Lecture 8: looking to the future

Prof. Mike Giles

mike.giles@maths.ox.ac.uk

Oxford University Mathematical Institute

Keeping up-to-date

Important in scientific computing to keep an eye on what is happening with both hardware and software

(I am self-taught through reading lots of blogs and websites, as well as academic papers on scientific computing)

Remember: at times the business aspects are as important as the technical in thinking about how things are developing

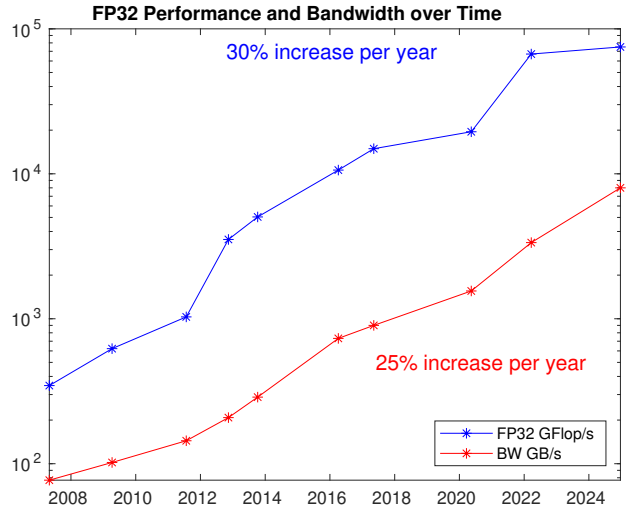
Feb 2025 market capitalization (i.e. company value)

- NVIDIA: $ 3.18 trn
- AMD: $ 175 bn
- Intel: $ 82 bn

10 years ago the order would have been reversed!

