Lecture 10
Future Directions

Prof Wes Armour

wes.armour@eng.ox.ac.uk
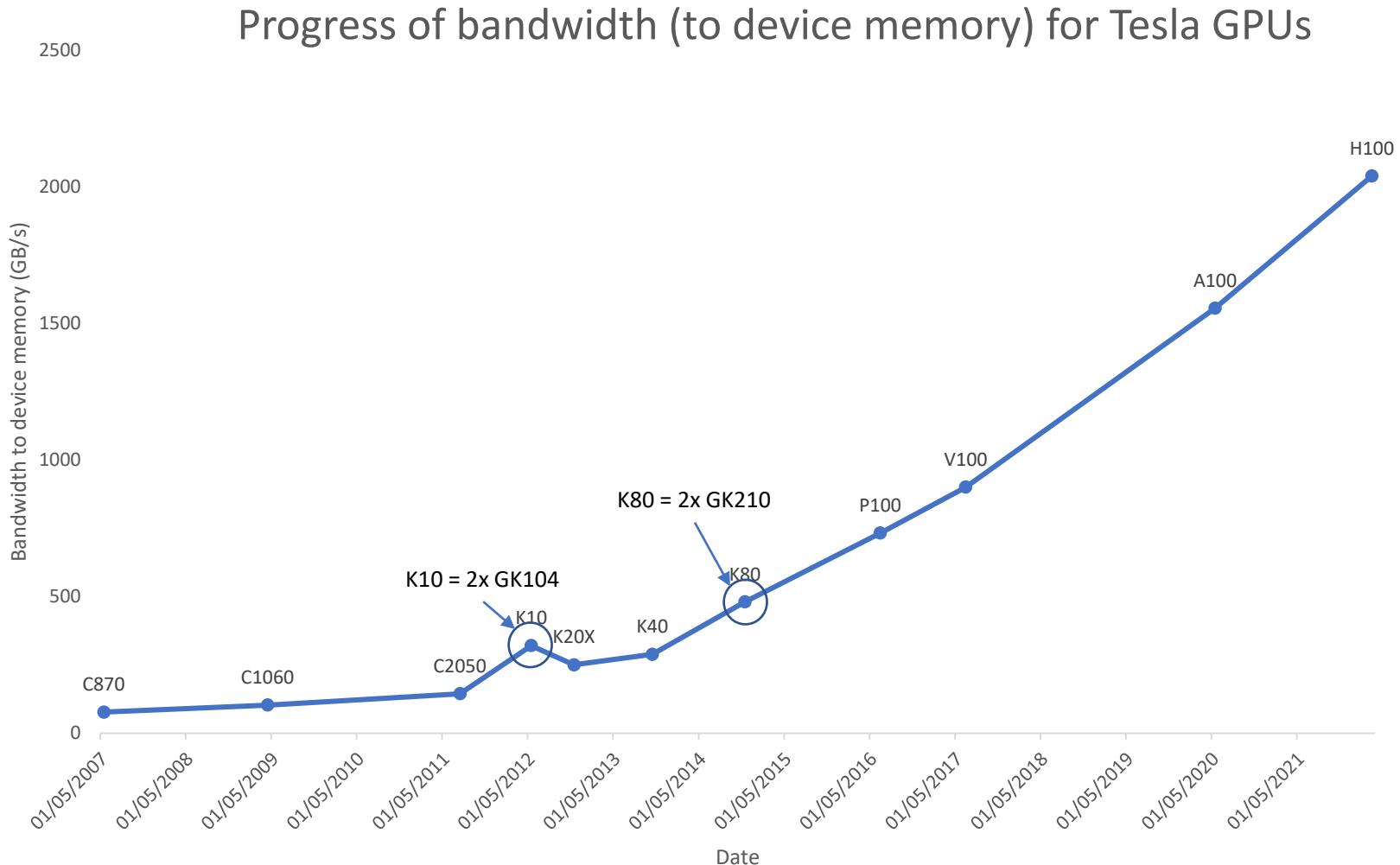
Oxford e-Research Centre

Department of Engineering Science

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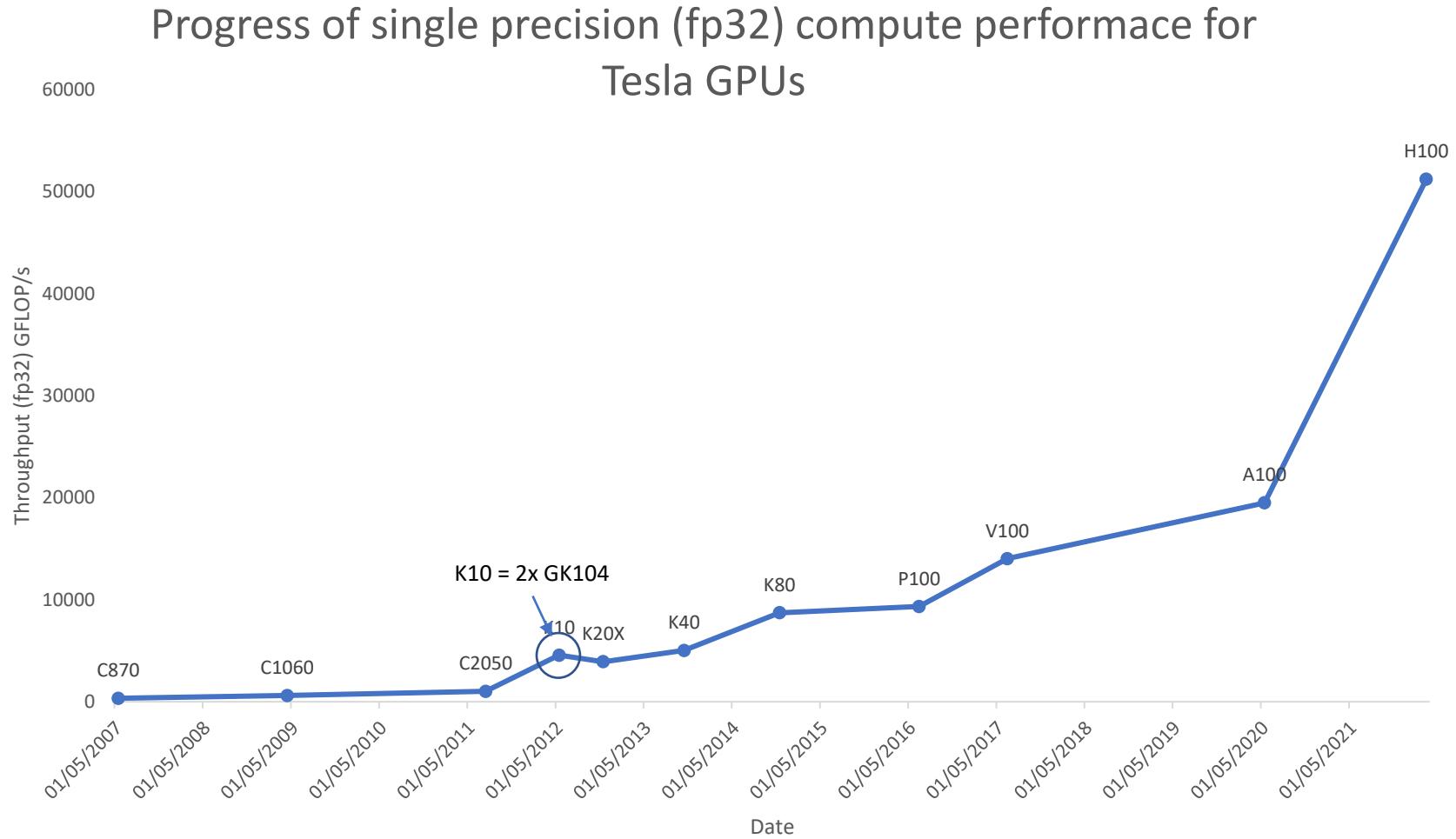
In this final lecture we will look at the current landscape of accelerated computing.

