Lecture 9
Future Directions

Prof Wes Armour

wes.armour@eng.ox.ac.uk

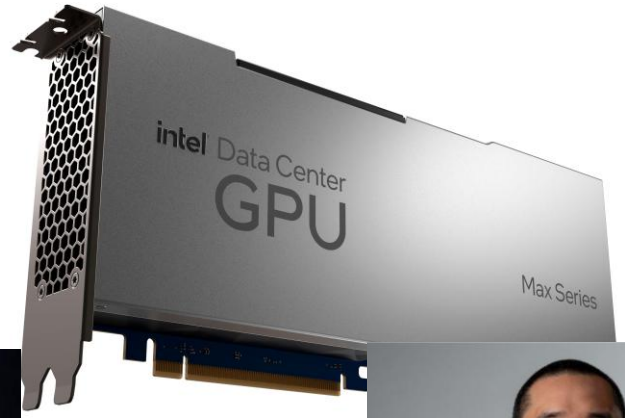
Oxford e-Research Centre

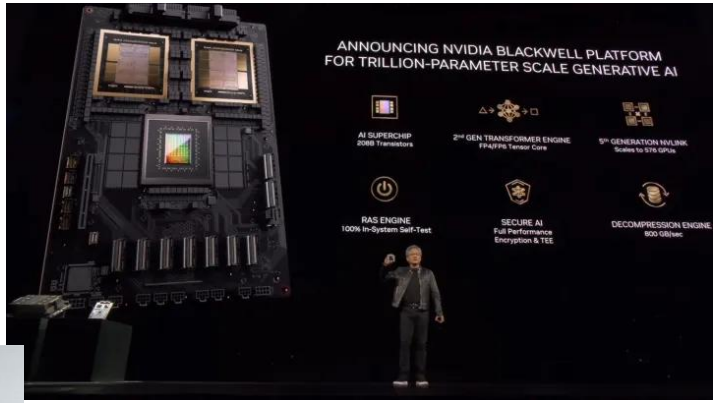
Department of Engineering Science

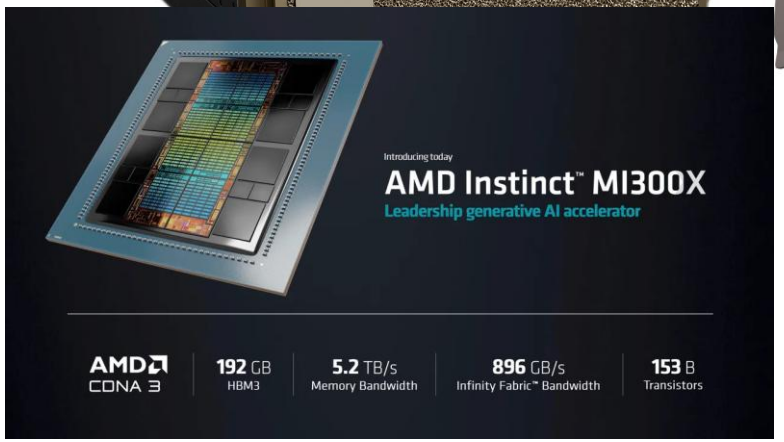
Learning outcomes

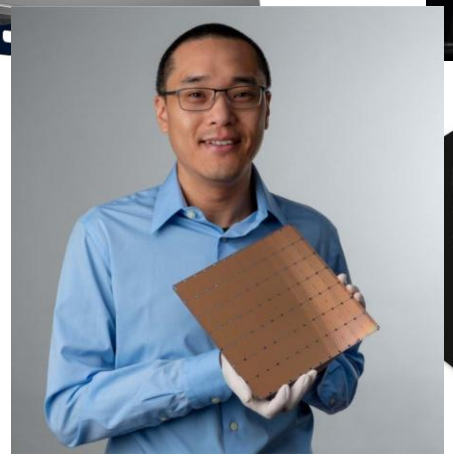
In this lecture we will look at the current landscape of accelerated computing.

We will look at hardware and software trends and potential future directions for accelerated computing.

