Data Science
Survival Skills

Hardware relevance in data science

Agenda

- Why you should even bother
- CPUs
- GPUs
- TPUs
- Hardware accelerators

A computer from the inside

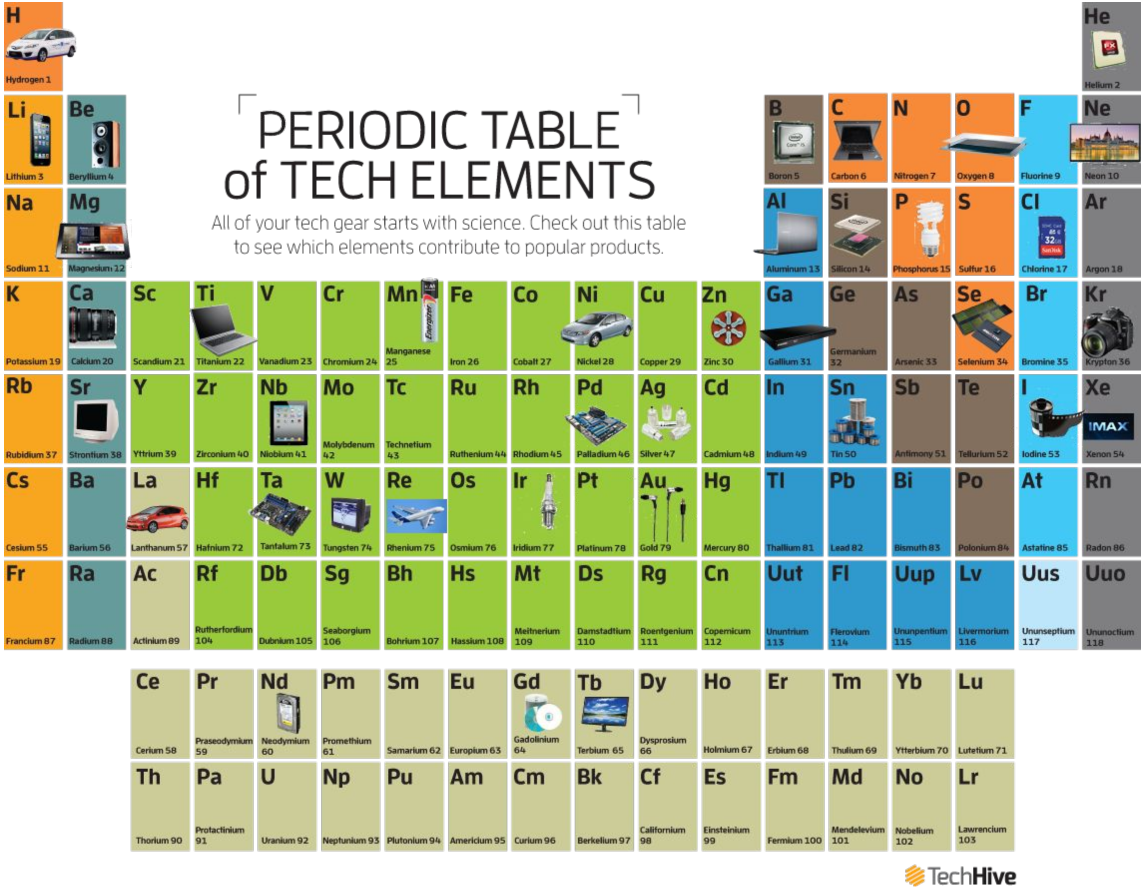
BACK

FRONT

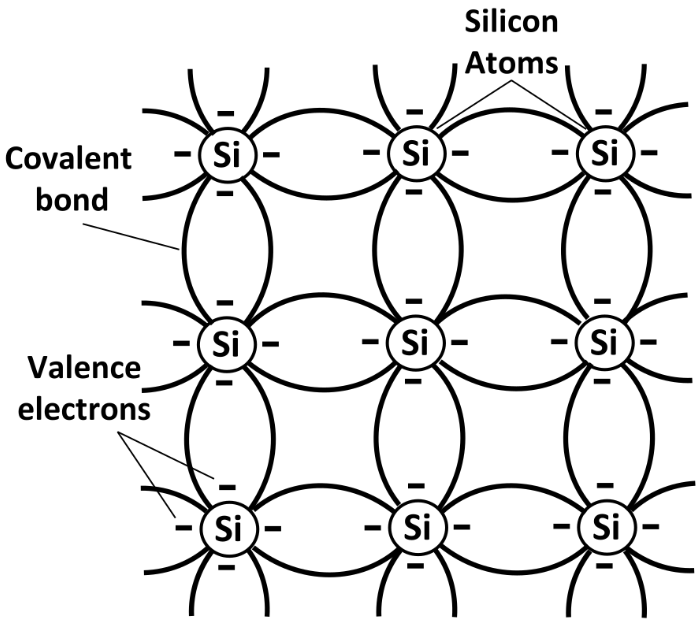
Periodic table

Group → Period ↓	1	2		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1 H																		2 He
2	3 Li	4 Be												5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg												13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca		21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr		39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	*	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	**	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Nh	114 Fl	115 Mc	116 Lv	117 Ts	118 Og
			*	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb		
			**	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No		

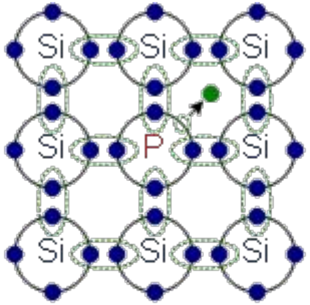
Periodic table of technology



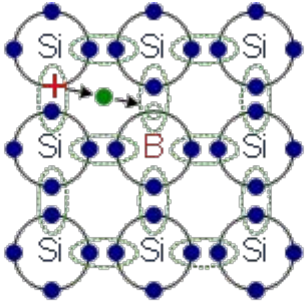
A computer is based on chemistry and physics!



Doping

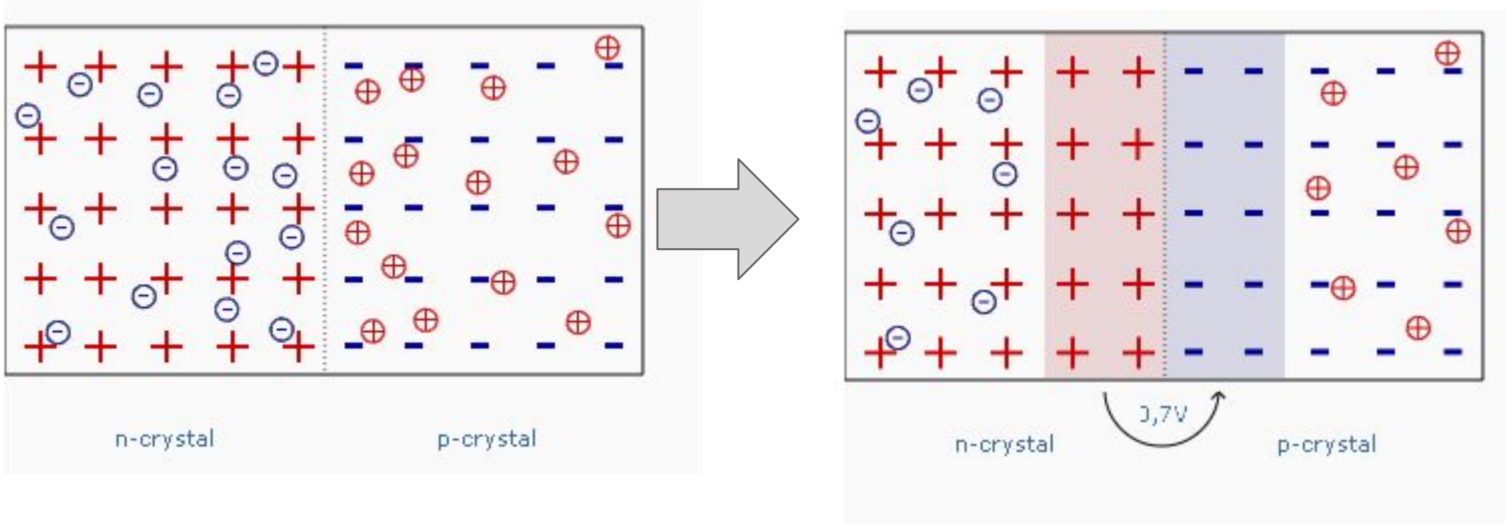


The phosphorus atom donates ist fifth valence electron. It acts as a free charge carrier.

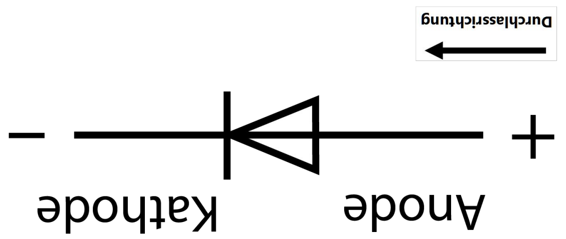


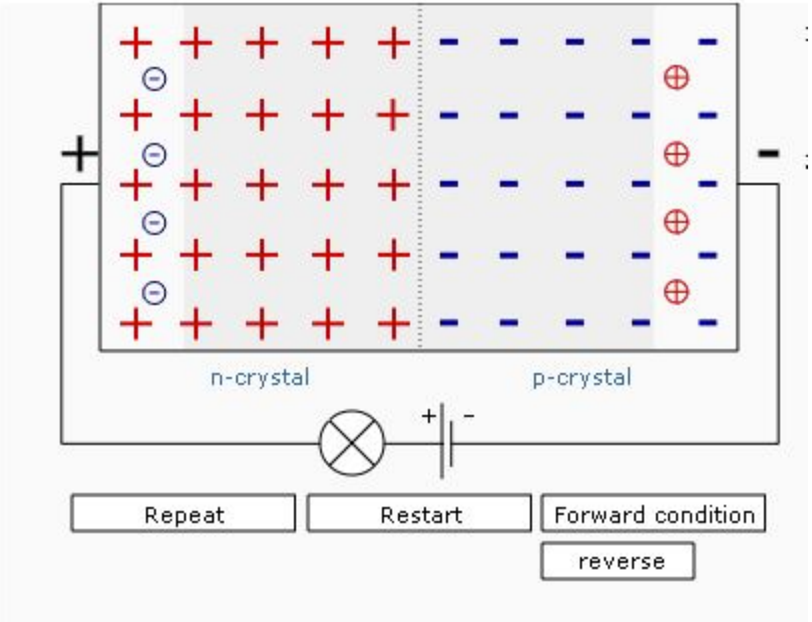
The free place on the boron atom is filled with an electron. Therefore a new hole („defect electron") is generated. This holes move in the opposite direction to the electrons