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

Compute

Progress of single precision (fp32) compute performace for Tesla GPUs

# Efficiency

GFLOP/s per Watt
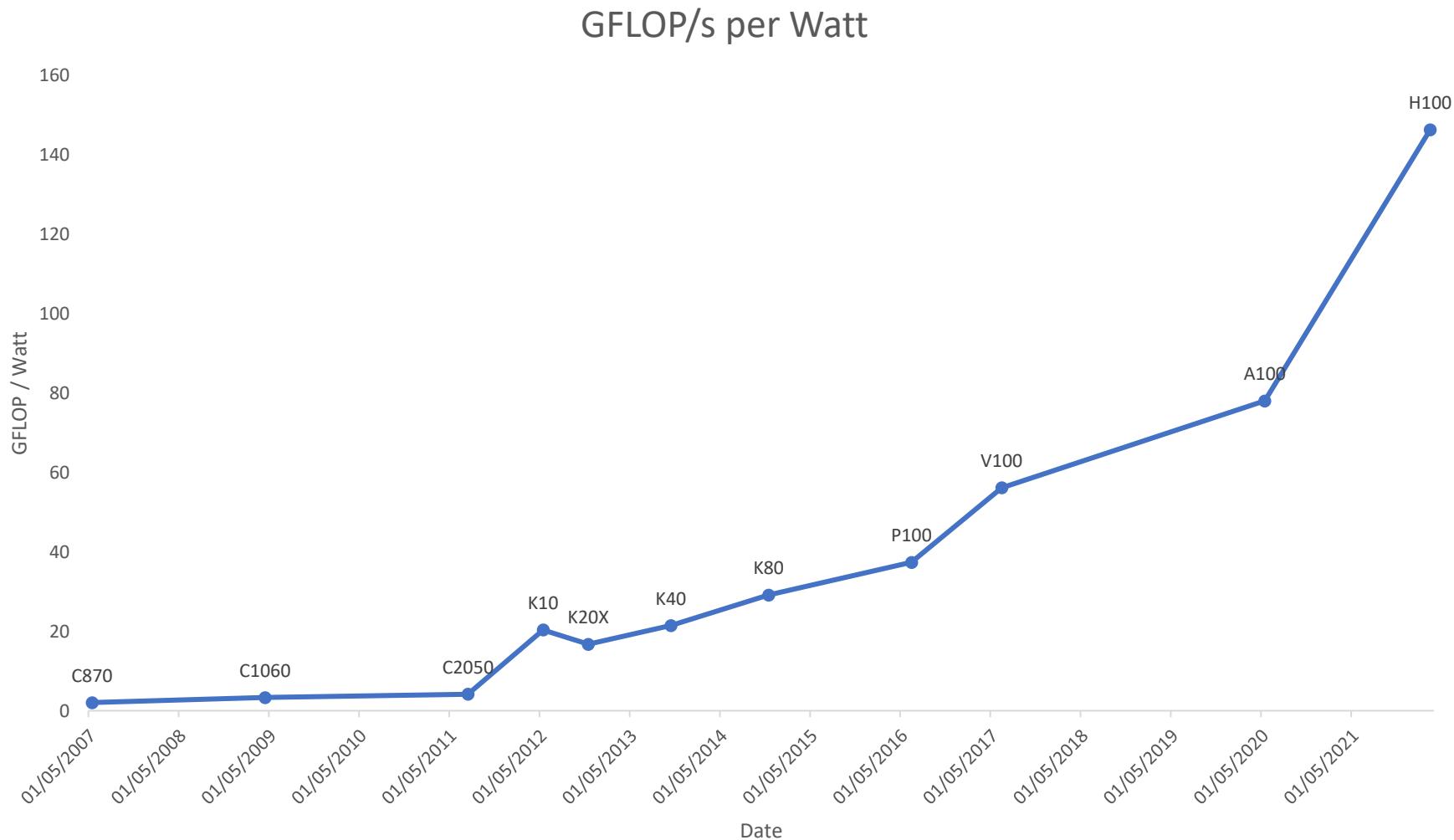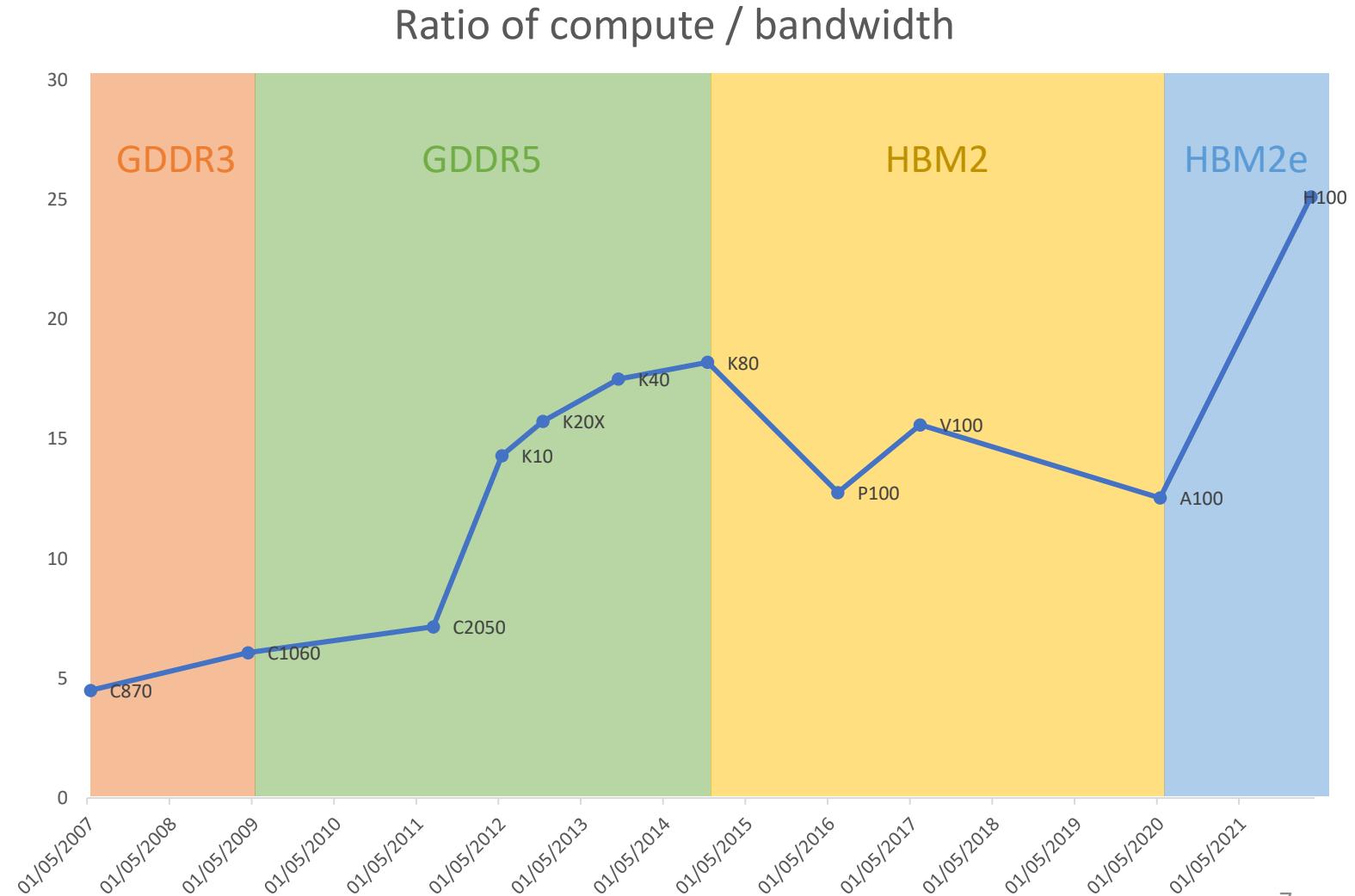
Ratio of compute / bandwidth