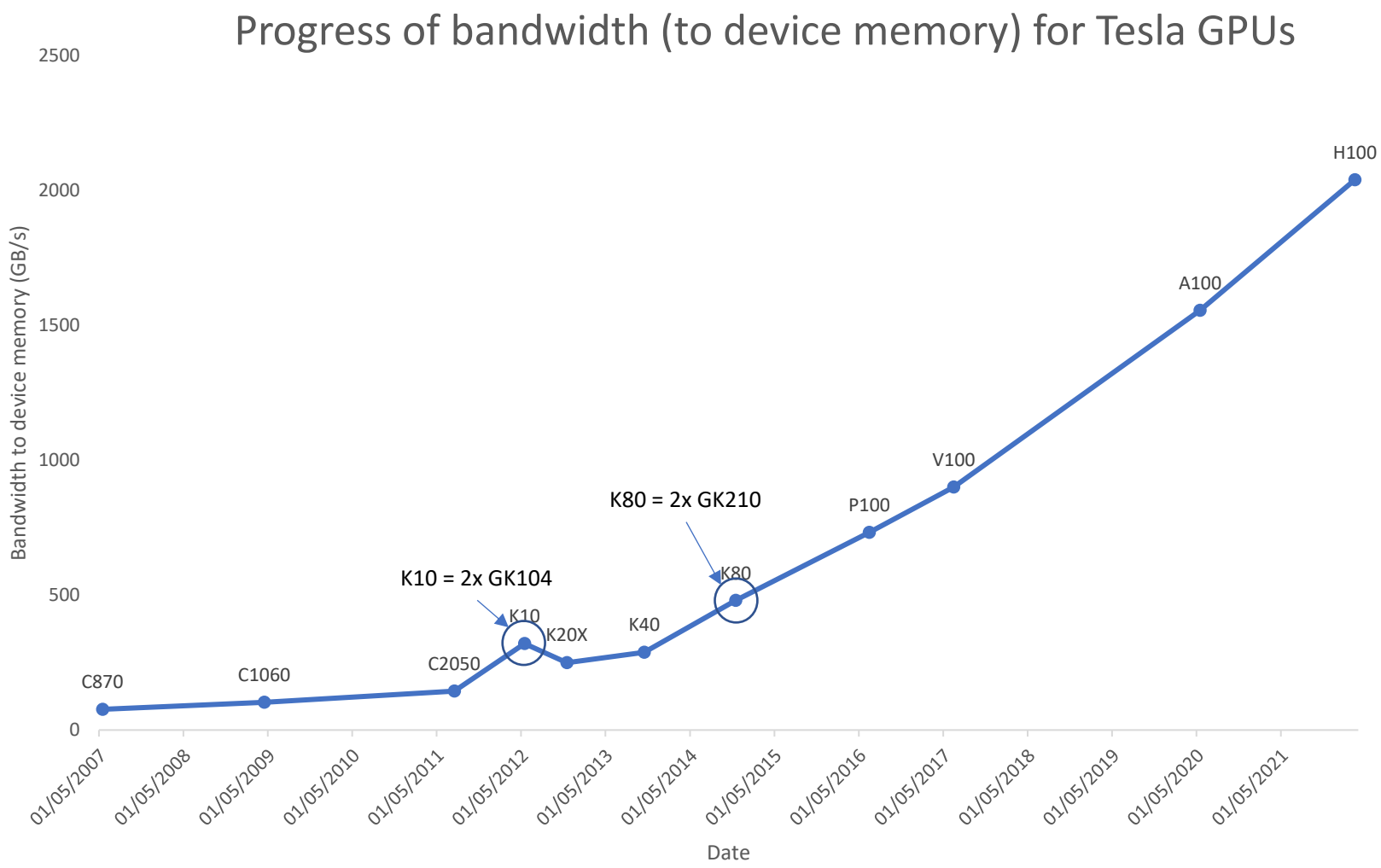




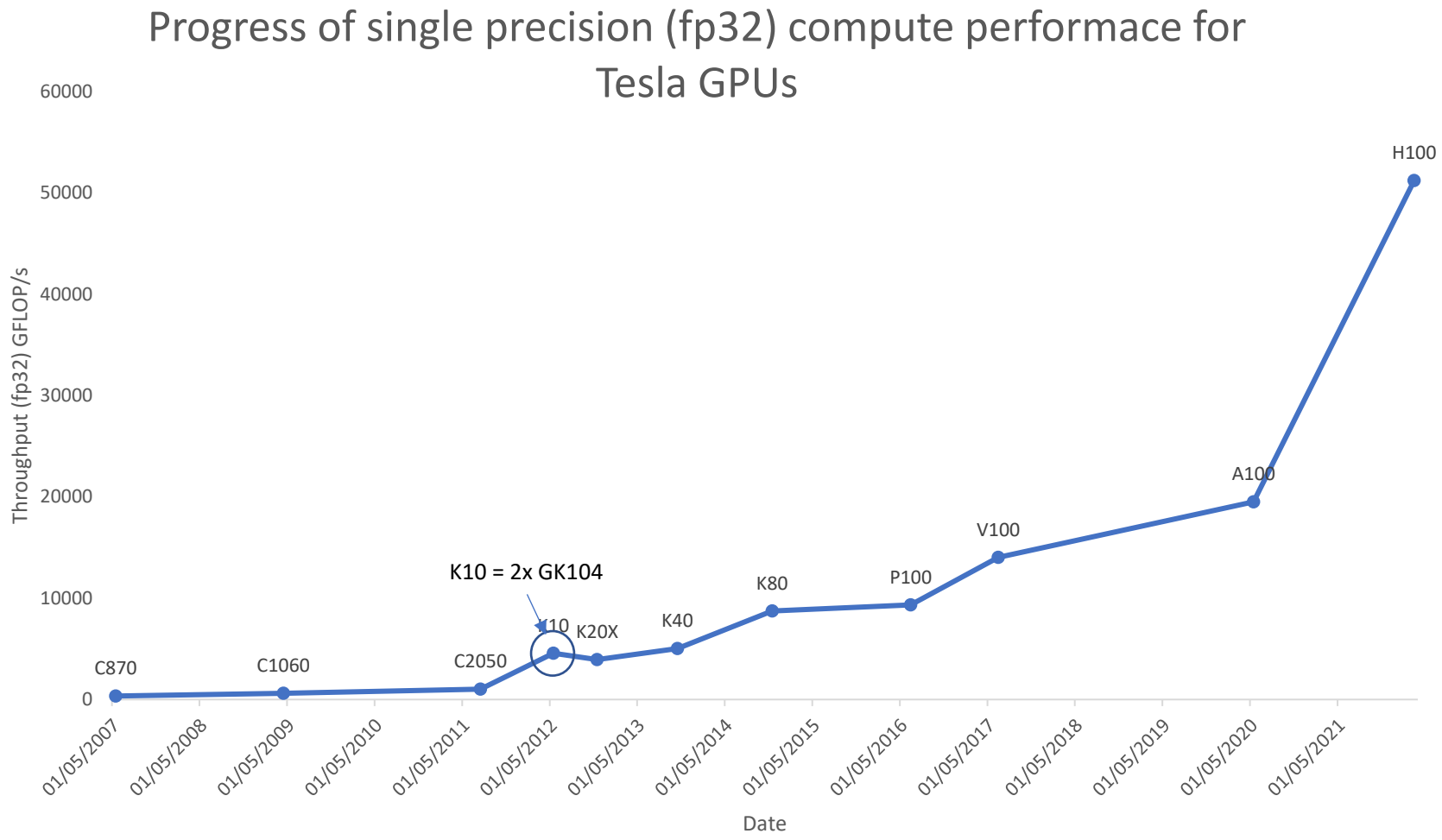




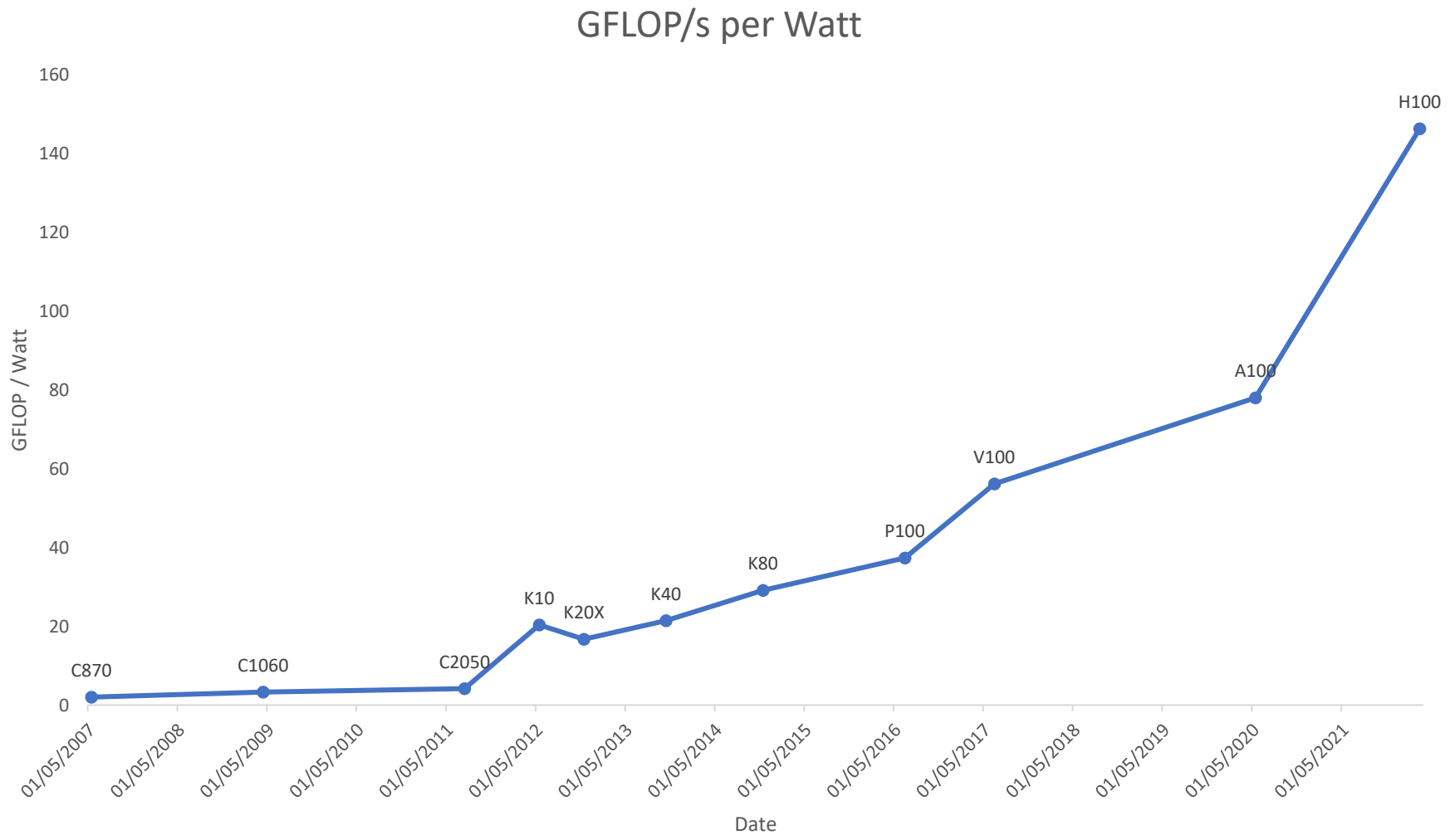
Bandwidth


Progress of bandwidth (to device memory) for Tesla GPUs
Bandwidth to device memory (GB/s)
2500
2000
1500
1000
500
0
C870
C1060
C2050
K10
K20X
K40
K80
P100
V100
A100
H100
K10 = 2x GK104
K80 = 2x GK210
01/05/2007
01/05/2008
01/05/2009
01/05/2010
01/05/2011
01/05/2012
01/05/2013
01/05/2014
01/05/2015
01/05/2016
01/05/2017
01/05/2018
01/05/2019
01/05/2020
01/05/2021
Date

Compute


Progress of single precision (fp32) compute performace for Tesla GPUs
Throughput (fp32) GFLOP/s
60000
50000
40000
30000
20000
10000
0
C870
C1060
C2050
K10 = 2x GK104
K10
K20X
K40
K80
P100
V100
A100
H100
01/05/2007
01/05/2008
01/05/2009
01/05/2010
01/05/2011
01/05/2012
01/05/2013
01/05/2014
01/05/2015
01/05/2016
01/05/2017
01/05/2018
01/05/2019
01/05/2020
01/05/2021
Date

Efficiency


GFLOP/s per Watt
GFLOP / Watt
160
140
120
100
80
60
40
20
0
C870
C1060
C2050
K10
K20X
K40
K80
P100
V100
A100
H100
01/05/2007
01/05/2008
01/05/2009
01/05/2010
01/05/2011
01/05/2012
01/05/2013
01/05/2014
01/05/2015
01/05/2016
01/05/2017
01/05/2018
01/05/2019
01/05/2020
01/05/2021
Date

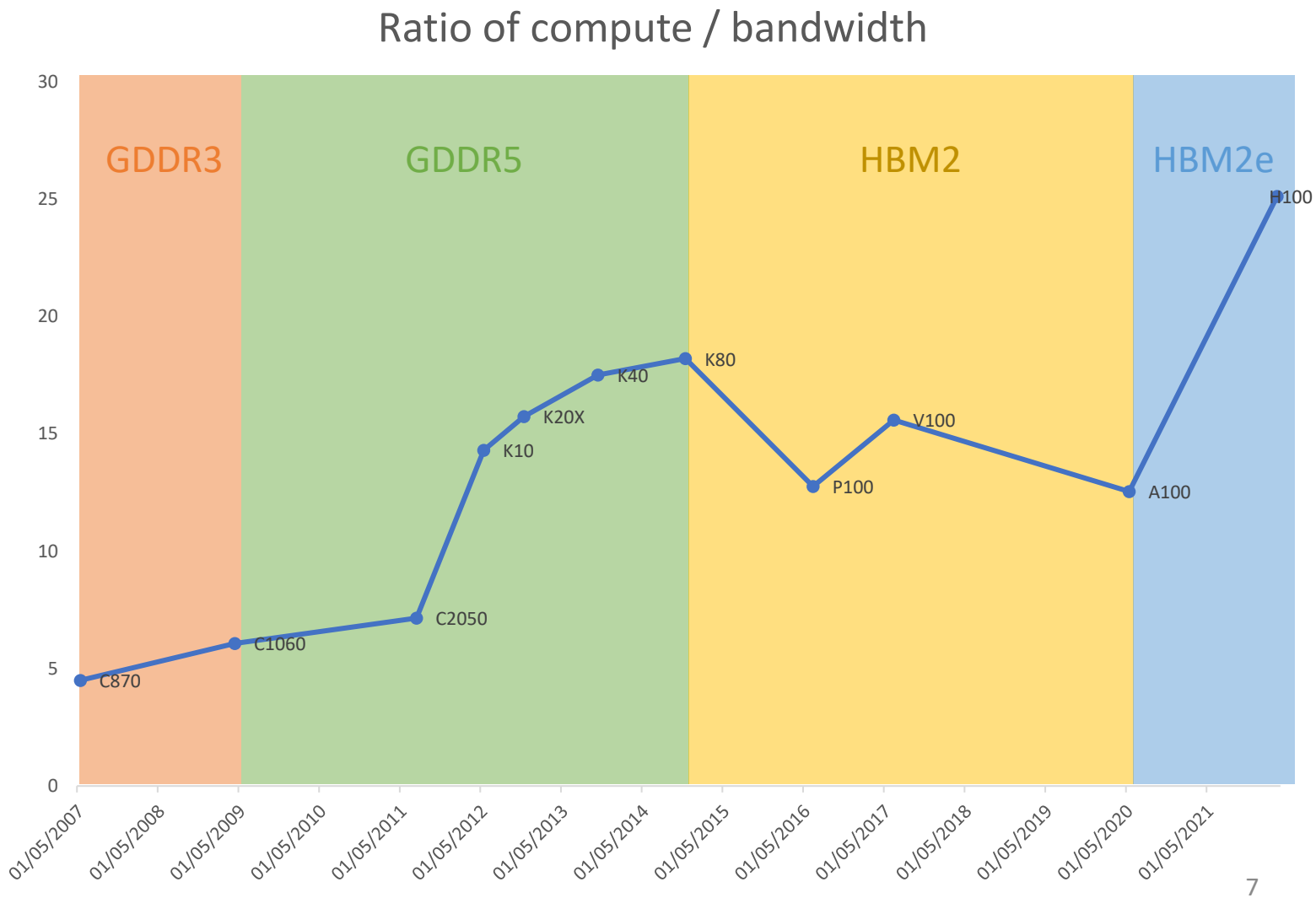
Ratio of compute / bandwidth

The ratio of compute / bandwidth often called **arithmetic intensity** or **operational intensity** ($\boldsymbol{I}$) tells us how many floating point operations we can perform in the time it takes to move each byte of data from device memory.

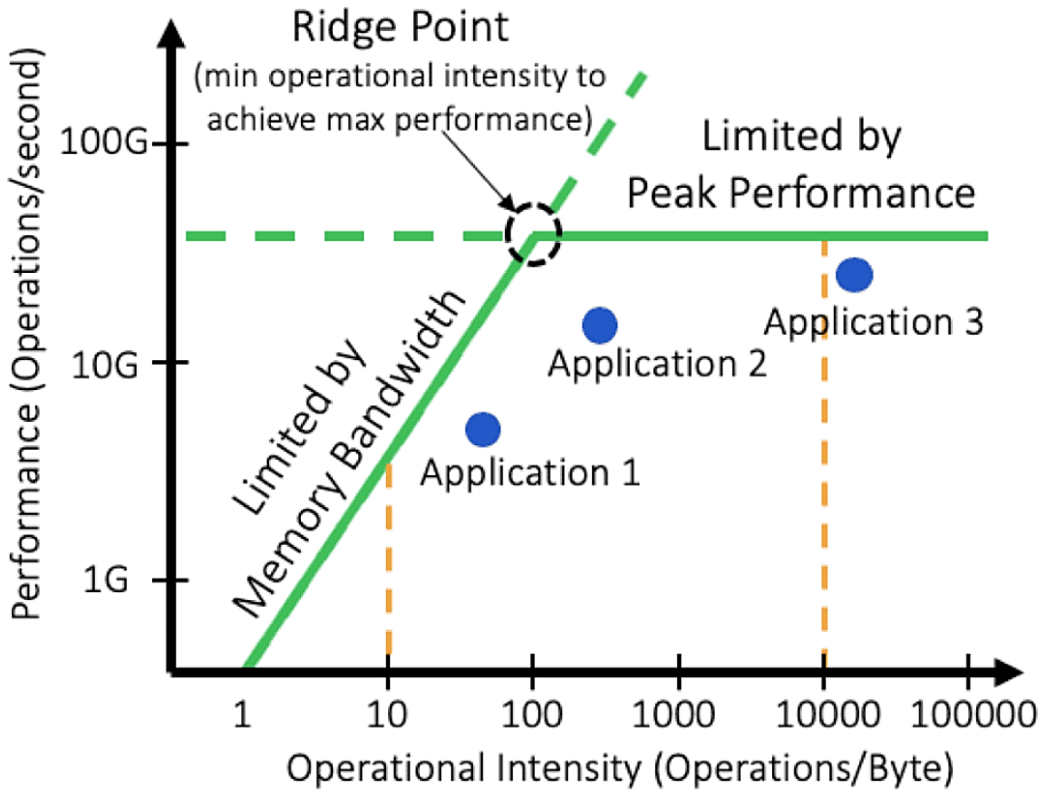
$$I = \frac{W}{Q} = \frac{Work}{Memory\ traffic} = \frac{FLOPs}{Byte}$$

Ratio of (peak) compute / bandwidth

We can see from the plot on the right that, although the introduction of new memory technologies reduces operational intensity for short periods, **the overall trend is for it to increase**.



Roofline model



The roofline model tells us, for given hardware, whether our application will be bandwidth bound or compute bound.