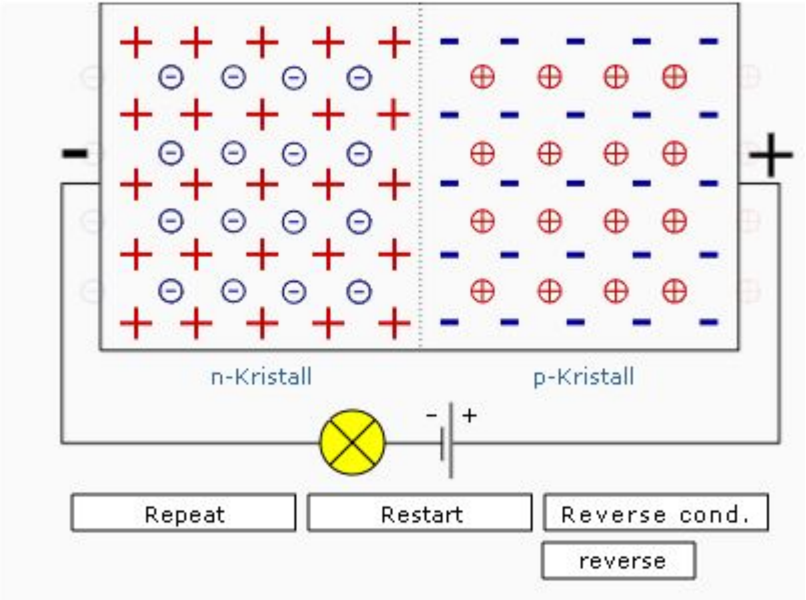
A diode



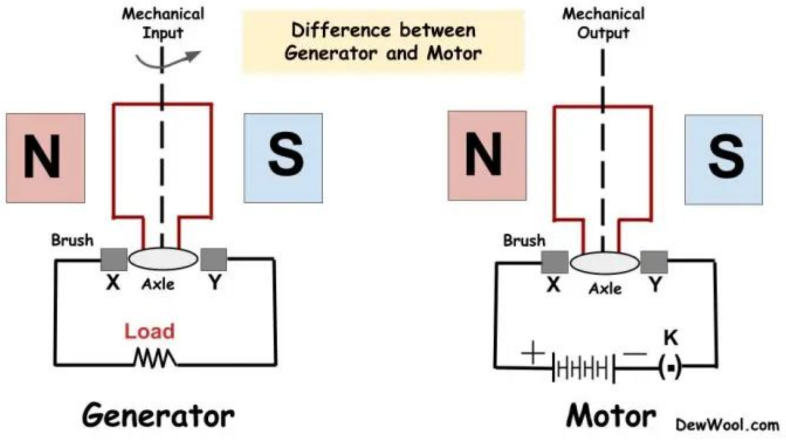
A diode


Durchlassrichtung
Kathode
Anode

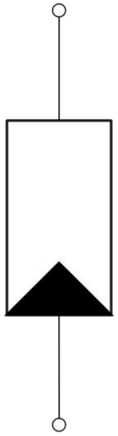

n-crystal
p-crystal
Repeat
Restart
Forward condition
reverse

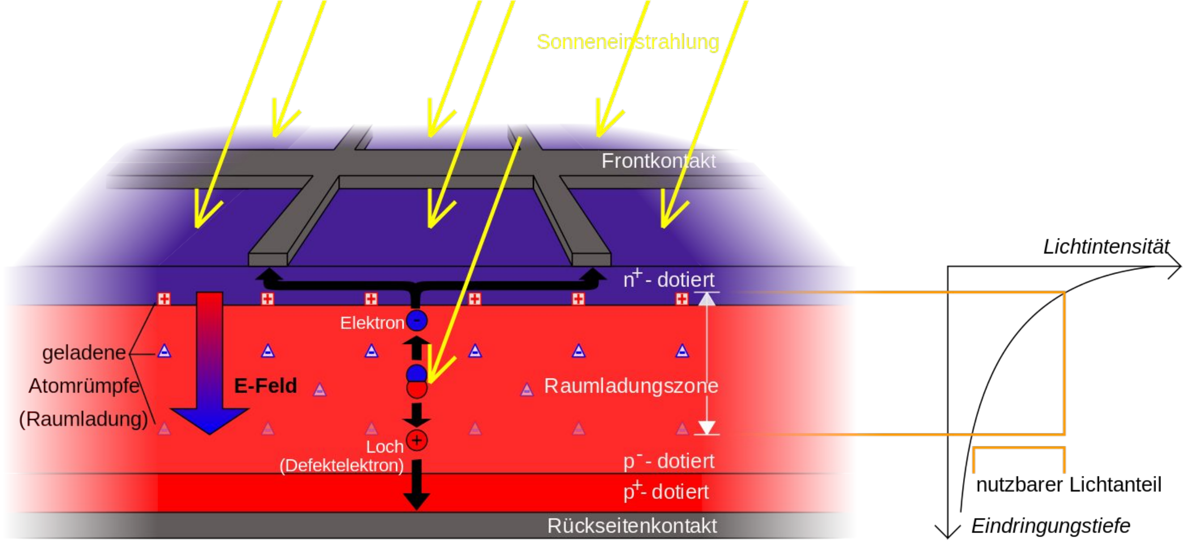

n-Kristall
p-Kristall
Repeat
Restart
Reverse cond.
reverse

A solar cell



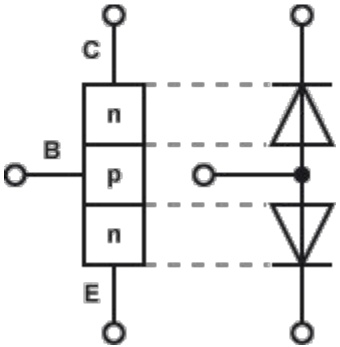
Mechanical Input
Difference between Generator and Motor
Mechanical Output
N
S
Brush
X
Axle
Y
Load
Generator
N
S
Brush
X
Axle
Y
K
Motor
DewWool.com

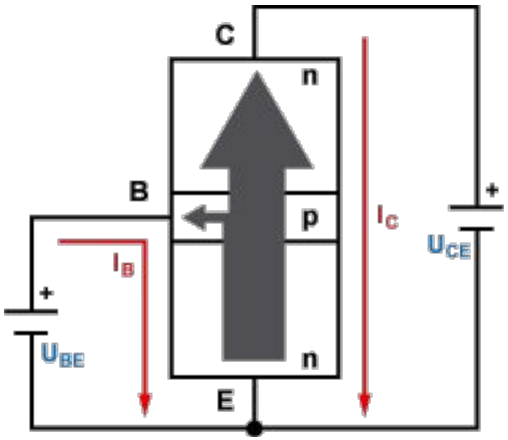


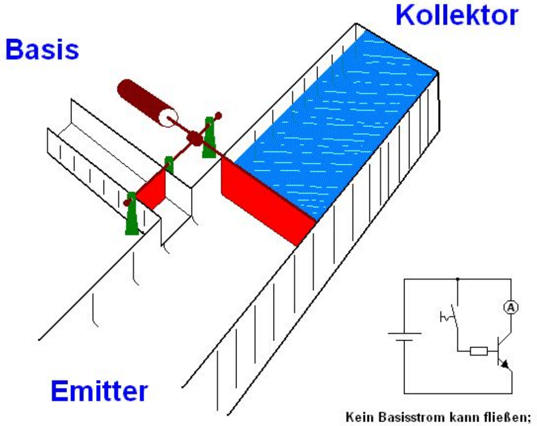


Sonneneinstrahlung
Frontkontakt
n^+- dotiert
Elektron
geladene
Atomrümpfe
(Raumladung)
E-Feld
Raumladungszone
Loch
(Defektelektron)
p^-- dotiert
p^+- dotiert
Rückseitenkontakt
Lichtintensität
nutzbarer Lichtanteil
Eindringungstiefe

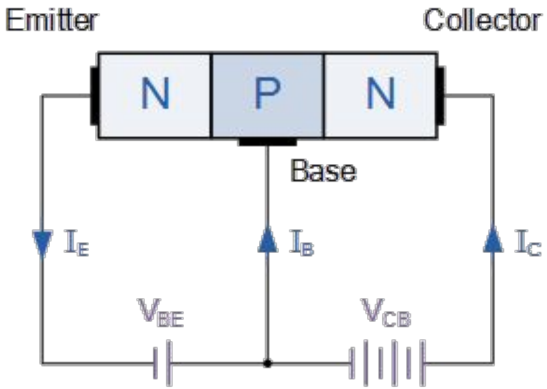
A transistor

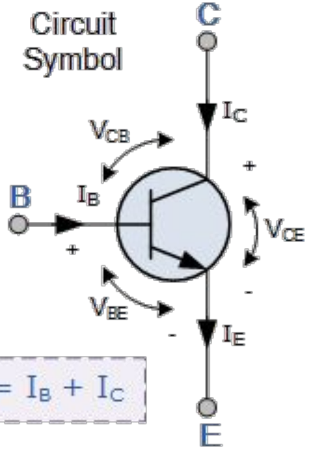






Kein Basisstrom kann fließen;
Der Transistor ist gesperrt.