The ratio of compute / bandwidth often called **arithmetic intensity** or **operational intensity** (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.
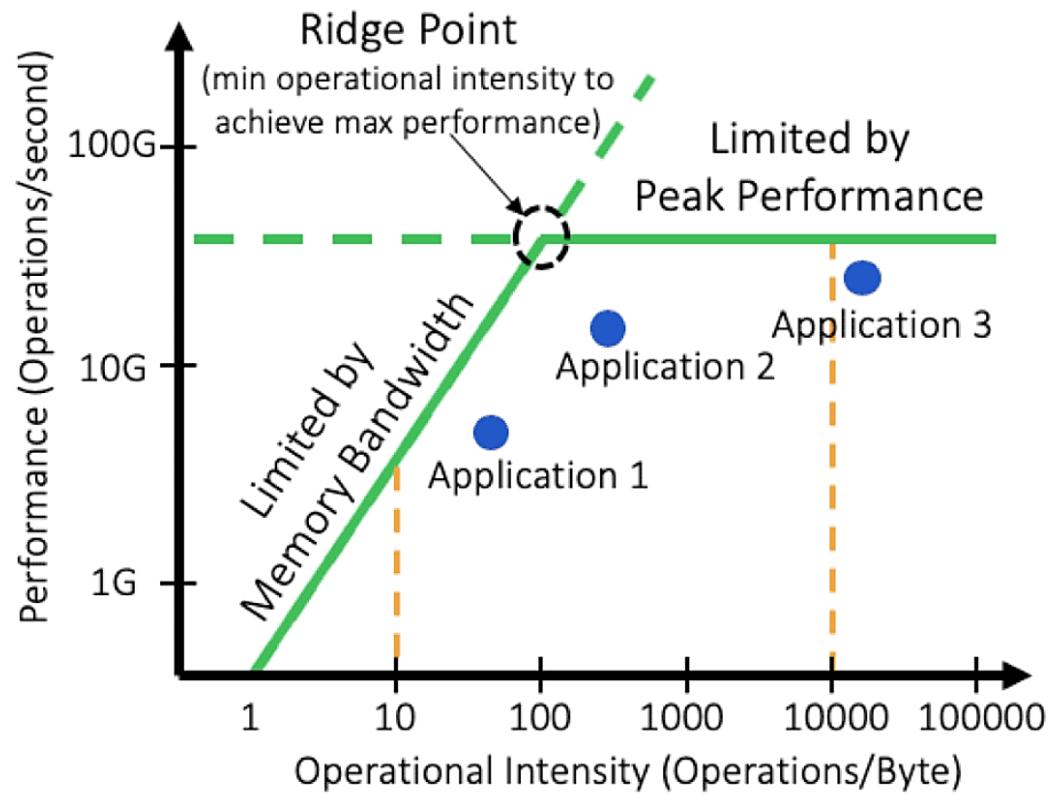
$$I = \frac{W}{Q} = \frac{Work}{Memory\ traffic} = \frac{FLOPs}{Byte}$$

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**.

Ratio of compute / bandwidth

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.

$$\frac{1}{16} \quad \begin{array}{|c|c|c|} \hline 1 & 2 & 1 \\ \hline 2 & 4 & 2 \\ \hline 1 & 2 & 1 \\ \hline \end{array}$$

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.

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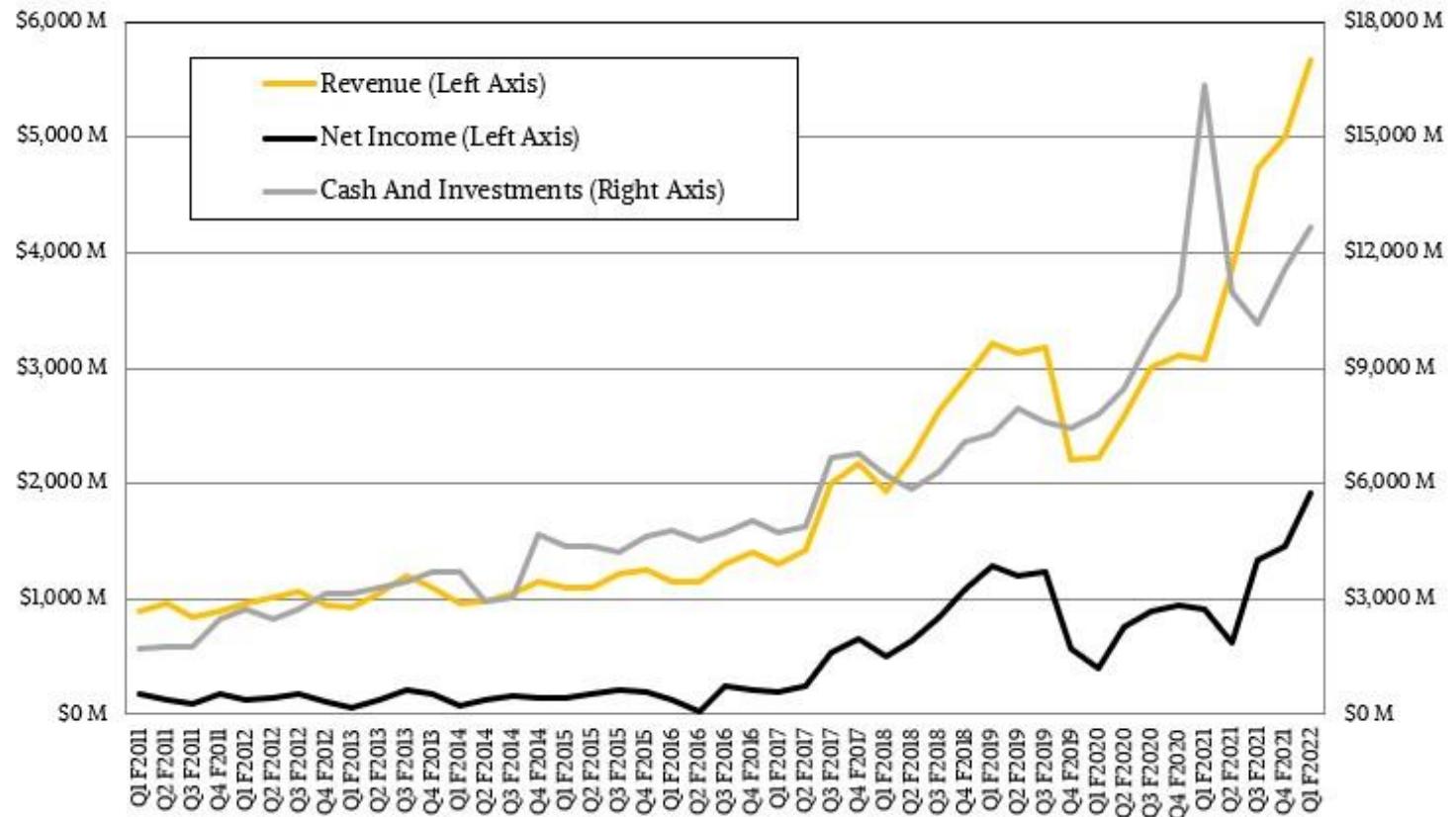
The growing cost of owning NVIDIA

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/

Multi-GPU computing

Multi-GPU computing exists at all scales, from cheaper workstations using PCIe, to more expensive Quadro / Titan products using fewer NVLink, to high-end NVIDIA DGX servers.