Reminder - recompute not transfer

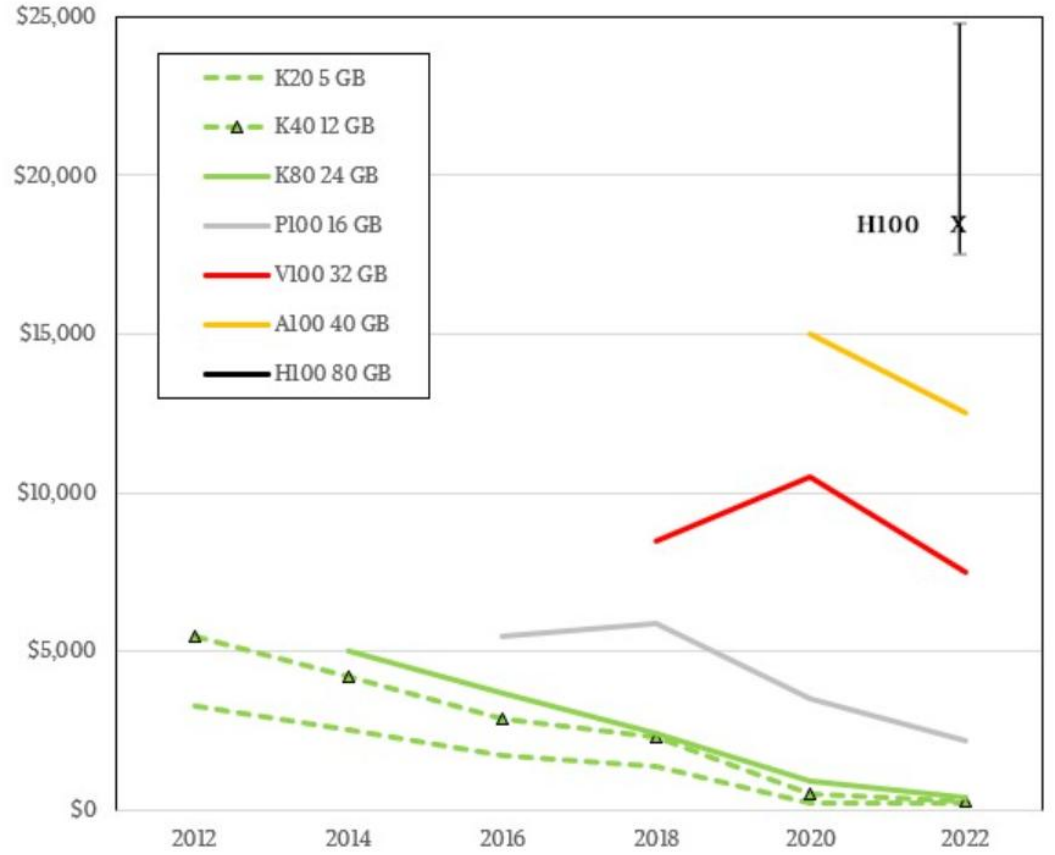
Given the fact that we can now perform so many FLOPs per byte that we move from device memory (e.g. ~100 for a single float on H100), it is worth considering whether it is more efficient to recompute values rather than transferring them.

$$\frac{1}{16}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

The growing cost of owning NVIDIA

Due to market dominance, commercial interest and the boom in AI, the cost of NVIDIA GPUs has increased significantly.

The plot on the right comes from nextplatform and shows the successive increase in launch price for different generations of GPUs.



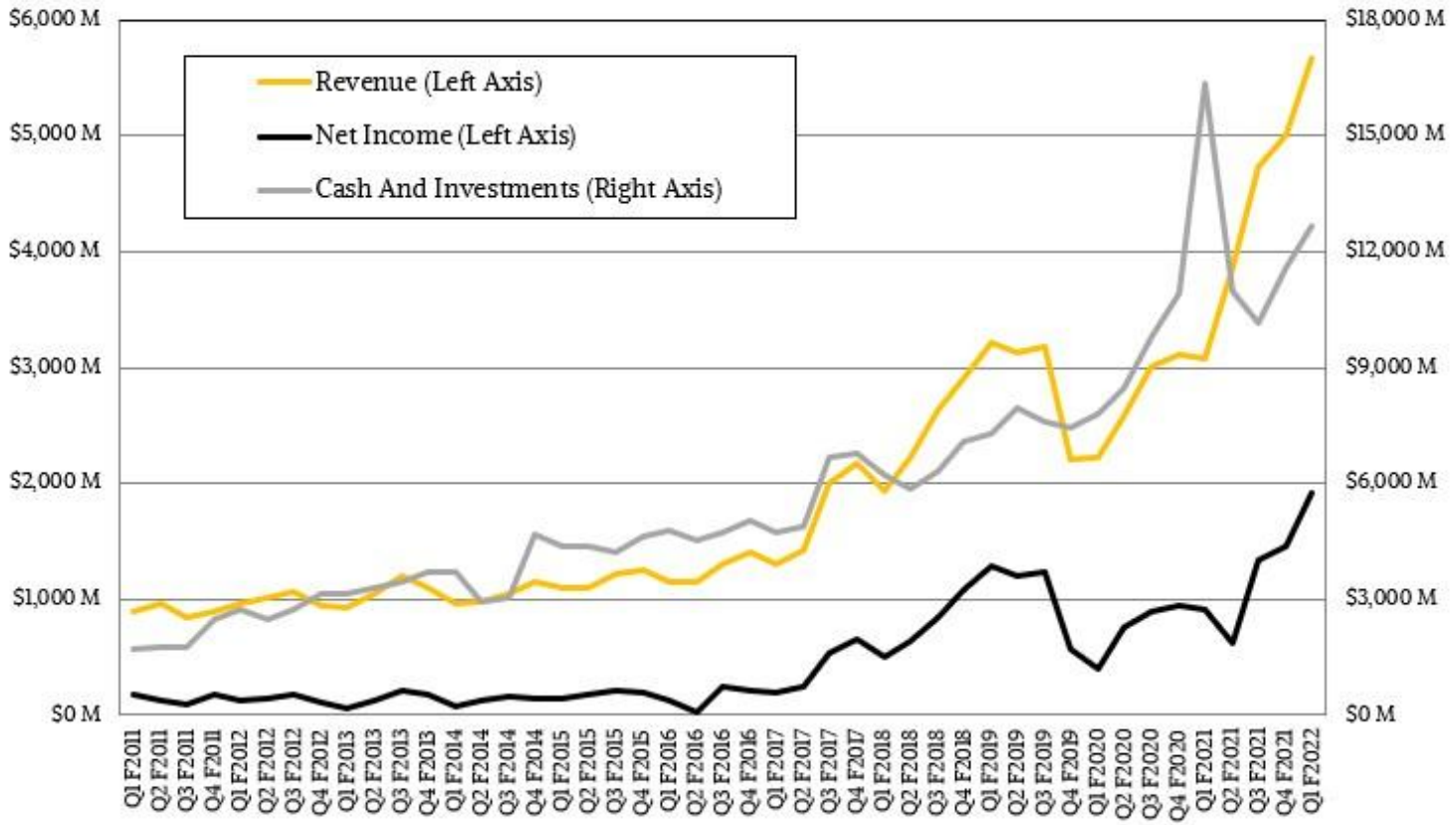
Timothy Prickett Morgan

https://www.nextplatform.com/2022/05/09/how-much-of-a-premium-will-nvidia-charge-for-hopper-gpus/

The growing cost of owning NVIDIA

Here we see the growth in NVIDIA's revenue over the last decade, again we see in the last few years near exponential growth.

So, what does this mean in terms of GPU availability and total cost of ownership.



https://www.nextplatform.com/2021/05/26/nvidias-next-major-wave-of-ai-revenues/

Timothy Prickett Morgan

Current prices – economy

Either a gaming GPU or “Pro” card (but beware of very slow fp64!!)

RTX 5090 / RTX 6000 Pro

- Built on GB202
- 32/96 GB of GDDR7 (6000 Pro has ECC)
- 170 SMs / ~20,000 “cuda cores”
- 105 TFLOP/s of fp32 (1.6 TFLOP/s of fp64)
- PCIe 5.0 (128 GB/s)

Estimated cost
£2K for 5090 founders edition
£8K-£9K for 6000 Pro

Power consumption 600W

Current prices – cloud

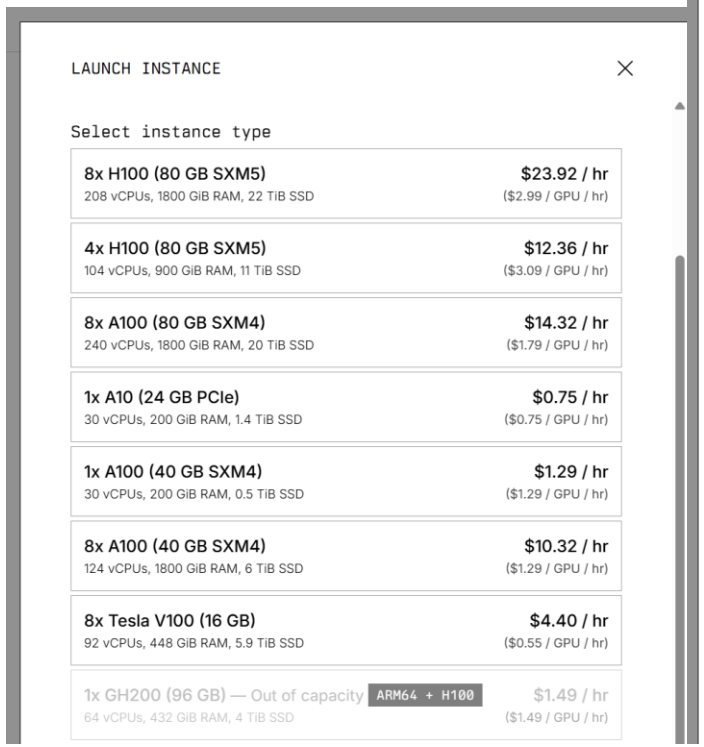
Recall that a DGX B200 costs about $700K to own (SCAN list price was £500K).

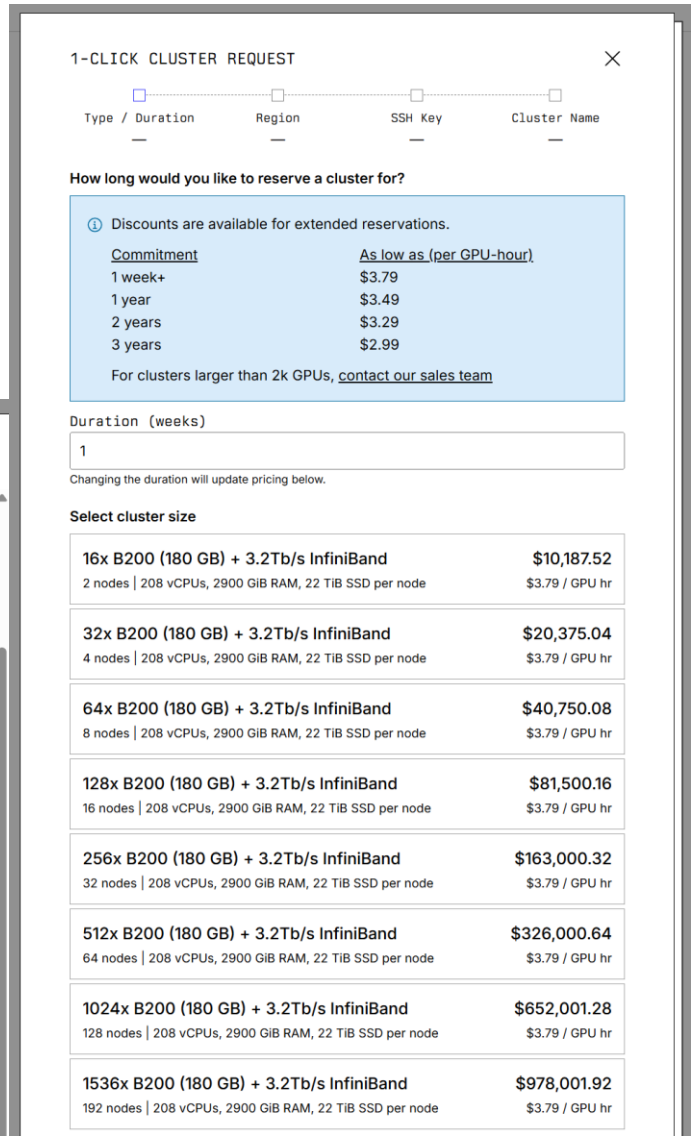
So 16x B200 (or 2x DGX nodes) about $1.4M.

Currently on Lambda to access B200 technology you must commit to at least 16x B200s and 1 week rental (~$10K).

Lower down the food chain you can access A100 for about $1.

For similar nodes to the ones that you have used this week for practical sessions, just $4.40 per hour.





Current prices – top end

Currently the highest performing multi-GPU solution is GB200 NVL72

- Built on GB200 “superchips” each having
  1x Grace CPU (72 Neoverse cores) and
  2x Blackwell GPUs
- A NVL72 rack has 72x NVIDIA GB200 “superchips”.

- 5760 TFLOPs of fp32 per rack.
- 13.4TB of HBM3e.
- NVLink 1.8TB/s per GPU (3.6TB/s per superchip).
- 130TB/s NVLink aggregate per rack.