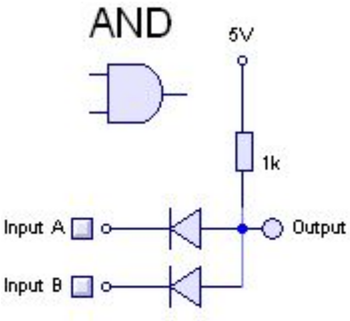


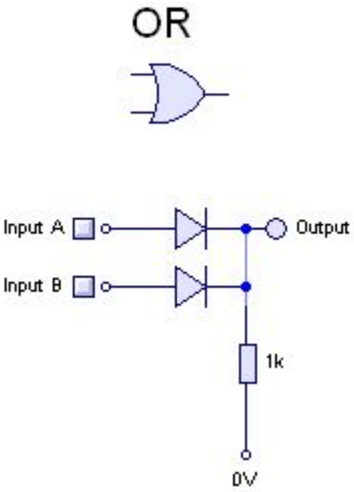
$$I_E = I_B + I_C$$

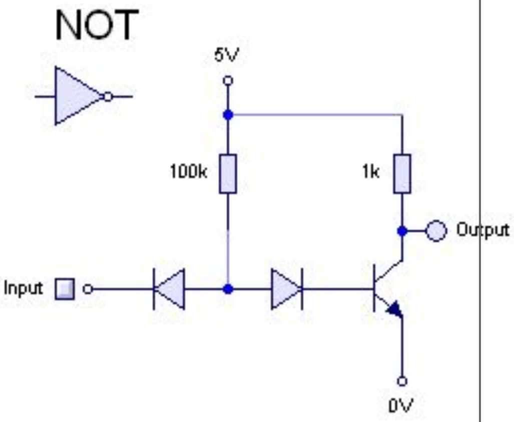
Logic gates

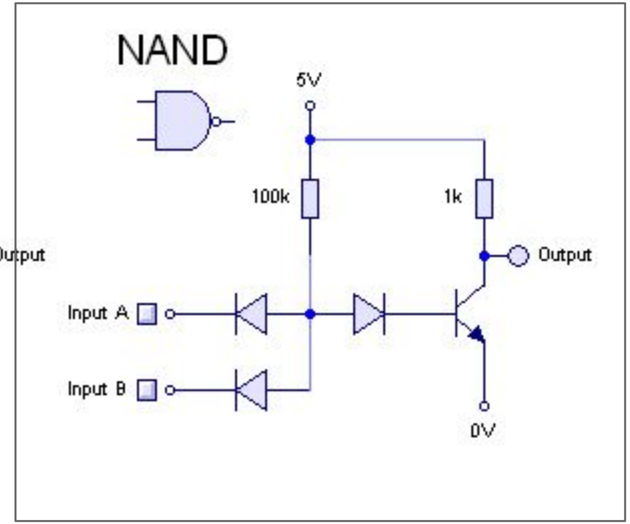
Drawn on black board

Transistor can perform computations



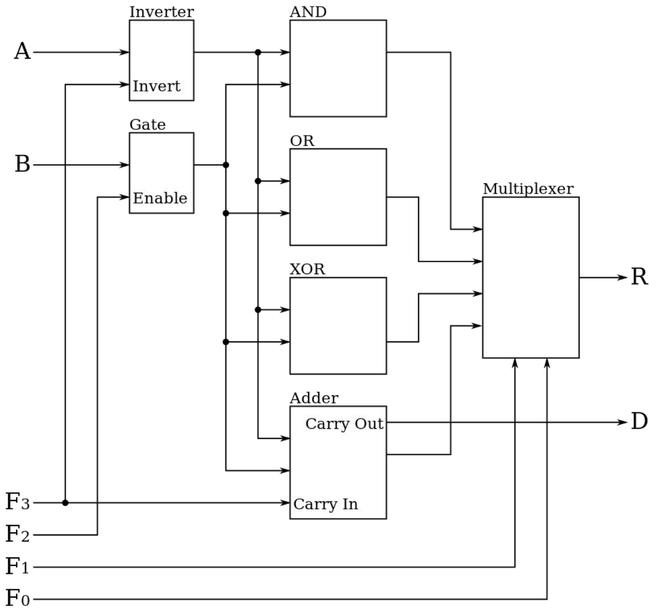


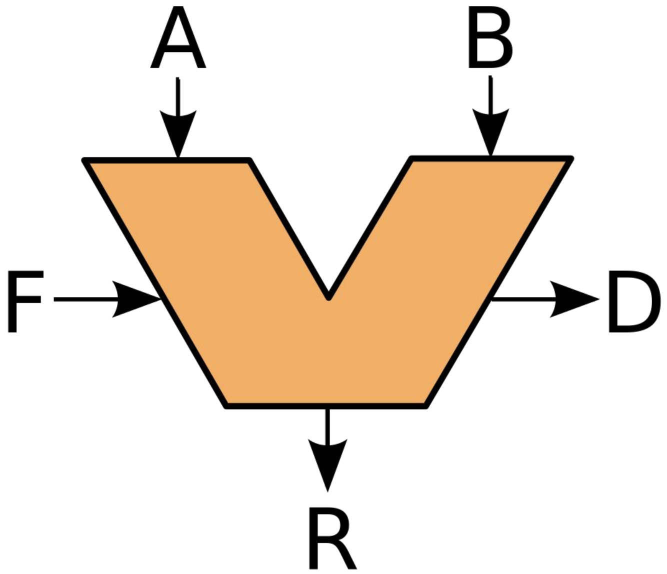




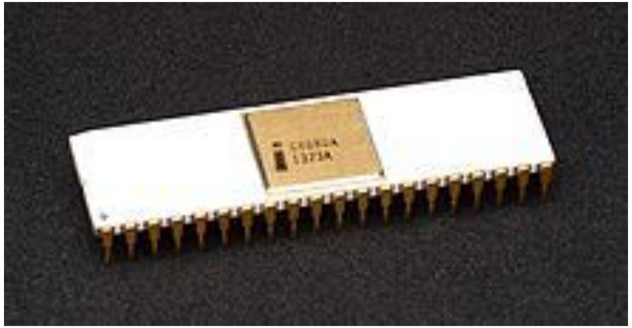
NAND is logic complete, i.e. can be used to gain every other function!

Arithmetic Logic Unit (ALU)





A modern CPU (central processing unit)

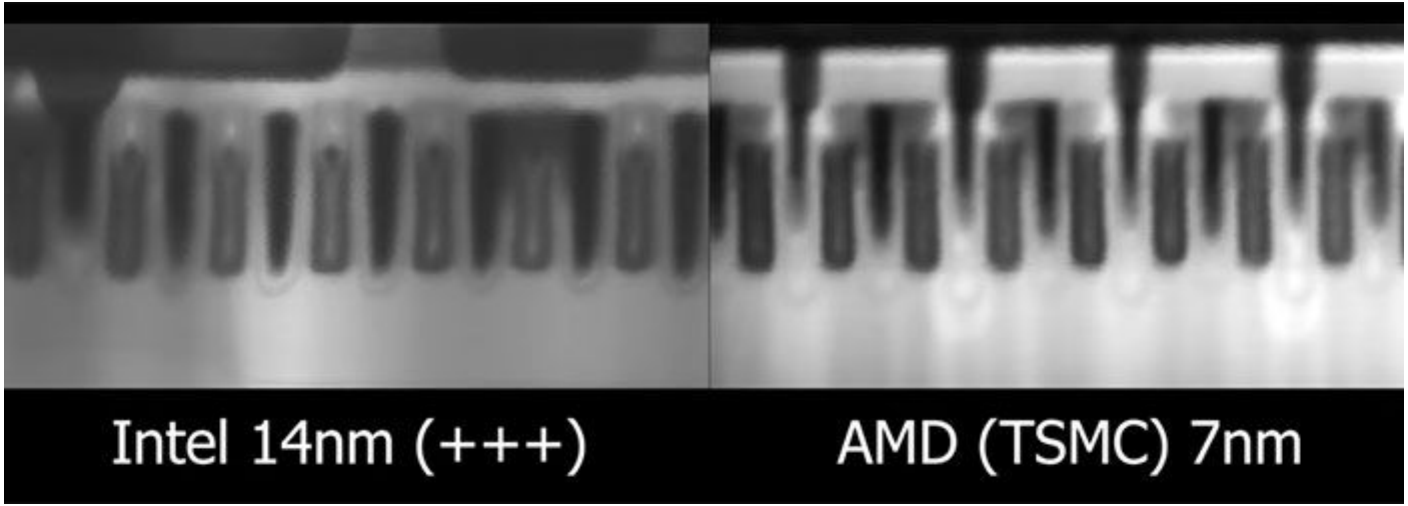


Intel's 8 bit processor (8080, 1974),
6000 Transistors, 6 um per transistor

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

Computations using transistors



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

Nowadays CPU

Desktop Ryzen 9 [Bearbeiten | Quelltext bearbeiten]

Modellnummer	Ordering Part Number: Tray	Ordering Part Number: PIB	Mikro-architektur	Fab	Sockel	Kerne/ Threads	Cache L1	Cache L2	Cache L3	Prozessortakt / Maximaltakt [anm 2]	Speicher-Controller [anm 1]	TDP	Einführungsdatum
Ryzen 9 3900	100-000000070	–	Zen 2	TSMC 7FF	AM4	12 / 24	768 KB	6 MB	64 MB	3,1 / 4,3 GHz	Dual-Channel DDR4-3200	65 W	24. Sep. 2019
Ryzen 9 Pro 3900	100-000000072	–											30. Sep. 2019
Ryzen 9 3900X	100-000000023	100-100000023BOX								3,8 / 4,6 GHz		105 W	7. Juli 201
Ryzen 9 3900XT	100-100000277	100-100000277WOF								3,8 / 4,7 GHz			7. Juli 202
Ryzen 9 3950X	100-000000051	100-100000051WOF				16 / 32	1 MB	8 MB		3,5 / 4,7 GHz			25. Nov. 2019
Ryzen 9 5900	100-000000062	-	Zen 3		AM4	12 / 24	768 KB	6 MB	64 MB	3,0 / 4,7 GHz	Dual-Channel DDR4-3200	65W	13. April 2021
Ryzen 9 5900X	100-000000061	100-100000061WOF								3,7 / 4,8 GHz		105 W	5. Nov. 20
								IB		3,4 / 4,9 GHz			

AMD

Lagernd

CUSTOMERS' CHOICE

AMD Ryzen 7 5800X 8x 3.80GHz So.AM4 WOF

€ 359,-*

inkl. 19% USt + Versandkosten

★ Mindstar Highlight ★

Lagernd

Artnr: 74531

AMD Ryzen 9 5950X 16x 3.40GHz So.AM4 WOF

€ 739,-*

inkl. 19% USt + Versandkosten

über 7.080 verkauft

Lagernd

CUSTOMERS' CHOICE

AMD Ryzen 7 3800X 8x 3.90GHz So.AM4 BOX

€ 289,-*

inkl. 19% USt + Versandkosten

über 23.490 verkauft

Lagernd

Artnr: 75251

Intel Core i9 11900K 8x 3.50GHz So.1200 WOF

€ 519,-*

inkl. 19% USt + Versandkosten

über 570 verkauft

Lagernd

Artnr: 74143

Intel Core i7 10700K 8x 3.80GHz So.1200 WOF

€ 322,-*

inkl. 19% USt + Versandkosten

über 7.640 verkauft

Lag

Artnr:

Intel Core i5 1

So.12

€ 158,97*

inkl. 19% USt + Versandkosten

über 3.000 verkauft

€ 427,76*

inkl. 19% USt + Versandkosten

über 2.900 verkauft

€ 248,97*

inkl. 19% USt + Versandkosten

über 1.200 verkauft

Die Modelle der Rocket Lake-Generation (komplett)

Modell	Release	Cores (Threads)	PCIe Lanes	Takt: Base	Takt: Turbo	TDP	L3-Cache	Bemerkungen
Core™ i9-11900K	Mrz. 21	8 (16)	20	3,5 GHz	5,2 GHz	125 W	16 MB	
Core™ i7-11700K		8 (16)		3,6 GHz	5,0 GHz		16 MB	

CPU

Ryzen Threadripper 3960X	–	100000010WOF				24 / 48	1,5 MB	12 MB	128 MB	3,8 / 4,5 GHz			25. Nov. 2019
Ryzen Threadripper 3970X	–	100000011WOF			sTRX4	32 / 64	2 MB	16 MB		3,7 / 4,5 GHz	Quad-Channel DDR4-3200		
Ryzen Threadripper 3990X		100000163WOF				64 / 128	4 MB	32 MB	256 MB	2,9 / 4,3 GHz			7. Feb. 2020
Ryzen Threadripper Pro 3945WX	100-000000168	- (OEM only)	Zen 2	TSMC 7FF		12 / 24	768 KB	6 MB	64 MB	4,0 / 4,3 GHz		280 W	
Ryzen Threadripper Pro 3955WX	100-000000167	- (OEM only)				16 / 32	1 MB	8 MB	64 MB	3,9 / 4,3 GHz			
Ryzen Threadripper Pro 3975WX	100-000000086	- (OEM only)			WRX80	32 / 64	2 MB	16 MB	128 MB	3,5 / 4,2 GHz	Octo-Channel DDR4-3200		14. Juli 2020
Ryzen Threadripper Pro 3995WX	100-000000087	- (OEM only)				64 / 128	4 MB	32 MB	256 MB	2,7 / 4,2 GHz			

AMD Ryzen Threadripper PRO 3995WX, 64C/128T, 2.70-4.20GHz, boxed ohne Kühler (100-100000087WOF)

jetzt bewerten!

Info beim Hersteller

Sockel	sWRX8 (LGA)
Codename	Castle Peak
Grafik	nein
TDP	280W
Kerne	64
Threads	128
Basistakt	2.70GHz
Turbotakt	4.20GHz
SMT	ja
Speichercontroller	Octa Channel PC4-25600L (DDR4-3200)
ECC-Unterstützung	ja
Freier Multiplikator	nein

Alle Produkteigenschaften anzeigen

Aktueller Preisbereich

€ 5079,00 bis **€ 7392,29**

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?!

- Mobile vs. desktop
- Frequency (clock rate, how fast it runs)
- Thermal power design (TPD) in Watts -> how much power your PC consumes/heat it produces
- L2/L3 cache (fast cache for operations, up to several MBs)
- On board GPU (only for checking emails...)
- General purpose!

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 (simulteanous 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 nodes** 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.