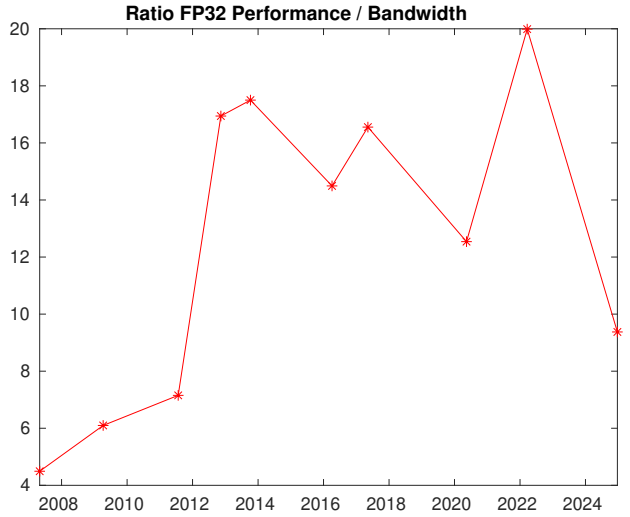
Hardware trends

NVIDIA high-end GPU performance and bandwidth



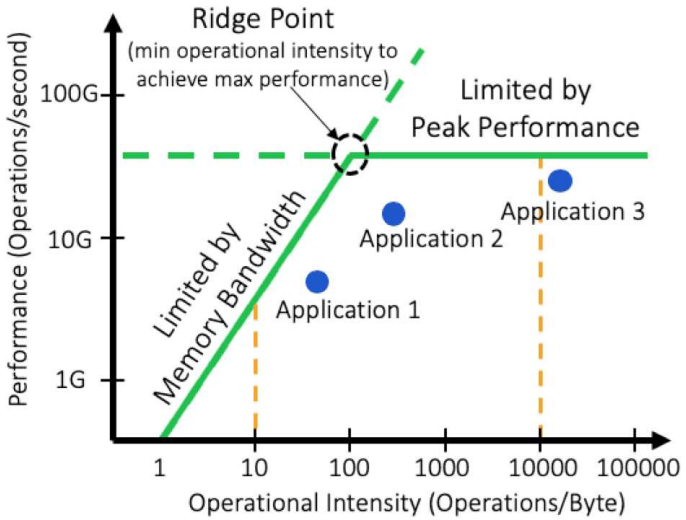
Hardware trends

Compute / bandwidth ratio



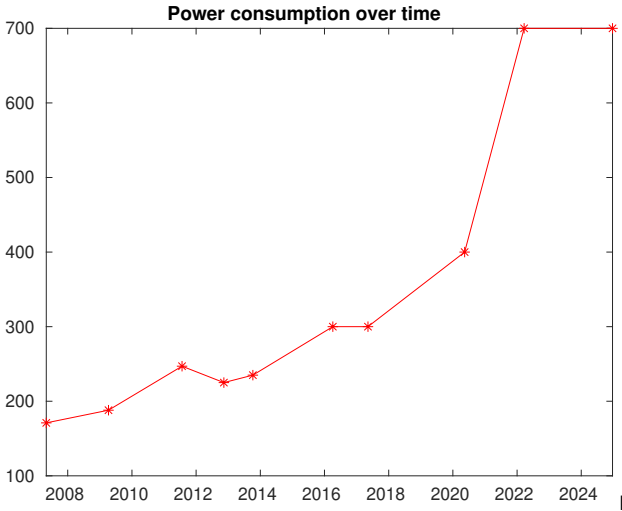
Hardware trends

Roofline model (image copyright Rambus Inc.)



Hardware trends

Increasing energy consumption by NVIDIA GPUs – moving to chilled-water cooling blocks



NVIDIA

NVIDIA

Ampere came out in 2020:

- A100 with 108 SMs, 40-80 GB HBM2 memory
- wide range of “tensor core” capabilities
- NVIDIA DGX A100 Deep Learning server
  `https://www.nvidia.com/en-us/data-center/dgx-a100/`
  - 8 NVIDIA A100 GPUs, each with 80GB HBM2
  - 2 $\times$ 64-core AMD “Rome” CPUs
  - 2 TB DDR4 memory, 30 TB SSD
  - 600GB/s NVlink interconnect between the GPUs

NVIDIA

- Hopper came out in 2023:
  - H100 for HPC
  - 228-264 SMs
  - 80GB HBM3 memory
  - 40MB L2 cache
  - NVlink improvements – up to 50% faster, 900GB/s
  - PCIe v5.0 – $2\times$ improvement
- Grace CPU also arrived in 2023:
  - Arm-based
  - up to 72 cores
  - 550GB/s bandwidth to LPDDR5X memory
  - 900GB/s NVlink connection to Hopper GPU in GB100 “superchip”

NVIDIA

- A Hopper refresh came out in 2024:
  - H100NVL – two GPUs on one card
  - 2 $\times$ 96GB HBM3 memory
  - 2 $\times$ 3.9TB/sec bandwidth
  - aimed particularly at LLMs needing lots of memory and bandwidth
- Late 2024, the Blackwell B100 came out:
  - 192 GB memory, 8192-bit bus, 8TB/s bandwidth
  - modest improvement in compute except for low-precision tensor cores
  - persistent rumours of over-heating

NVIDIA

Current status:

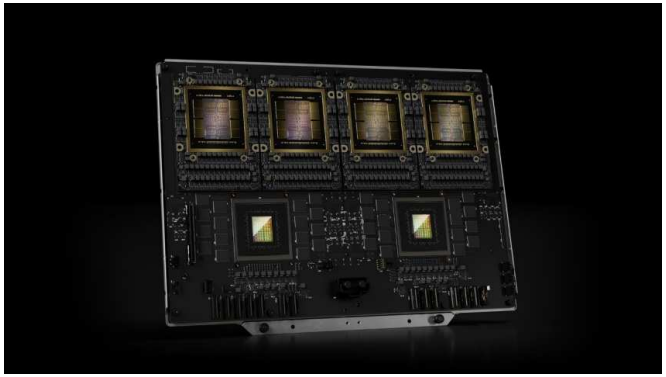
- big AI companies are competing to buy huge numbers (10,000+) of Hopper H100 and Blackwell B100 GPUs – many orders are worth over $1bn
- supply is limited, prices have become inflated, and it’s very difficult for academics to get any
- emergence of Grace CPU is significant – gives NVIDIA freedom to design their own combined CPU/GPU offerings with high bandwith interconnect