Estimated cost $3M

Power consumption ~120KW

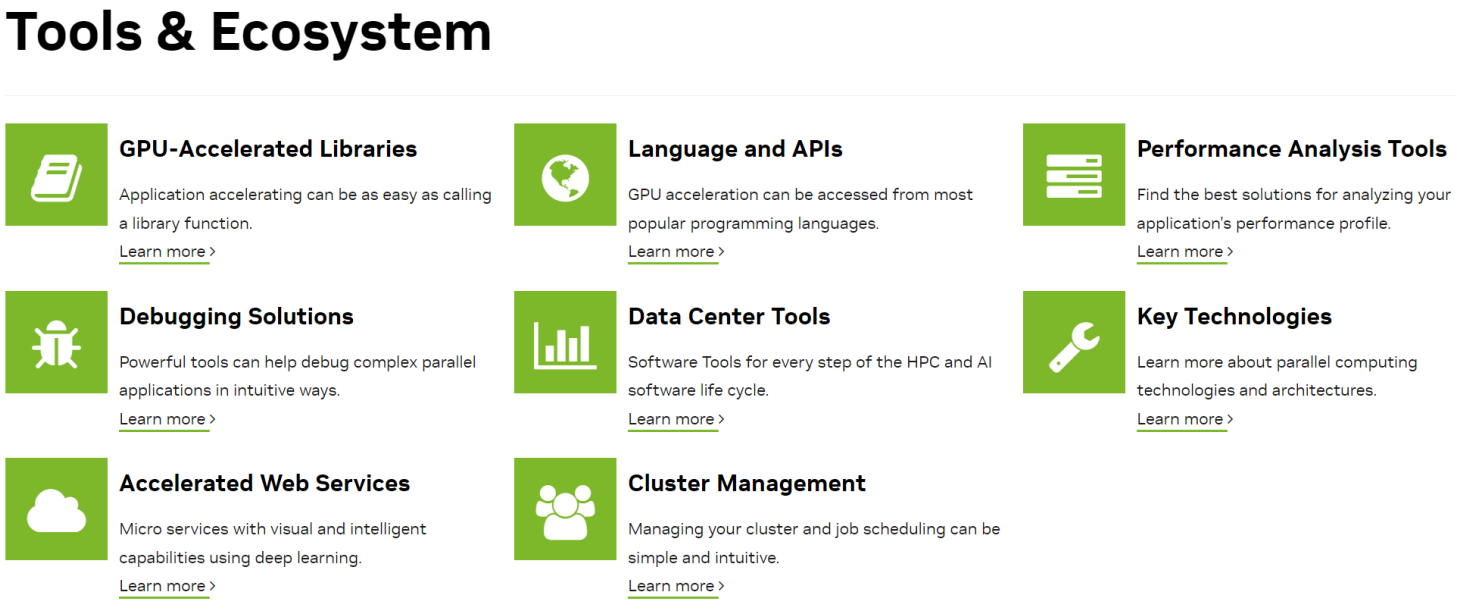
Ease of use

Even though we see the **cost of hardware (the CapEx) increasing significantly** (at the moment), **total cost of ownership should also consider the operating costs** (OpEx) and any upfront costs in adopting GPU technologies.

Hopefully during **this week you will have developed a feel for how easy it will be for you to gain GPU acceleration in your projects** / codes.

NVIDIAs rich software ecosystem makes it *relatively* easy to adopt GPU technology into your codes.

This helps to minimise development time needed to port an existing project to use GPUs.



So, we should buy NVIDIA, right?

It’s getting harder, why?

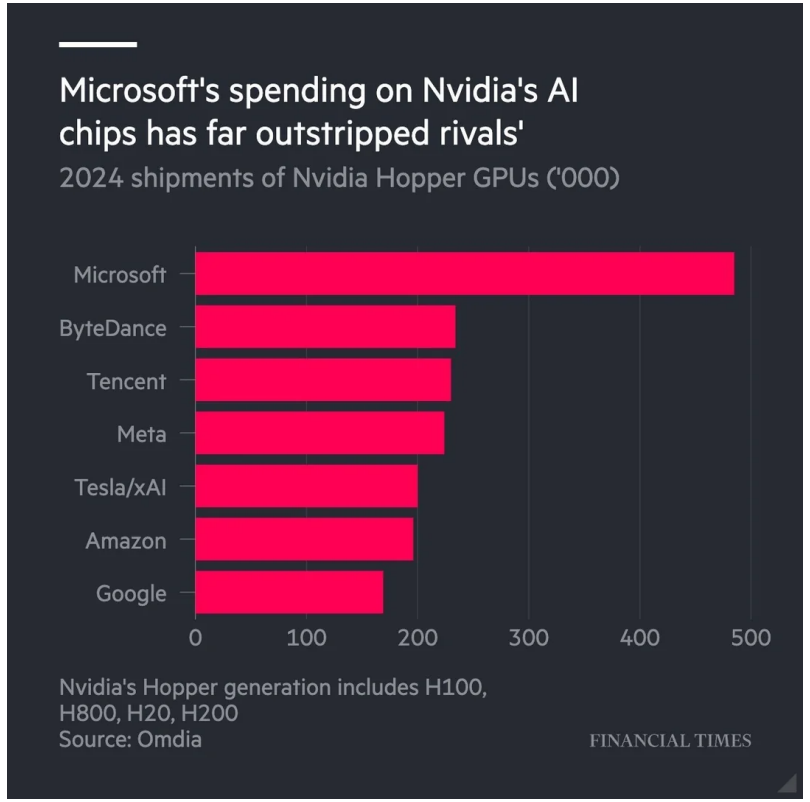
To begin with, the interest in training LLMs using NVIDIA GPUs generated a supply and demand issue.

Now we have useful LLMs even more GPUs are needed to serve them to us.

In 2024 Meta aimed to buy 600,000 H100 GPUs:
https://www.instagram.com/reel/C2QARHJR1sZ/?utm_source=ig_embed&ig_rid=74e68412-c1ea-4b5f-8e39-d7c0f41d16be
Reports indicate that they acquired 225,000.

Last year Microsoft acquired nearly 500,000 H100 GPUs…
https://www.windowscentral.com/microsoft/microsoft-reportedly-acquired-the-most-nvidia-gpus-compared-to-its-rivals-including-google-and-meta-for-its-ai-projects-translating-to-485-000-chips-and-usd31-billion-in-expenditure

This month, July 2025, NVIDIA became only company in the world to have market cap greater than $4 trillion! (June 2024 they hit $3 trillion).



Look at the world's largest machines, past and future trends - June 2021

The top500 lists the worlds fastest computers. A new list is produced in June and November each year.

Looking back, just 3 years ago, three out of the five fastest computers in the world were powered by NVIDIA GPUS.

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
2	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
3	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94.64	125.71	7,438
4	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93.01	125.44	15,371
5	**Perlmutter** - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE DOE/SC/LBNL/NERSC United States	706,304	64.59	89.79	2,528

Look at the world's largest machines, past and future trends - June 2025

Four years later…

New machines, taking the top places in the top500, including the worlds first exaflop machine, are based on AMD, not NVIDIA.

Does this just reflect the trend of NVIDIA going all in on AI/ML?

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**El Capitan** - HPE Cray EX255a, AMD 4th Gen EPYC 24C 1.8GHz, AMD Instinct MI300A, Slingshot-11, TOSS, HPE DOE/NNSA/LLNL United States	11,039,616	1,742.00	2,746.38	29,581
2	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Cray OS, HPE DOE/SC/Oak Ridge National Laboratory United States	9,066,176	1,353.00	2,055.72	24,607
3	**Aurora** - HPE Cray EX - Intel Exascale Compute Blade, Xeon CPU Max 9470 52C 2.4GHz, Intel Data Center GPU Max, Slingshot-11, Intel DOE/SC/Argonne National Laboratory United States	9,264,128	1,012.00	1,980.01	38,698
4	**JUPITER Booster** - BullSequana XH3000, GH Superchip 72C 3GHz, NVIDIA GH200 Superchip, Quad-Rail NVIDIA InfiniBand NDR200, RedHat Enterprise Linux, EVIDEN EuroHPC/FZJ Germany	4,801,344	793.40	930.00	13,088
5	**Eagle** - Microsoft NDv5, Xeon Platinum 8480C 48C 2GHz, NVIDIA H100, NVIDIA Infiniband NDR, Microsoft Azure Microsoft Azure United States	2,073,600	561.20	846.84	

El Capitan

Hosted at Lawrence Livermore National Laboratory, El Capitan is currently the world's fastest supercomputer and one of three ExaFLOP machines in the world. Meaning it can achieve over 10^{18} floating point operations per second.

Costing approximately $600M, it was delivered in partnership with AMD and HPE (Cray) and uses the Cray EX Shasta architecture.

https://asc.llnl.gov/exascale/el-capitan

The road to El Capitan: rtec.pdf
Early science: 2024.12.1.pdf

El Capitan – Fun facts

El Capitan can (theoretically) achieve $2.7\text{x}10^{18}$Floating Point Operations Per Second.

To put that figure into context…

If you travelled back in time 2.7 quintillion seconds, you'd arrive over 70 billion years before the Big Bang.

Or if every person on the planet worked on a single addition and multiply, every second of every day, around the clock, it would take 8 years to do what El Capitan can do in a single second.

Credit: Lawrence Livermore National Laboratory

https://hpc.llnl.gov/sites/default/files/El-Cap-Fun-Facts.pdf

El Capitan – Compute configuration

The HPE Cray EX rack is a **liquid cooled and blade-based system**. This allows for very high density in a small footprint.

The EX4000 cabinet is a sealed unit that **uses closed-loop cooling to ensure minimal heat is exhausted into the data centre**.

Both Atos and Lenovo have similar technology.

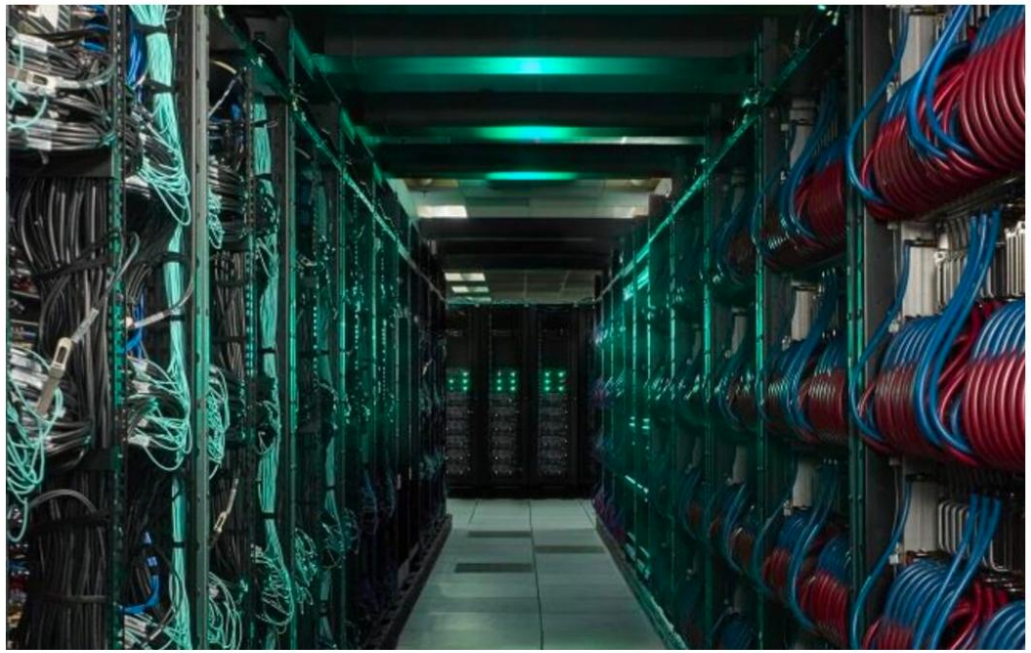
All solutions use direct attached liquid cooled cold plates to remove heat from compute components.

This allows densities of up to 250KW per rack.

https://www.hpe.com/psnow/doc/a00094635enw

El Capitan – Specs

- 11,136 nodes in liquid-cooled Cray EX racks.
- 44,544 AMD Instinct MI300A GPUs.
- APU so CPU and GPU share an internal on-chip coherent interconnect.
- HPE Slingshot interconnect.
- Each rack holds 64 blades; each blade has two nodes.
- A node consists of 4x MI300A, so 4x AMD "Genoa" CPUs (96 cores per node) and 4x AMD MI300 GPUs, with a total of 512 GB RAM per node.
- Peak power 35 Megawatts.