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

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

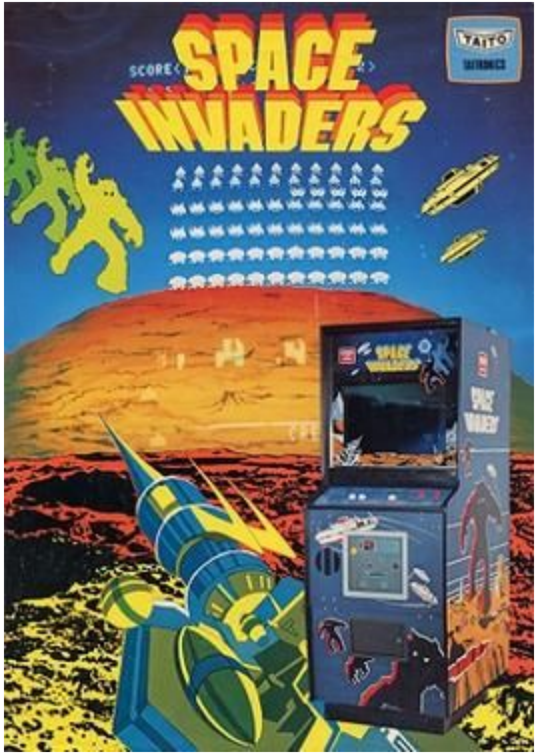
€ 4.896,74*

inkl. 19% USt + Versandkosten

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

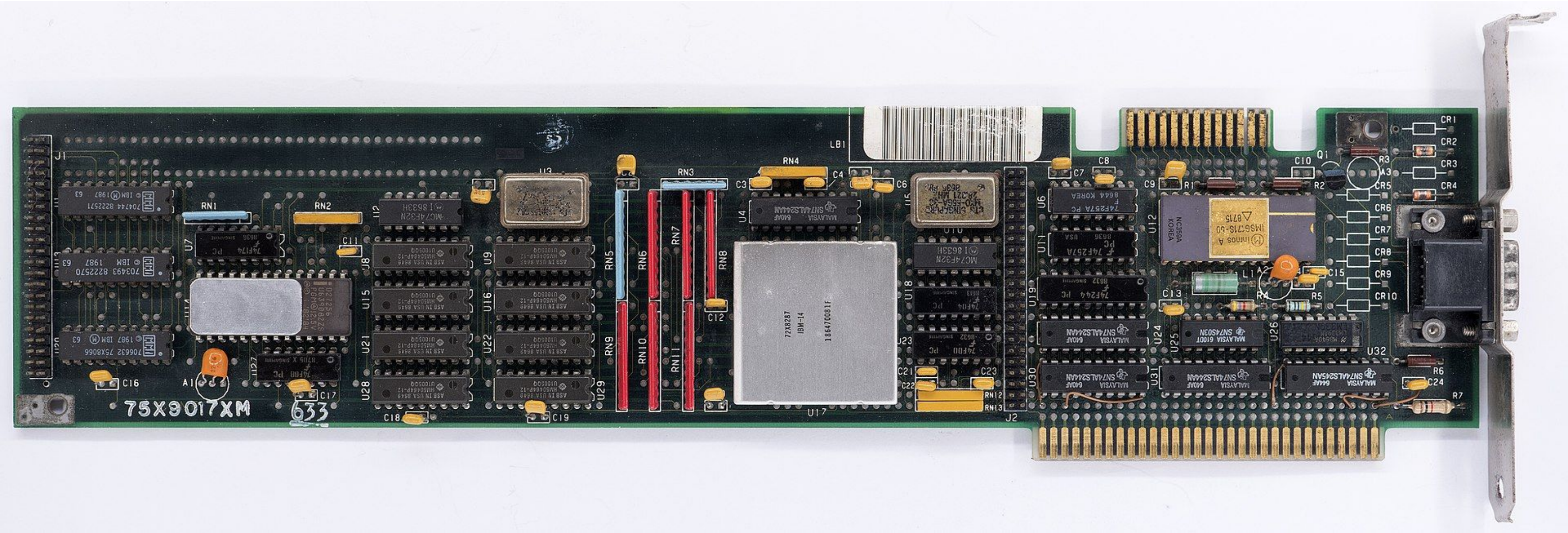
The GPU

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

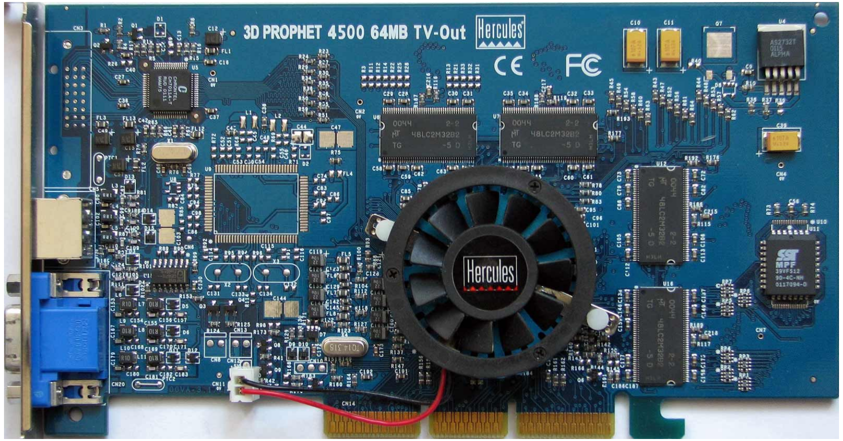


IBM's VGA (Video Graphics Array)

Introduced 1987 ➔ VGA resolution is 640x480



Where I spent my pocket money...



Hercules Kyro II

NVIDIA GeForce4 Ti 4200-8X

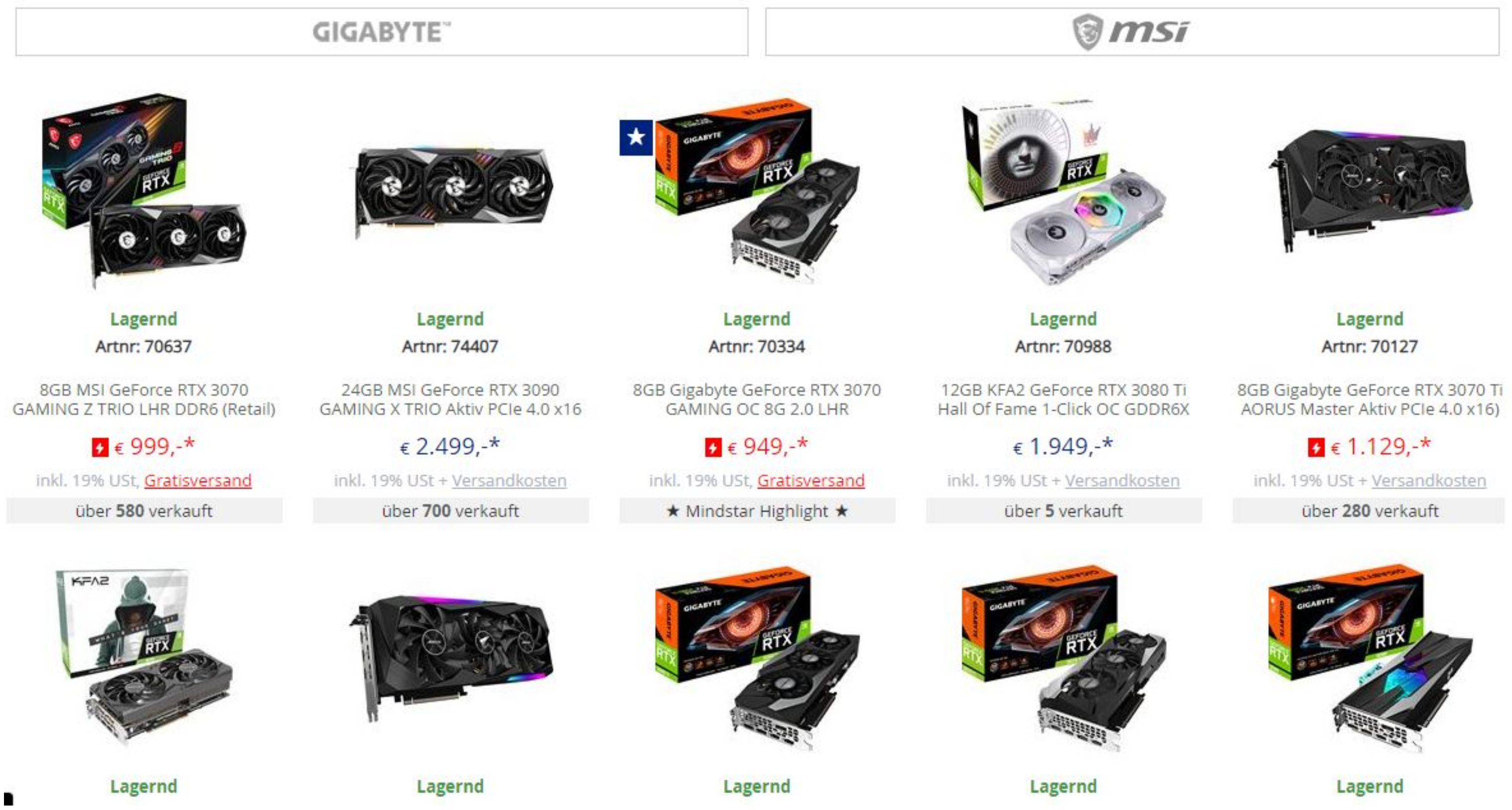
NV28	4	2	8	4	64 MB	SDR	128 bit
GRAPHICS PROCESSOR	PIXEL SHADERS	VERTEX SHADERS	TMUS	ROPS	MEMORY SIZE	MEMORY TYPE	BUS WIDTH

