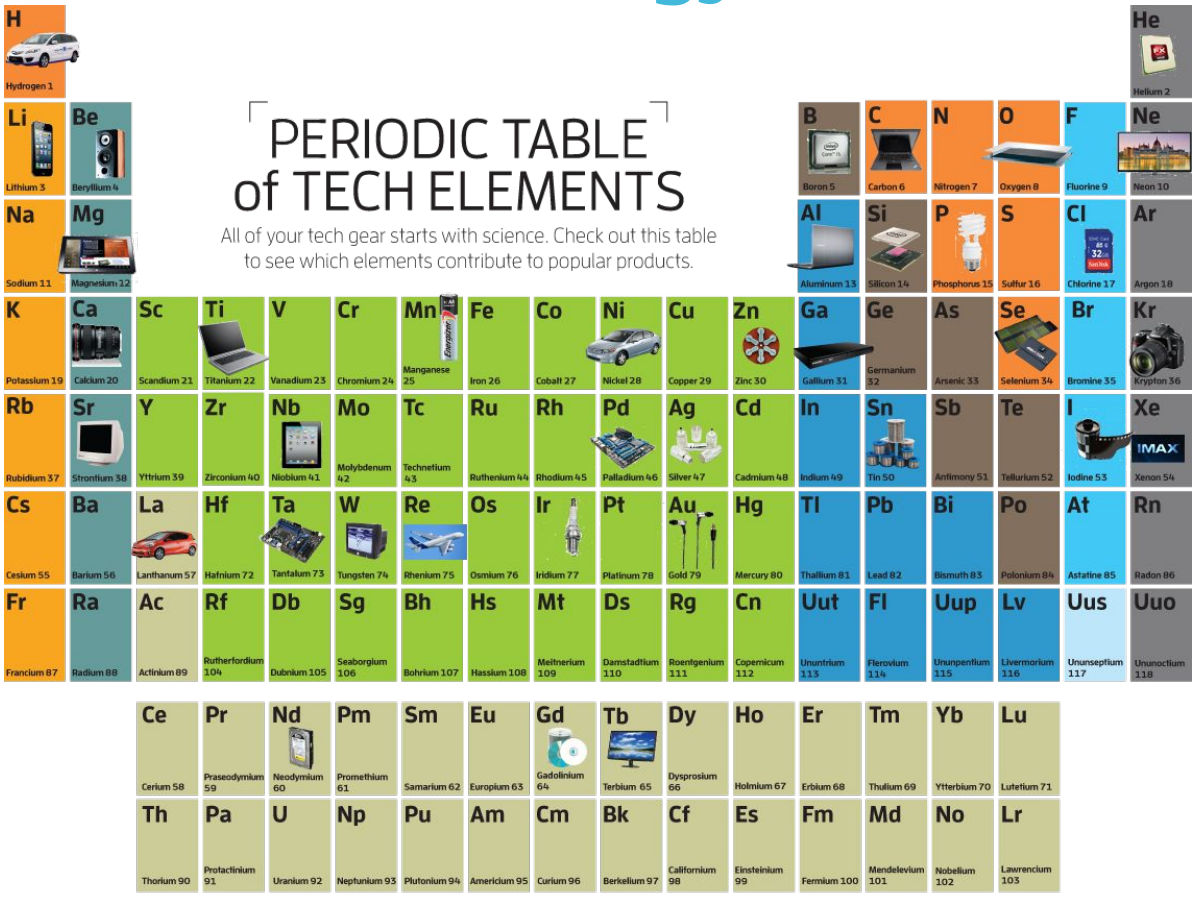
BACK

FRONT

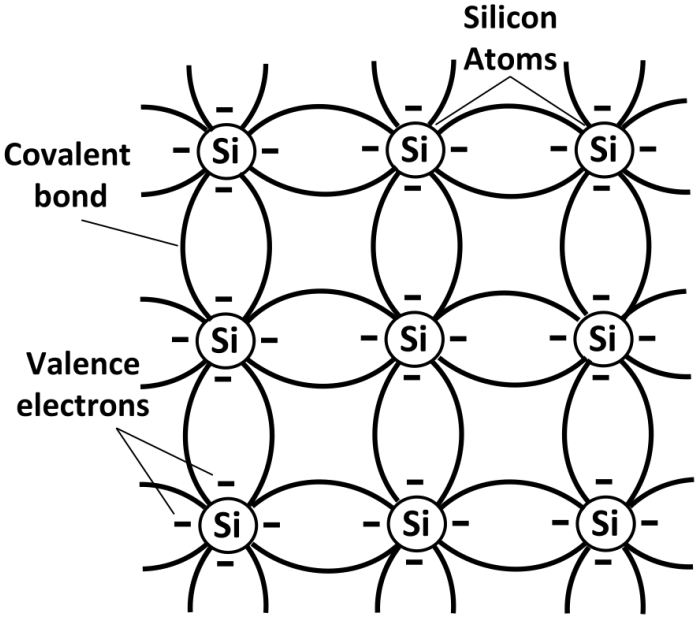
Periodic table



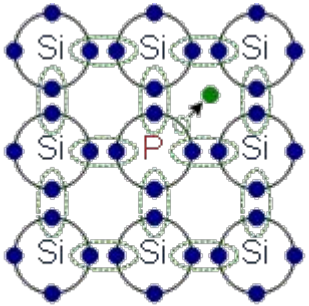
Periodic table of technology



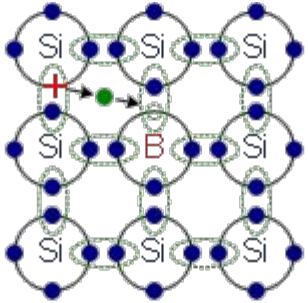
A computer is based on chemistry and physics!



Doping

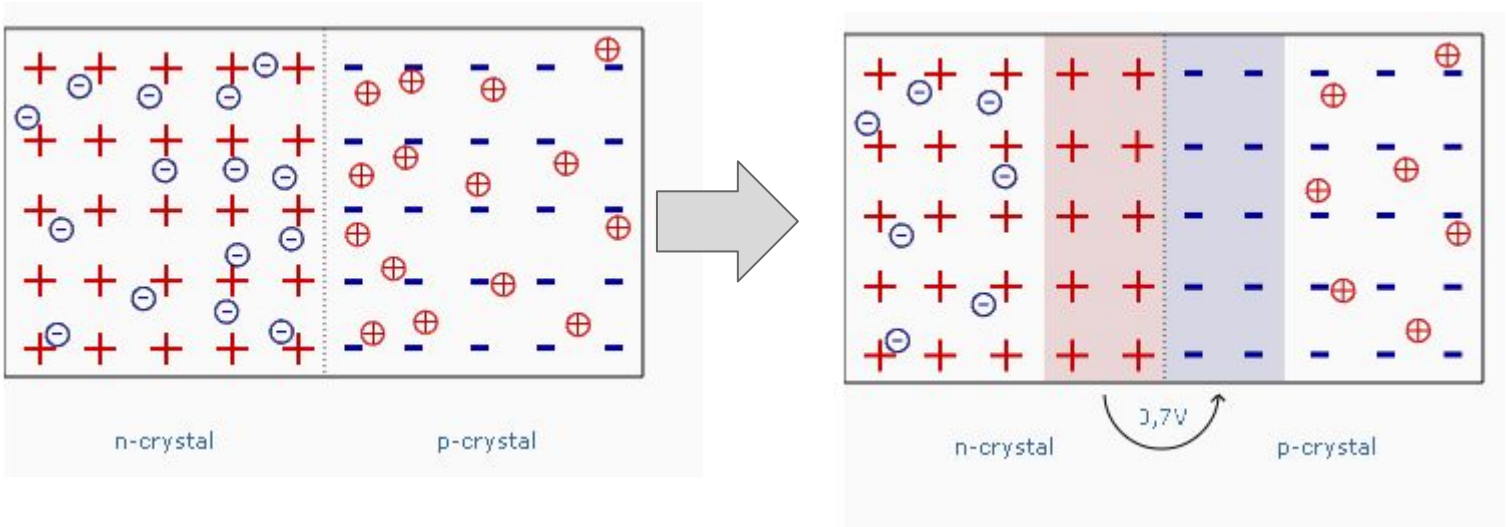


The phosphorus atom donates ist fifth valence electron. It acts as a free charge carrier.

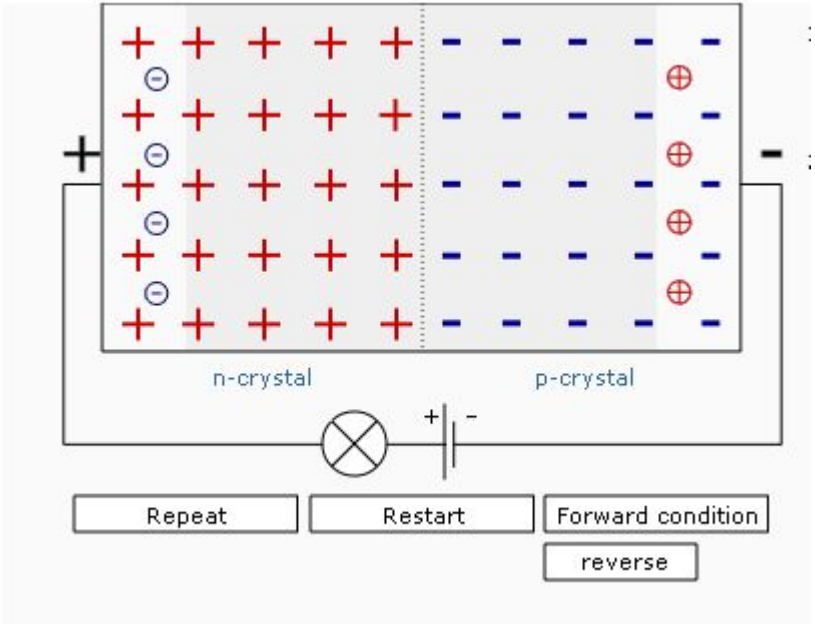


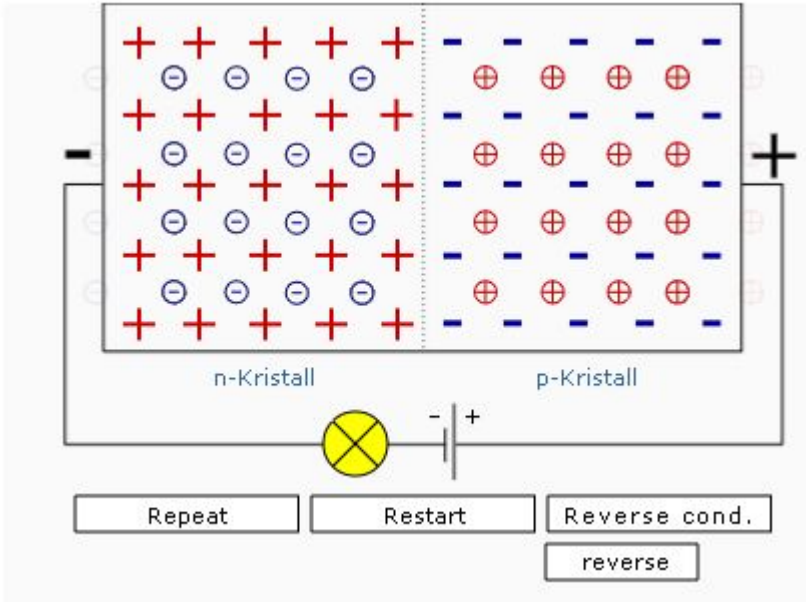
The free place on the boron atom is filled with an electron. Therefore a new hole („defect electron") is generated. This holes move in the opposite direction to the electrons

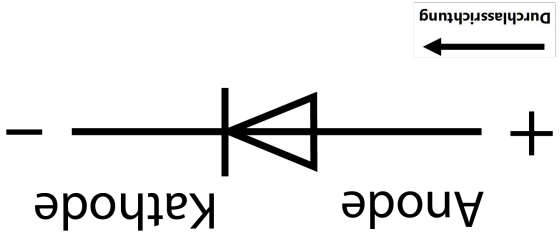
A diode



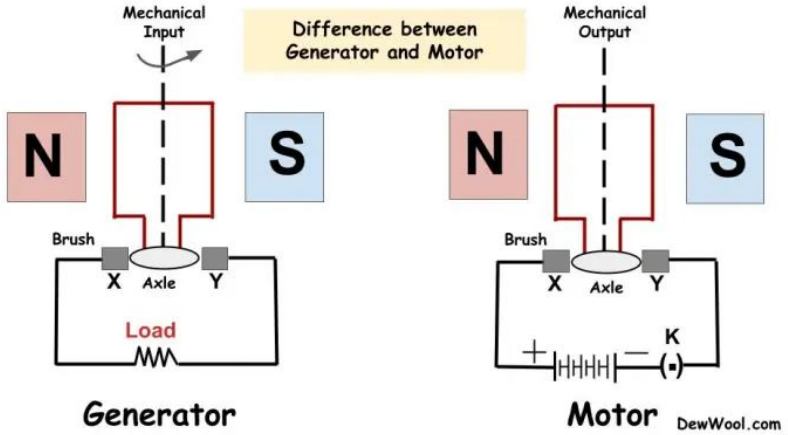
A diode

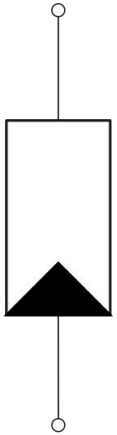

n-crystal
p-crystal
+
-
Repeat
Restart
Forward condition
reverse

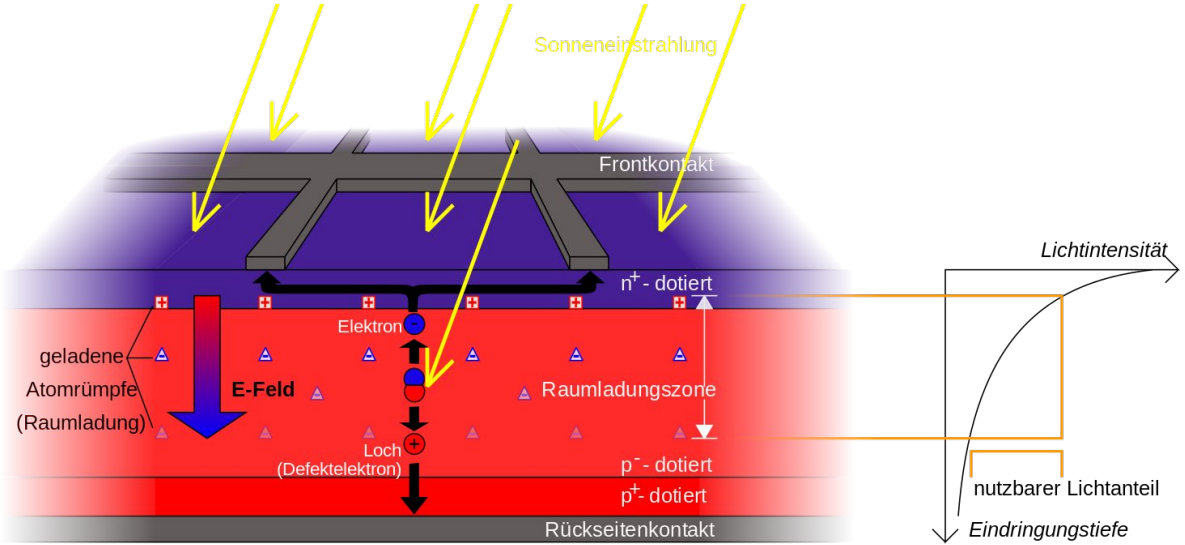

n-Kristall
p-Kristall
-
+
Repeat
Restart
Reverse cond.
reverse


Durchlassrichtung
Anode
Kathode
+
-

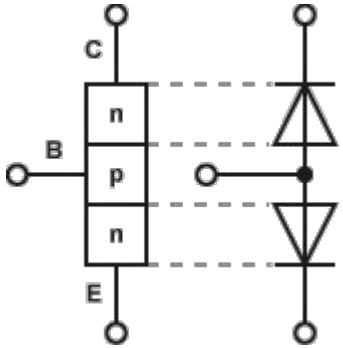
A solar cell

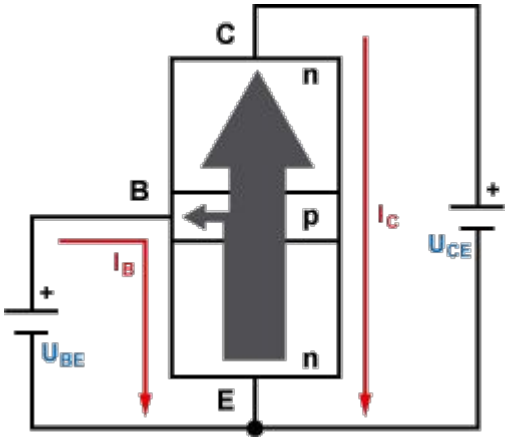

Mechanical Input
Difference between Generator and Motor
Mechanical Output
N
S
Brush
X
Axle
Y
Load
Generator
K
Motor
DewWool.com