LLNL El Capitan Aisle

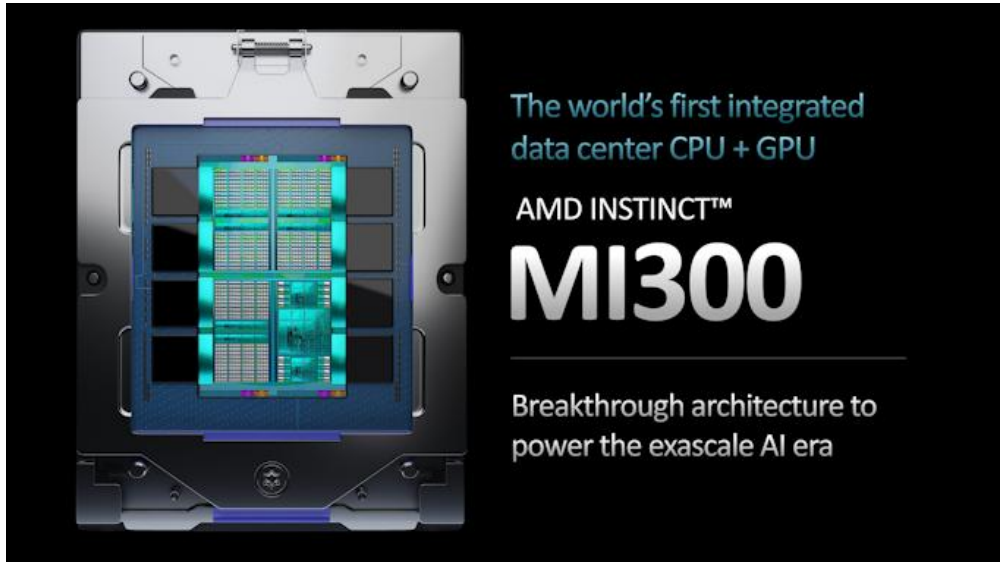
https://hpc.llnl.gov/documentation/user-guides/using-el-capitan-systems/hardware-overview#ats4-comparison

AMD as a solution? Hardware

We see from the change in the top500, **AMD GPUs are now gaining traction in HPC and scientific computing**.

This is because when the **total cost of ownership was considered for both El Capitan and Frontier**, it was decided that **AMD GPUs would be more cost effective.**

DoE spent approximately 1/3 of their budget on hardware, the other 2/3 was on software porting and running costs.



The new MI300A used in El Capitan.

https://www.amd.com/en/products/specifications/professional-graphics/4476,19496

AMD as a solution? Cost vs performance

Currently, for a reasonable server expect to pay:

8x H200 server £250K+
8x MI300X server £200K

AMD claims the MI325X is up to 2.4x faster than H200 for a range of representative scientific codes and up to 1.3x for training.

The MI300 is designed to rival the H100, MI325 to rival H200.

AMDs roadmap has MI325, MI350 and MI400 out to 2026

YMMV…

Our AMD experience (Ask Yishun)

Per single GPU (for model training) MI300 is competitive with H100.

For HPC, things look far better (and in line with AMD claims), the MI300 was built for HPC workloads after all…

Once you try to train on multiple GPUs things look bad.

The AMD software ecosystem is far behind Nvidia.

The hardware has been buggy.

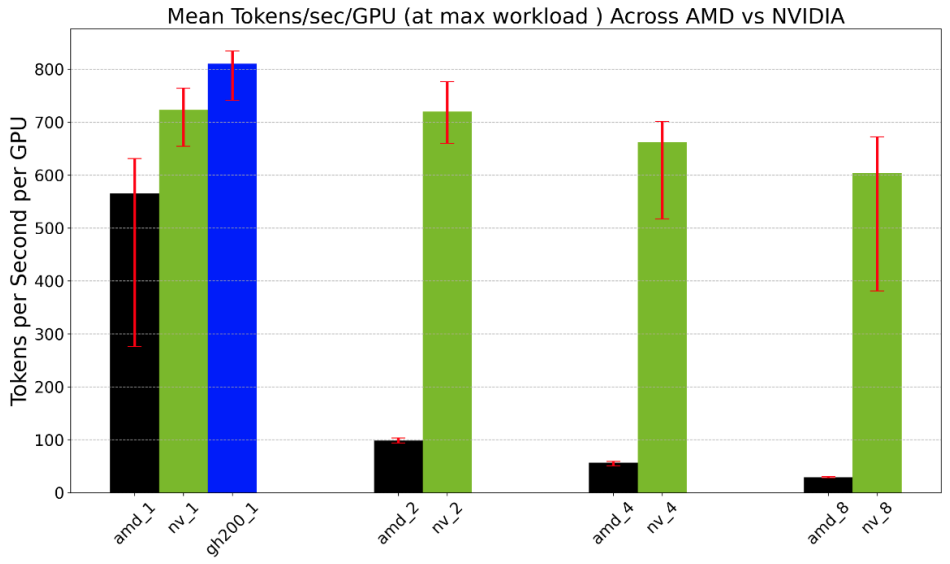


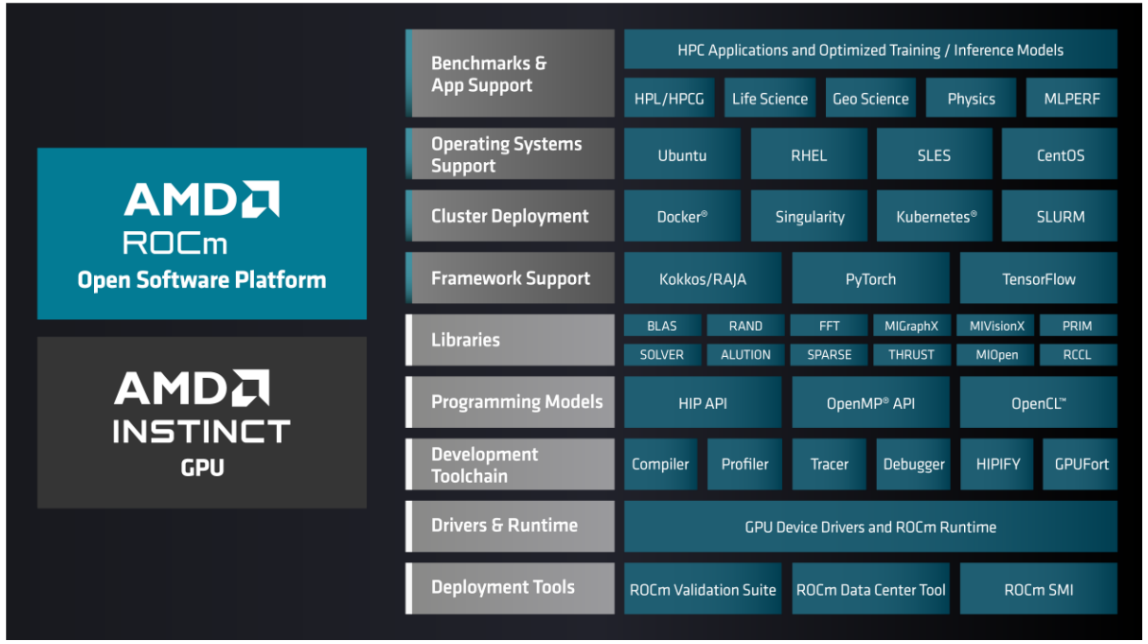
Figure 2: Mean tokens per second per GPU for AMD and NVIDIA under maximum workload when fine-tuning LLaMA-3.2-Vision-Instruct. "_n" denotes the number of GPUs used within a single node.

Software - Radeon Open Compute Platform (ROCm)

AMD ROCm

One of the reasons **NVIDIA has been so dominant in the HPC space is its software ecosystem** and its ability to run on basic gaming cards (GeForce), to prosumer (Titan) to high end data centre cards (Tesla).

AMD now has a similar, growing (and in some parts rather familiar) software ecosystem called HIP/ROCm.



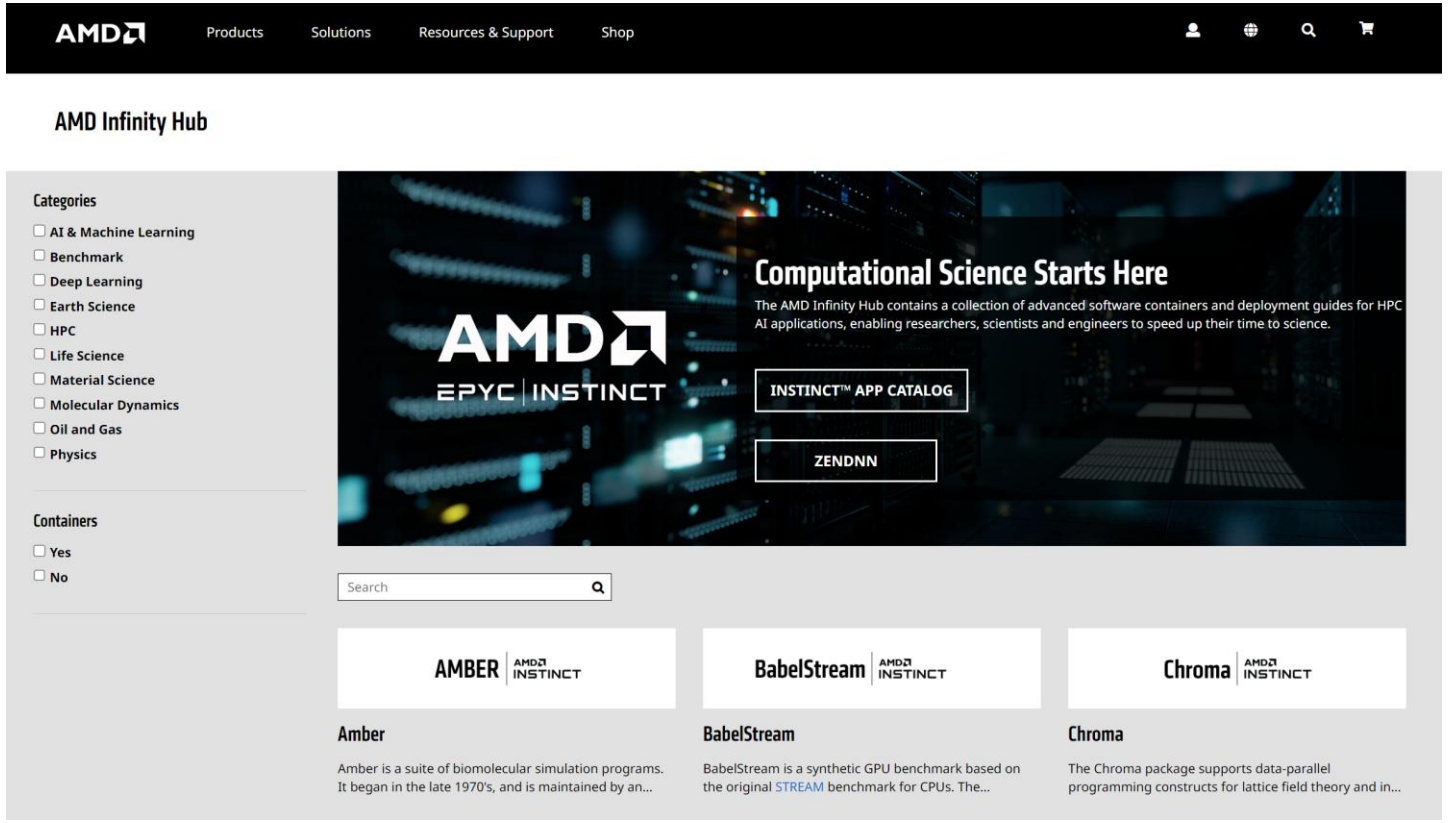
https://rocm.docs.amd.com/en/latest/how-to/rocm-for-hpc/index.html

Software - Infinity hub

Have a growing number of leading packages optimised of Instinct. For example:

- Amber
- Gromacs
- Chroma
- QUDA
- CP2K
- PyTorch

Largely driven by DoE contracts.



https://www.amd.com/en/developer/resources/infinity-hub.html

https://www.amd.com/content/dam/amd/en/documents/resources/gpu-accelerated-applications-catalog.pdf

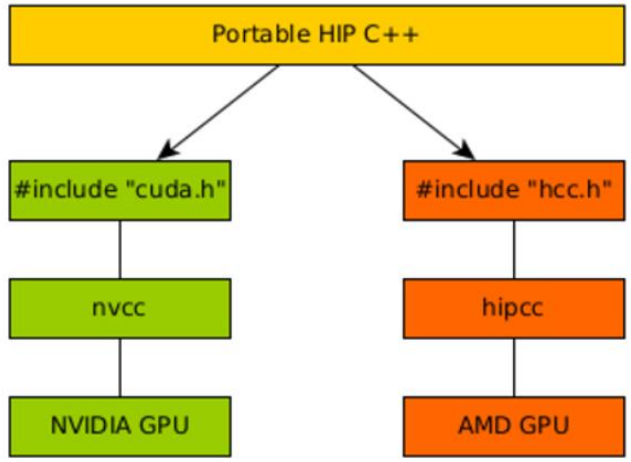
Heterogeneous-Compute Interface for Portability (HIP)

HIP is AMDs "version" of CUDA, it's a Kernel Language that looks, in many parts, similar to CUDA.