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

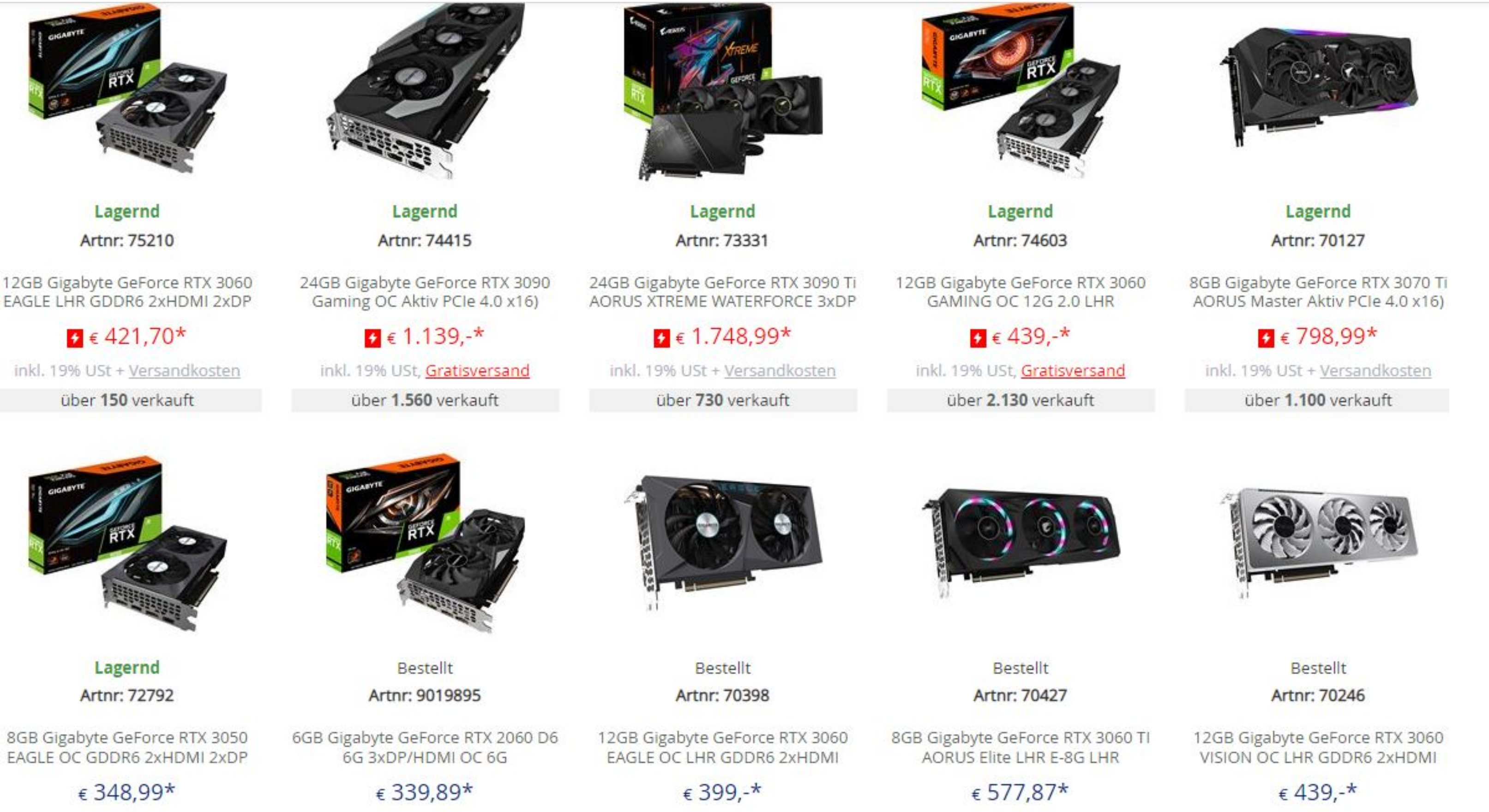
GPUs these days

Screenshot from WS 2021/22



GPUs these days

Screenshot from WS 2022/23



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

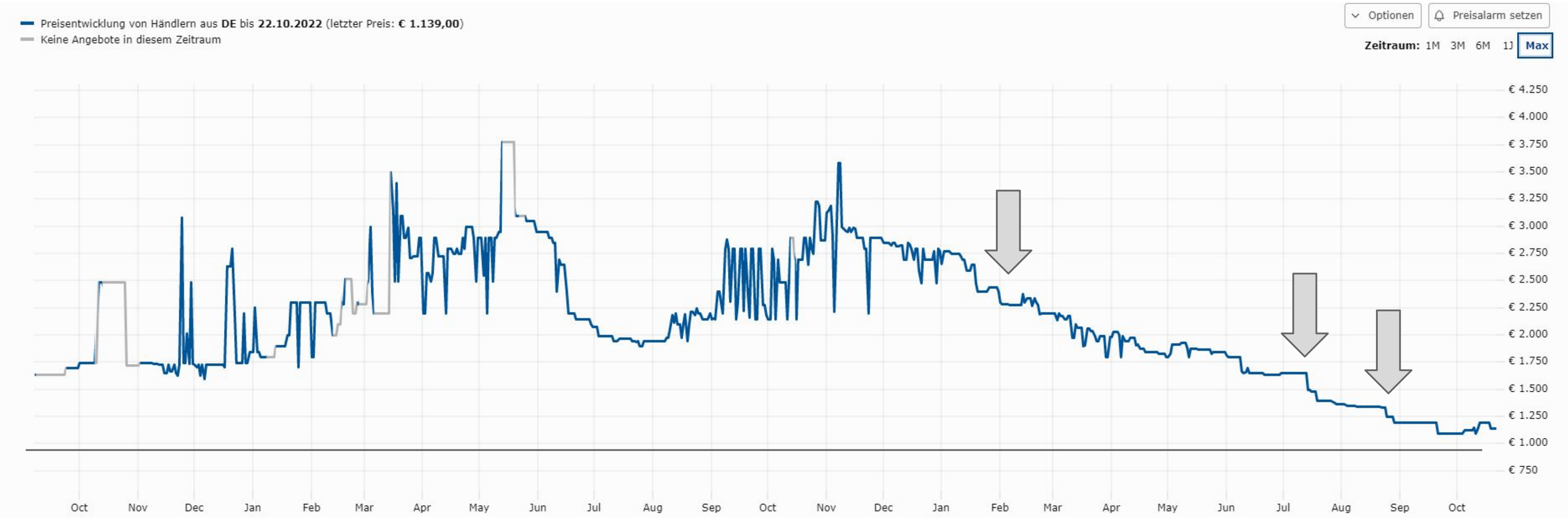
Artnr: 9113760

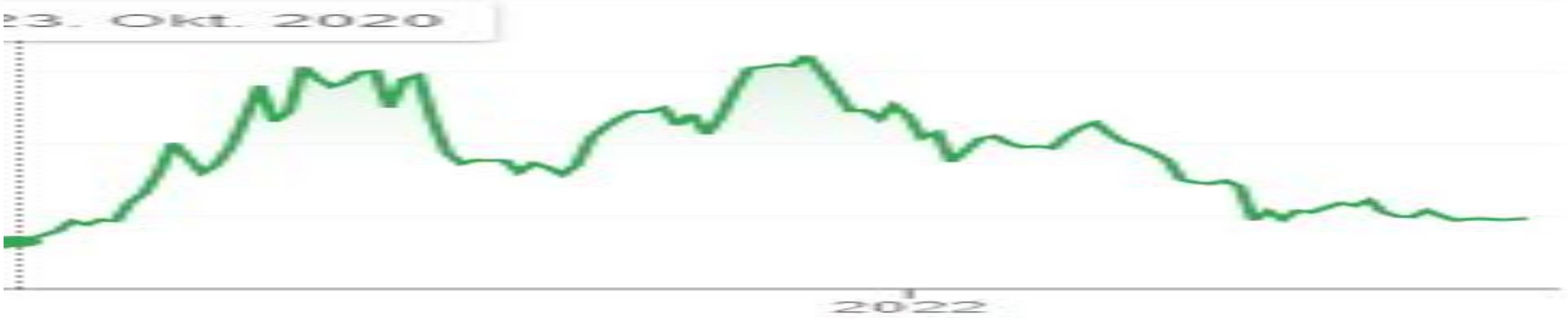
12GB MSI Geforce RTX 4070 VENTUS OC Aktiv PCIe 4.0 x16

€ 608,57*

GPU price

RTX 3090 UVP: **1500 EUR**





Bitcoin etc mining

GPU terminology

Usage specific GPU[edit]

Most GPUs are designed for a specific usage, real-time 3D graphics or other mass calculations:

1. Gaming
    - GeForce GTX, RTX
    - Nvidia Titan
    - Radeon HD, R5, R7, R9, RX, Vega and Navi series
    - Radeon VII
2. Cloud Gaming
    - Nvidia GRID
    - Radeon Sky
3. Workstation
    - Nvidia Quadro
    - Nvidia RTX
    - AMD FirePro
    - AMD Radeon Pro
4. Cloud Workstation
    - Nvidia Tesla
    - AMD FireStream
5. Artificial Intelligence training and Cloud
    - Nvidia Tesla
    - AMD Radeon Instinct
6. Automated/Driverless car
    - Nvidia Drive PX

AMD vs NVIDIA?!

NVIDIA CUDA

NVIDIA.COM

cuDNN

OpenCL™

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

What do GPUs do?

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

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)

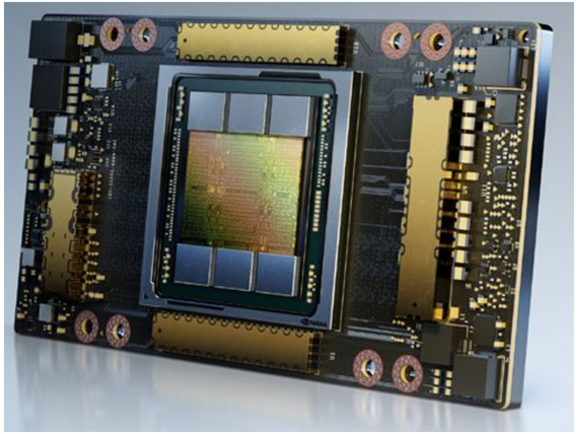
Graphics Card	
Release Date:	Sep 1st, 2020
Availability:	Sep 24th, 2020
Generation:	GeForce 30
Predecessor:	GeForce 20
Production:	Active
Launch Price:	1,499 USD
Current Price:	Amazon / Newegg
Bus Interface:	PCIe 4.0 x16
Reviews:	56 in our database

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

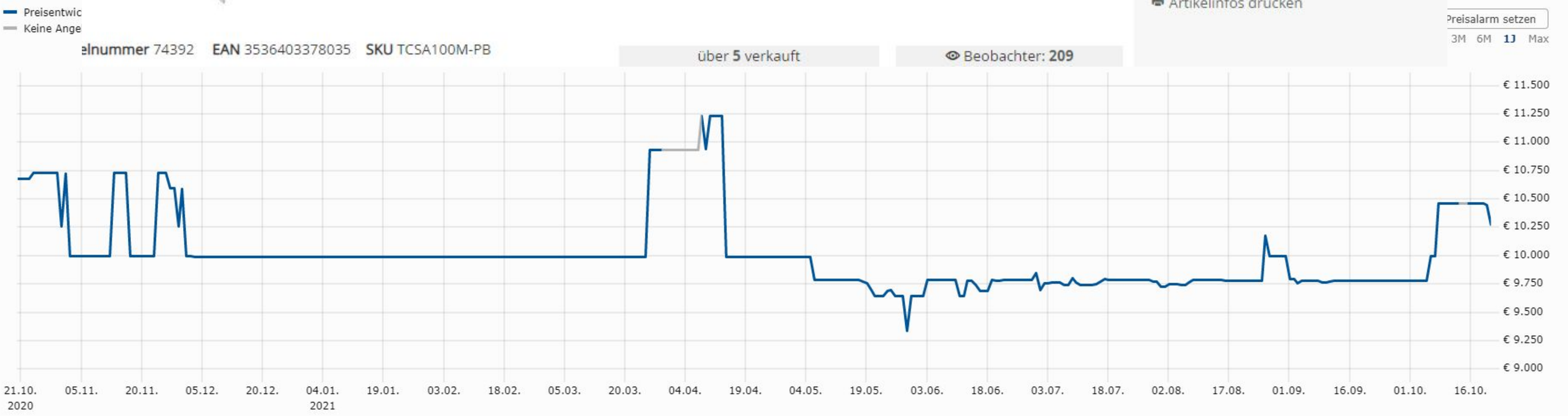
- PaketShop-Lieferung möglich
- Versandkosten sparen
- Feedback zum Artikel
- Artikelinfos drucken