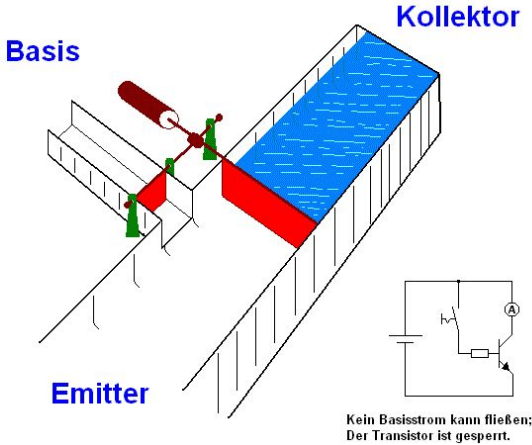


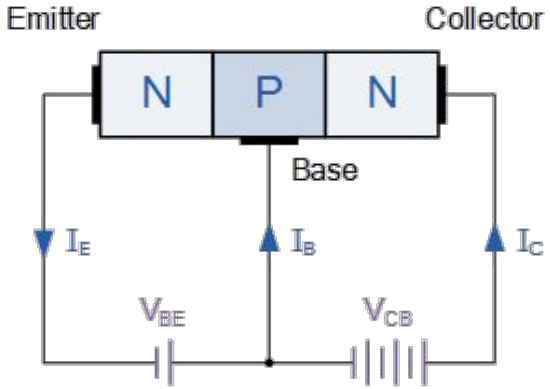

Sonneneinstrahlung
Frontkontakt
n+ - dotiert
Elektron
geladene
Atomrümpfe
(Raumladung)
E-Feld
Raumladungszone
Loch
(Defektelektron)
p- - dotiert
p+ - dotiert
Rückseitenkontakt
Lichtintensität
nutzbarer Lichtanteil
Eindringungstiefe

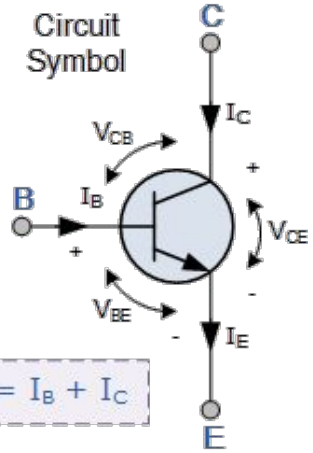
A transistor









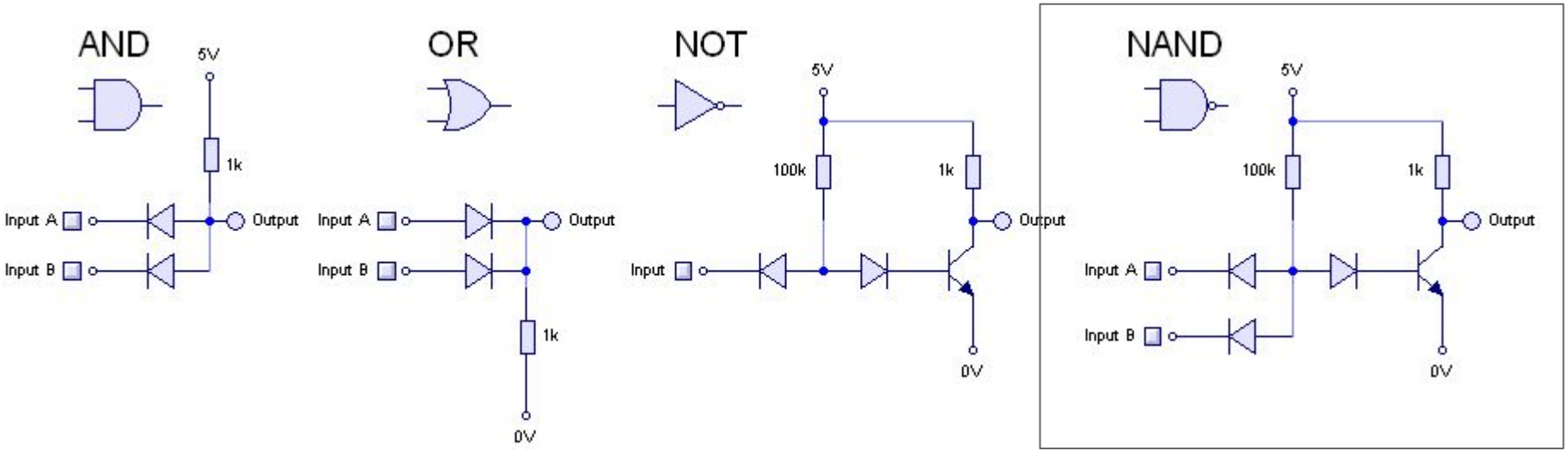


$I_E = I_B + I_C$

Logic gates

Drawn on black board

Transistor can perform computations



NAND is logic complete, i.e. can be used to gain every other function!

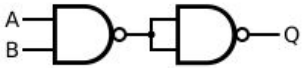
NOR as well

Why is it logic complete?

Desired AND Gate

Q = A AND **B**

NAND Construction



= (**A** NAND **B**) NAND (**A** NAND **B**)

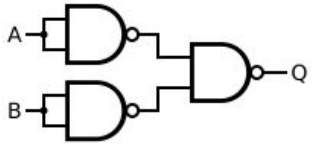
Truth Table

Input A	Input B	Output Q
0	0	0
0	1	0
1	0	0
1	1	1

Desired OR Gate

Q = A OR **B**

NAND Construction

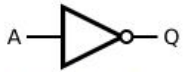


= (**A** NAND **A**) NAND (**B** NAND **B**)

Truth Table

Input A	Input B	Output Q
0	0	0
0	1	1
1	0	1
1	1	1

Desired NOT Gate



Q = NOT(A)

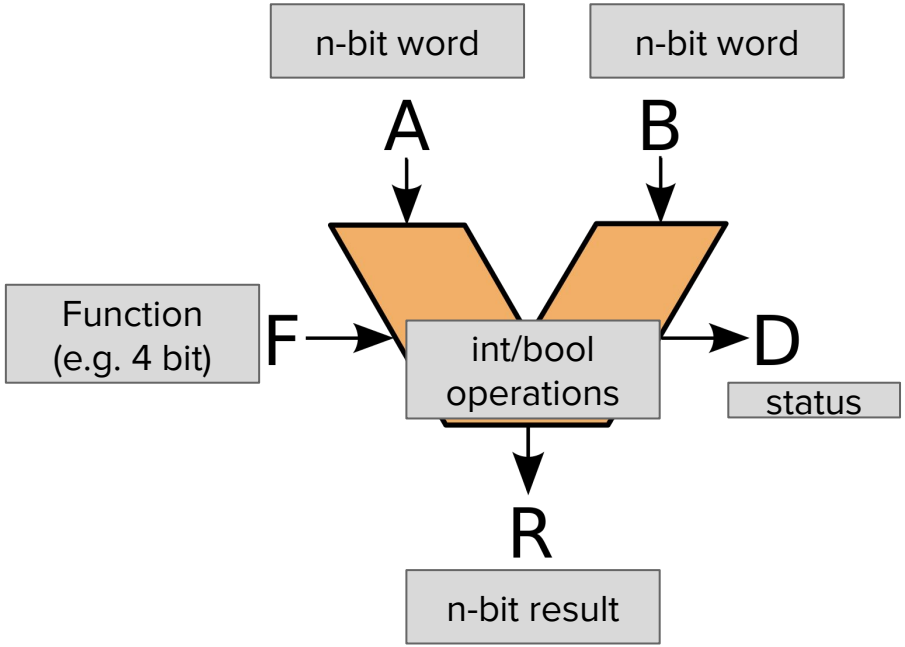
NAND Construction

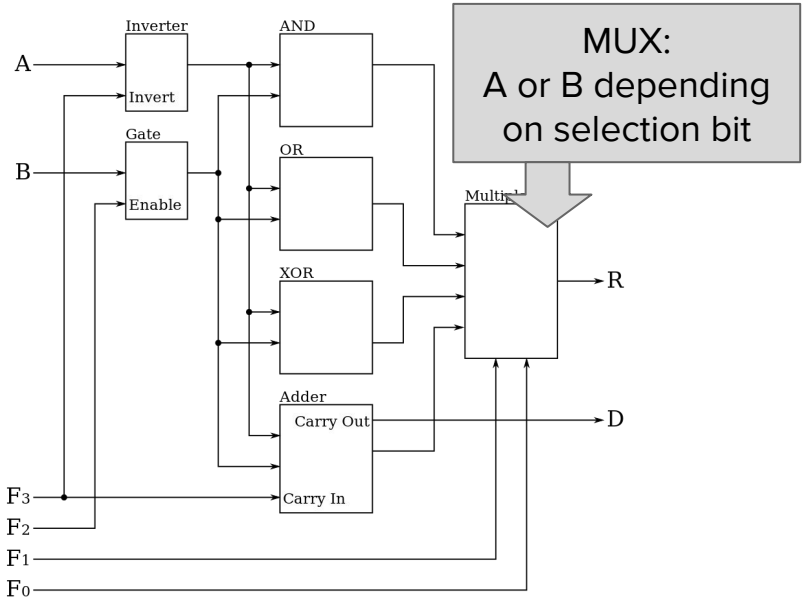
= **A** NAND **A**

Truth Table

Input A	Output Q
0	1
1	0

Arithmetic Logic Unit (ALU)

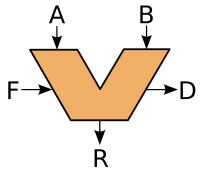




All arithmetic can be used (+,-,*,/) and logic gates; control bits define what operations is performed

https://www.youtube.com/watch?v=IWhuXhDWqtI

From ALU to FPU



Steuertabelle für *n*-Bit ALU

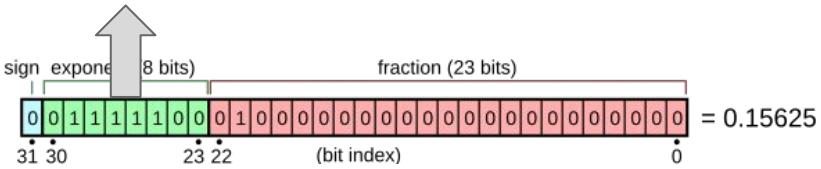
F_3	F_2	F_1	F_0	R
0	0	0	0	0
0	0	0	1	A
1	0	0	1	NOT A
0	1	0	0	A AND B
0	1	0	1	A OR B
0	1	1	0	A XOR B
0	1	1	1	A + B
1	1	1	1	B – A

Floating point number

FPUs

FP32 - IEEE 754

$$12.345 = \underbrace{12345}_{\text{significand}} \times \underbrace{10^{\overbrace{-3}^{\text{exponent}}}}_{\text{base}}$$

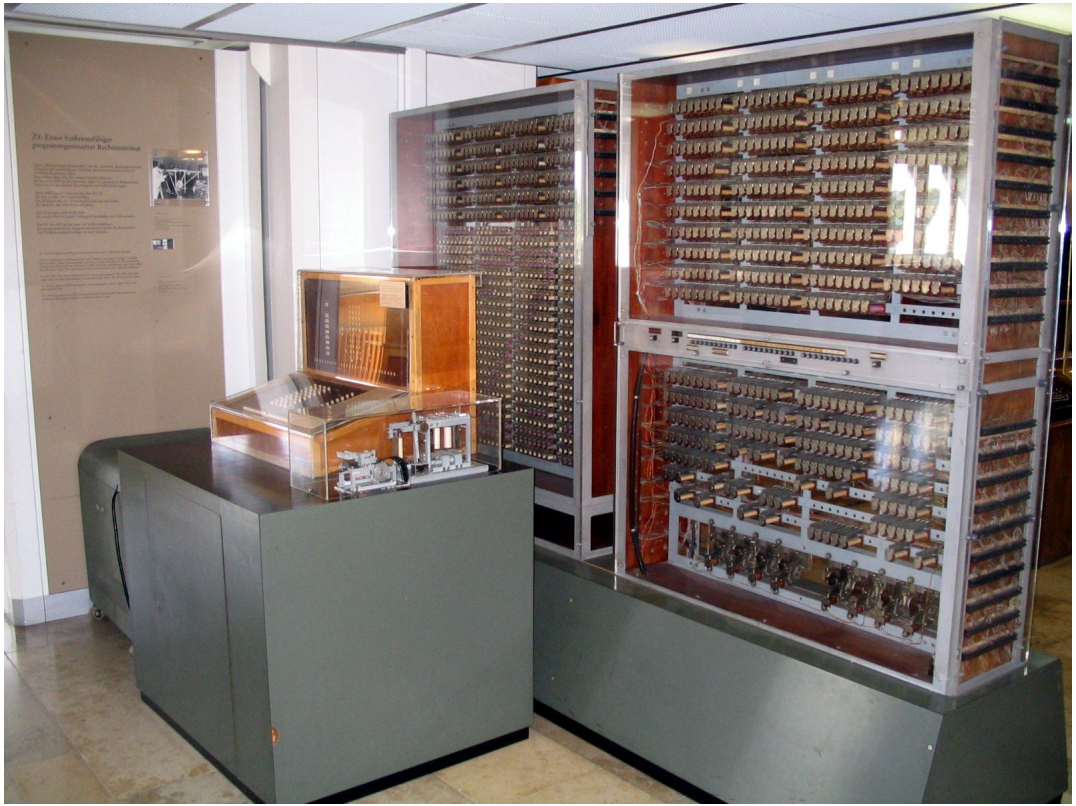


$$(-1)^{b_{31}} \times 2^{(b_{30}b_{29}\ldots b_{23})_2 - 127} \times (1.b_{22}b_{21}\ldots b_0)_2,$$

$$\text{value} = (-1)^{\text{sign}} \times 2^{(E-127)} \times \left(1 + \sum_{i=1}^{23} b_{23-i}2^{-i}\right).$$

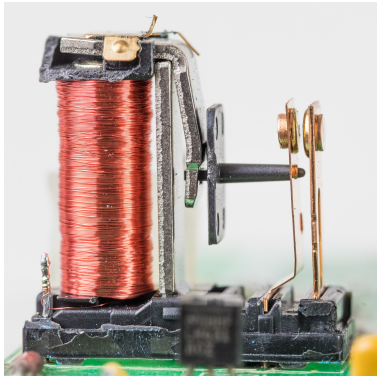
```
bit 23 = 1
bit 22 = 0.5
bit 21 = 0.25
bit 20 = 0.125
bit 19 = 0.0625
bit 18 = 0.03125
bit 17 = 0.015625
.
.
bit 6 = 0.00000762939453125
bit 5 = 0.000003814697265625
bit 4 = 0.0000019073486328125
bit 3 = 0.00000095367431640625
bit 2 = 0.000000476837158203125
bit 1 = 0.0000002384185791015625
bit 0 = 0.00000011920928955078125
```