It **aims** to allow you **to create applications that are portable**, so when you write in HIP, your code will be able to run not only AMD GPUs, but NVIDIA also (at least that's the aim, just like OpenCL…).

AMD Claim:

- HIP has little (or no) performance impact compared to coding directly in CUDA.
- HIP allows coding in a single-source C/C++ programming language.
- The HIPIFY tools automatically convert most source from CUDA to HIP.
- Developers can specialize for the platform (CUDA or AMD) to tune for performance or handle tricky cases.



https://github.com/ROCm-Developer-Tools/HIP

https://www.youtube.com/watch?v=hSwgh-BXx3E

https://www.lumi-supercomputer.eu/preparing-codes-for-lumi-converting-cuda-applications-to-hip/

Heterogeneous-Compute Interface for Portability (HIP)

Let's look at some HIP (the main() code)…

```
…
char* inputBuffer;
char* outputBuffer;

hipMalloc((void**)&inputBuffer, (strlength + 1) * sizeof(char));
hipMalloc((void**)&outputBuffer, (strlength + 1) * sizeof(char));

hipMemcpy(inputBuffer, input, (strlength + 1) * sizeof(char), hipMemcpyHostToDevice);

hipLaunchKernelGGL(helloworld,
          dim3(1),
          dim3(strlength),
          0, 0,
          inputBuffer ,outputBuffer );

hipMemcpy(output, outputBuffer,(strlength + 1) * sizeof(char), hipMemcpyDeviceToHost);

hipFree(inputBuffer);
hipFree(outputBuffer);
…
```

https://github.com/ROCm-Developer-Tools/HIP-Examples/blob/master/HIP-Examples-Applications/HelloWorld/HelloWorld.cpp

Heterogeneous-Compute Interface for Portability (HIP)

Let's look at some HIP (the kernel code)…

```
__global__ void helloworld(char* in, char* out)
{
          int num = hipThreadIdx_x + hipBlockDim_x * hipBlockIdx_x;
          out[num] = in[num] + 1;
}
```

It all looks rather familiar, almost like someone has done a global "find `cuda` replace with `hip`"…

https://github.com/ROCm-Developer-Tools/HIP-Examples/blob/master/HIP-Examples-Applications/HelloWorld/HelloWorld.cpp

HIPIFY

HIPIFY is a set of scripts that will (try) to translate your CUDA source code into HIP auto~~matically~~magically for you.

The scripts are based on perl and clang.

Jack tried to take our AstroAccelerate code base (admittedly it is large and in parts quite complicated) and use HIPIFY to generate an AMD executable code.

He wasn't able (through no fault of his own!!).

When Jack emailed support he was pointed to the git repo and asked to raise an issue.

So some work to do before this is truly automagical.

Supported CUDA APIs

- Runtime API
- Driver API
- cuComplex API
- Device API
- RTC API
- cuBLAS
- cuRAND
- cuDNN
- cuFFT
- cuSPARSE
- CUB

https://github.com/ROCm-Developer-Tools/HIPIFY

What about Intel?

Whilst Intel didn't invent the idea of a coprocessor, they did popularise it with the x87, dedicated to accelerating and adding functionality for floating point computations.

Since then Intel have had several failed attempts at entering the accelerator computing market.

- i860
- Larabee
- Xeon Phi (MIC)

In 2018 Intel revived the idea of a GPGPU accelerator and this has now become the Intel Xe (eXascaler for everyone).

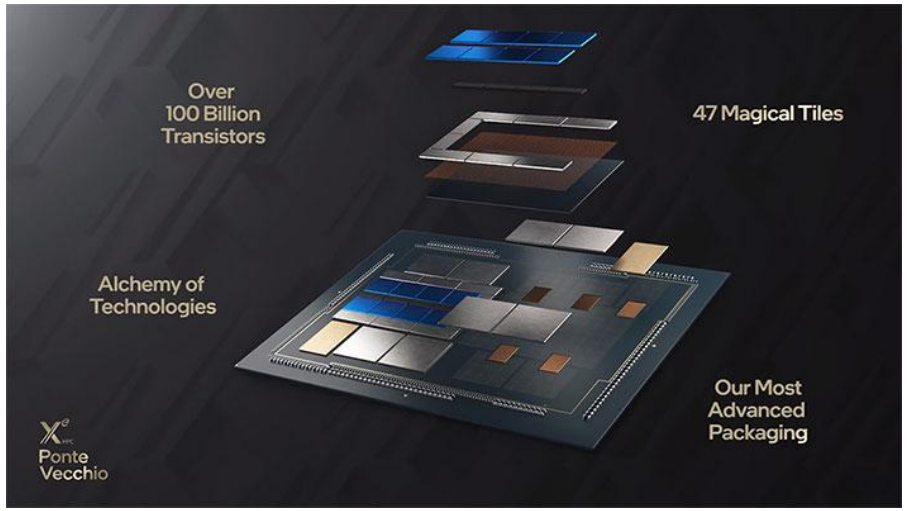
Intel Xe-HPC - Ponte Vecchio GPU

The **Ponte Vecchio GPU is used in Aurora.**

Intel specs are: 45 TFLOP/s, 5 TB/s bandwidth and 2 TB/s connectivity (I think this is Xe Link).

Tests show that **for some applications this reaches about 80% of the performance of an A100**.

We should keep in mind though, that the A100 is four years old, NVIDIAs current flagship GPU is the H100, soon to be replaced by Blackwell.



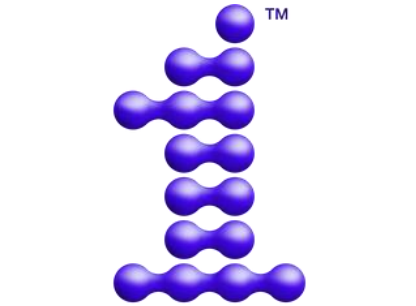
OneAPI

OneAPI is Intel's answer to HIP. It is an open standard and aims to deliver a single unified API that **can be used across all of its products from FPGAs to GPUs to CPUs.**

It aims to go further that just Intel products. **OneAPI has some functionality for both NVIDIA and AMD GPUs (via Codeplay plugins).**

This work is part of Intel’s plan to make oneAPI the preferred alternative for heterogeneous, parallel programming.

One ring to rule them all…



https://www.intel.com/content/www/us/en/developer/tools/oneapi/code-samples.html#gs.2twqvx

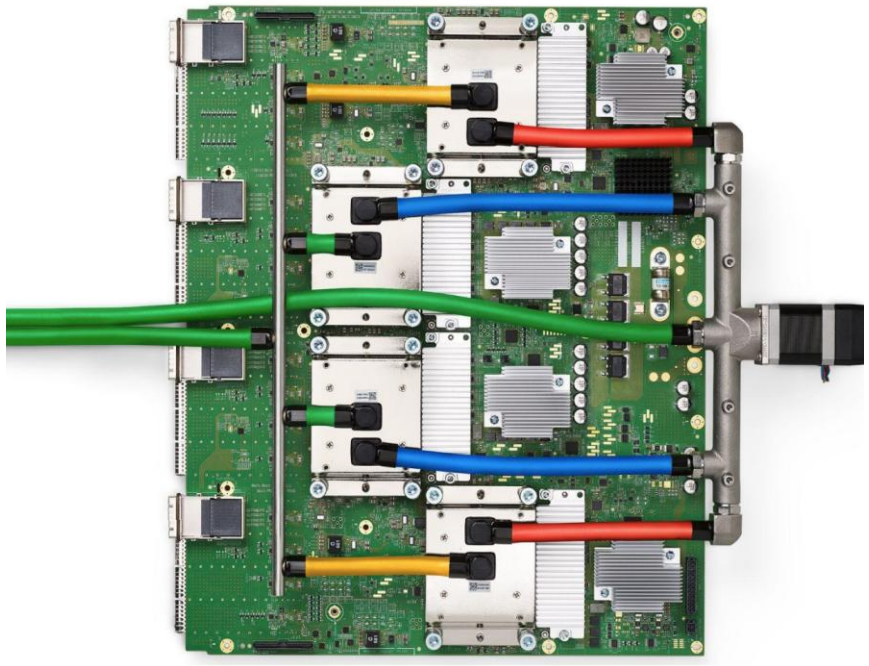
https://www.eejournal.com/article/intels-latest-version-of-oneapi-takes-advantage-of-new-intel-xeon-improvements-supports-amd-and-nvidia/?cid=org&source=linkedin_organic_cmd&campid=ww_23_oneapi&content=art-idz_&linkId=100000207031089

Google – Tensor Processing Unit

Google offer the TPU. This only suited to AI/ML training, but again, it's not general purpose in the way a CPU or GPU is.

TPUs can be accessed through google cloud. **To use them you write your code in TensorFlow / Torch or JAX and it's compiled to use TPU acceleration**.

TPUs are application specific integrated circuits (ASICs) that focus on the acceleration of matrix operations (performing similar operations to NVIDIAs tensor cores).



https://arxiv.org/ftp/arxiv/papers/2304/2304.01433.pdf

https://cloud.google.com/tpu

https://cloud.google.com/tpu/docs/system-architecture-tpu-vm

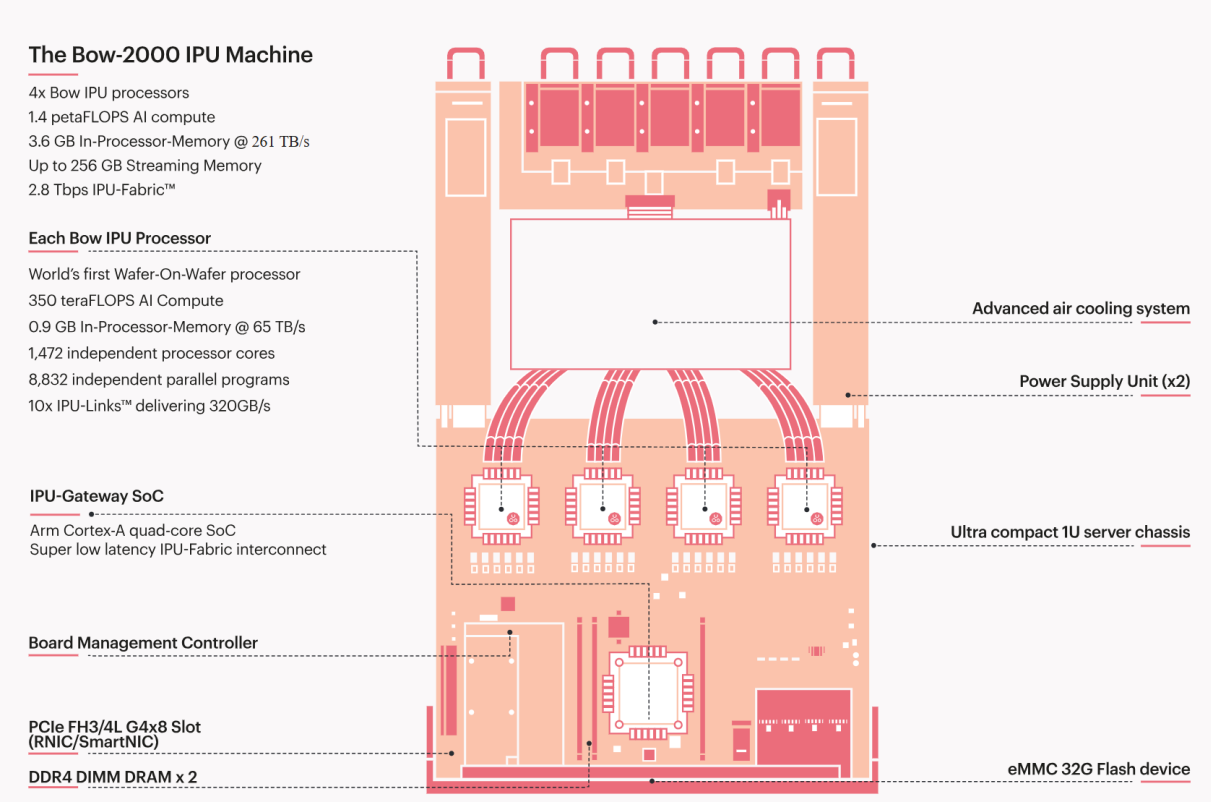
Graphcore

Graphcore produce the Colossus Intelligent Processing Unit.

The Mark 2 IPU was released in 2020. The system design is aimed at sparse problems and has a memory system that is ideal for large AI models.

The Mark 3 IPU is still in development, aiming to double the performance of the Mark 2 IPU. However, “BoW” was announced in 2022 that improved the performance of GC200 (Mk2) to 350 TFLOPs of “AI Compute”.

For **certain application spaces** graphcore products are more than competitive with NVIDIA GPUs.