Single workstation / server:
- a big enclosure for good cooling!
- up to 4 high-end cards in 16x PCIe v4 slots – up to 16GB/s interconnect.
- 2x high-end CPUs.
- 2-3kW power consumption – not one for the office!
- £12K-£18K

NVIDIA DGX H100 Deep Learning server:
- 8 NVIDIA GH100 GPUs, each with 80GB HBM2.
- 2× 56-core Intel Xeons (Platinum 8480C 2.0 GHz).
- 2 TB RAM memory, 8x 3.84TB NVMe.
- 900GB/s NVlink interconnect between the GPUs.
- £???,??? (DGX A100 currently costs ~£350K, launch price was £200K)

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.

Tools & Ecosystem

GPU-Accelerated Libraries

Application accelerating can be as easy as calling a library function.

Learn more ›

Debugging Solutions

Powerful tools can help debug complex parallel applications in intuitive ways.

Learn more ›

Accelerated Web Services

Micro services with visual and intelligent capabilities using deep learning.

Learn more ›

Language and APIs

GPU acceleration can be accessed from most popular programming languages.

Learn more ›

Data Center Tools

Software Tools for every step of the HPC and AI software life cycle.

Learn more ›

Cluster Management

Managing your cluster and job scheduling can be simple and intuitive.

Learn more ›

Performance Analysis Tools

Find the best solutions for analyzing your application's performance profile.

Learn more ›

Key Technologies

Learn more about parallel computing technologies and architectures.

Learn more ›

So we should buy NVIDIA right?

If you want to buy **A100 GPUs** in quantity you will be faced with an **8 month lead time**. For **H100 it is worse, currently about 12 months**.

Why? Because there is so much interest in training LLMs / foundational models using NVIDIA GPUs it has generated a supply and demand issue.

AI start-ups have access to significant funds, allowing them to buy lots of GPUs at a premium price…

"Inflection AI, a new startup found by the former head of deep mind and backed by Microsoft and Nvidia, last week raised $1.3 billion…"

https://www.tomshardware.com/news/startup-builds-supercomputer-with-22000-nvidias-h100-compute-gpus

150 Miles of Optical Fiber | 2,112 60mm Fans | 70K CFM | 40K lbs | 1 GPU

22,000 H100s…

https://wccftech.com/inflection-ai-develops-supercomputer-equipped-with-22000-nvidia-h100-ai-gpus/

Look at the worlds 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 2 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

15

Look at the worlds largest machines, past and future trends - June 2023

TOP500
The List.

Two years later...

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

AMD

arm

AMD

NVIDIA.

NVIDIA.

Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,699,904	1,194.00	1,679.82	22,703
2	**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
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	2,220,288	309.10	428.70	6,016
4	**Leonardo** - BullSequana XH2000, Xeon Platinum 8358 32C 2.6GHz, NVIDIA A100 SXM4 64 GB, Quad-rail NVIDIA HDR100 Infiniband, Atos EuroHPC/CINECA Italy	1,824,768	238.70	304.47	7,404
5	**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

16

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

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

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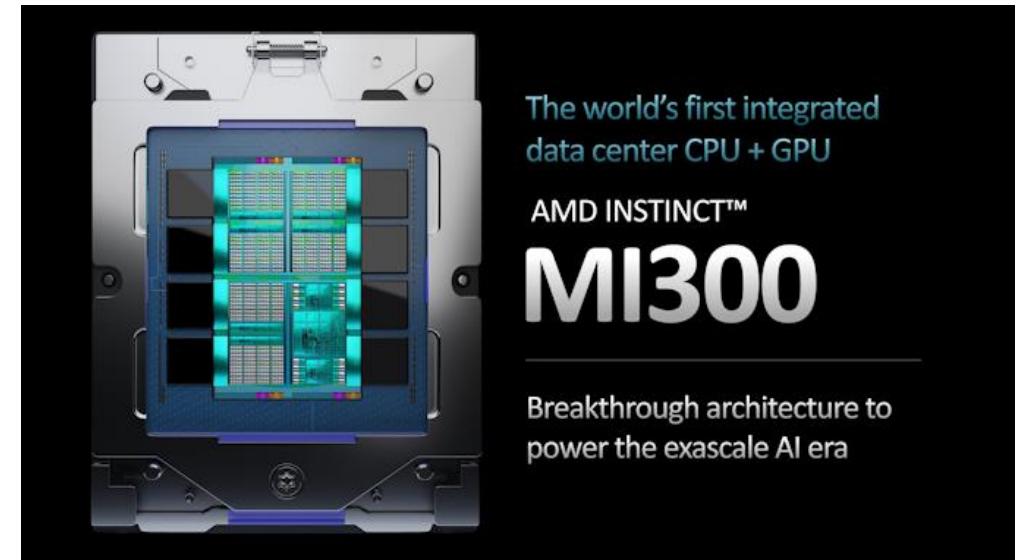
AMD as a solution? Hardware

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 Frontier and LUMI**, 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.

A bit more on the MI250X that Frontier uses: 2x 64GB of HBM2e, 3.2TB/s bandwidth, 48TFLOP/s (fp32 and fp64) and 500 Watts TDP.

The world's first integrated data center CPU + GPU

AMD INSTINCT™ MI300

Breakthrough architecture to power the exascale AI era

The forthcoming MI300 will be used in the 2 Eflop El Capitan machine

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:

2x A100 server £25K
2x H100 server £37K
2x MI250X server £22K

AMD claims the MI250X is between 1.5x and 2.5x faster than A100 for a range of representative scientific codes (it should be though, it's 18 months newer).

YMMV…

LAMMPS   Molecular Dynamics

Classical molecular dynamics package.

Application	Metric	Test Modules	Bigger is Better	4xMI250	4xA100 (SXM)	MI250/A100
LAMMPS	ATOM-Time Steps/s	Reaxff	Yes	19,482,180.48	8,850,000	Up to 2.2x[2]

LSMS   Physics

Locally Self-Consistent Multiple Scattering (LSMS) is an order-N approach to the calculation of the electronic structure of large systems.

Application	Metric	Test Modules	Bigger is Better	1xMI250	1xA100 (SXM)	MI250/A100
LSMS	ATOM-Interactions/s	FePt54	Yes	3.95E_09	2.44E+09	Up to 1.6x[3]

MILC   Physics

Lattice Quantum Chromodynamics (LQCD) codes simulate how elemental particles are formed and bound by the "strong force" to create larger particles like protons and neutrons.

Application	Metric	Test Modules	Bigger is Better	1xMI250	1xA100(SXM)	MI250/A100
MILC	Total Time (Sec)	Apex Medium	No	1,604.6	2,262	Up to 1.4x[1]

OpenMM   Molecular Dynamics

High performance, customizable molecular simulation. Times in seconds to complete 8 identical simulations on one GPU.

Application	Metric	Test Modules	Bigger is Better	1xMI250	1xA100 (SXM)	MI250/A100
OpenMM	Total Time (Sec) / 10,000 steps	amoebakg	No	387	921	Up to 2.4x[4]

Software - Radeon Open Compute Platform (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 ROCm.