  (also part of ARM breakthrough into the server market? hyperscalers are responsible for 50% of the server market, and 50% of their CPUs are Arm-based)

NVIDIA

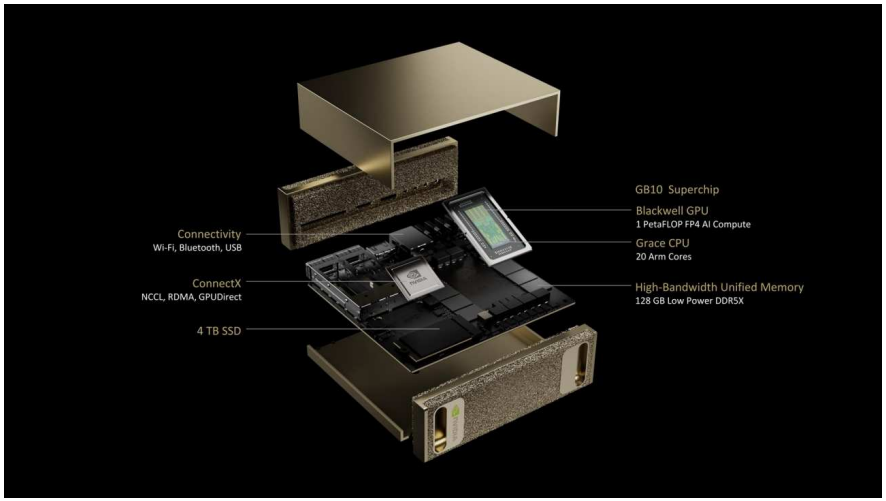
Coming later this year:

- GB200 superchip combining 1 Grace CPU / 2 B200s
- GB200 NVL72 rack with 36 GB200s (72 B200 GPUs) connected by an NVLINK switch
- also GB200 NVL4 card with 2 Grace CPUs / 4 B200s



NVIDIA

Coming later this year – NVIDIA DIGITS system:



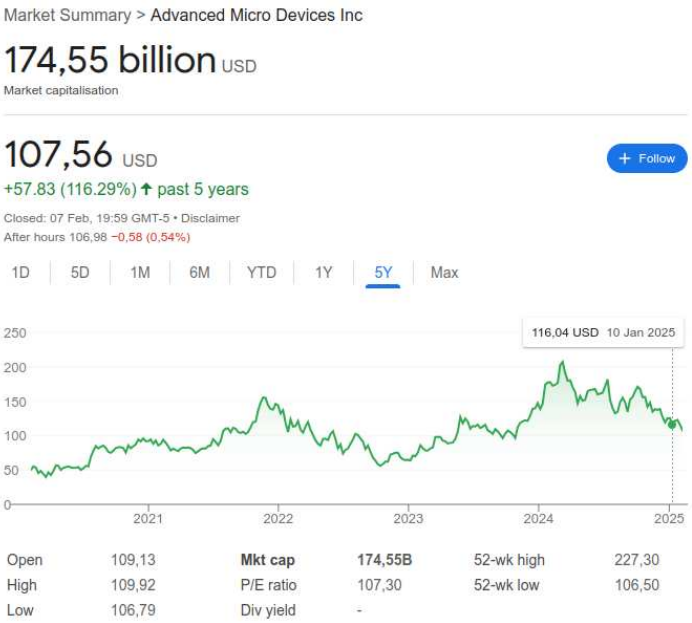
NVIDIA

"A Grace-Blackwell AI Supercomputer on your desk":

- complete small desktop system – $3000 in US
- GB10 combined CPU/GPU superchip
- 20 Arm cores in CPU with 128GB of DDR5X memory
- 1 PFlop FP4 (?!) Blackwell GPU – 7% of B100
- 4 TB SSD and Linux operating system

I want one for Christmas!