elnummer 74392 EAN 3536403378035 SKU TCSA100M-PB

über 5 verkauft

Beobachter: 209

Preisent



Floating point precision

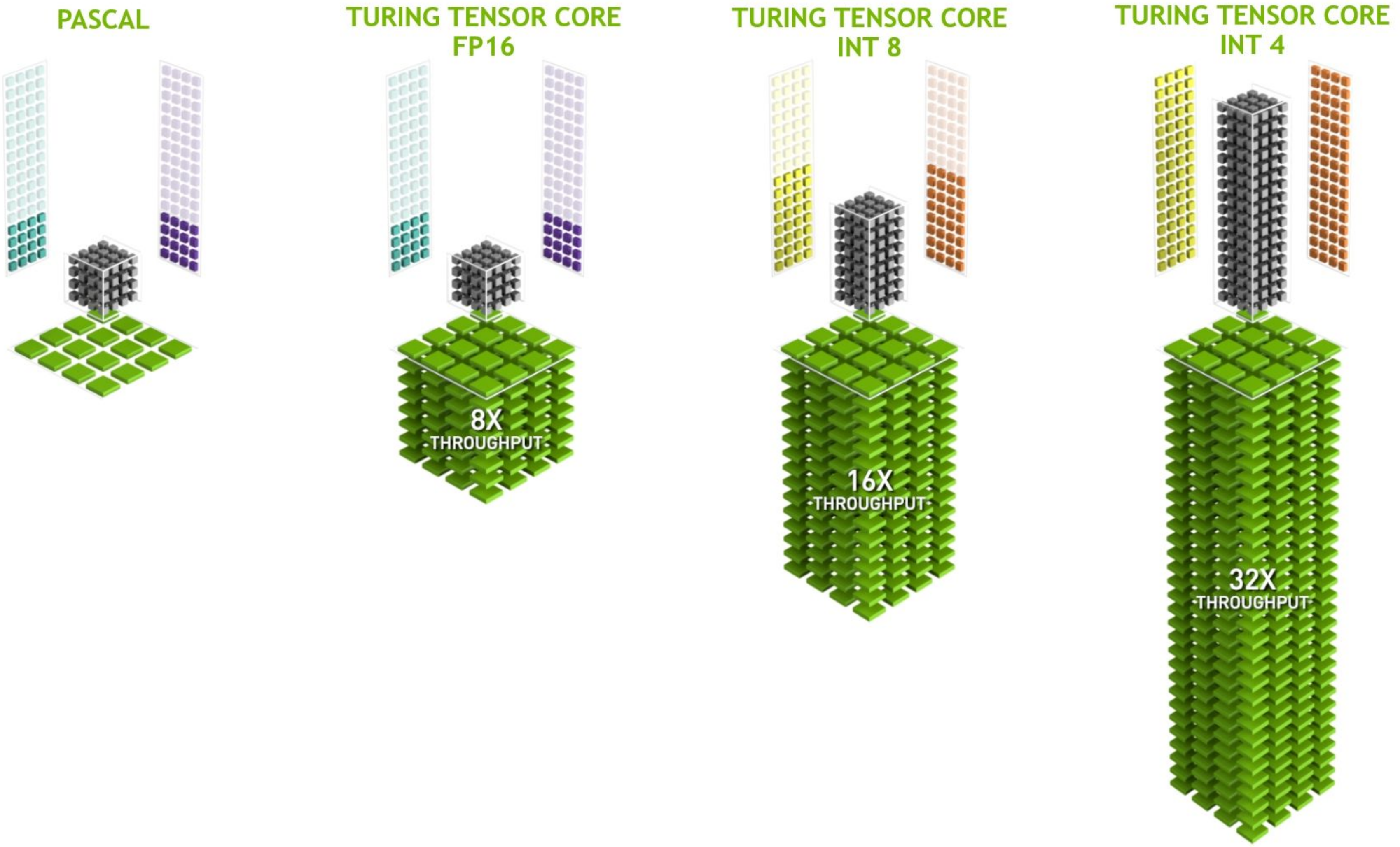
INT8

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

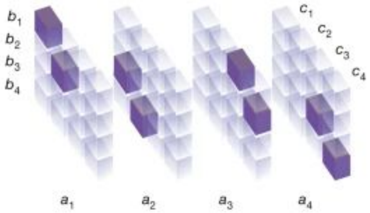

PASCAL
TURING TENSOR CORE
FP16
8X
THROUGHPUT
TURING TENSOR CORE
INT 8
16X
THROUGHPUT
TURING TENSOR CORE
INT 4
32X
THROUGHPUT

AlphaTensor

Volker Strassen

a

$$\begin{pmatrix} c_1 & c_2 \\ c_3 & c_4 \end{pmatrix} = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \cdot \begin{pmatrix} b_1 & b_2 \\ b_3 & b_4 \end{pmatrix}$$



b

$m_1 = (a_1 + a_4)(b_1 + b_4)$

$m_2 = (a_3 + a_4)\, b_1$

$m_3 = a_1\,(b_2 - b_4)$

$m_4 = a_4\,(b_3 - b_1)$

$m_5 = (a_1 + a_2)\, b_4$

$m_6 = (a_3 - a_1)(b_1 + b_2)$

$m_7 = (a_2 - a_4)(b_3 + b_4)$

$c_1 = m_1 + m_4 - m_5 + m_7$

$c_2 = m_3 + m_5$

$c_3 = m_2 + m_4$

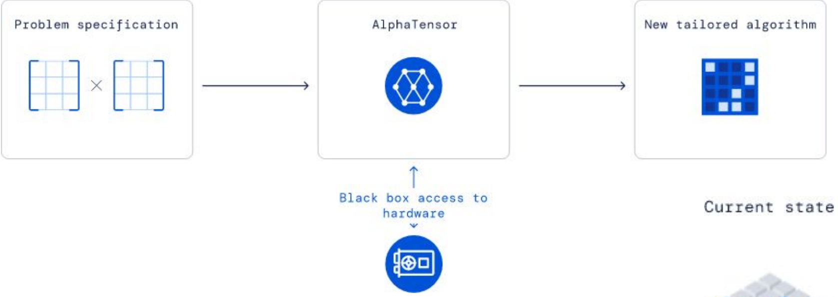
$c_4 = m_1 - m_2 + m_3 + m_6$

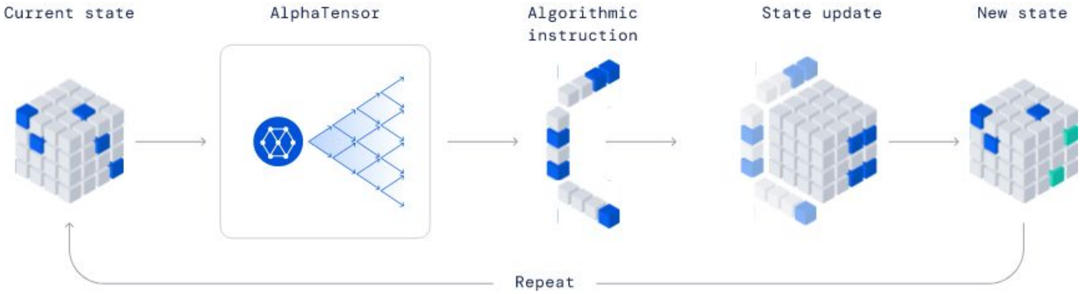
c

$$\mathbf{U} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & -1 \end{pmatrix}$$

$$\mathbf{V} = \begin{pmatrix} 1 & 1 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & -1 & 0 & 1 & 0 & 1 \end{pmatrix}$$

$$\mathbf{W} = \begin{pmatrix} 1 & 0 & 0 & 1 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$





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

GPU memory

Short: the more, the better (larger models, bigger batch sizes, ...)

How can I fit +24GB models into 10GB memory?

It is a bit contradictory that I just said if you want to train big models, you need lots of memory, but we have been struggling with big models a lot since the onslaught of BERT and solutions exists to train 24 GB models in 10 GB memory. If you do not have the money or what to avoid cooling/power issues of the RTX 3090, you can get RTX 3080 and just accept that you need do some extra programming by adding memory-saving techniques. There are enough techniques to make it work, and they are becoming more and more commonplace.

Here just a list of common techniques:

- FP16/BF16 training (apex)
- Gradient checkpointing (only store some of the activations and recompute them in the backward pass)
- GPU-to-CPU Memory Swapping (swap layers not needed to the CPU; swap them back in just-in-time for backprop)
- Model Parallelism (each GPU holds a part of each layer; supported by fairseq)
- Pipeline parallelism (each GPU hols a couple of layers of the network)
- ZeRO parallelism (each GPU holds partial layers)
- 3D parallelism (Model + pipeline + ZeRO)
- CPU Optimizer state (store and update Adam/Momentum on the CPU while the next GPU forward pass is happening)

https://timdettmers.com/2020/09/07/which-gpu-for-deep-learning/

The best GPU for Deep Learning

https://timdettmers.com/2020/09/07/which-gpu-for-deep-learning/

TL;DR advice

Best GPU overall: RTX 3080 and RTX 3090.

GPUs to avoid (as an individual): Any Tesla card; any Quadro card; any Founders Edition card; Titan RTX, Titan V, Titan XP.

Cost-efficient but expensive: RTX 3080.

Cost-efficient and cheaper: RTX 3070, RTX 2060 Super

I have little money: Buy used cards. Hierarchy: RTX 2070 ($400), RTX 2060 ($300), GTX 1070 ($220), GTX 1070 Ti ($230), GTX 1650 Super ($190), GTX 980 Ti (6GB $150).

I have almost no money: There are a lot of startups that promo their clouds: Use free cloud credits and switch companies accounts until you can afford a GPU.

I do Kaggle: RTX 3070.

I am a competitive computer vision, pretraining, or machine translation researcher: 4x RTX 3090. Wait until working builds with good cooling, and enough power are confirmed (I will update this blog post).

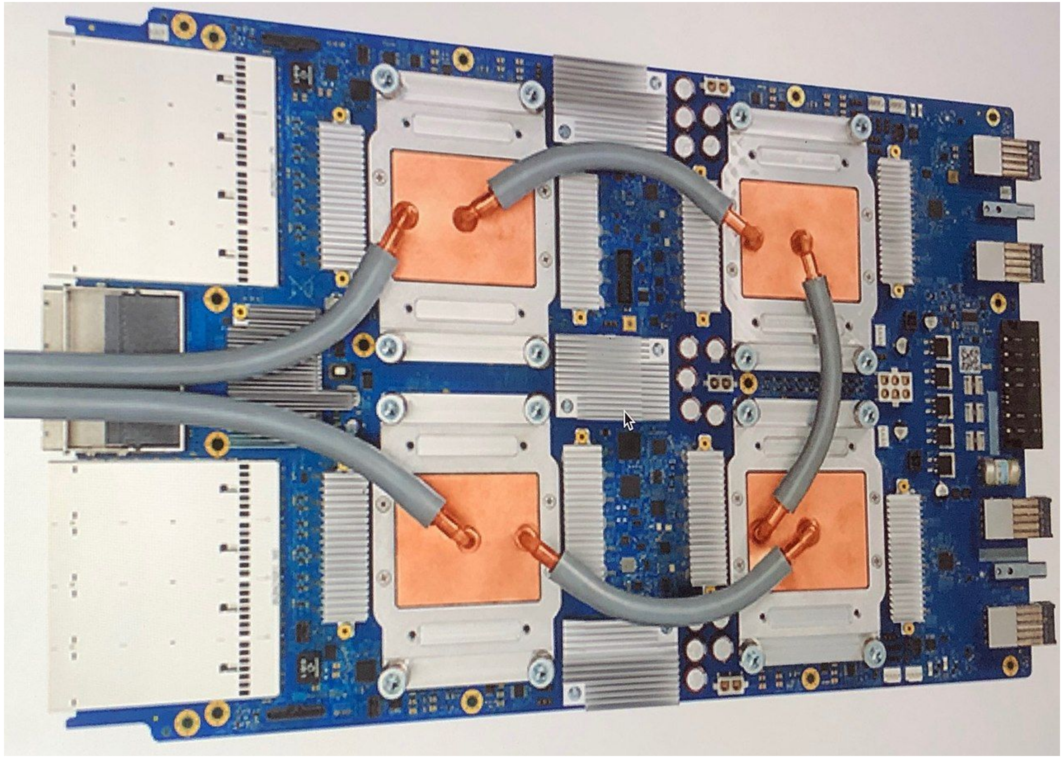
I am an NLP researcher: If you do not work on machine translation, language modeling, or pretraining of any kind, an RTX 3080 will be sufficient and cost-effective.

What do we have in the lab?