Deployment Tools	ROCm Validation Suite	ROCm Data Center Tool		ROCm SMI
Libraries	rocBLAS · rocRAND · rocFFT · MIGraphX · MIVisionX · rocPRIM			
	rocSOLVER · rocALUTION · rocSPARSE · rocTHRUST · MIOpen · RCCL			
Compilers & Tools	Compiler · Profiler · Debugger · hipify · TENSILE			
Programming Models	OpenMP® API · HIP API · OpenCL™			
Drivers/Runtimes	RedHat®, SLES & Ubuntu® Device Drivers and Run-Time			

https://www.amd.com/en/graphics/servers-solutions-rocm

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/technologies/infinity-hub

https://www.amd.com/system/files/documents/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 automatically 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
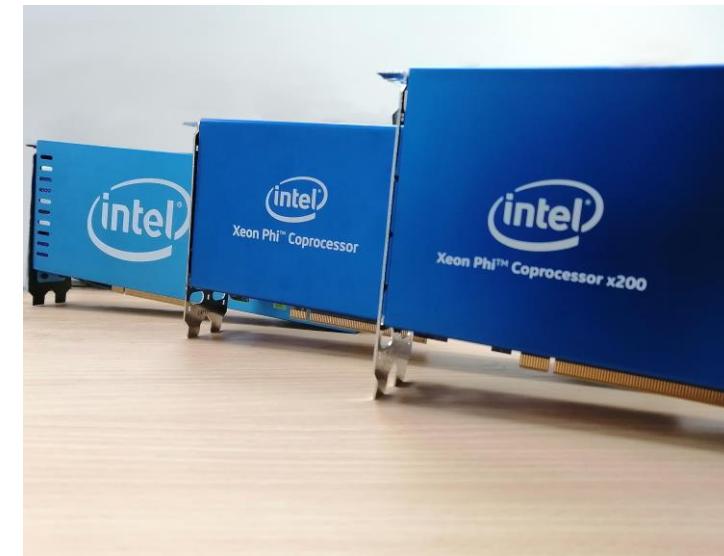
What about Intel?

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).

By D2theW - Own work, CC BY-SA 4.0,
https://commons.wikimedia.org/w/index.php?curid=114913985
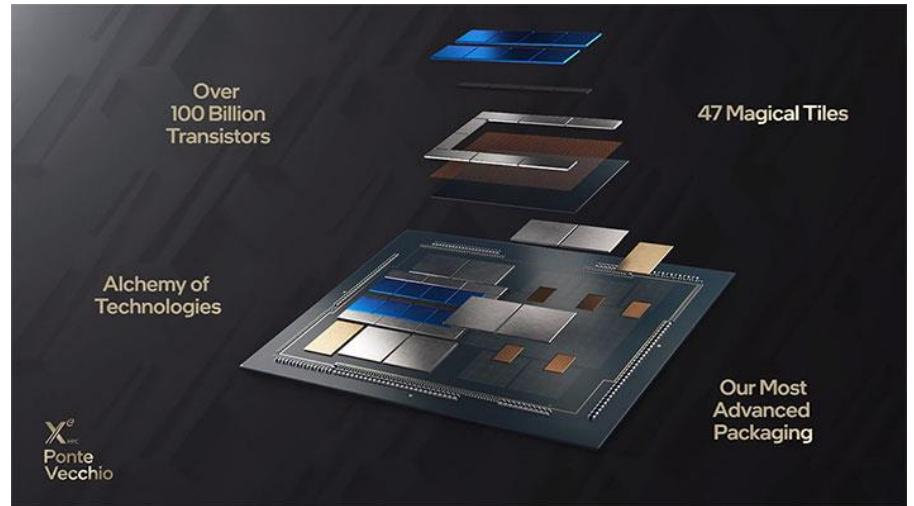
Intel Xe-HPC - Ponte Vecchio GPU

The new **Ponte Vecchio GPU is only just reaching the market.**

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

Tests on development silicon (very tentatively) indicate that **for some applications this reaches about 80% of the performance of an A100**.

We should keep in mind though, that the A100 is two years older than this, NVIDIAs current flagship GPU is the H100.
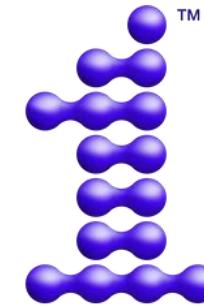
OneAPI

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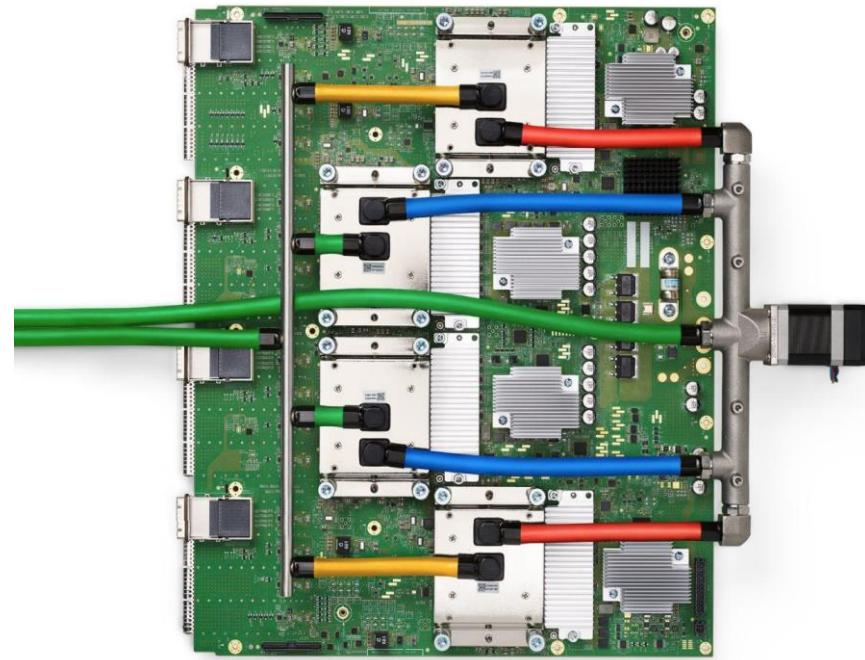
https://www.eejournal.com/article/intels-latest-version-of-oneapi-takes-advantage-of-new-intel-xeon-improvements-supports-amd-and-nvidia/?cid=org&source=linkedin_organic_cmd&campid=ww_23_oneapi&content=art-idz_&linkId=100000207031089

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.