Zuse Z3



First Universal PC of The world (1941) by Konrad Zuse (Berlin)

Built using ~8k Relais

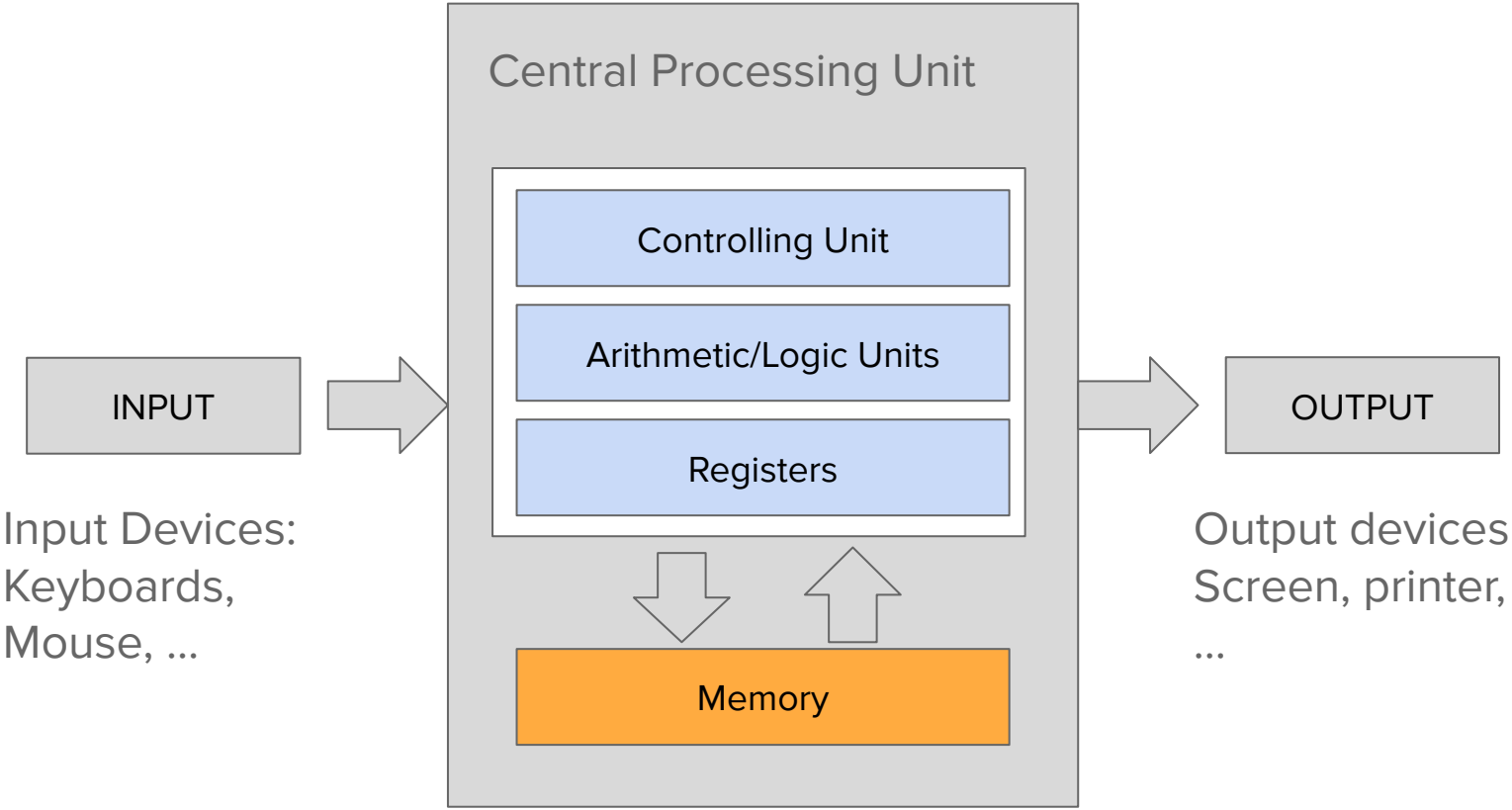


FPUs

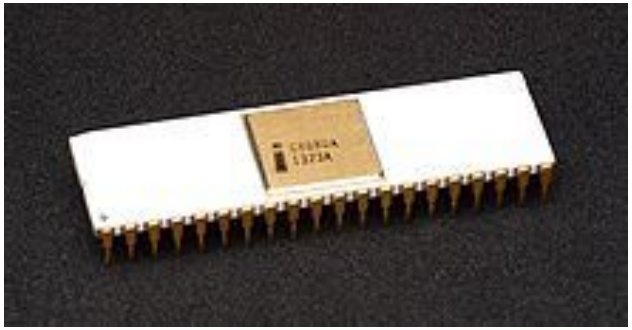
You have three ways to integrate a floating-point arithmetic;

- Software-wise using a floating point arithmetic library
- Using Add-On hardware (co processor)
- Integrated into the main processor

Von Neumann architecture



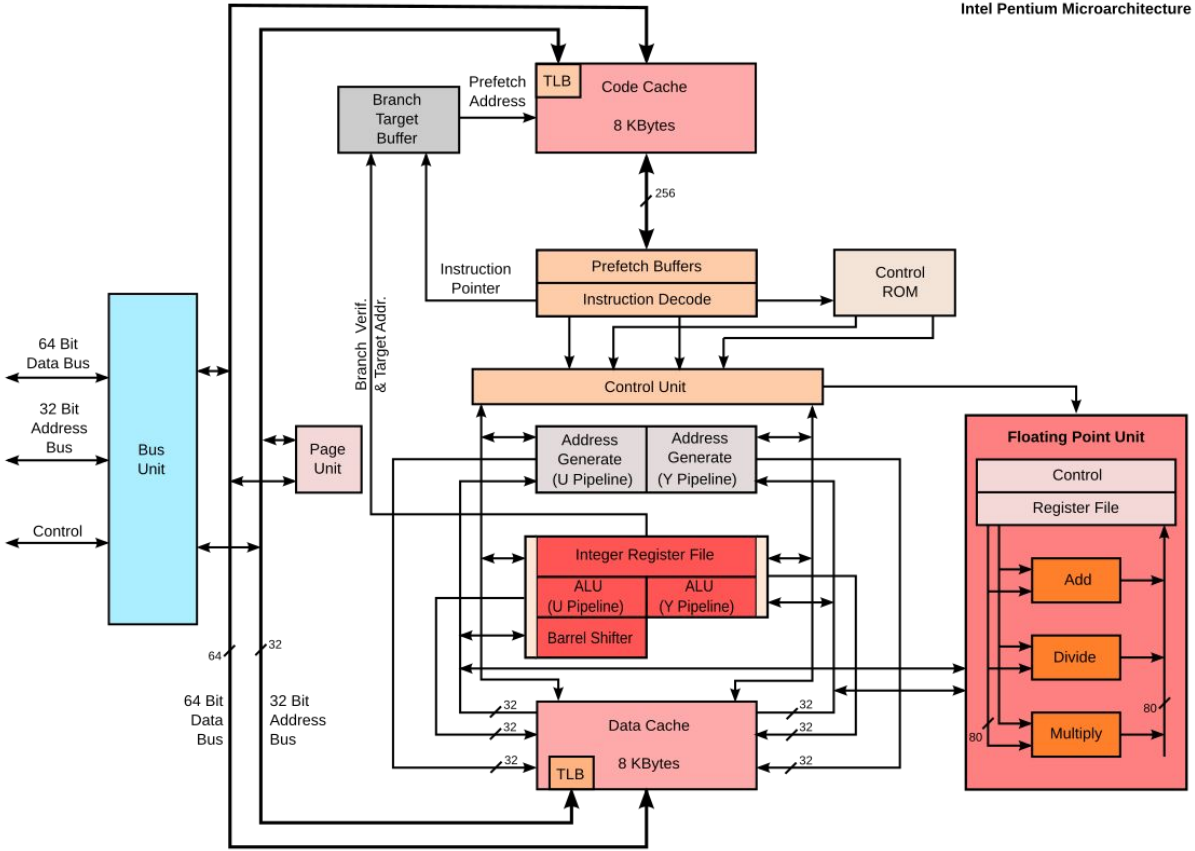
A modern CPU (central processing unit)



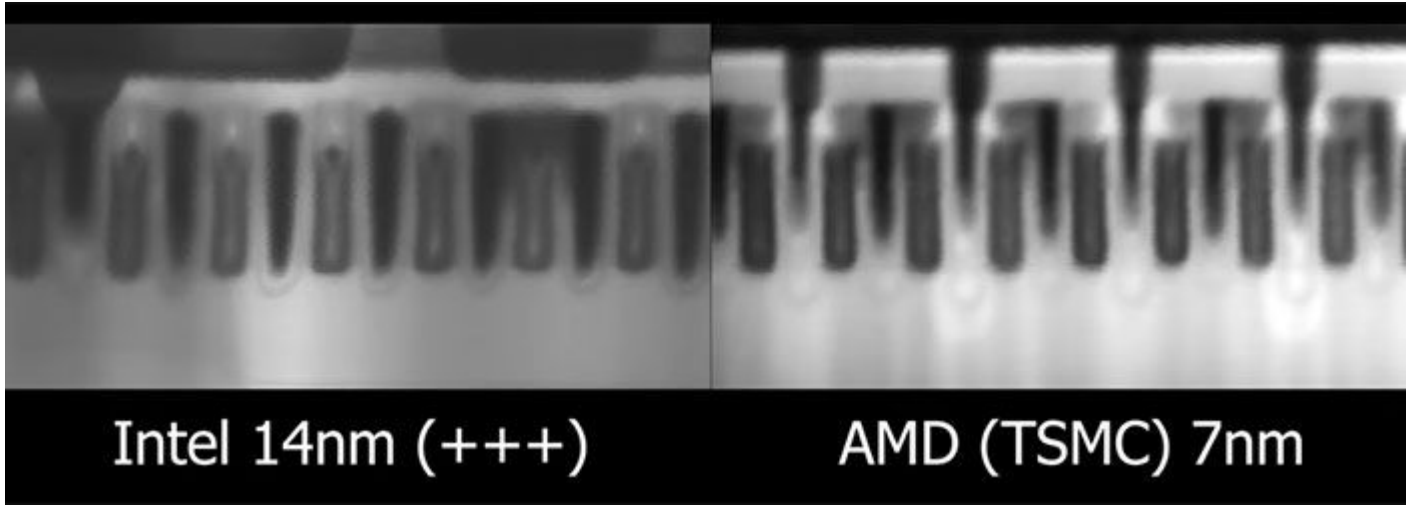
Intel's 8 bit processor (8080, 1974),
6000 Transistors, 6 um per transistor

60 MHz
0.8 um (transistor)
3.1M transistors
1993

AMD Athlon
First consumer CPU > 1 GHz
End 90s, beginning 2000s
0.25 um - 0.13 um


Intel Pentium Microarchitecture
Branch Target Buffer
Prefetch Address
TLB
Code Cache
8 KBytes
256
Instruction Pointer
Prefetch Buffers
Instruction Decode
Control ROM
Branch Verif. & Target Addr.
64 Bit Data Bus
32 Bit Address Bus
Control
Bus Unit
Page Unit
Control Unit
Address Generate (U Pipeline)
Address Generate (Y Pipeline)
Floating Point Unit
Control
Register File
Integer Register File
ALU (U Pipeline)
ALU (Y Pipeline)
Barrel Shifter
Add
Divide
Multiply
64
32
64 Bit Data Bus
32 Bit Address Bus
Data Cache
8 KBytes
TLB
80

Computations using transistors



https://www.youtube.com/watch?v=1kQUXpZpLXI

Single-Core vs. Multi-Cores

Multi-Core Processors: A New Way Forward and Challenges

Abinash Roy, Jingye Xu and Masud H. Chowdhury

Department of Electrical and Computer Engineering, University of Illinois at Chicago
Chicago, IL 60607, USA

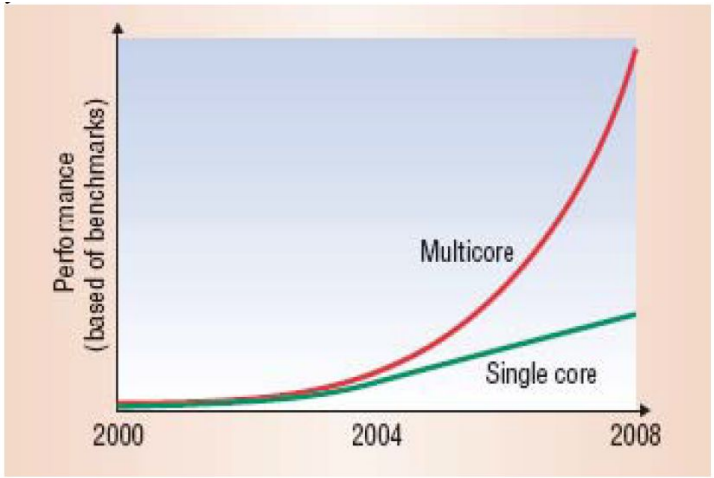
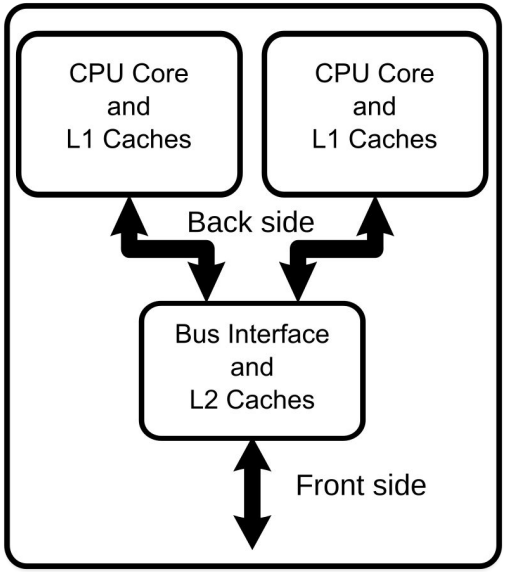


Figure 1. Performance comparison between a single core and multi-core processor



We talk about this in more detail end of the semester...

Nowadays CPU

Lagernd

Artnr: 76971

AMD Ryzen 5 7600X3D 6x 4.10GHz
So.AM5 WOF

€ 309,-*

über **890** verkauft

Lagernd

Artnr: 75431

AMD Ryzen 7 7800X3D 8x 4.20GHz
So.AM5 WOF

€ 447,-*

über **74.520** verkauft

Lagernd

Artnr: 9107257

AMD Ryzen 5 7500F 6x 3.70GHz
So.AM5 TRAY

€ 146,89*

über **13.240** verkauft

Lagernd

Artnr: 74025

AMD Ryzen 5 7600X 6x 4.70GHz
So.AM5 WOF

€ 203,89*

über **18.960** verkauft

Lagernd

Artnr: 76645

AMD Ryzen 7 5700X3D 8x 3.00GHz
So.AM4 WOF

€ 199,-*

über **3.430** verkauft

Lagernd

Artnr: 74032

AMD Ryzen 9 7950X 16x 4.50GHz
So.AM5 WOF

€ 470,25*

über **6.420** verkauft

Lagernd

Artnr: 74028

AMD Ryzen 7 7700X 8x 4.50GHz
So.AM5 WOF

€ 289,69*

über **13.930** verkauft

Lagernd

Artnr: 75355

AMD Ryzen 5 7600 6x 3.80GHz
So.AM5 BOX

€ 185,25*

über **10.070** verkauft

Lagernd

Artnr: 76764

AMD Ryzen 5 8600G 6x 4.30GHz
So.AM5 BOX

€ 166,88*

über **1.270** verkauft

Lagernd

Artnr: 76799

AMD Ryzen 5 5600GT 6x 3.60GHz
So.AM4 BOX

€ 119,86*

über **930** verkauft

CPU - Power Horses

AMD Ryzen Threadripper PRO 7995WX, 96C/192T, 2.50-5.10GHz, tray

100-000000884

★★★★★ jetzt bewerten!