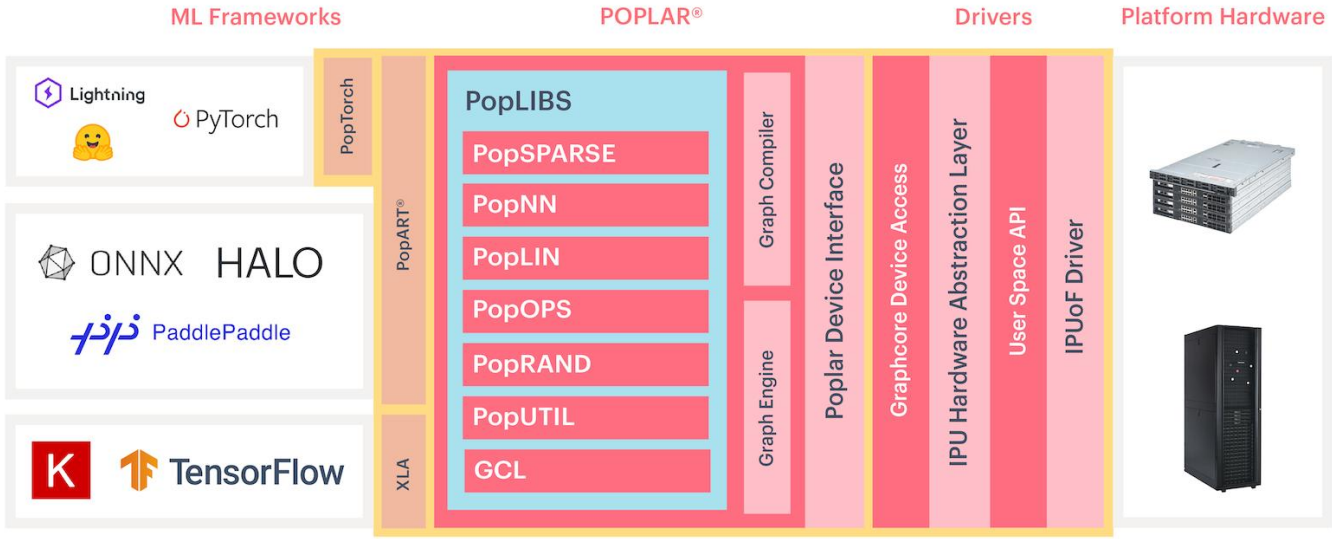
Graphcore - software

Graphcore have a **software stack called Poplar.**

This will take **code written using TensorFlow, PyTorch and Keras and generate code to run on the IPU.**

But be aware – the IPUs cannot do anything else. They are designed specifically for AI/ML training and work really well in areas such as NLP where models need large memory capacity close to the compute.

Softbank have agreed to acquire Graphcore for around $500 million. The deal is under review…



https://docs.graphcore.ai/projects/ipu-overview/en/latest/programming_tools.html

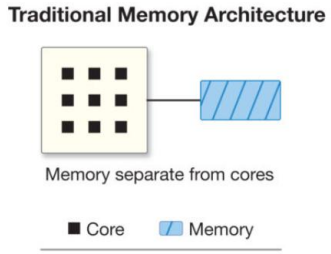
Cerebras

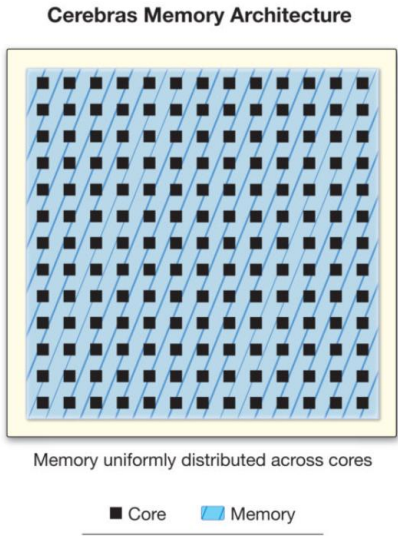
Cerebras produce **wafer level processors**. - Quite amazing.

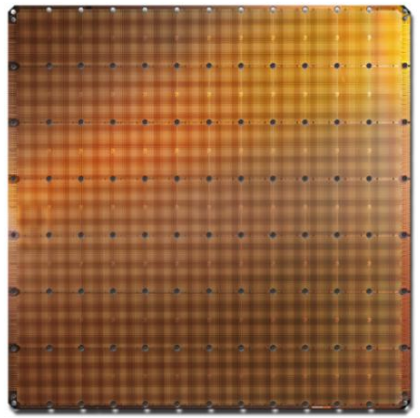
In terms of **software**, Cerebras has a similar approach to Graphcore. It has the **Csoft environment**. It too integrates Torch and TensorFlow to produce code that runs on the WSE-3 platform.

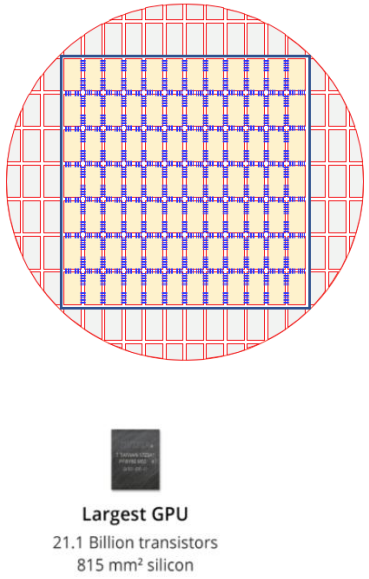
It also has a SDK to allow developers to write custom kernels.

I haven't seen a good comparison to other technology as yet.







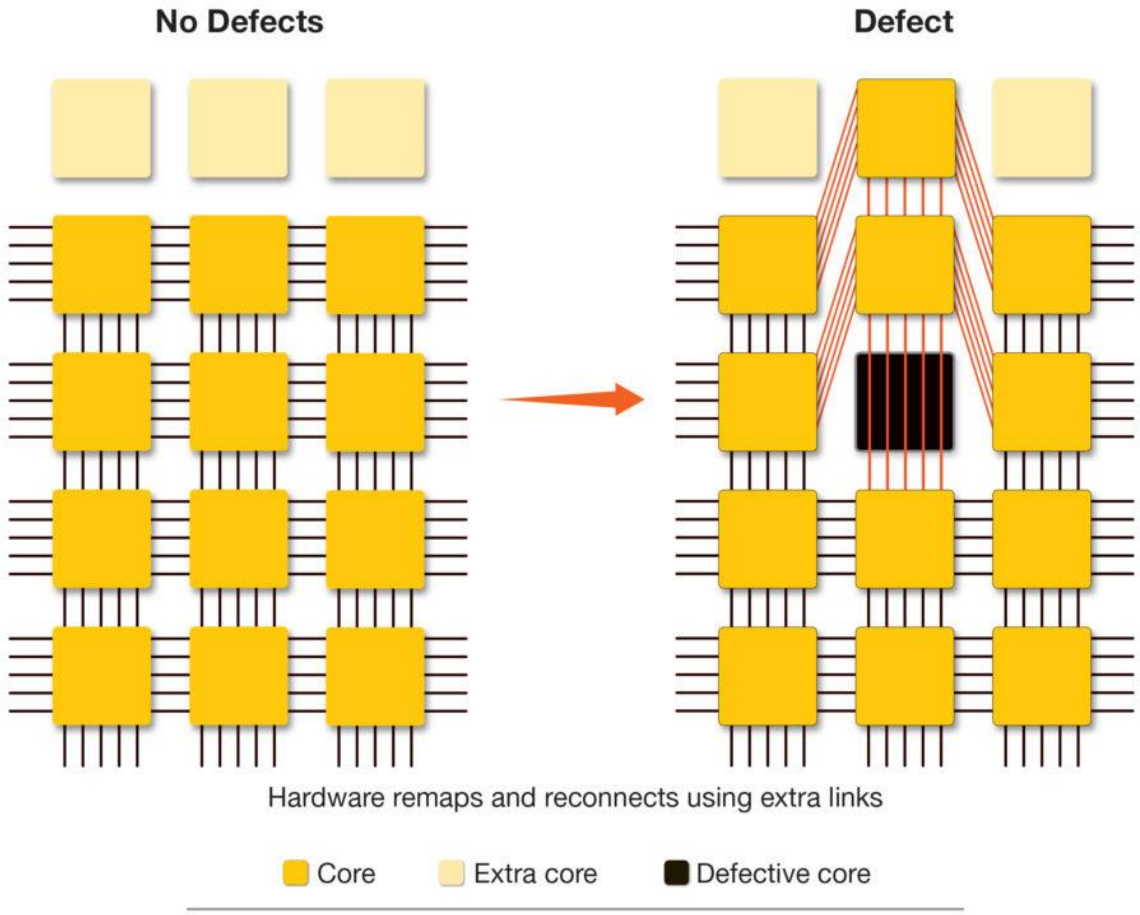


Cerebras

From a 12 inch wafer Cerebras produce a single processor (NVIDIA would get about 60 H100).

For those interested –
TMSC can now produce 50-80K wafers per month (capacity has doubled year on year for the last few years)*

To ensure high yield, defective cores are identified the time of manufacturing and then the interconnect between cores is configured to avoid defective cores. Then added for that chip.



*CoWoS, TSMC has capacity to produce ~15M wafers per year.

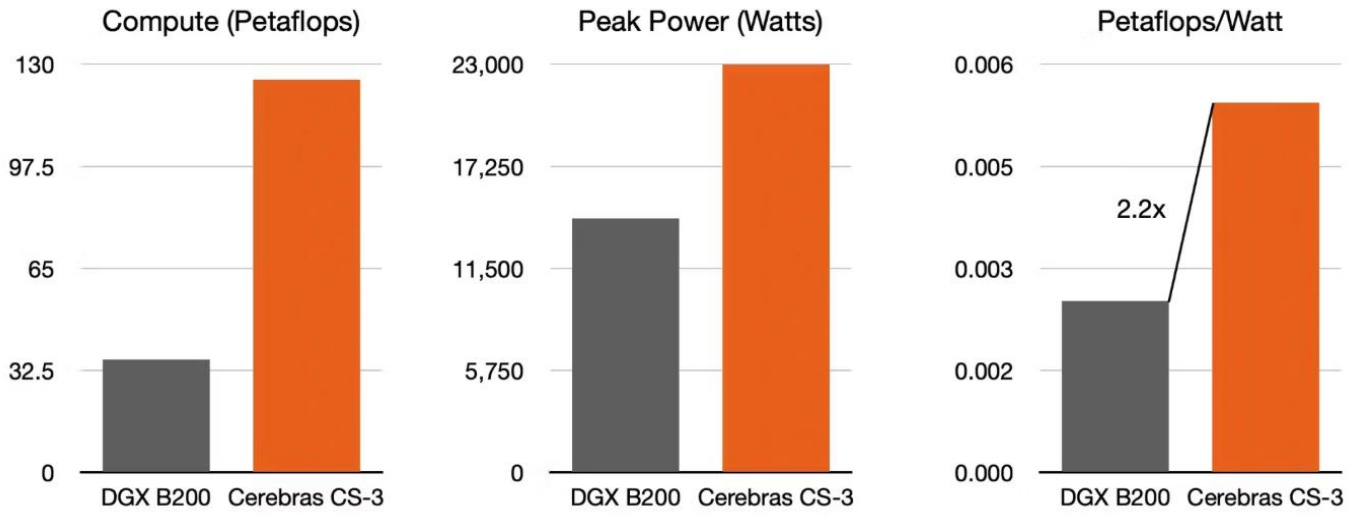
https://www.cerebras.net/blog/wafer-scale-processors-the-time-has-come/#:~:text=A%20wafer%20is%20a%20circle,called%20scribe%20lines%20between%20them.

Cerebras uptake

Cerebras hardware seems to be gaining traction, they have supplied many large organisations (Hugging face, Perplexity, Minstrel, DoE, GSK…)

The Condor Galaxy 3 system delivers 8 exaFLOPs of compute using 64 CS-3 systems.