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

COLOSSUS MK2
the worlds most complex processor

59.4Bn transistors, TSMC 7nm @ 823mm^2

250TFlops AI-Float | 900MB In-Processor-Memory

1472 independent processor cores

8832 separate parallel threads

>8x step-up in system performance vs Mk1

GC200 - IPU

|	NVIDIA DGX-A100 (8x A100)	GRAPHCORE 8x M2000	
FP32 compute	156TFLOP	2PFLOP	>12x
AI compute	2.5PFLOP[1]	8PFLOP[2]	>3x
AI Memory	320GB[3]	3.6TB[4]	>10x
System Price	$199,000 MSRP	$259,600 MSRP	
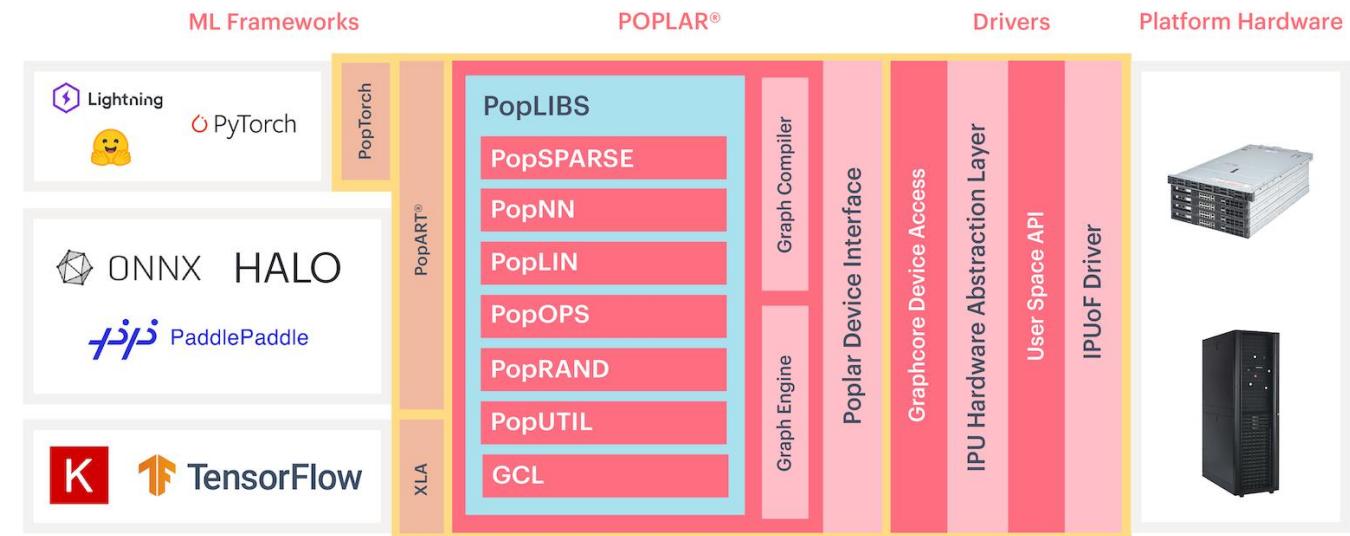
NOTES:
[1] Actual figure for TF32/FP16. NVIDIA 8xA100 5PFlop reference is for 50% sparsity which includes Pflops for operations that aren't run
[2] Graphcore AI Float with IEEE FP16.16 multiply.accumulate and IEEE FP16.SR 16bit float with stochastic rounding, with equivalent accuracy performance as FP32
[3] 40GB HBM memory on A100 modules *8 modules per DGX-A100 system
[4] IPU-Exchange Memory which includes attached DRAM and IPU In-processor-memory with 100x bandwidth vs. HBM memory sub-system

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.

Cerebras

Cerebras produce **wafer level processors**. - Quite amazing.
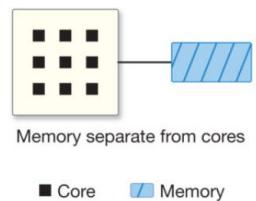
In terms of **software**, Cerebras has a similar approach to Graphcore. It has the **Csoft environment**. It too integrates Torch and TensorFlow to produce code that runs on the CS-2 platform.

**It also has a SDK to allow developers to write custom kernels.**
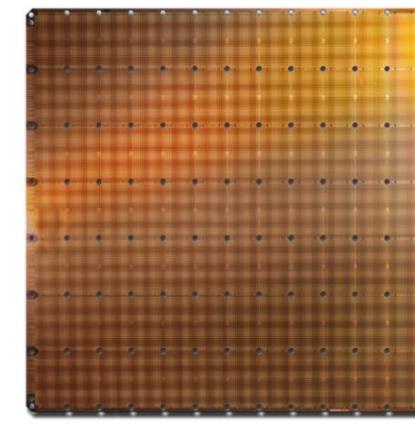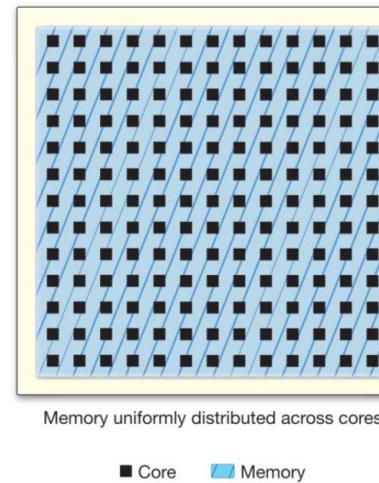
I haven't seen a good comparison to other technology as yet.


Cerebras Memory Architecture

Traditional Memory Architecture

Memory separate from cores

■ Core   ▨ Memory

Memory uniformly distributed across cores

■ Core   ▨ Memory

Cerebras WSE
1.2 Trillion transistors
46,225 mm² silicon
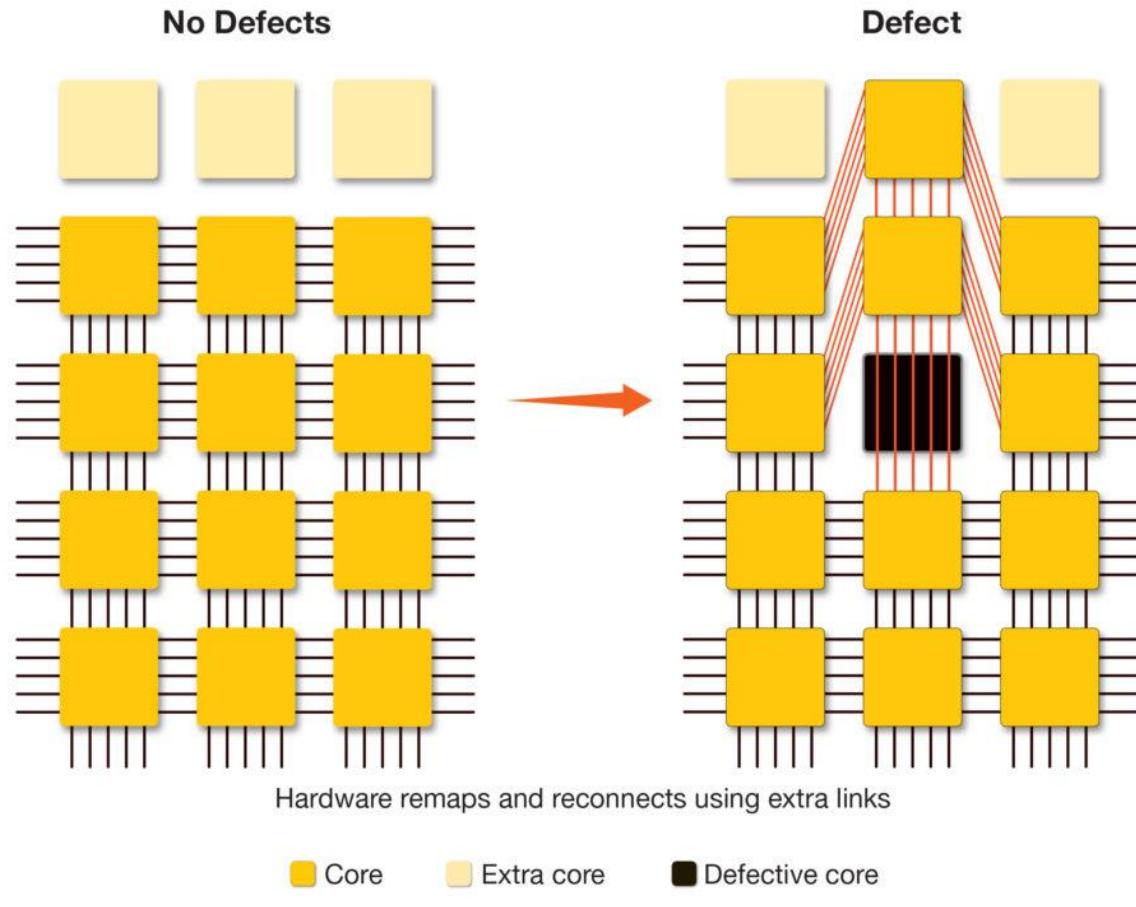
Largest GPU
21.1 Billion transistors
815 mm² silicon

Cerebras

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

For those interested –
TMSC can produce ~ 8K wafers per month*

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.