- RTX 2080 Ti (1x)
- Titan RTX (1x)
- RTX 3090 (3x)
- RTX A4000 (2x)
- RTX A5000 (2x)

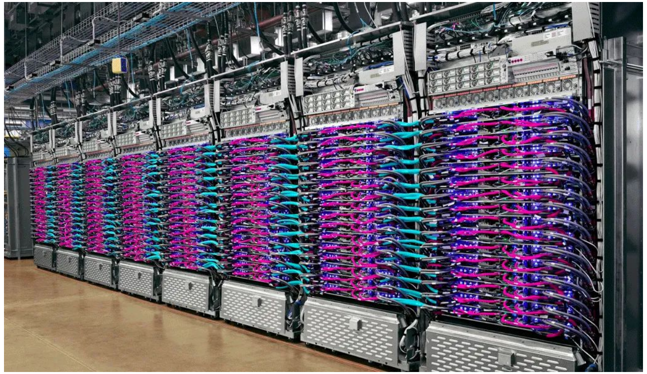
Application Specific ICs (integrated circuits) ASIC

TPUs

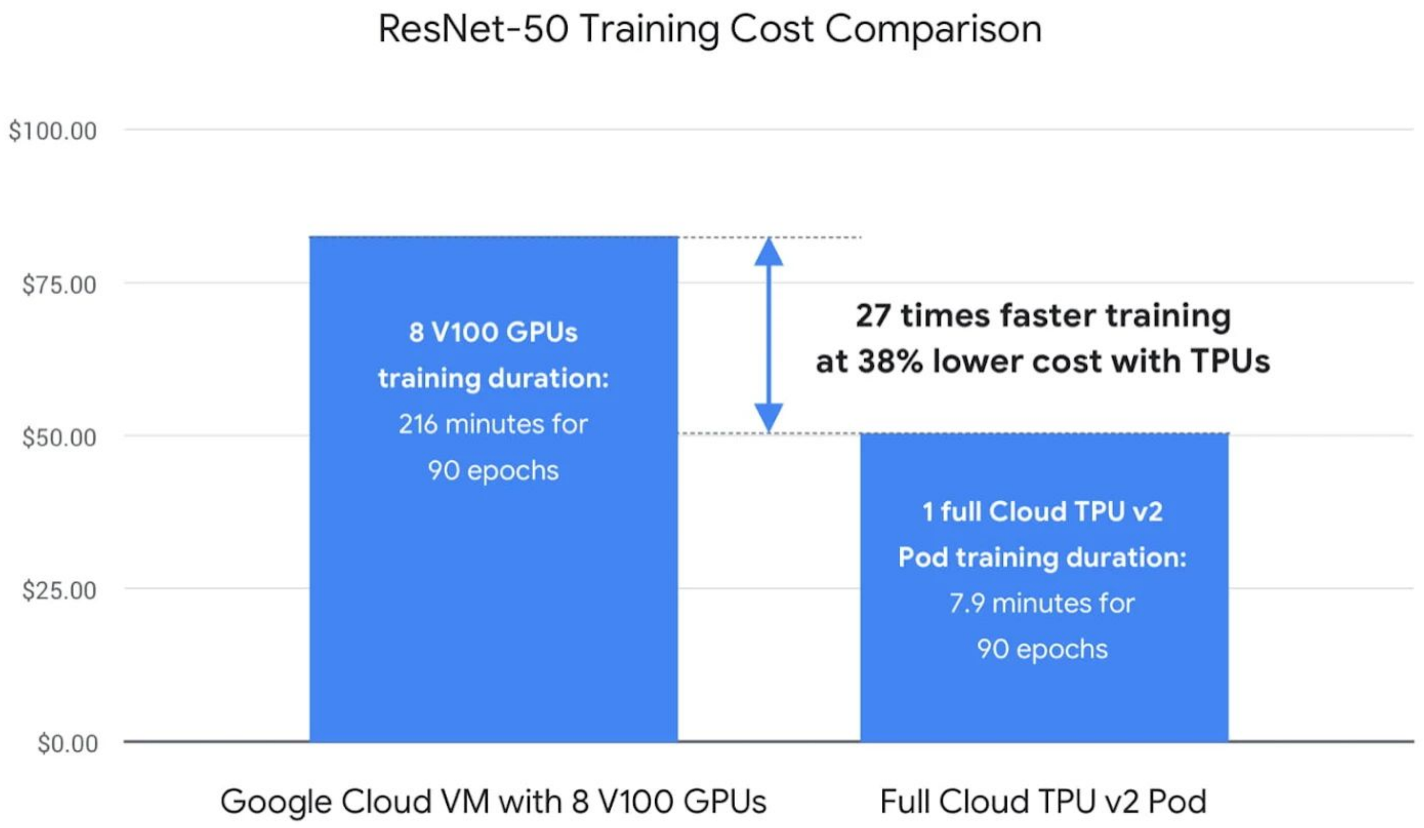


Products[13] [edit]

	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

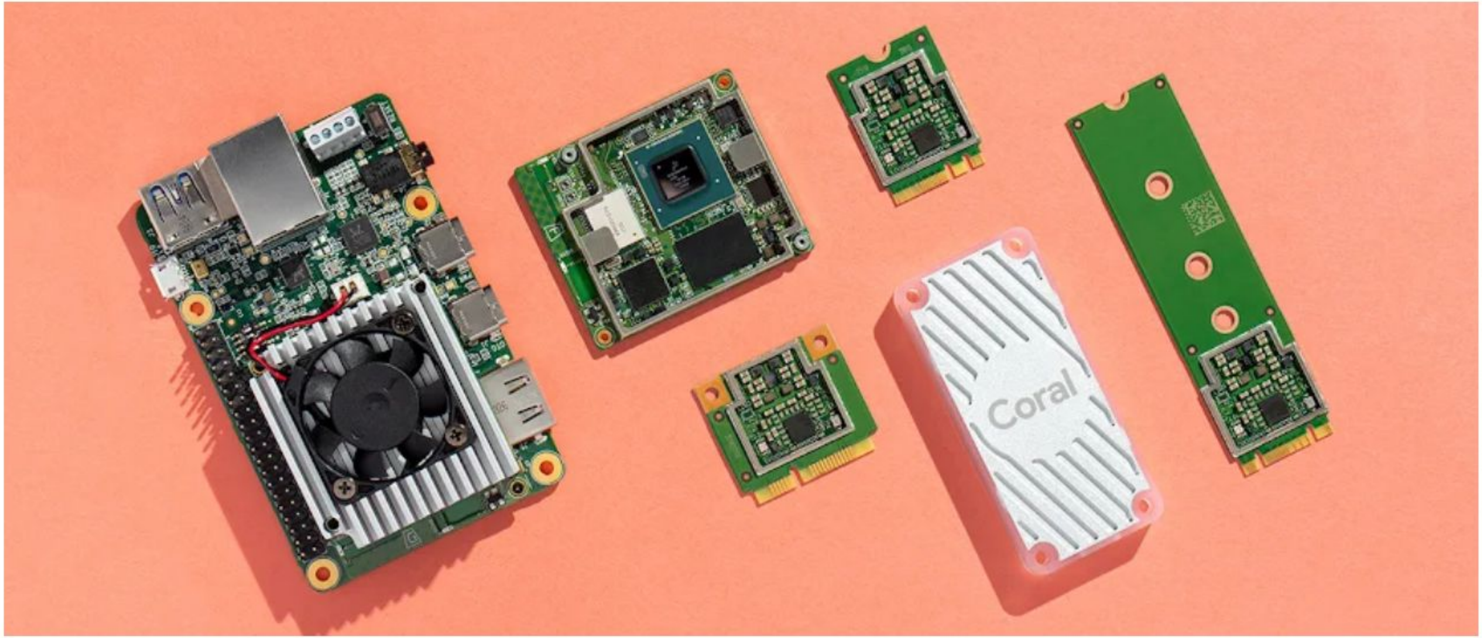
TPU type (v2)	
v2-8	
v2-32	
v2-128	
v2-256	
v2-512	
TPU type (v3)	
v3-8	
v3-32	
v3-64	
v3-128	
v3-256	256
v3-512	512
v3-1024	1024
v3-2048	2048

US | EUROF

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

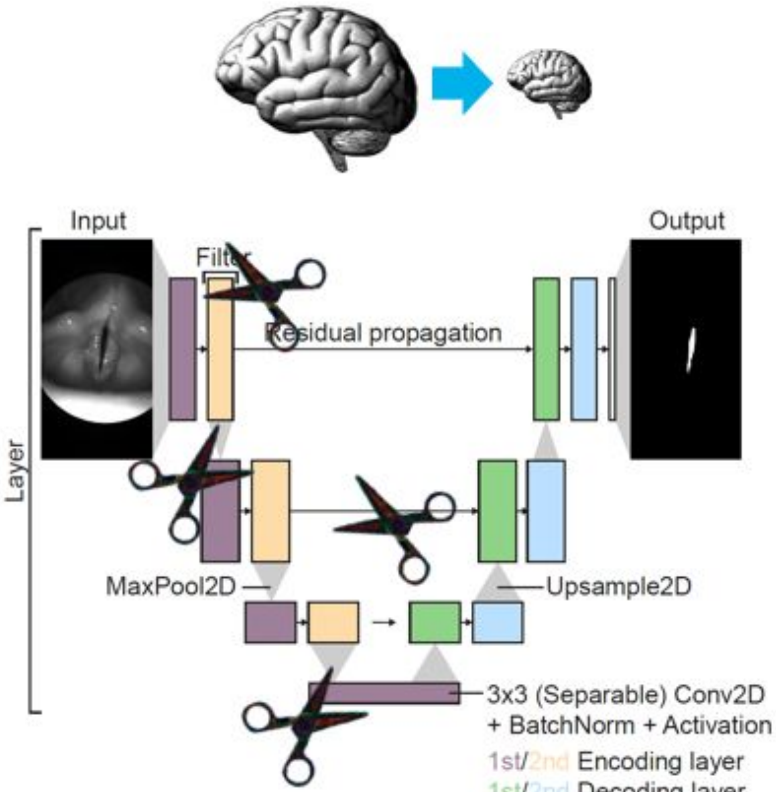


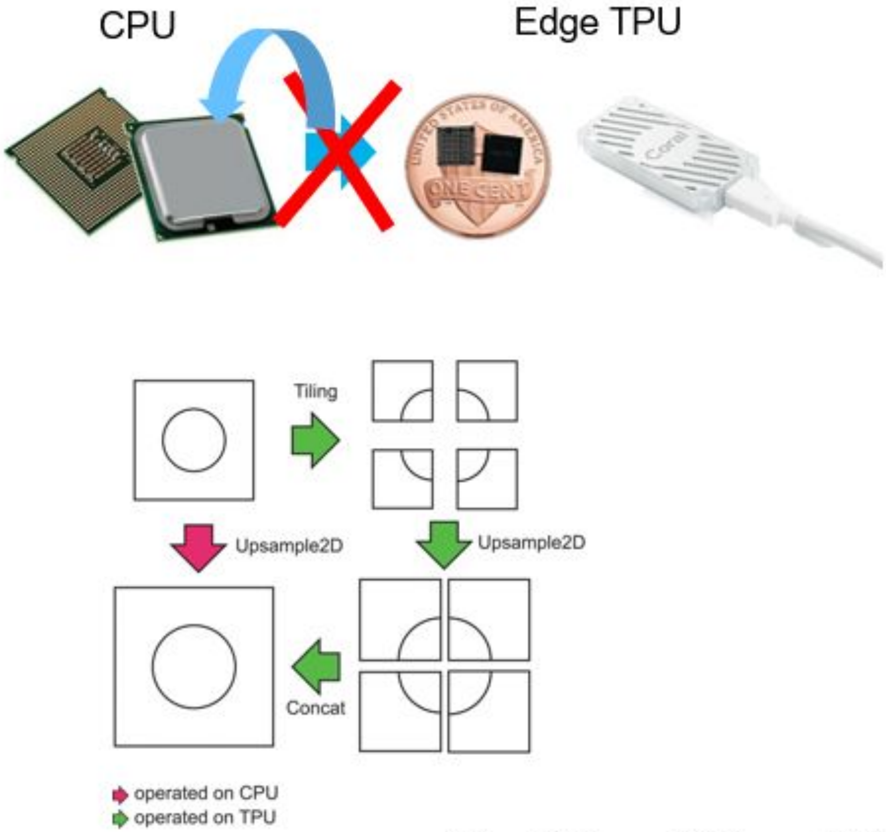
Inference boost

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?!?!