Access to Cerebras Cloud here:
https://www.cerebras.net/product-cloud/



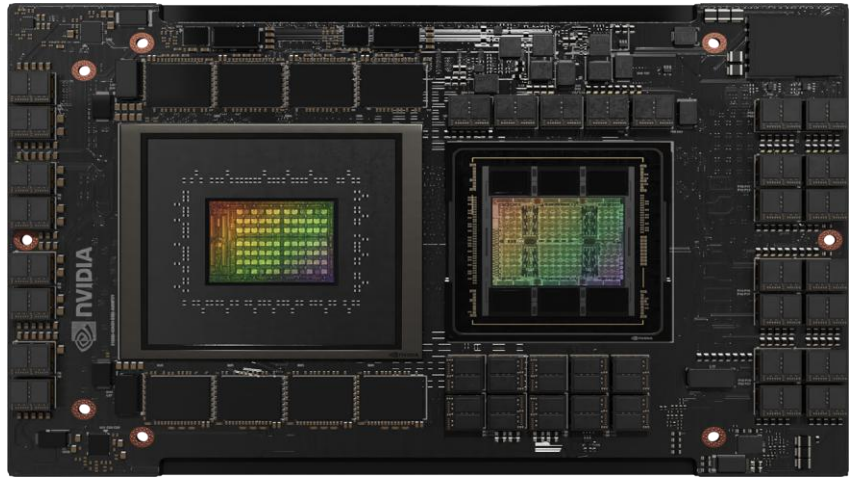
https://www.cerebras.ai/blog/cerebras-cs-3-vs-nvidia-b200-2024-ai-accelerators-compared

NVIDIA – Grace-Hopper

Grace-Hopper was NVIDIAs answer to the likes of Cerebras and Graphcore. The “Superchip” combines a Grace CPU and a Hopper GPU using NVLink C2C to deliver a CPU+GPU coherent memory model. The fruition of project Denver begun by NVIDIA in (Circa) 2014.

This kind of design will be crucial in progressing exascale computing in the years to come.

Whitepaper:
https://resources.nvidia.com/en-us-grace-cpu/nvidia-grace-hopper



NVIDIA – Grace-Blackwell

NVIDIA Grace + Blackwell:

- 1x Grace CPU + 2x Blackwell GPUs
- 72x Arm Neoverse V2 cores (4×128-bit SIMD units per core).
- Up to 480 GB of LPDDR5X memory (512 GB/s of memory bandwidth).
- Up to 64x PCIe Gen5 lanes (Gen6 for GB300 superchip).
- NVIDIA NVLink-C2C - Up to 1.8 TB/s total bandwidth.
- Unified address space.
- NVLink Switch System connects 72x NVIDIA Grace Blackwell Superchips using NVLink 5.
- Each GPU can address all HBM3E memory of all superchips in the network, **for up to 20TB of GPU addressable memory in NVL72**.

DGX GB300 AI Supercomputer

The DGX GB300 was announced at GTC in March 2025.

NVLink 5 Switch can connect up to 576 Grace-Blackwell GPUs (within a single NVLink domain).

Per Rack (NVL72):

- **Single 40 terabyte unified memory space**.
- **130 TB/s of aggregate bandwidth**.
- **1.4 exaFLOPS of FP4 AI performance**.

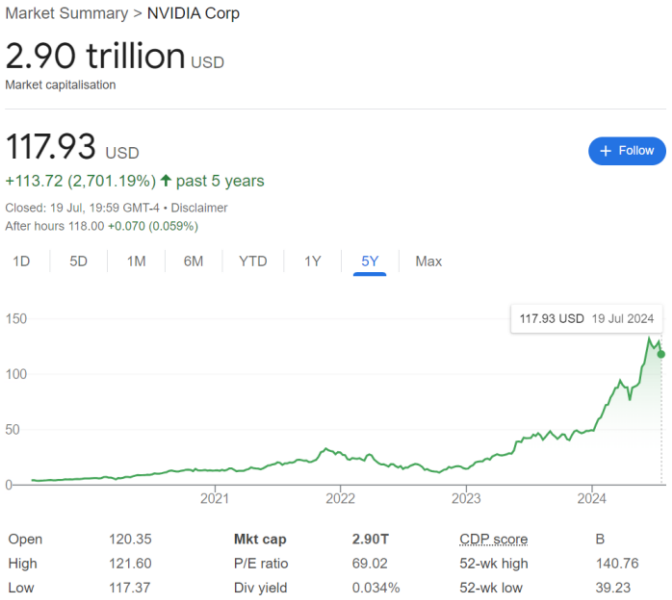
Whilst aimed at AI, this is a general purpose machine and so could be used for other areas of scientific computing.

The future?

NVIDIA's value continues to grow…

2023…

2024…

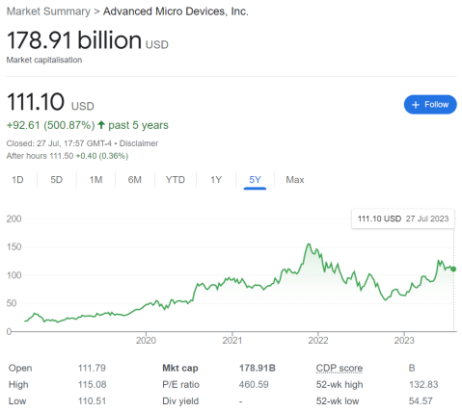


This year…

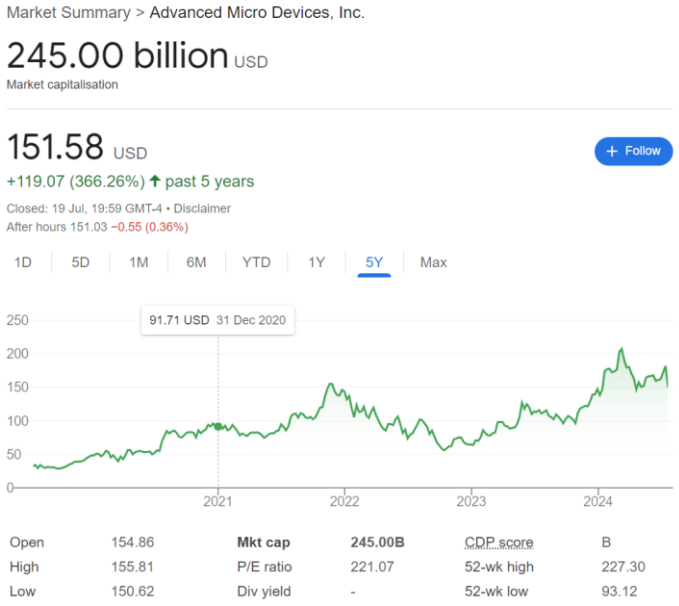
The future?

AMD, even with the success of El Capitan and Frontier, is still an order of magnitude behind.

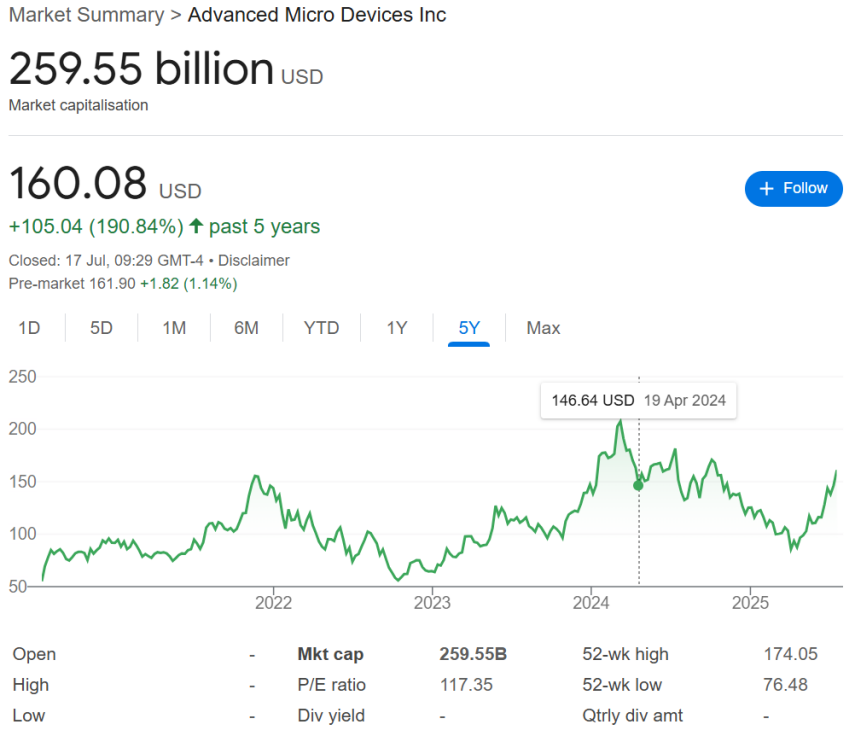
2023...



2024...



This year...



The future?

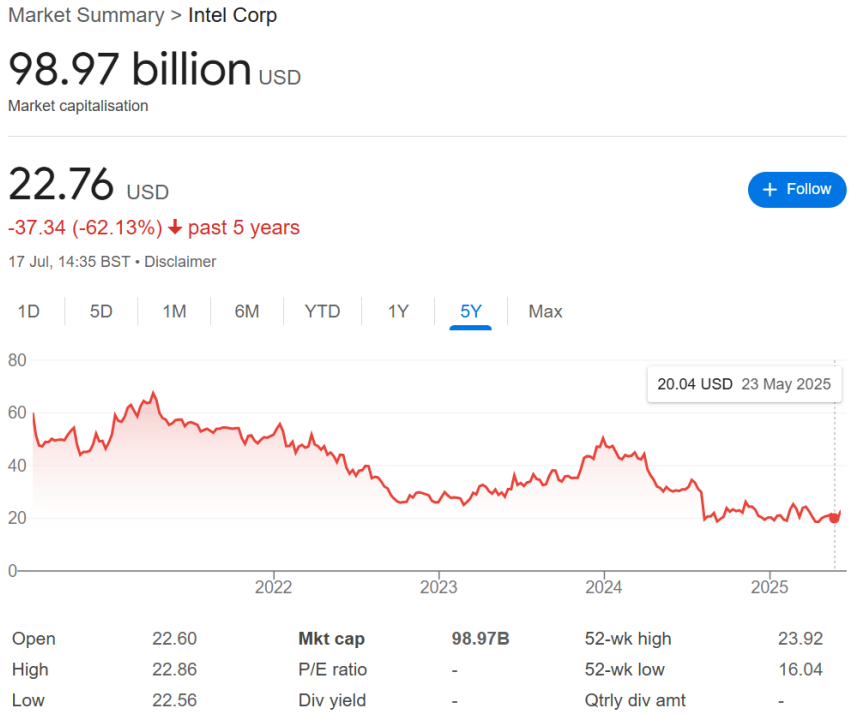
Intel, finally delivered Aurora, originally contracted to be completed by 2018.

Recently Intel's CEO said that they are now too far behind to deliver and AI chip.

So maybe they will go back to their roots and concentrate on HPC?

2024...

This year...



The future?

AMD and NVIDIA now seem to be fighting over the same market – AI.

The thirst for compute in this area has driven prices up significantly (~3x) in the last few years.

TMSC can only make so many chips and the cost of making silicone is the cost, that gets passed onto you, the consumer, no matter which technology you buy.

However, for scientific computing maybe AMD are pushing ahead…

ıore National Laboratory

Summary

This lecture has looked at some present alternatives to NVIDIA and CUDA. We've also taken a look at some up-coming technologies, both software and hardware that might we worth watching out for over the coming years.

Lots of what you have learnt this week is transferable!

Also keep an eye on Mike's computing webpage here:

https://people.maths.ox.ac.uk/gilesm/computing.html

Frontier – The worlds first Exaflop machine

Hosted at the Oak Ridge Leadership Computing Facility (OLCF) Tennessee, **Frontier is the worlds only ExaFLOP supercomputer.**

It was delivered in partnership with HPE (Cray) and was also the worlds "greenest" supercomputer when it became operational in May 2022.
https://www.top500.org/lists/green500/2022/06/

Great presentation by Bronson Messer
(Director of Science):

http://www.phys.utk.edu/archives/colloquium/2022/10-03-messer.pdf

By OLCF at ORNL - https://www.flickr.com/photos/olcf/52117623843/, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid=119231238

Frontier – Compute configuration

The HPE Cray EX rack is a **liquid cooled and blade-based system**. This allows for very high density in a small footprint.

The EX4000 cabinet is a sealed unit that **uses closed-loop cooling to ensure minimal heat is exhausted into the data centre**.

Both Atos and Lenovo have similar technology.

All solutions use direct attached liquid cooled cold plates to remove heat from compute components.

This allows densities of up to 250KW per rack.

https://www.hpe.com/psnow/doc/a00094635enw

Frontier – Specs

- 9472 AMD Epyc "Trento" 64 core 2 GHz CPUs.
- 37888 Radeon Instinct MI250X GPUs.
- HPE Slingshot interconnect.
- Frontier is liquid-cooled, allowing 5x the density of an air-cooled architecture.
- Each rack holds 64 blades, each blade has two nodes.
- A node consists of one CPU, 4x GPUs (each having 128GB memory), 512 GB RAM and 4TB of flash memory.
- 21 Megawatts

https://docs.olcf.ornl.gov/systems/frontier_user_guide.html