No Defects   **Defect**

Hardware remaps and reconnects using extra links

Core   Extra core   Defective core

*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 and G42

Cerebras have very recently announced that they will supply G42 (UAE AI company) 3x Condor Galaxy systems as part of $100M deal.
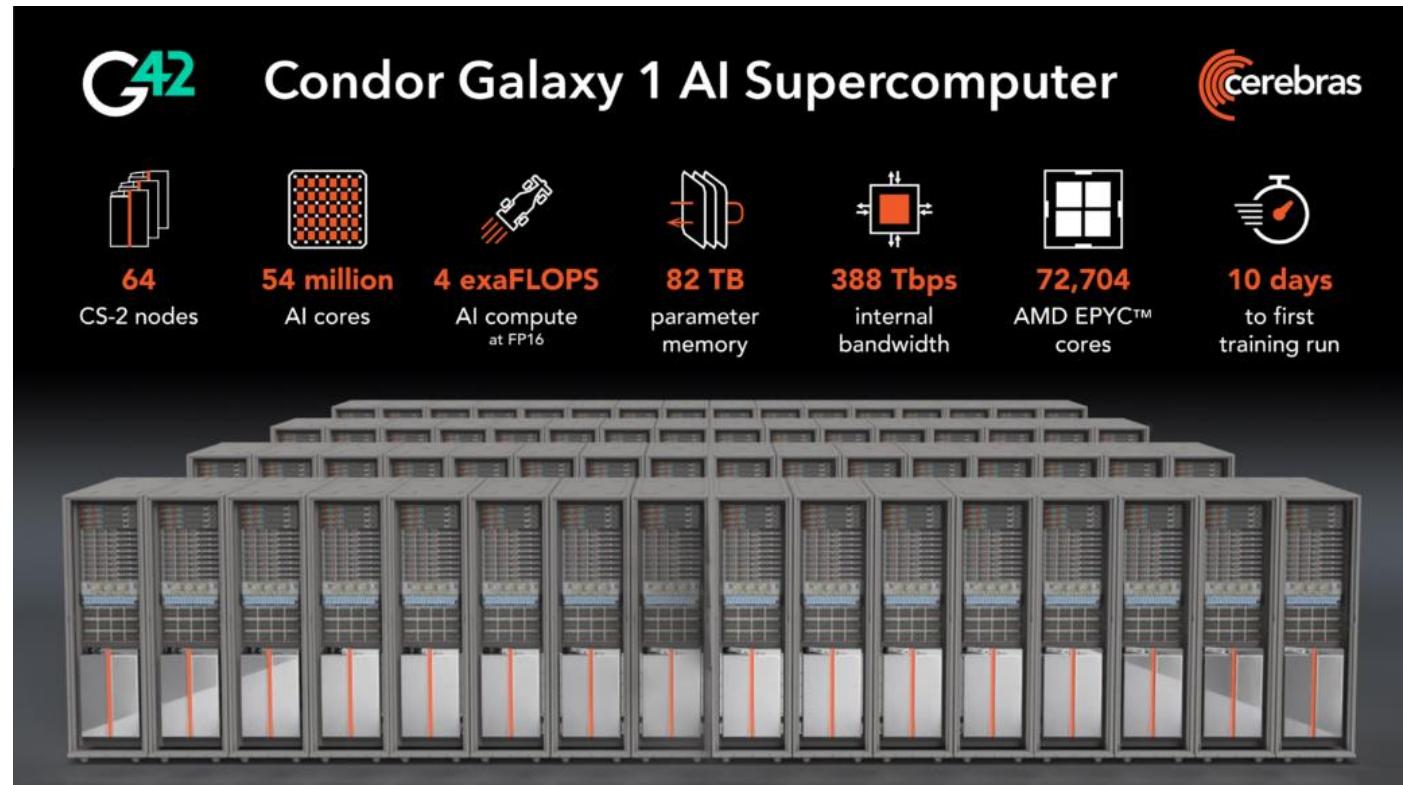
A single CG system is capable of 4 EFLOP (fp16), has 54M Cerebras cores and 82TB of memory.

More on CG here:
https://www.condorgalaxy.ai/

Access to CG-1 here:
https://www.cerebras.net/product-cloud/
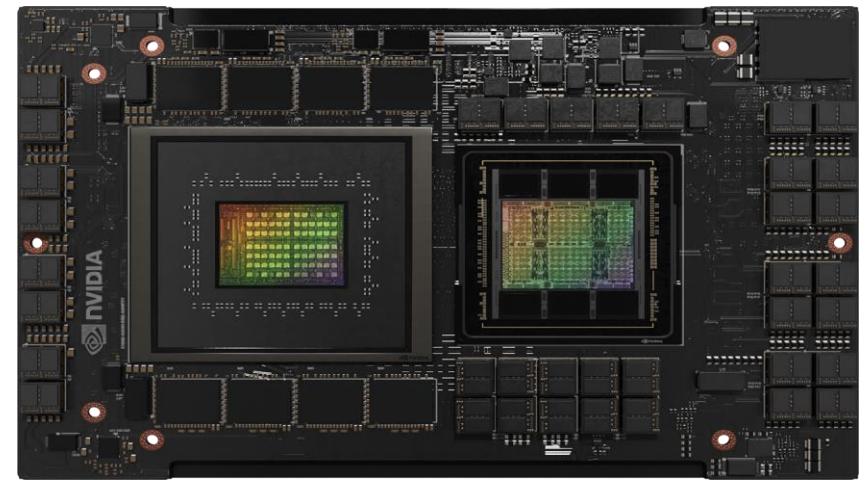
NVIDIA – Grace-Hopper

Grace-Hopper is 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
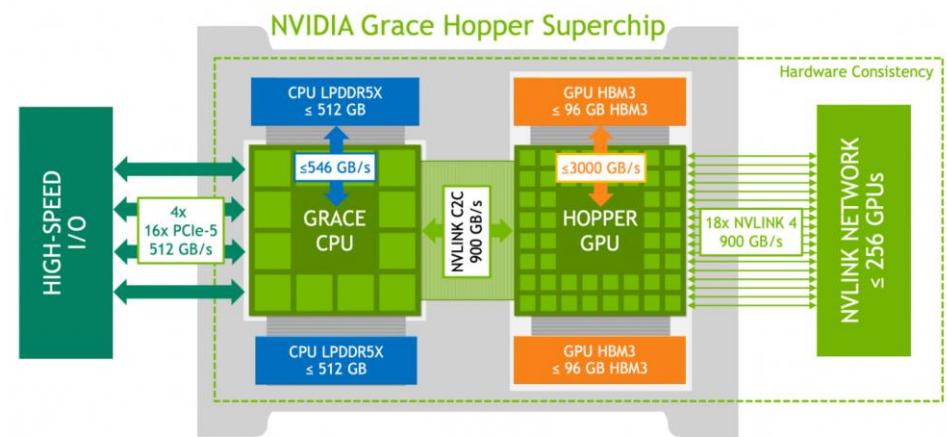
NVIDIA – Grace-Hopper

NVIDIA Grace + Hopper:

- 72x Arm Neoverse V2 cores (4×128-bit SIMD units per core).
- Up to 117 MB of L3 Cache.
- Up to 512 GB of LPDDR5X memory (546 GB/s of memory bandwidth).
- Up to 64x PCIe Gen5 lanes.
- NVIDIA Scalable Coherency Fabric (SCF) mesh and distributed cache with up to 3.2 TB/s memory bandwidth.
- NVIDIA Hopper GPU.
- NVIDIA NVLink-C2C - Up to 900 GB/s total bandwidth.
- Unified address space - each Hopper GPU can address up to 608 GB of memory within a superchip.
- NVIDIA NVLink Switch System connects up to 256x NVIDIA Grace Hopper Superchips using NVLink 4.
- Each NVLink-connected Hopper GPU can address all HBM3 and LPDDR5X memory of all superchips in the network, **for up to 150 TB of GPU addressable memory**.