Info beim Hersteller

Alle 12 Varianten anzeigen — ab **€ 1718,79** — 170 Angebote

Kerne	96 (96C)
Threads	192
Turbotakt	5.10GHz
Basistakt	2.50GHz
TDP	350W
Grafik	nein
Sockel	AMD sTR5 (LGA4844)
Chipsatz-Eignung	TRX50, WRX90
Codename	Storm Peak
Architektur	Zen 4
Fertigung	TSMC 5nm (CPU), TSMC 6nm (I/O)
L2-Cache	96MB (96x 1MB)

Aktueller Preisbereich

€ 10845,78 bis **€ 12945,56**

Preisentwicklung 1W 1M 3M 6M **1J**

Preisentwicklung öffnen
Preisalarm setzen
Zur Wunschliste hinzufügen
Zur Vergleichsliste hinzufügen
Feedback senden

Server CPUs

Verfügbar

Artnr: 8972690

AMD Epyc 7F32 8x 3.70GHz So.SP3 TRAY

€ 2.109,47*

inkl. 19% USt + Versandkosten

Verfügbar

Artnr: 8978613

AMD Epyc 7F52 16x 3.50GHz So.SP3 TRAY

€ 3.426,95*

inkl. 19% USt + Versandkosten

über **5** verkauft

Verfügbar

Artnr: 8984108

AMD Epyc 7H12 64x 2.60GHz So.SP3 TRAY

€ 5.900,86*

inkl. 19% USt + Versandkosten

Verfügbar

Artnr: 8847113

AMD Epyc 7401P 24x 2.00GHz So.SP3 TRAY

€ 1.064,86*

inkl. 19% USt + Versandkosten

über **10** verkauft

Verfügbar

Artnr: 9029943

AMD EPYC MILAN 48-CORE 7643 2.3GHZ

€ 5.421,50*

inkl. 19% USt + Versandkosten

CPU - what for?!

- **General-purpose processors:** Used for a wide variety of tasks, including data preprocessing, model management, and running less parallelized algorithms.
- **Best for sequential tasks:** Ideal for tasks that require strong single-threaded performance or that do not benefit much from parallelization.
- **Data wrangling and preprocessing:** Often used for data cleaning, feature extraction, and data transformation since these tasks may involve logic-heavy operations.
- **Model development:** Useful for running traditional machine learning models (e.g., linear regression, decision trees, SVMs) that do not require heavy parallelization.
- **Small-to-medium datasets:** Efficient for computations that do not need specialized hardware, or for smaller datasets that fit in memory.
- **Versatility:** Can be used for general system tasks, running Jupyter notebooks, and handling mixed workloads (combining I/O and computation).

One core vs multiple cores

- Single core performance on Intel mostly better than on AMD
  ➜ most code you write
- Multiple core performance (AMD > Intel)
  ➜ multithreading (simultaneous runs of programs without stopping the other one)
  ➜ multiprocessing (computations distributed across cores)
  ⇒ Later lectures!
- In general, in a scientific environment it depends ON YOUR TASK how many cores you need.

Fluid dynamics: lots of cores ➜ cpu cluster

Deep learning: lots of ???

Our High Performance Cluster (HPC)

Cluster “Fritz” @ RRZE

...

- **992 compute node**s with direct liquid cooling (DLC), each with two **Intel Xeon Platinum 8360Y “Ice Lake”** processors (36 cores per chip) running at a base frequency of 2.4 GHz and 54 MB Shared L3 cache per chip, **256 GB of DDR4-RAM**.

...

Intel Xeon Platinum 8360Y “IceLake” Processor

Hyperthreading (SMT) is disabled; sub-NUMA clustering (Cluster-on-Die, CoD) is activated. This results in 4 NUMA domains with 18 cores each per compute node.

The processor can be operated in 3 modes; in Fritz it’s running in its default mode with 36 cores and 250 W TDP.

Launch Date	Q2’21
Lithography	10 nm
Total Cores (Threads)	36 (72 – SMT is disabled on Fritz)
Max Turbo Frequency (non-AVX code)	3.50 GHz (significantly lower for heavy AVX2/AVX512 workload)
Processor Base Frequency (non-AVX code)	2.40 GHz (significantly lower for heavy AVX2/AVX512 workload)
Last level cache (L3)	54 MB
# of UPI Links	3
TDP	250 W
Memory Channels & Memory Type	8 channels DDR4 @ 3200 per socket (in Fritz: 16x 16 GB DRR4-3200 per node)
Instruction Set Extensions	Intel SSE4.2, Intel AVX, Intel AVX2, Intel AVX-512
# of AVX-512 FMA Units	2

https://hpc.fau.de/systems-services/documentation-instructions/clusters/fritz-cluster/

Alex cluster

FAU's **Alex cluster** (system integrator: Megware) is a high-performance compute resource with Nvidia GPGPU accelerators and partially high speed interconnect. It is intended for single and multi GPGPU workloads, e.g. from molecular dynamics, or machine learning. Alex serves for both, FAU's basic Tier3 resources as well as NHR's project resources.

- **2 front end nodes**, each with two AMD EPYC 7713 "Milan" processors (64 cores per chip) running at 2.0 GHz with 256 MB Shared L3 cache per chip, 512 GB of RAM, and 100 GbE connection to RRZE's network backbone but no GPGPUs.
- **20 GPGPU nodes**, each with two AMD EPYC 7713 "Milan" processors (64 cores per chip) running at 2.0 GHz with 256 MB Shared L3 cache per chip, 1,024 GB of DDR4-RAM, **eight Nvidia A100 (each 40 GB HBM2 @ 1,555 GB/s; HGX board with NVLink; 9.7 TFlop/s in FP64 or 19.5 TFlop/s in FP32)**, two HDR200 Infiniband HCAs, 25 GbE, and 14 TB on local NVMe SSDs.
- **15 GPGPU nodes**, each with two AMD EPYC 7713 "Milan" processors (64 cores per chip) running at 2.0 GHz with 256 MB Shared L3 cache per chip, 2,048 GB of DDR4-RAM, **eight Nvidia A100 (each 80 GB HBM2 @ 2,039 GB/s; HGX board with NVLink; 9.7 TFlop/s in FP64 or 19.5 TFlop/s in FP32)**, two HDR200 Infiniband HCAs, 25 GbE, and 14 TB on local NVMe SSDs.
- **38 GPGPU nodes**, each with two AMD EPYC 7713 "Milan" processors (64 cores per chip) running at 2.0 GHz with 256 MB Shared L3Cache per chip, 512 GB of DDR4-RAM, **eight Nvidia A40 (each with 48 GB DDR6 @ 696 GB/s; 37.42 TFlop/s in FP32)**, 25 GbE, and 7 TB on local NVMe SSDs.

Verfügbar

Artnr: 9035623

AMD EPYC MILAN 64-CORE 7713P 2.0GHZ tray

€ 5.642,59*

inkl. 19% USt + Versandkosten

über **5** verkauft

Verfügbar

Artnr: 9019893

AMD Epyc 7713 64x 2.00GHz So.SP3 TRAY

€ 4.896,74*

inkl. 19% USt + Versandkosten

AMD EPYC 7713 "Milan" Processor

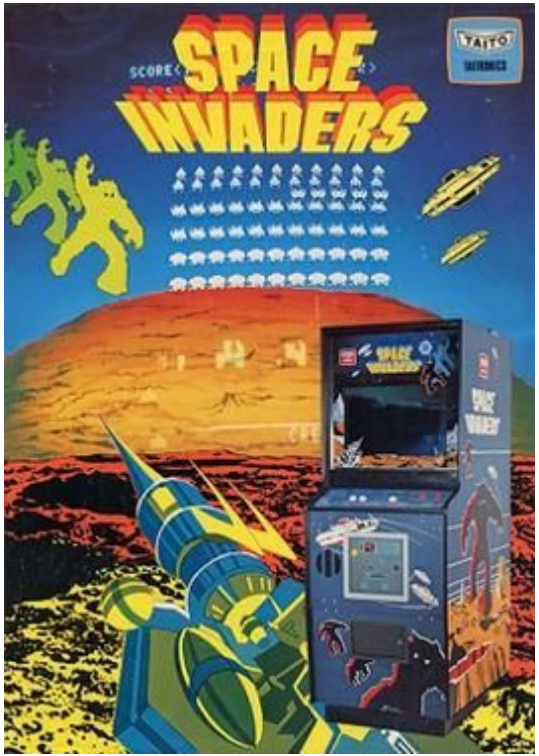
Each node has two processor chips. The specs per processor chip are as follows:

- # of CPU Cores: 64
- # of Threads: 128 – *hyperthreading (SMT) is disabled on Alex for security reasons; thus, threads and physical cores are identical*
- Max. Boost Clock: Up to 3.675 GHz
- Base Clock: 2.0 GHz
- Default TDP: 225W; AMD Configurable TDP (cTDP): 225-240W
- Total L3 Cache: 256MB
- System Memory Type: DDR4 @ 3,200 MHz
- Memory Channels: 8 – *these can be arranged in 1-4 ccNUMA domains ("NPS" setting); Alex is running with NPS=4*
- Theoretical per Socket Mem BW: 204.8 GB/s

https://hpc.fau.de/systems-services/documentation-instructions/clusters/alex-cluster/

The GPU

Graphics Processing Unit
➜ started in the 1970s to play video games



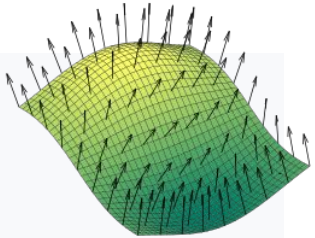
Increasing Complexity of Graphics:

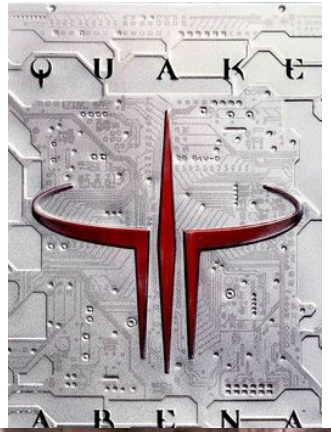
- 80s/90s: video games need **real-time rendering** of increasingly complex graphics.
- CPUs struggled to efficiently render the highly parallelizable tasks of **drawing pixels**, **textures**, and **shading** needed for 3D environments.
- Processing **hundreds of thousands of pixels** (later millions) per frame, which CPUs couldn't handle without significant slowdowns. Dedicated hardware to perform **parallel processing**.
- The transition from 2D to **3D rendering** in the 1990s further increased computational demands. 3D graphics involved intensive calculations like matrix operations, vector transformations, and lighting

People become creative...

$$\frac{1}{\sqrt{x}},$$

Quake III (1999) Fast Inverse Square Root Function



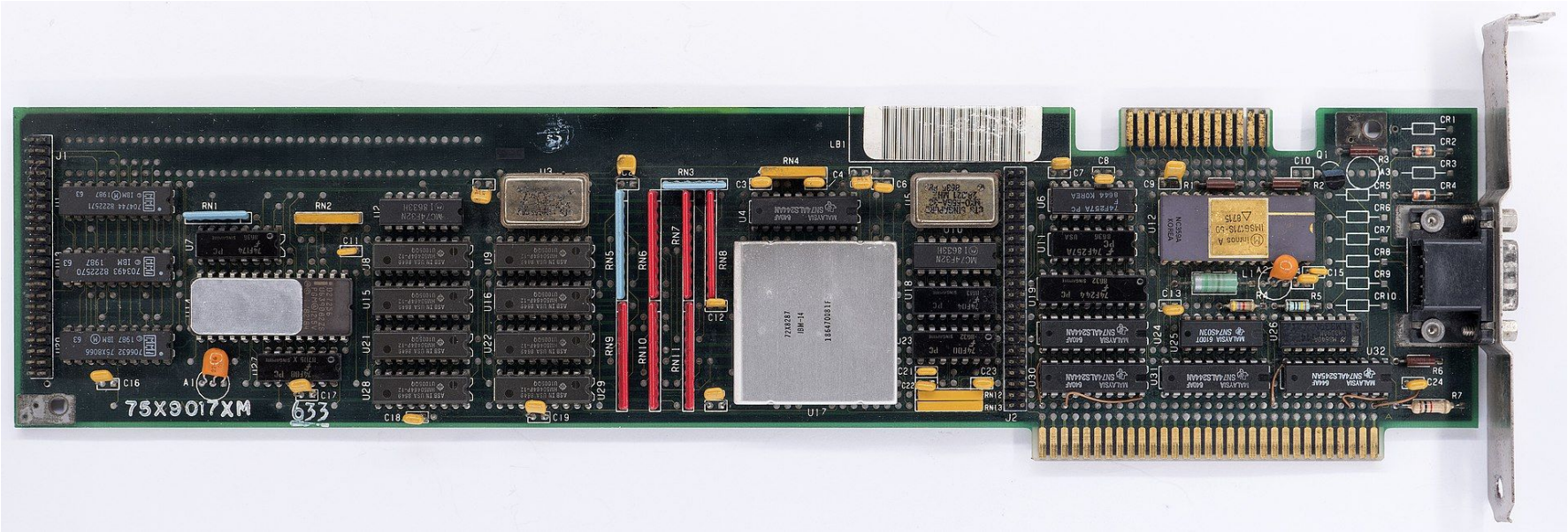



```
float Q_rsqrt( float number )
{
    long i;
    float x2, y;
    const float threehalfs = 1.5F;

    x2 = number * 0.5F;
    y  = number;
    i  = * ( long * ) &y;                       // evil floating point bit
level hacking
    i  = 0x5f3759df - ( i >> 1 );               // what the fuck?
    y  = * ( float * ) &i;
    y  = y * ( threehalfs - ( x2 * y * y ) );   // 1st iteration
//  y  = y * ( threehalfs - ( x2 * y * y ) );   // 2nd iteration, this can be
removed

    return y;
}
```

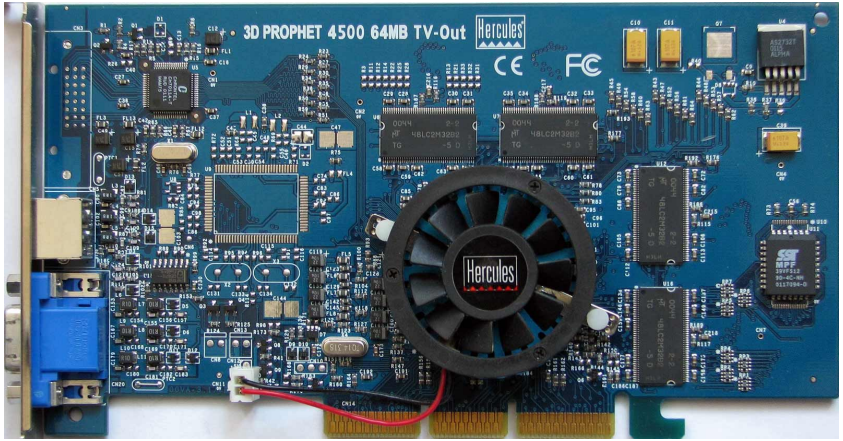
Further watching: https://www.youtube.com/watch?v=p8u_k2LIZyo

IBM's VGA (Video Graphics Array)

Introduced 1987 → VGA resolution is 640x480



Where I spent my pocket money...