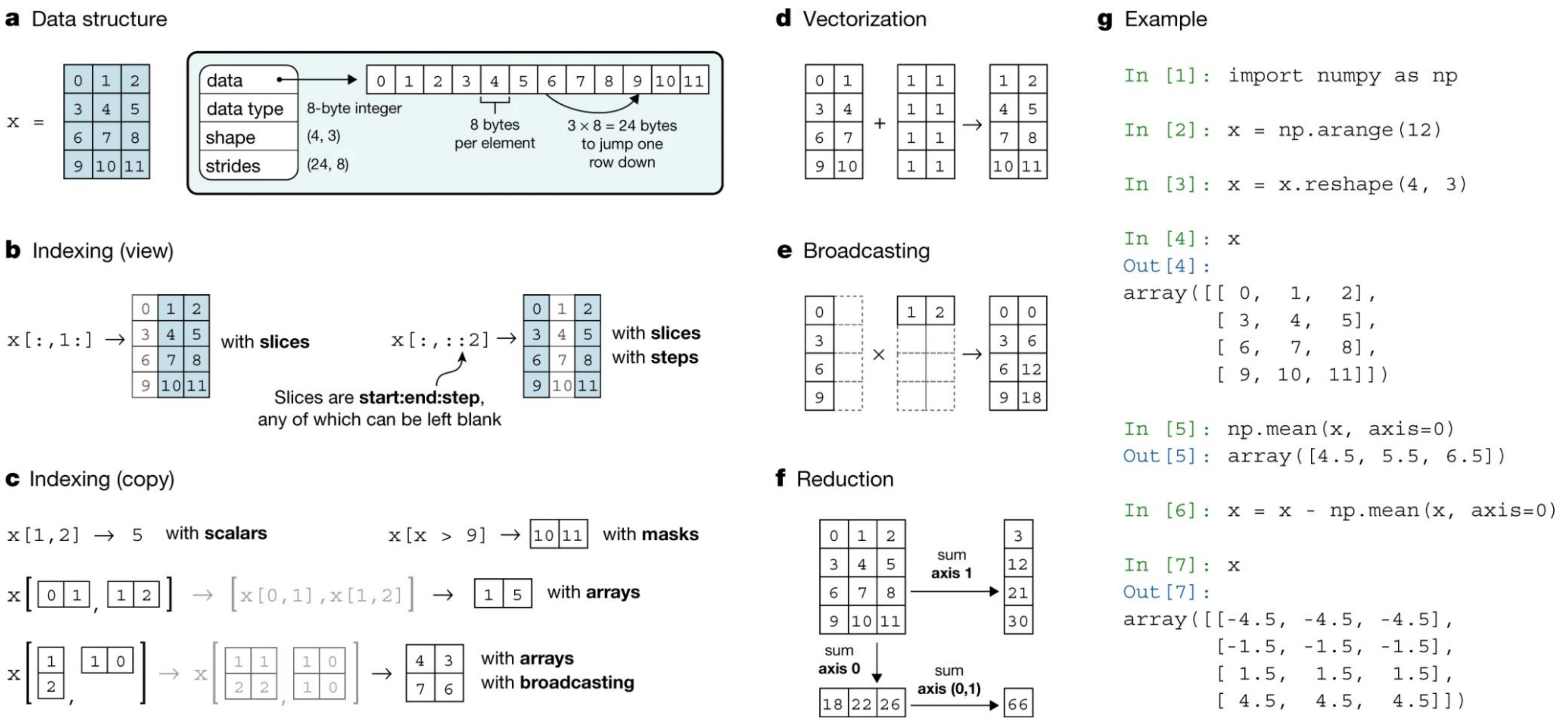
Programming on CPUs and GPUs

Review Article | Open Access | Published: 16 September 2020

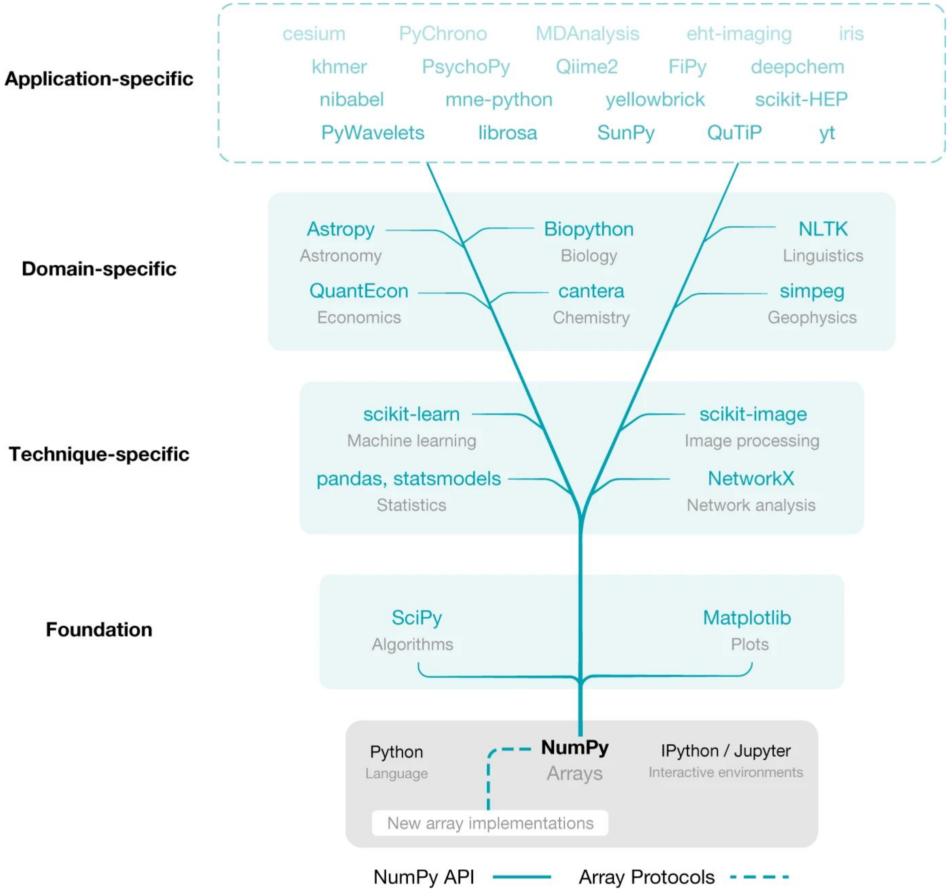
Array programming with NumPy

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

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



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

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.

CuPy

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.

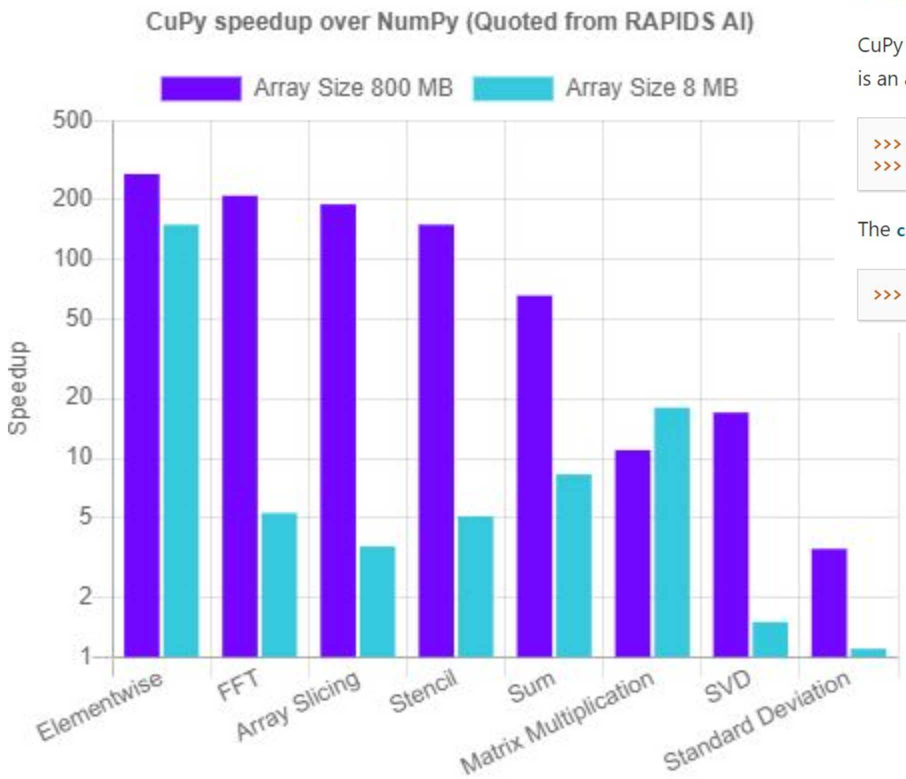
cuML is fully open source, and the RAPIDS team welcomes new and seasoned contributors, users and hobbyists! Thank you for your wonderful support!

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
```

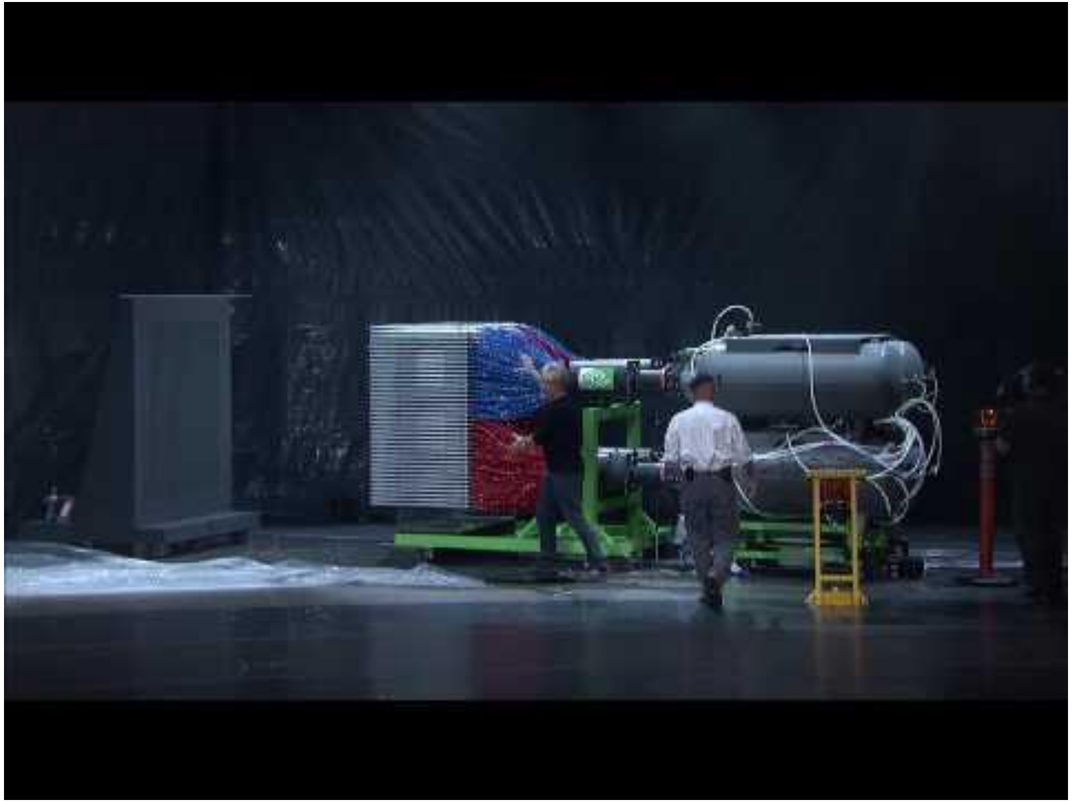
The **`cupy.ndarray`** class is at the core of `CuPy` and is a replacement class for `NumPy`'s **`numpy.ndarray`**.

```
>>> 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

We will record the action how to build a computer and upload it to fau.tv.

Homework

Homework

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