DGX GH200 AI Supercomputer

The DGX GH200 was announced at Computex May 2023. It's NVIDIAs answer to the likes of Cerebras.

It connects 256 Grace-Hopper "superchips" via NVLink.

- **Single 144 terabytes GPU memory space**.

- **900 GB/s GPU-to-GPU bandwidth**.

- **1 exaFLOPS of FP8 AI performance**.

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

https://nvidianews.nvidia.com/news/nvidia-announces-dgx-gh200-ai-supercomputer

https://resources.nvidia.com/en-us-dgx-gh200/nvidia-dgx-gh200-datasheet-web-us

The future?

NVIDIAs value continues to grow

Market Summary > NVIDIA Corp

1.13 trillion USD
Market capitalisation

459.00 USD

+395.97 (628.22%) ↑ past 5 years

Closed: 27 Jul, 17:59 GMT-4 • Disclaimer
After hours 460.26 +1.26 (0.27%)

+ Follow

| 1D | 5D | 1M | 6M | YTD | 1Y | 5Y | Max |

459.00 USD 27 Jul 2023

Open	465.19	**Mkt cap**	**1.13T**	CDP score	B
High	473.95	P/E ratio	238.54	52-wk high	480.88
Low	457.50	Div yield	0.035%	52-wk low	108.13

The future?

AMDs, even with the success of Frontier is some way behind.

Market Summary > Advanced Micro Devices, Inc.

178.91 billion USD
Market capitalisation
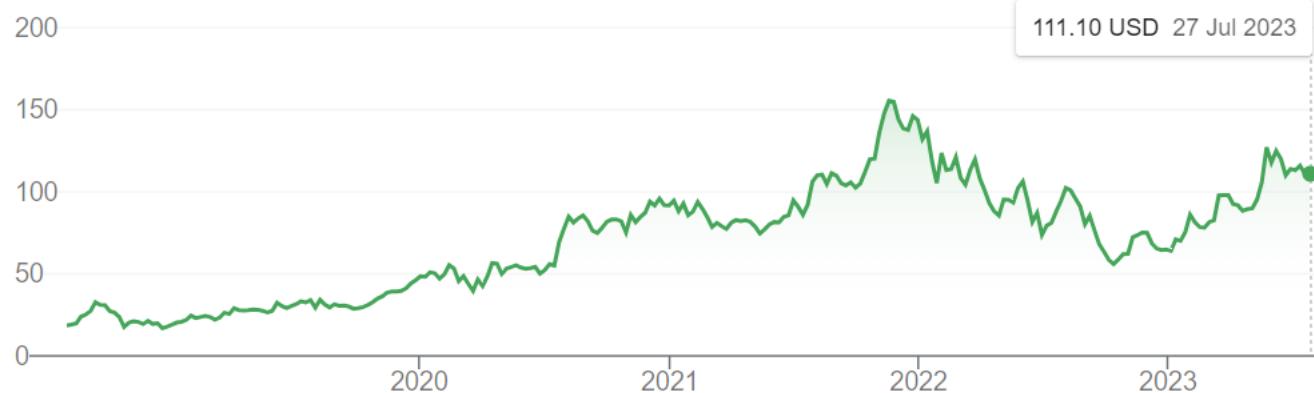
111.10 USD

+92.61 (500.87%) ↑ past 5 years

Closed: 27 Jul, 17:57 GMT-4 • Disclaimer
After hours 111.50 +0.40 (0.36%)

+ Follow

| 1D | 5D | 1M | 6M | YTD | 1Y | 5Y | Max |

111.10 USD 27 Jul 2023

Open	111.79	**Mkt cap**	**178.91B**	CDP score	B
High	115.08	P/E ratio	460.59	52-wk high	132.83
Low	110.51	Div yield	-	52-wk low	54.57

The future?

Intel, failed to deliver Aurora, originally contracted to be completed by 2018, Intel are hoping to deliver later this year. Should be 2 ExaFLOP machine, each node will have 2x Intel Sapphire Rapids (CPU) and 6x Ponte Vecchio GPUs.

Out of the three Intel is worth the least!

Market Summary > Intel Corporation

144.11 billion USD
Market capitalisation

34.55 USD

+ Follow
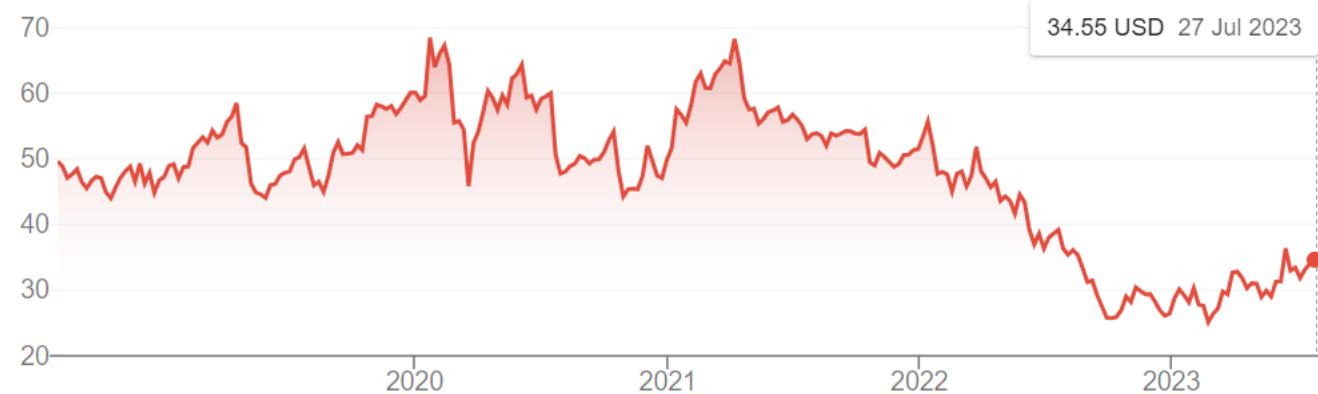
-15.08 (-30.38%) ↓ past 5 years

Closed: 27 Jul, 18:02 GMT-4 • Disclaimer
After hours 37.30 +2.75 (7.96%)

| 1D | 5D | 1M | 6M | YTD | 1Y | 5Y | Max |

34.55 USD 27 Jul 2023

	Open	34.82	**Mkt cap**	**144.11B**	CDP score	B
	High	35.03	P/E ratio	-	52-wk high	40.42
	Low	34.11	Div yield	1.45%	52-wk low	24.59

The future?

It's likely due to the cost of NVIDIA and shortage of supply that AMD will get a growing fraction of the accelerator market, especially given that they seem to be following (very closely!) NVIDIAs strategy – a great software ecosystem.

The HPE El Capitan supercomputer, due to be commissioned in Q4 2023 is an upcoming exascale supercomputer, hosted at the Lawrence Livermore, will be a 2+ ExaFLOP supercomputer and will displace Frontier as the world's fastest supercomputer.

- It's based on… AMD.

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 Mikes computing webpage here:

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