Hercules Kyro II

NVIDIA GeForce4 Ti 4200-8X

NV28	4	2	8	4	64 MB	SDR	128 bit
GRAPHICS PROCESSOR	PIXEL SHADERS	VERTEX SHADERS	TMUS	ROPS	MEMORY SIZE	MEMORY TYPE	BUS WIDTH

Recommer

System requirements

- Windows 98/Me/2000/XP, Linux 2.2+ or Mac OS X 10.2.6+
- Pentium III or AMD Athlon 1.0 GHz processor
- 128 MB RAM minimum (256 MB recommended)
- GeForce 2 MX or Radeon with 32 MB RAM (64 MB video memory recommended)
- 3 GB HDD space (3.5 GB free recommended)
- DirectX 8.1 or OpenGL 1.2

GPUs these days

Screenshot from WS 2021/22

GIGABYTE™ | msi

Lagernd
Artnr: 70637

8GB MSI GeForce RTX 3070 GAMING Z TRIO LHR DDR6 (Retail)

€ 999,-*

inkl. 19% USt, Gratisversand

über **580** verkauft

Lagernd
Artnr: 74407

24GB MSI GeForce RTX 3090 GAMING X TRIO Aktiv PCIe 4.0 x16

€ 2.499,-*

inkl. 19% USt + Versandkosten

über **700** verkauft

Lagernd
Artnr: 70334

8GB Gigabyte GeForce RTX 3070 GAMING OC 8G 2.0 LHR

€ 949,-*

inkl. 19% USt, Gratisversand

★ Mindstar Highlight ★

Lagernd
Artnr: 70988

12GB KFA2 GeForce RTX 3080 Ti Hall Of Fame 1-Click OC GDDR6X

€ 1.949,-*

inkl. 19% USt + Versandkosten

über **5** verkauft

Lagernd
Artnr: 70127

8GB Gigabyte GeForce RTX 3070 Ti AORUS Master Aktiv PCIe 4.0 x16)

€ 1.129,-*

inkl. 19% USt + Versandkosten

über **280** verkauft

Lagernd

Lagernd

Lagernd

Lagernd

Lagernd

GPUs these days

Screenshot from WS 2022/23

Lagernd

Artnr: 75210

12GB Gigabyte GeForce RTX 3060 EAGLE LHR GDDR6 2xHDMI 2xDP

€ 421,70*

inkl. 19% USt + Versandkosten

über **150** verkauft

Lagernd

Artnr: 74415

24GB Gigabyte GeForce RTX 3090 Gaming OC Aktiv PCIe 4.0 x16)

€ 1.139,-*

inkl. 19% USt, Gratisversand

über **1.560** verkauft

Lagernd

Artnr: 73331

24GB Gigabyte GeForce RTX 3090 Ti AORUS XTREME WATERFORCE 3xDP

€ 1.748,99*

inkl. 19% USt + Versandkosten

über **730** verkauft

Lagernd

Artnr: 74603

12GB Gigabyte GeForce RTX 3060 GAMING OC 12G 2.0 LHR

€ 439,-*

inkl. 19% USt, Gratisversand

über **2.130** verkauft

Lagernd

Artnr: 70127

8GB Gigabyte GeForce RTX 3070 Ti AORUS Master Aktiv PCIe 4.0 x16)

€ 798,99*

inkl. 19% USt + Versandkosten

über **1.100** verkauft

Lagernd

Artnr: 72792

8GB Gigabyte GeForce RTX 3050 EAGLE OC GDDR6 2xHDMI 2xDP

€ 348,99*

Bestellt

Artnr: 9019895

6GB Gigabyte GeForce RTX 2060 D6 6G 3xDP/HDMI OC 6G

€ 339,89*

Bestellt

Artnr: 70398

12GB Gigabyte GeForce RTX 3060 EAGLE OC LHR GDDR6 2xHDMI

€ 399,-*

Bestellt

Artnr: 70427

8GB Gigabyte GeForce RTX 3060 TI AORUS Elite LHR E-8G LHR

€ 577,87*

Bestellt

Artnr: 70246

12GB Gigabyte GeForce RTX 3060 VISION OC LHR GDDR6 2xHDMI

€ 439,-*

GPUs these days (Oct 2023)

Lagernd
Artnr: 75097

12GB MSI GeForce RTX 3060 Ventus 2X OC Aktiv PCIe 4.0 x16 (Retail)

€ 297,-*

inkl. 19% USt + Versandkosten

über **7.700** verkauft

Lagernd
Artnr: 76099

8GB MSI GeForce RTX 4060 Ventus 2X Black OC Aktiv PCIe 4.0 x16 (x8)

€ 328,96*

inkl. 19% USt + Versandkosten

über **870** verkauft

Lagernd
Artnr: 9108835

12GB MSI GeForce RTX 4070 Ti Ventus 3X E OC Aktiv PCIe 4.0 x16

€ 878,-*

inkl. 19% USt + Versandkosten

über **220** verkauft

Lagernd
Artnr: 76251

16GB MSI GeForce RTX 4060 Ti Ventus 2X Black OC Aktiv PCIe 4.0

€ 474,-*

inkl. 19% USt + Versandkosten

über **290** verkauft

Lagernd
Artnr: 76338

8GB MSI GeForce RTX 4060 Gaming X NV Edition Aktiv PCIe 3.0 x16 (x8)

€ 365,-*

inkl. 19% USt + Versandkosten

über **10** verkauft

Lagernd
Artnr: 74852

16GB MSI GeForce RTX 4080 Suprim X Aktiv PCIe 4.0 x16 (Retail)

€ 1.299,-*

Lagernd
Artnr: 75973

8GB MSI GeForce RTX 4060 Ti Ventus 3X Aktiv PCIe 4.0 x16 (x8)

€ 464,26*

Lagernd
Artnr: 9108833

12GB MSI GeForce RTX 4070 Ti Gaming X Slim Aktiv PCIe 4.0 x16

€ 894,99*

Lagernd
Artnr: 76180

16GB MSI GeForce RTX 4060 Ti Gaming X Aktiv PCIe 4.0 x16 (x8)

€ 514,26*

Lagernd
Artnr: 9113760

12GB MSI Geforce RTX 4070 VENTUS OC Aktiv PCIe 4.0 x16

€ 608,57*

GPU these days (2024/2025)

ASRock | GIGABYTE | msi | PowerColor | SAPPHIRE | XFX

Lagernd
Artnr: 77011
16GB MSI GeForce RTX 4070 Ti SUPER Gaming Slim Stalker2
€ 944,-*
★ Mindstar Highlight ★

Lagernd
Artnr: 76938
12GB MSI GeForce RTX 4070 SUPER Ventus 3X OC Aktiv PCIe 4.0 x16
€ 653,-*
★ Mindstar Highlight ★

Lagernd
Artnr: 76181
16GB MSI GeForce RTX 4060 Ti Gaming X Slim Aktiv PCIe 4.0 x16
€ 504,-*
★ Mindstar Highlight ★

Lagernd
Artnr: 76741
16GB MSI GeForce RTX 4070 Ti SUPER Ventus 2X OC Aktiv PCIe 4.0
€ 823,-*
★ Mindstar Highlight ★

Lagernd
Artnr: 9124429
6GB MSI GeForce RTX 3050 Ventus 2X 6G OC Aktiv PCIe 4.0 x16 (x8)
€ 187,89*
über **170** verkauft

Lagernd
Artnr: 76655
12GB MSI GeForce RTX 4070 SUPER Ventus 2X OC Aktiv PCIe 4.0 x16
€ 620,-*
★ Mindstar Highlight ★

Lagernd
Artnr: 9109880
8GB MSI GeForce RTX 4060 Ventus 2X White OC Aktiv PCIe 4.0 x16 (x8)
€ 314,-*
★ Mindstar Highlight ★

Lagernd
Artnr: 75097
12GB MSI GeForce RTX 3060 Ventus 2X OC Aktiv PCIe 4.0 x16 (Retail)
€ 283,-*
★ Mindstar Highlight ★

Lagernd
Artnr: 9080681
8GB MSI GeForce RTX 3060 Ventus 2X OC Aktiv PCIe 4.0 x16 (Retail)
€ 266,64*
über **350** verkauft

Lagernd
Artnr: 75832
8GB MSI GeForce RTX 4060 Ti Gaming X Aktiv PCIe 4.0 x16 (x8)
€ 405,-*
★ Mindstar Highlight ★

What do GPUs do?

The First True GPU – NVIDIA GeForce 256 (1999):

- **NVIDIA** is credited with creating the first true GPU with the release of the **GeForce 256** in 1999. NVIDIA defined it as a **Graphics Processing Unit** capable of **hardware transform and lighting (T&L)**, which offloaded not just rasterization but also the mathematical transformations required to display 3D objects (such as moving, rotating, and lighting them) from the CPU.
- This GPU was a significant leap forward as it combined:
  - **Vertex processing**: Handling transformations and lighting calculations on 3D models.
  - **Pixel processing**: Managing the rasterization of 3D models into 2D images.
- It also introduced the concept of **parallel processing**,
  with multiple cores working simultaneously on different parts of the image.

SUMMARY

- Highly parallel processing
- Dedicated ICs for video encoding and decoding
- Ray tracing/rendering
- Matrix multiplications!

GPGPUs (General-Purpose Computing on GPUs)

- Around 2006, researchers realized that the parallel architecture of GPUs could be leveraged for **general-purpose computing** beyond graphics (i.e., **GPGPU**). This led to using GPUs for tasks like matrix operations, simulations, and eventually, **deep learning**.

- NVIDIA capitalized on this trend with the release of **CUDA** (Compute Unified Device Architecture) in 2006, which allowed developers to program GPUs for non-graphics applications.

- This flexibility unlocked the GPU's potential for fields like data science, scientific simulations, and neural network training, **making GPUs essential for modern AI workloads.**

NVIDIA

PROFILE

Jensen Huang

CEO & President, NVIDIA

$120.4B Real Time Net Worth
as of 10/20/24
#11 in the world today

11th richest person on earth

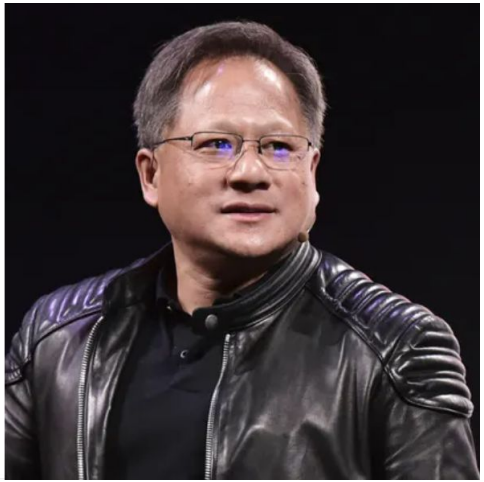
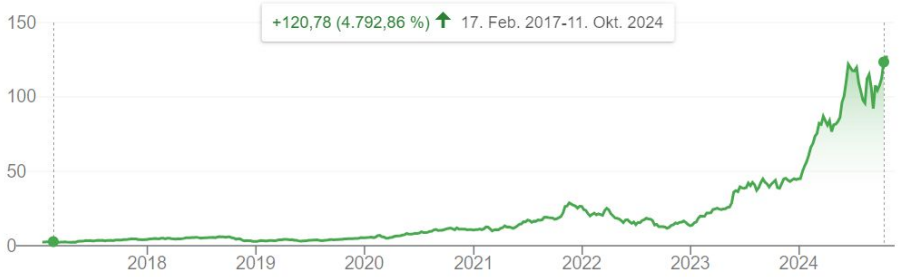


PHOTO BY MANDEL NGAN/AFP/GETTY IMAGES

126,90 EUR
+124,46 (5.100,82 %) ↑ immer
18. Okt., 17:35 MESZ • Haftungsausschluss

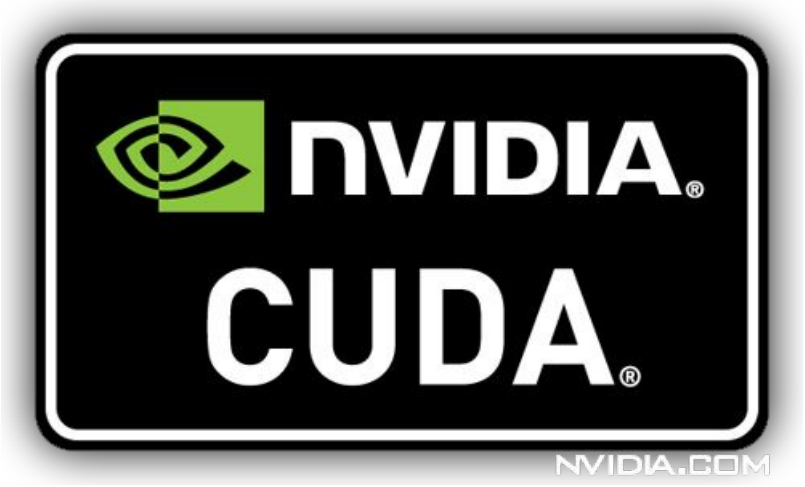
+ Folgen

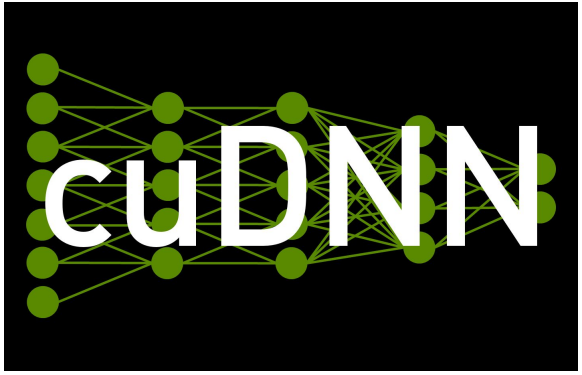
1 T. | 5 T. | 1 M. | 6 M. | YTD | 1 J. | 5 J. | Max.

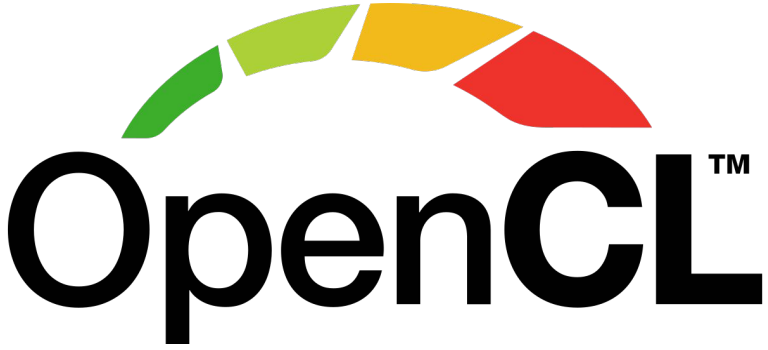


Founded 1993;
IPO 1999 w/ 12 USD

AMD vs NVIDIA?!







General purpose, but 30% slower on NVIDIA GPUs...