AMD



Top500

Top 5 on Top500 list, November 2024:

- #1 El Capitan (DoE/LLNL, USA)
  - HPE: 44,000 AMD MI300A GPUs
- #2 Frontier (DoE/ORNL, USA)
  - HPE: 40,000 AMD MI250X GPUs
- #3 Aurora (DoE/ALC, USA)
  - Intel/HPE: 54,000 Intel Max GPUs
- #4 Eagle (Microsoft Azure)
  - Microsoft: NVIDIA H100 GPUs
- #5 HPC6 (Eni, Italy)
  - HPE: 14,000 AMD MI250X GPUs

AMD

El Capitan: #1 supercomputer based on Linpack performance

- sited at Lawrence Livermore National Laboratory (DoE)
- 1.7 Exaflops, 30 MW
- system from HPE; CPUs and GPUs from AMD
- 11,136 compute nodes, each with 4 MI300A GPUs

AMD

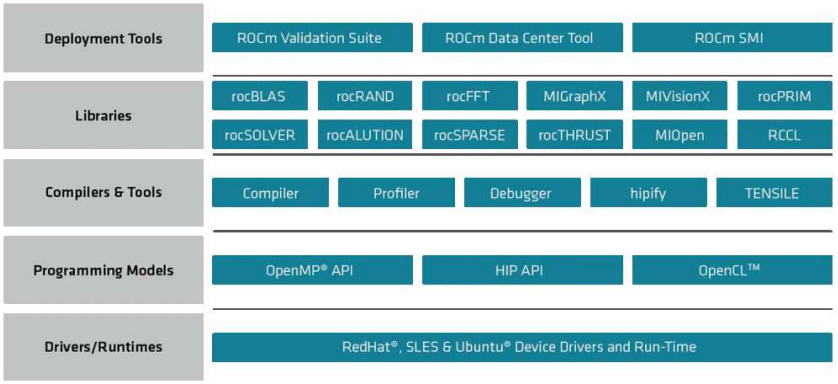
- over past decade AMD has had excellent CPUs and GPUs (and pioneered chiplet packaging) but has not invested enough in software – that is changing
- hired lots of software specialists in the past 2 years, including many of the NAG team responsible for ACML (AMD's version of Intel's MKL libraries)
- "Genoa" Zen4 EPYC CPUs:
  - up to 64 cores with vector units and 384MB L3
  - now getting about 20% share of server market
- Frontier has previous generation "Trento" Zen3 EPYC CPUs

AMD

- Instinct GPUs:
  - MI250X has 220 Compute Units, 128 GB HBM2e, 3.2 TB/s: comparable to A100 GPU for PyTorch
  - MI300A has 228 Compute Units, 912 Matrix Cores, 128 GB HBM2e, 5.3 TB/s: comparable to H100?
  - MI325X has 304 Compute Units, 1216 Matrix Cores, 256 GB HBM3e, 6 TB/s: comparable to B100?
  - programmed using AMD's ROCm (similar to CUDA) with extensive library support
  - portability provided through HIP (Heterogeneous computing Interface for Portability) with compilation to either CUDA or AMD's ROCm

AMD

AMD's ROCm eco-system:



AMD

AMD's HIP – some example 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);
```


AMD

Now for some kernel code:

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

Can see why it is fairly easy for AMD's HIPIFY tool to convert most simple CUDA code to HIP – this is another reason to avoid "exotic" CUDA features as much as possible.

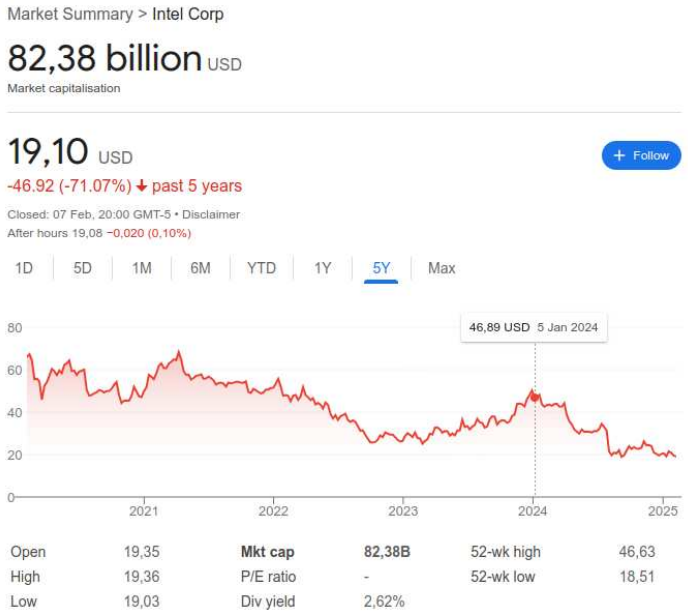
Warning: AMD GPUs have a warp size of 64, not 32, so use `warpSize` variable in your code rather than hard-coding a warp size of 32.

AMD

- ROCm and HIP look **very** similar to CUDA – probably required to win the major DoE and EU contracts
- pricing and availability of GPUs are both much better than NVIDIA currently, especially for academics

  (major AI companies are placing $1bn orders with NVIDIA so no GPUs left for us!)
- AMD's software eco-system is still maturing – will take at least another 5 years to get close to CUDA
- still, very good to see competition in the marketplace

Intel



Intel

- current “Granite Rapids” Xeon-SP CPUs:
  - up to 128 cores, each with one or two 512-bit AVX-512 vector units per core (512 bits = 16 floats)
  - up to 500MB L3 (shared), 2MB L2 per core
  - up to 600 GB/s memory bandwidth with new DDR5 MRDIMM memory
- “Ponte Vecchio” a.k.a. Data Center GPU Max:
  - 128 Xe cores, each with 8 $\times$ 256-bit vector units and 8 tensor cores
  - 408MB L2 cache, 128GB HBM2e with 8192-bit bus
  - development of “Rialto Bridge” successor ended; not clear what Intel will do now

Others

Special designs, solely for the needs of Machine Learning:

- Google: Tensor Processing Unit (TPU)
- Graphcore: Colossus Intelligent Processing Unit
- Cerebras: in-memory computing (lots of computing elements interspersed within a huge amount of memory in wafer-scale chips)

It seems unlikely that Google will get into the hardware business in a big way, and if any startup makes real progress they’ll be bought out by NVIDIA, AMD or Intel.

Outlook

My current software assessment:

- CUDA is dominant in HPC because of
  - ease-of-use
  - NVIDIA dominance of hardware, with huge sales in machine learning in particular
  - extensive library support
  - support for many different languages (Fortran, Python, R, MATLAB, etc.)
  - extensive eco-system of tools
- HIP is a real threat to that dominance by offering platform independence with compilation to both CUDA and AMD’s ROCm

Final thoughts

- NVIDIA holds a dominant market position, maybe hard to justify their huge market valuation but they’re the leader for a good reason – they have excellent hardware and software, and focussed early on the needs of AI/ML

  Original gaming market no longer significant, the auto market is the next big one they’re working on
- By addressing their software weakness, AMD is back in the game for both HPC and AI/ML – great to have competition again
- I remain unconvinced by Intel’s new hardware and software products, though traditional Xeon CPUs remain powerful and sell well
- Other vendors are unlikely to break through significantly