GPU specs (example RTX 3090)

Graphics Processor	
GPU Name:	GA102
GPU Variant:	GA102-300-A1
Architecture:	Ampere
Foundry:	Samsung
Process Size:	8 nm
Transistors:	28,300 million
Die Size:	628 mm²

Theoretical Performance	
Pixel Rate:	189.8 GPixel/s
Texture Rate:	556.0 GTexel/s
FP16 (half) performance:	35.58 TFLOPS (1:1)
FP32 (float) performance:	35.58 TFLOPS
FP64 (double) performance:	556.0 GFLOPS (1:64)

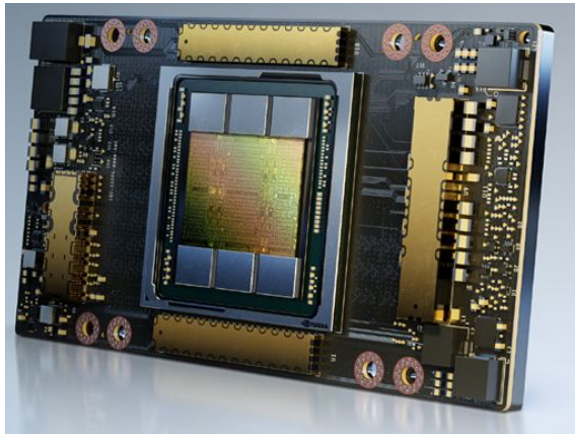
Graphics Card	
Release Date:	Sep 1st, 2020
Availability:	Sep 24th, 2020
Generation:	GeForce 30
Predecessor:	GeForce 20
Production:	Active
Launch Price:	1,499 USD
Current Price:	Amazon / Newegg
Bus Interface:	PCIe 4.0 x16
Reviews:	56 in our database

Memory	
Memory Size:	24 GB
Memory Type:	GDDR6X
Memory Bus:	384 bit
Bandwidth:	936.2 GB/s

Graphics Features	
DirectX:	12 Ultimate (12_2)
OpenGL:	4.6
OpenCL:	3.0
Vulkan:	1.2
CUDA:	8.6
Shader Model:	6.6

Render Config	
Shading Units:	10496
TMUs:	328
ROPs:	112
SM Count:	82
Tensor Cores:	328
RT Cores:	82
L1 Cache:	128 KB (per SM)
L2 Cache:	6 MB

The real power horses



OPTIMIZED SOFTWARE AND SERVICES FOR ENTERPRISE

NVIDIA CERTIFIED

EVERY DEEP LEARNING FRAMEWORK

mxnet PYTORCH Apache Spark TensorFlow

NVIDIA A100 TENSOR CORE GPU SPECIFICATIONS (SXM4 AND PCIE FORM FACTORS)

	A100 40GB PCIe	A100 80GB PCIe	A100 40GB SXM	A100 80GB SXM
FP64	9.7 TFLOPS			
FP64 Tensor Core	19.5 TFLOPS			
FP32	19.5 TFLOPS			
Tensor Float 32 (TF32)	156 TFLOPS | 312 TFLOPS*			
BFLOAT16 Tensor Core	312 TFLOPS | 624 TFLOPS*			
FP16 Tensor Core	312 TFLOPS | 624 TFLOPS*			
INT8 Tensor Core	624 TOPS | 1248 TOPS*			
GPU Memory	40GB HBM2	80GB HBM2e	40GB HBM2	80GB HBM2e
GPU Memory Bandwidth	1,555GB/s	1,935GB/s	1,555GB/s	2,039GB/s
Max Thermal Design Power (TDP)	250W	300W	400W	400W
Multi-Instance GPU	Up to 7 MIGs @ 5GB	Up to 7 MIGs @ 10GB	Up to 7 MIGs @ 5GB	Up to 7 MIGs @ 10GB
Form Factor	PCIe		SXM	

Alex cluster

FAU's **Alex cluster** (system integrator: Megware) is a high-performance compute resource with Nvidia GPGPU accelerators and partially high speed interconnect. It is intended for single and multi GPGPU workloads, e.g. from molecular dynamics, or machine learning. Alex serves for both, FAU's basic Tier3 resources as well as NHR's project resources.

- **2 front end nodes**, each with two AMD EPYC 7713 "Milan" processors (64 cores per chip) running at 2.0 GHz with 256 MB Shared L3 cache per chip, 512 GB of RAM, and 100 GbE connection to RRZE's network backbone but no GPGPUs.
- **20 GPGPU nodes**, each with two AMD EPYC 7713 "Milan" processors (64 cores per chip) running at 2.0 GHz with 256 MB Shared L3 cache per chip, 1,024 GB of DDR4-RAM, **eight Nvidia A100 (each 40 GB HBM2 @ 1,555 GB/s; HGX board with NVLink; 9.7 TFlop/s in FP64 or 19.5 TFlop/s in FP32)**, two HDR200 Infiniband HCAs, 25 GbE, and 14 TB on local NVMe SSDs.
- **15 GPGPU nodes**, each with two AMD EPYC 7713 "Milan" processors (64 cores per chip) running at 2.0 GHz with 256 MB Shared L3 cache per chip, 2,048 GB of DDR4-RAM, **eight Nvidia A100 (each 80 GB HBM2 @ 2,039 GB/s; HGX board with NVLink; 9.7 TFlop/s in FP64 or 19.5 TFlop/s in FP32)**, two HDR200 Infiniband HCAs, 25 GbE, and 14 TB on local NVMe SSDs.
- **38 GPGPU nodes**, each with two AMD EPYC 7713 "Milan" processors (64 cores per chip) running at 2.0 GHz with 256 MB Shared L3Cache per chip, 512 GB of DDR4-RAM, **eight Nvidia A40 (each with 48 GB DDR6 @ 696 GB/s; 37.42 TFlop/s in FP32)**, 25 GbE, and 7 TB on local NVMe SSDs.

So there is **a total of 304 Nvidia A40, 160 Nvidia A100/40GB, and 96 Nvidia A100/80GB GPGPUs**. The Nvidia A40 GPGPUs have a very high single precision floating point performance (even higher than an A100!) and are much less expensive than Nvidia A100 GPGPUs. All workloads which only require single precision floating point operations, like many molecular dynamics applications, thus, should target the Nvidia A40 GPGPUs.

https://hpc.fau.de/systems-services/documentation-instructions/clusters/alex-cluster/

Comparison A40 vs A100

	A40	A100 (SMX)
GPU architecture	Ampere; `SM_86`, `compute_86`	Ampere; `SM_80`, `compute_80`
GPU memory	48 GB GDDR6 with ECC (ECC disabled on Alex)	40GB HBM2 / 80 GB HBM2
Memory bandwidth	696 GB/s	1,555 GB/s / 2,039 GB/s
Interconnect interface	PCIe Gen4 31.5 GB/s (bidirectional)	NVLink: 600GB/s
CUDA Cores (Ampere generation)	10,752 (84 SMs)	6,912 (108 SMs)
RT Cores (2nd generation)	84	
Tensor Cores (3rd generation)	336	432
FP64 TFLOPS (non-Tensor)	0.5	9.7
FP64 Tensor TFLOPS		19.5
Peak FP32 TFLOPS (non-Tensor)	37.4	19.5
Peak TF32 Tensor TFLOPS	74.8	156
Peak FP16 Tensor TFLOPS with FP16 Accumulate	149.7	312
Peak BF16 Tensor TFLOPS with FP32 Accumulate	149.7	312
RT Core performance TFLOPS	73.1	?
Peak INT8 Tensor TOPS Peak INT 4 Tensor TOPS	299.3 598.7	624 1,248

A100

40GB PNY A100 PCIe, HBM2 (TCSA100M-PB)

Verfügbar
(Max. Bestellmenge: 2)

0 Bewertungen

nur € 10.420,36*

* inkl. 19% USt zzgl. Versandkosten / Lieferbeschränkungen

+ € 3,90* Geschenkverpackung

1 In den Warenkorb

Auf den Wunschzettel

- PaketShop-Lieferung möglich
- Versandkosten sparen
- Feedback zum Artikel
- Artikelinfos drucken

Preisent

Preisentwic
Keine Ange

elnummer 74392 **EAN** 3536403378035 **SKU** TCSA100M-PB

über **5** verkauft

Beobachter: **209**

Preisalarm setzen

3M 6M 1J Max

Floating point precision

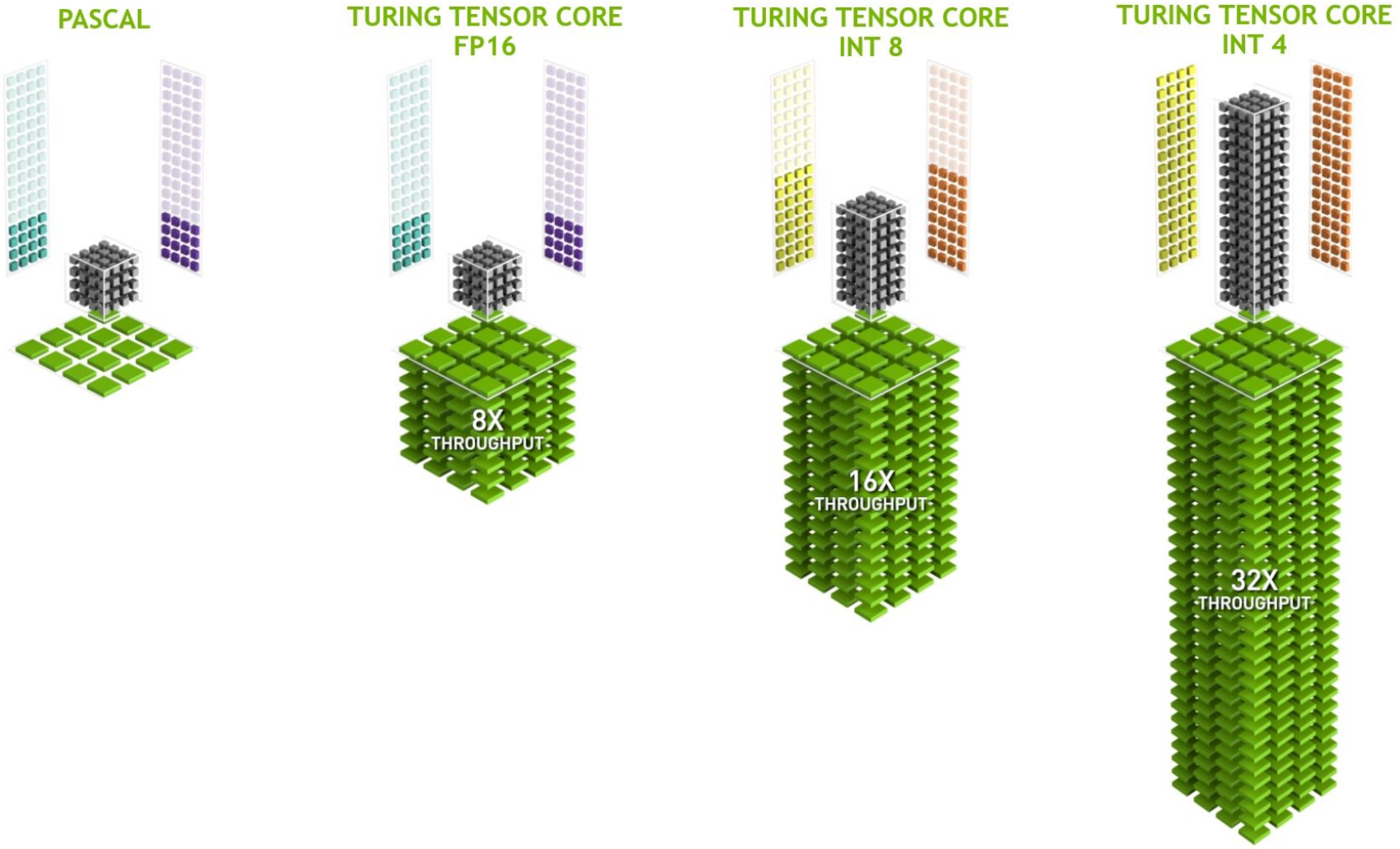
INT8

Table 2: ResNet-50 ImageNet validation set accuracy per math type

Math type	Multiply-add type	top-1 acc (%)	top-5 acc (%)
float32	FMA	76.130	92.862
(8, 1, 5, 5, 7) log	**ELMA**	**-0.90**	**-0.20**
(7, 1) posit	EMA	-4.63	-2.28
(8, 0) posit	EMA	-76.03	-92.36
(8, 1) posit	EMA	-0.87	-0.19
(8, 2) posit	EMA	-2.20	-0.85
(9, 1) posit	EMA	-0.30	-0.09
Jacob et al. [15]:			
float32	FMA	76.400	n/a
int8/32	MAC	-1.50	n/a
Migacz [23]:			
float32	FMA	73.230	91.180
int8/32	MAC	-0.20	-0.03

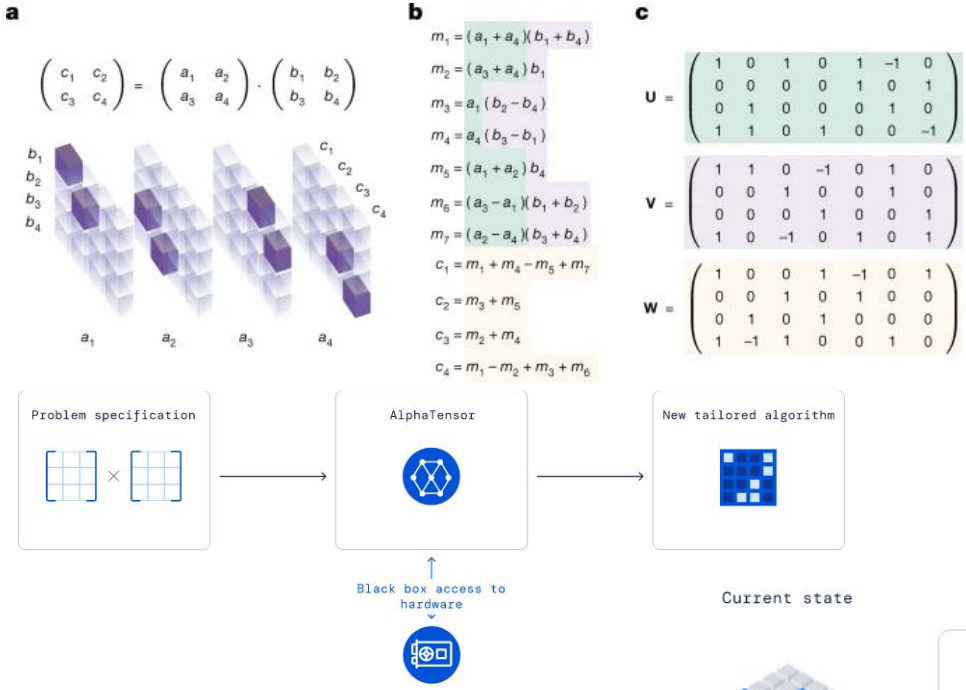
Rethinking floating point for deep learning

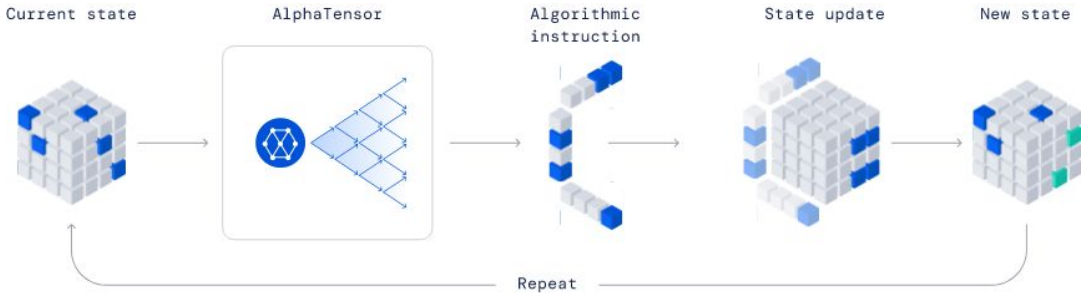
Jeff Johnson
Facebook AI Research
New York, NY
jhj@fb.com


PASCAL
TURING TENSOR CORE FP16
8X THROUGHPUT
TURING TENSOR CORE INT 8
16X THROUGHPUT
TURING TENSOR CORE INT 4
32X THROUGHPUT

AlphaTensor

Volker Strassen





Mixed precision

A: All models are suitable for AMP, although the speed-up may vary from model to model. The following table provides some examples of the speed-up for different models:

Table 3. Speed-ups FP32 and Mixed Precision models.

Model Script[1]	Framework	Data Set	FP32 Accuracy	Mixed Precision Accuracy	FP32 Throughput	Mixed Precision Throughput	Speed-up
BERT Q&A[2]	TensorFlow	SQuAD	90.83 Top 1%	90.99 Top 1%	66.65 sentences/sec	129.16 sentences/sec	1.94
SSD w/RN50[1]	TensorFlow	COCO 2017	0.268 mAP	0.269 mAP	569 images/sec	752 images/sec	1.32
GNMT[3]	PyTorch	WMT16 English to German	24.16 BLEU	24.22 BLEU	314,831 tokens/sec	738,521 tokens/sec	2.35
Neural Collaborative Filter[1]	PyTorch	MovieLens 20M	0.959 HR	0.960 HR	55,004,590 samples/sec	99,332,230 samples/sec	1.81
U-Net Industrial[1]	TensorFlow	DAGM 2007	0.965-0.988	0.960-0.988	445 images/sec	491 images/sec	1.10
ResNet-50 v1.5[1]	MXNet	ImageNet	76.67 Top 1%	76.49 Top 1%	2,957 images/sec	10,263 images/sec	3.47
Tacotron 2 / WaveGlow 1.0[1]	PyTorch	LJ Speech Dataset	0.3629/-6.1087	0.3645/-6.0258	10,843 tok/s 257,687 smp/s	12,742 tok/s 500,375 smp/s	1.18/1.94

Mixed precision computations

RESEARCH ARTICLE

Impact of Mixed Precision Techniques on Training and Inference Efficiency of Deep Neural Networks

MARION DÖRRICH, MINGCHENG FAN, AND ANDREAS M. KIST, (Member, IEEE)
Department of Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universiät Erlangen-Nürnberg, 91052 Erlangen, Germany
Corresponding author: Andreas M. Kist (andreas.kist@fau.de)

This work was supported by Kompetenznetzwerk für Technisch-Wissenschaftliches Hoch- und Höchstleistungsrechnen in Bayern (KONWIHR).

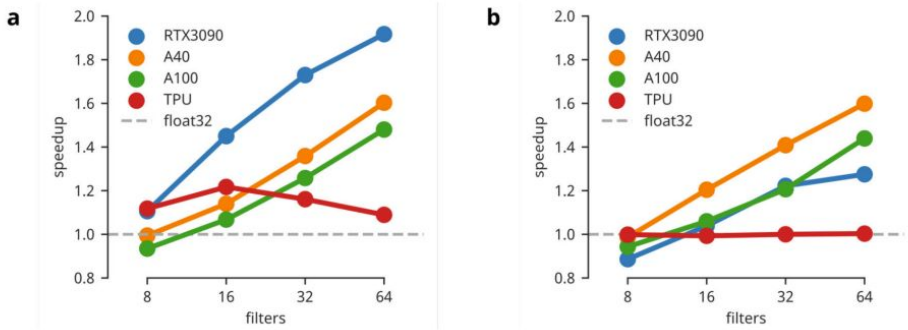
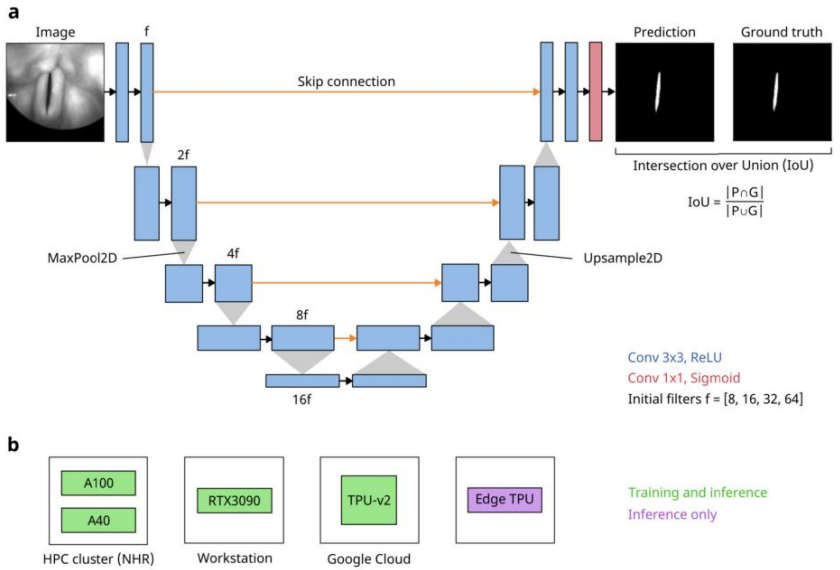
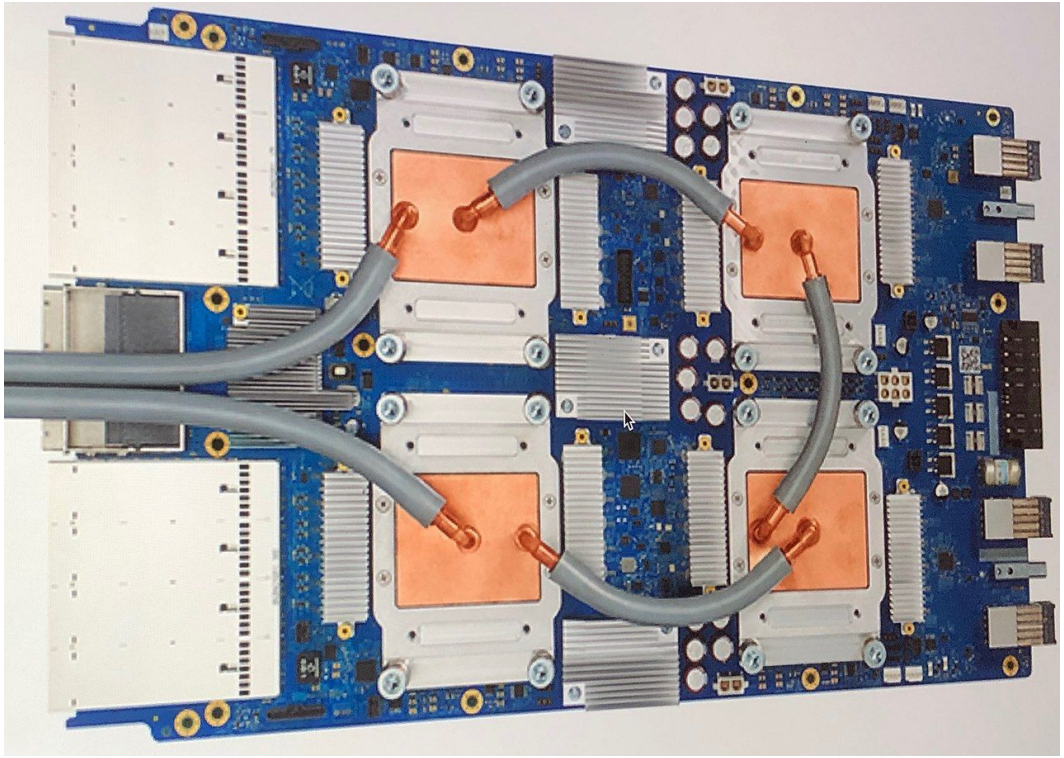


FIGURE 5. Training speedup (averaged over 4 folds) gained from mixed precision compared to the baseline, i.e. single precision training. (a) Using TensorFlow. (b) Using PyTorch.



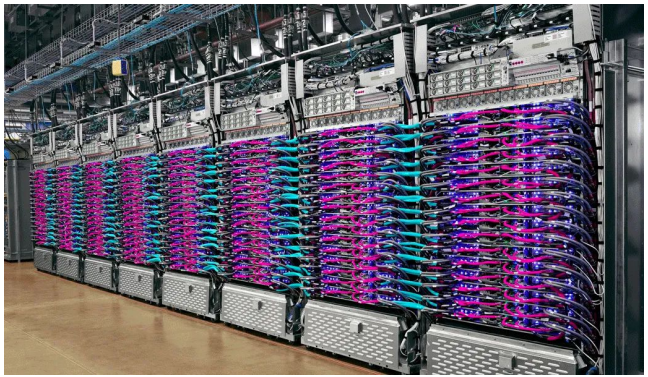
Application Specific ICs (integrated circuits)
ASIC

TPUs

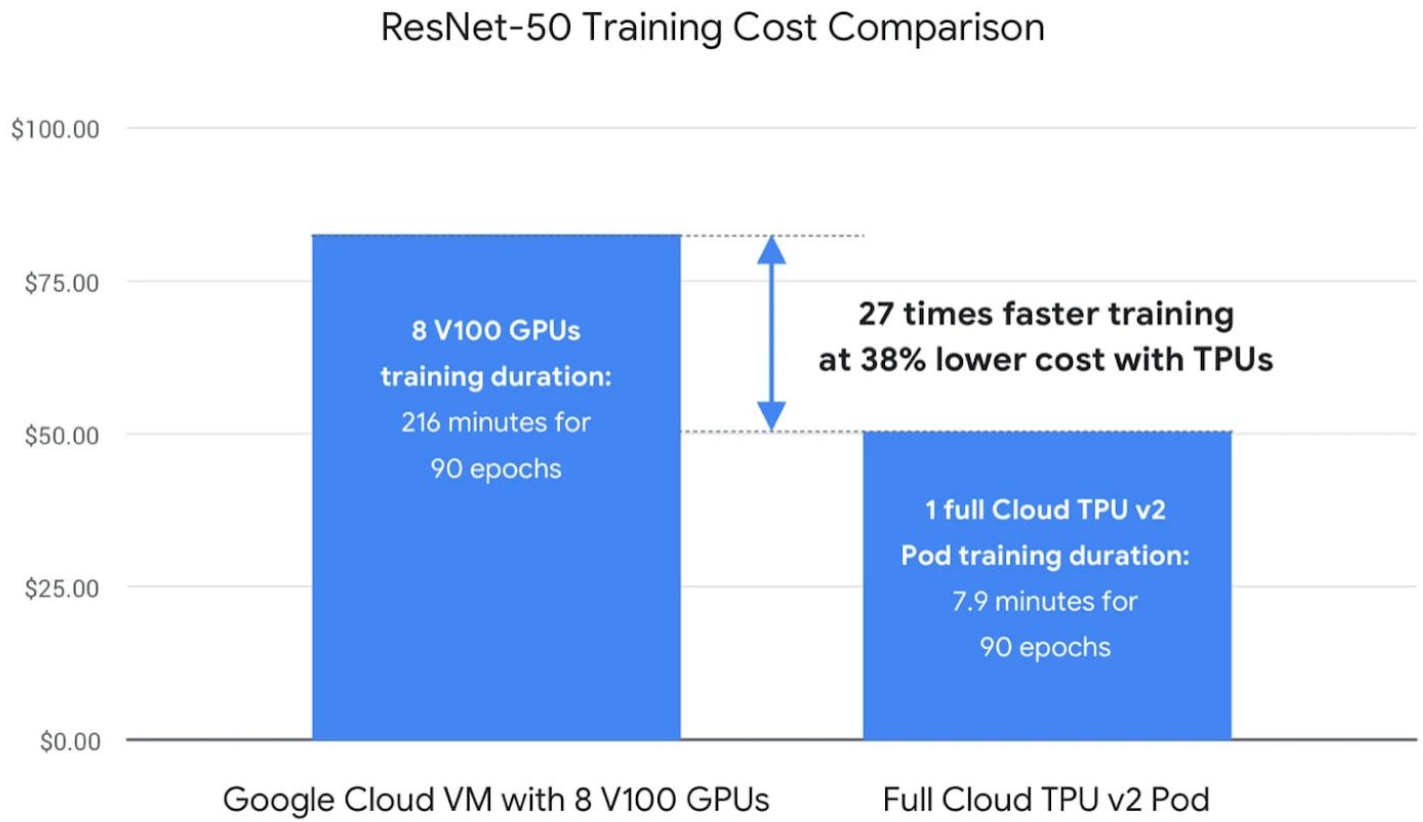


Products[13] [edit]

	TPUv1	TPUv2	TPUv3	TPUv4[14]	Edge v1
Date Introduced	2016	2017	2018	2021	2018
Process Node	28 nm	16 nm	16 nm	7 nm	
Die Size (mm2)	331	< 625	< 700	< 400	
On chip memory (MiB)	28	32	32	144	
Clock Speed (MHz)	700	700	940	1050	
Memory (GB)	8GB DDR3	16GB HBM	32GB HBM	8GB	
TDP(W)	75	280	450	175	2
TOPS	23	45	90	?	4



TPUs are faster (in some regard)



TPUs are

Cloud TPUs are available in the fol

US | Europe | Asia Pacific

TPU type (v2)	US	EUROP
v2-8		
v2-32		
v2-128		
v2-256		
v2-512		
TPU type (v3)		
v3-8		
v3-32		
v3-64		
v3-128		
v3-256		256
v3-512		512
v3-1024		1024
v3-2048		2048

Version

Cloud TI

Cloud TI

Cloud TPU v2 Pod	Evaluation Price / hr	1-yr Commitment Price (37% discount)
32-core Pod slice	$24 USD	$132,451 USD
128-core Pod slice	$96 USD	$529,805 USD
256-core Pod slice	$192 USD	$1,059,610 USD
512-core Pod slice	$384 USD	$2,119,219 USD
Cloud TPU v3 Pod	**Evaluation Price / hr**	**1-yr Commitment Price (37% discount)**
32-core Pod slice	$32 USD	$176,601 USD

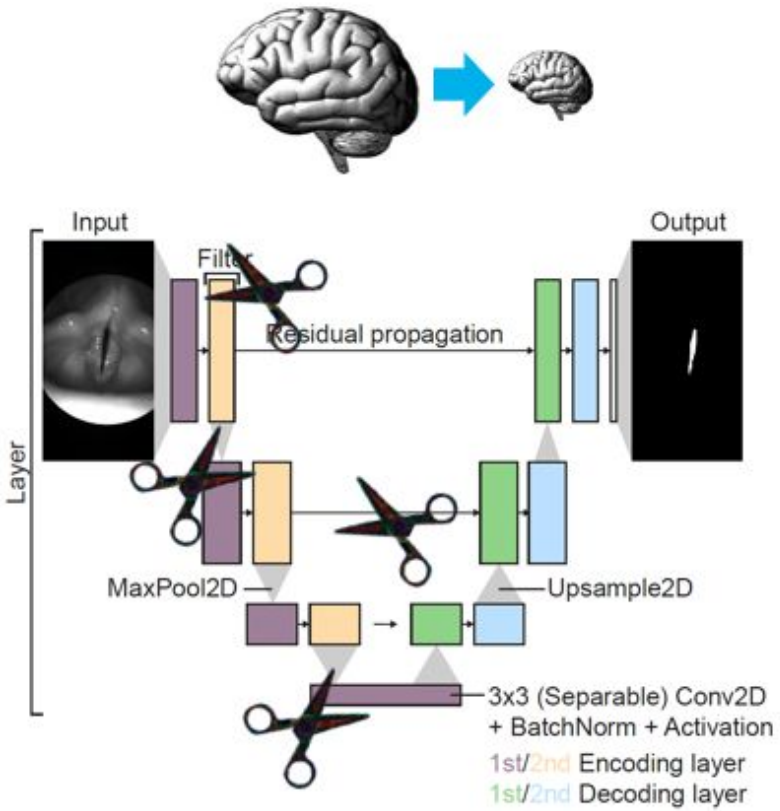
The Edge TPU

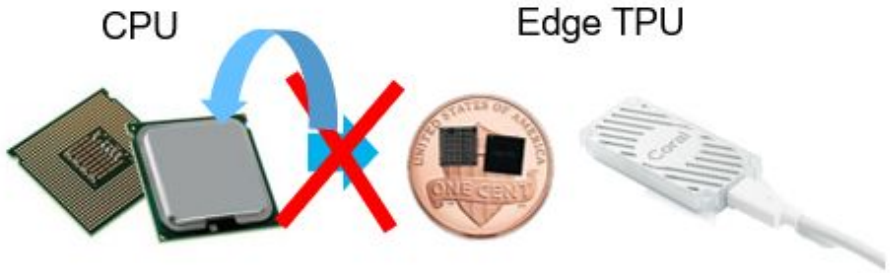
Inference boost

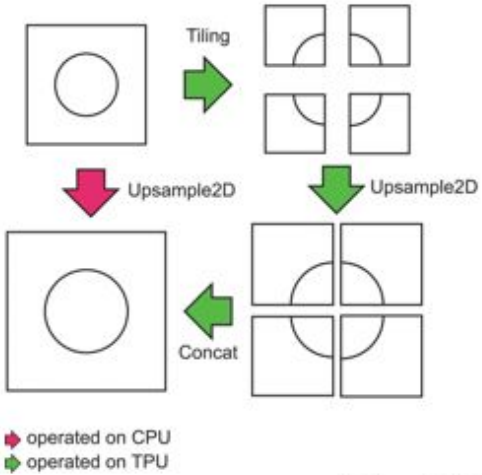
Table 1. Time per inference, in milliseconds (ms)

Model architecture	Desktop CPU [1]	Desktop CPU [1] + USB Accelerator (USB 3.0) *with Edge TPU*	Embedded CPU [2]	Dev Board [3] *with Edge TPU*
Unet Mv2 (128x128)	27.7	3.3	190.7	5.7
DeepLab V3 (513x513)	394	52	1139	241
DenseNet (224x224)	380	20	1032	25
Inception v1 (224x224)	90	3.4	392	4.1
Inception v4 (299x299)	700	85	3157	102
Inception-ResNet V2 (299x299)	753	57	2852	69
MobileNet v1 (224x224)	53	2.4	164	2.4
MobileNet v2 (224x224)	51	2.6	122	2.6
MobileNet v1 SSD (224x224)	109	6.5	353	11
MobileNet v2 SSD (224x224)	106	7.2	282	14
ResNet-50 V1 (299x299)	484	49	1763	56

Semantic segmentation on Edge TPUs







Kist and Döllinger, IEEE Access 2020

More AI hardware accelerators

Nvidia Jetson platform
(GPU based)
➜ Jetson SDK

VPU (Vision Processing unit) ➜ OpenVINO

FPGAs

Field-programmable gate arrays

Circuit can be programmed

Amazing for dedicated tasks

Deep learning:

Xilinx Alveo U50/U250...

Summary

- CPUs
  - Quick prototyping that requires maximum flexibility
  - Simple models that do not take long to train
  - Small models with small effective batch sizes
  - Models that are dominated by custom TensorFlow operations written in C++
  - Models that are limited by available I/O or the networking bandwidth of the host system
- GPUs
  - Models for which source does not exist or is too onerous to change
  - Models with a significant number of custom TensorFlow operations that must run at least partially on CPUs
  - Models with TensorFlow ops that are not available on Cloud TPU (see the list of available TensorFlow ops)
  - Medium-to-large models with larger effective batch sizes
- TPUs
  - Models dominated by matrix computations
  - Models with no custom TensorFlow operations inside the main training loop
  - Models that train for weeks or months
  - Larger and very large models with very large effective batch sizes

How do I use CPU/GPU/TPUs?!?!

Programming on CPUs and GPUs

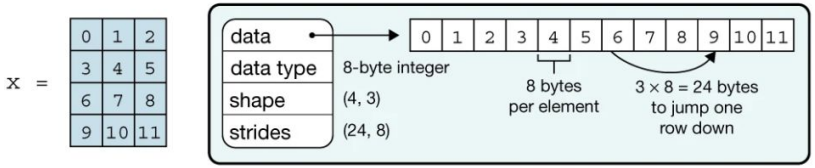
Review Article | Open Access | Published: 16 September 2020

Array programming with NumPy

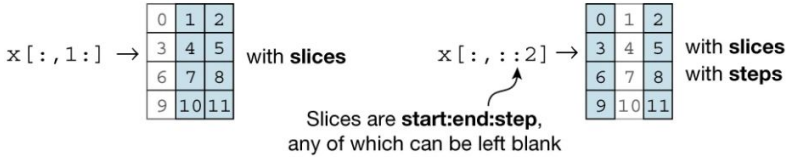
Charles R. Harris, K. Jarrod Millman, [...] Travis E. Oliphant

Nature **585**, 357–362 (2020) | Cite this article

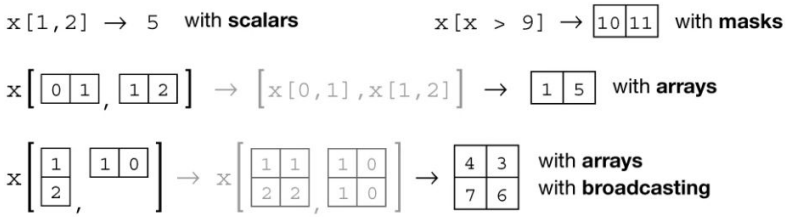
a Data structure



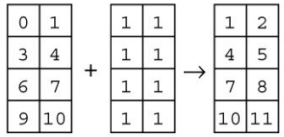
b Indexing (view)



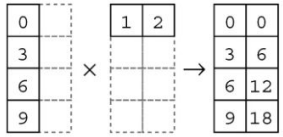
c Indexing (copy)



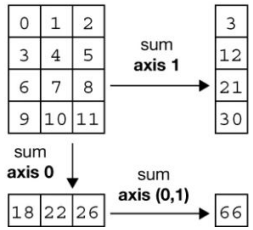
d Vectorization



e Broadcasting



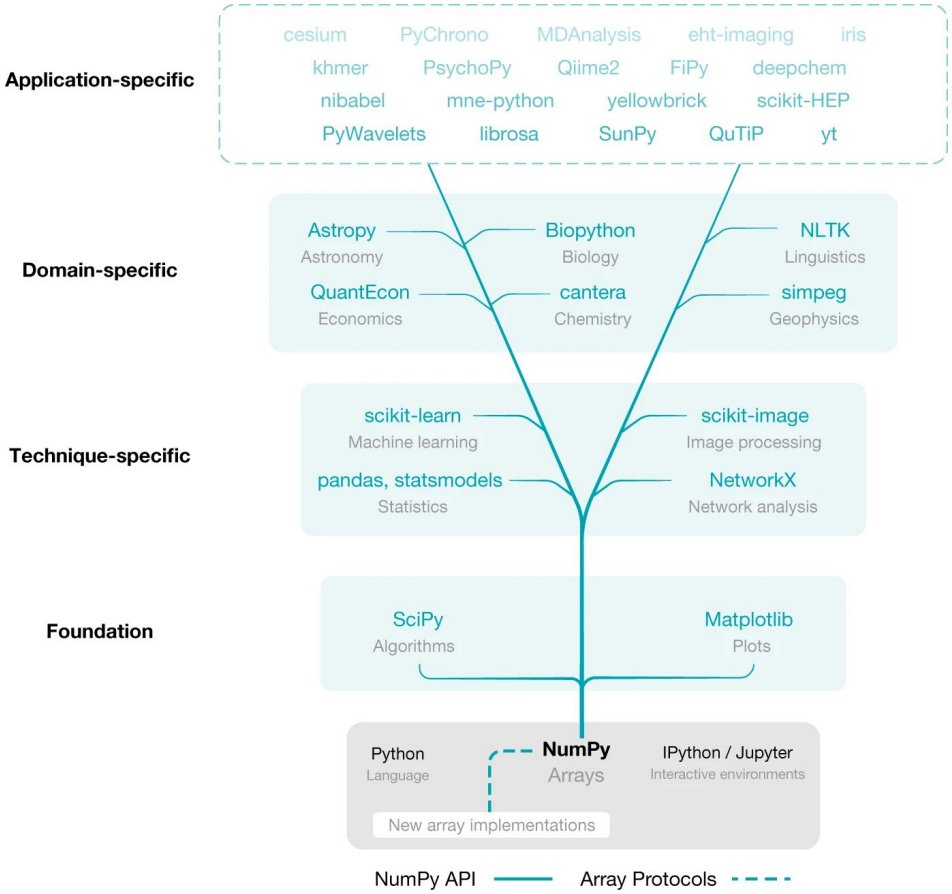
f Reduction



g Example

```
In [1]: import numpy as np

In [2]: x = np.arange(12)

In [3]: x = x.reshape(4, 3)

In [4]: x
Out[4]:
array([[ 0,  1,  2],
       [ 3,  4,  5],
       [ 6,  7,  8],
       [ 9, 10, 11]])

In [5]: np.mean(x, axis=0)
Out[5]: array([4.5, 5.5, 6.5])

In [6]: x = x - np.mean(x, axis=0)

In [7]: x
Out[7]:
array([[-4.5, -4.5, -4.5],
       [-1.5, -1.5, -1.5],
       [ 1.5,  1.5,  1.5],
       [ 4.5,  4.5,  4.5]])
```

What libraries are using numpy?



… and GPUs?

GPU-Accelerated Computing with Python

NVIDIA's CUDA Python provides a driver and runtime API for existing toolkits and libraries to simplify GPU-based accelerated processing. Python is one of the most popular programming languages for science, engineering, data analytics, and deep learning applications. However, as an interpreted language, it's been considered too slow for high-performance computing.

Numba—a Python compiler from Anaconda that can compile Python code for execution on CUDA®-capable GPUs—provides Python developers with an easy entry into GPU-accelerated computing and for using increasingly sophisticated CUDA code with a minimum of new syntax and jargon. With CUDA Python and Numba, you get the best of both worlds: rapid iterative development with Python combined with the speed of a compiled language targeting both CPUs and NVIDIA GPUs.

Welcome to cuML's documentation!

cuML is a suite of fast, GPU-accelerated machine learning algorithms designed for data science and analytical tasks. Our API mirrors Sklearn's, and we provide practitioners with the easy fit-predict-transform paradigm without ever having to program on a GPU.

As data gets larger, algorithms running on a CPU becomes slow and cumbersome. RAPIDS provides users a streamlined approach where data is intially loaded in the GPU, and compute tasks can be performed on it directly.

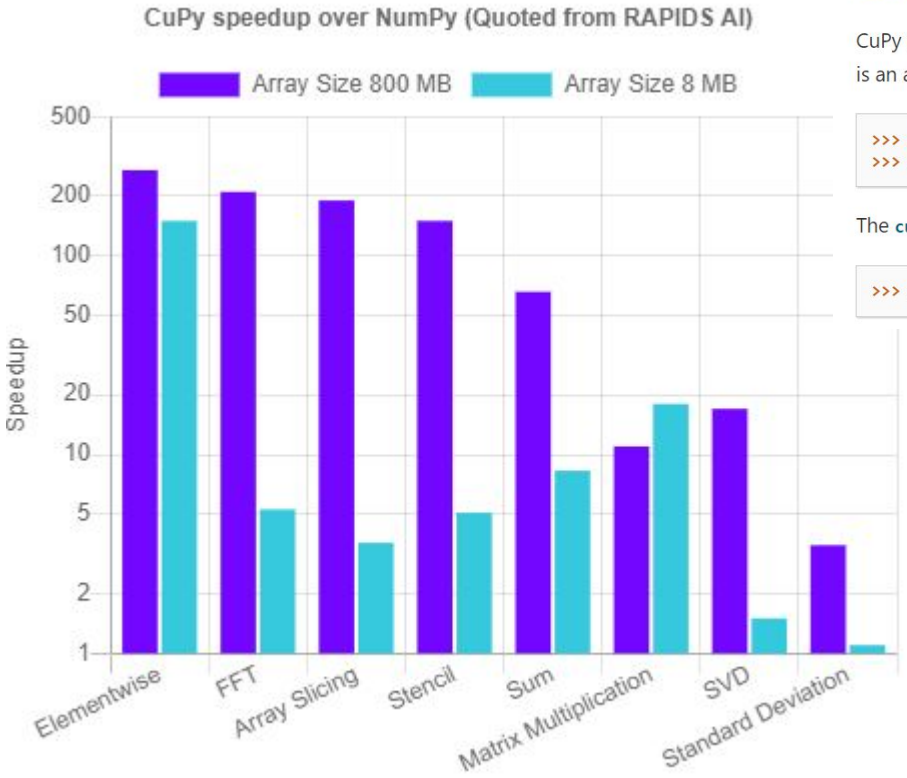
cuML is fully open source, and the RAPIDS team welcomes new and seasoned contributors, users and hobbyists! Thank you for your wonderful support!

An installation requirement for cuML is that your system must be Linux-like. Support for Windows is possible in the near future.

Welcome to cuDF's documentation!

cuDF is a Python GPU DataFrame library (built on the Apache Arrow columnar memory format) for loading, joining, aggregating, filtering, and otherwise manipulating data. cuDF also provides a pandas-like API that will be familiar to data engineers & data scientists, so they can use it to easily accelerate their workflows without going into the details of CUDA programming.

CuPY



Basics of cupy.ndarray

CuPy is a GPU array backend that implements a subset of NumPy interface. In the following code, `cp` is an abbreviation of `cupy`, following the standard convention of abbreviating `numpy` as `np`:

```
>>> import numpy as np
>>> import cupy as cp
```

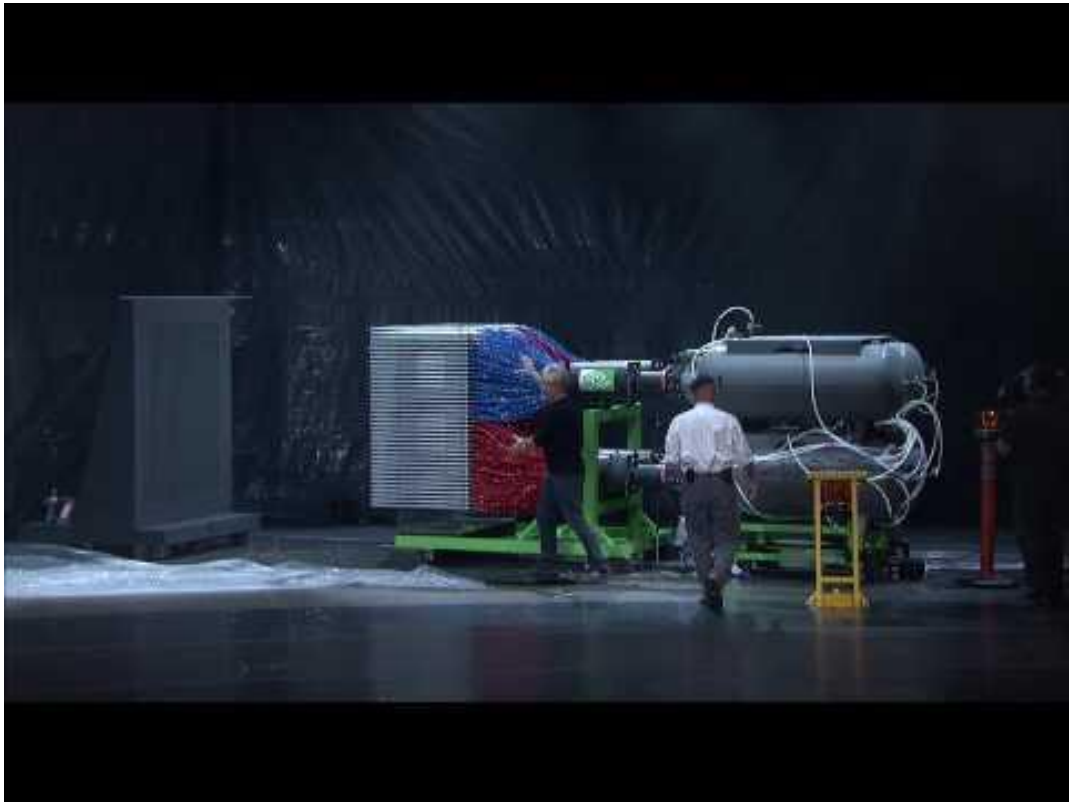
The **`cupy.ndarray`** class is at the core of `CuPy` and is a replacement class for `NumPy`'s **`numpy.ndarray`**.

```
>>> x_gpu = cp.array([1, 2, 3])
```

- Routines (SciPy)
  - Discrete Fourier transforms (**`cupyx.scipy.fft`**)
  - Legacy discrete fourier transforms (**`cupyx.scipy.fftpack`**)
  - Linear algebra (**`cupyx.scipy.linalg`**)
  - Multidimensional image processing (**`cupyx.scipy.ndimage`**)
  - Sparse matrices (**`cupyx.scipy.sparse`**)
  - Special functions (**`cupyx.scipy.special`**)
  - Signal processing (**`cupyx.scipy.signal`**)
  - Statistical functions (**`cupyx.scipy.stats`**)

The last slide

- Comparison CPU and GPU in 1:30



Exercise

Building a PC

Please watch our OER “how to buy and build your own PC” to be uploaded to StudOn/FAU.TV

Homework

Homework → Deadline: 03.11.2024 23:59

1) Create a slide that shows and explains the components of a PC.

2) Select a CUDA-based research paper, read it and explain why it is of interest to you.