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

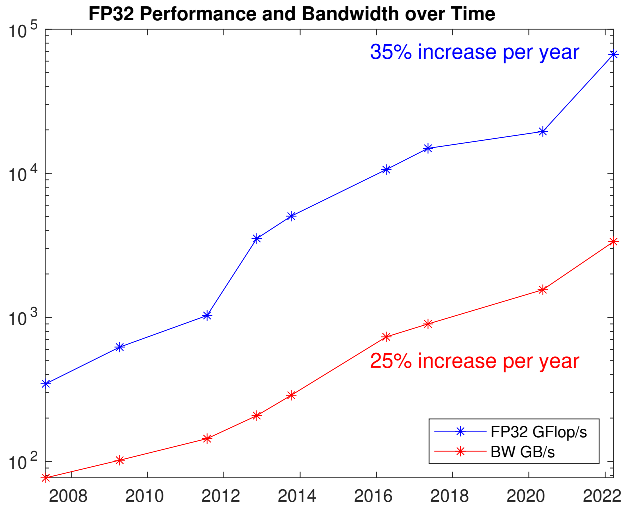
Current market capitalization (i.e. company value)

- NVIDIA: $ 1150 bn
- AMD: $ 187 bn
- Intel: $ 148 bn

10 years ago the order would have been reversed!

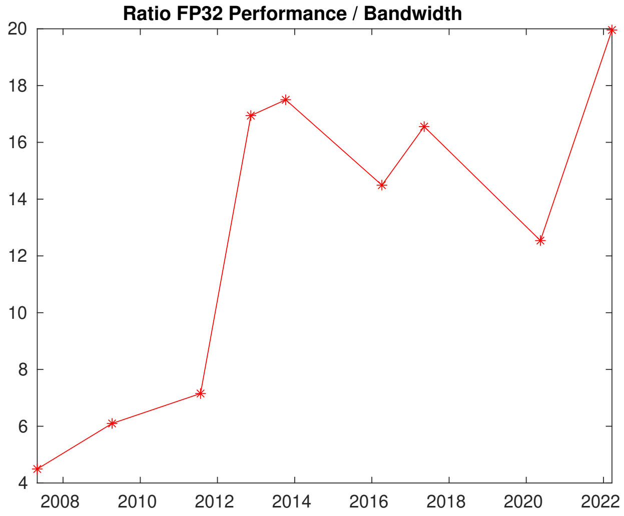
Hardware trends

NVIDIA high-end GPU performance and bandwidth



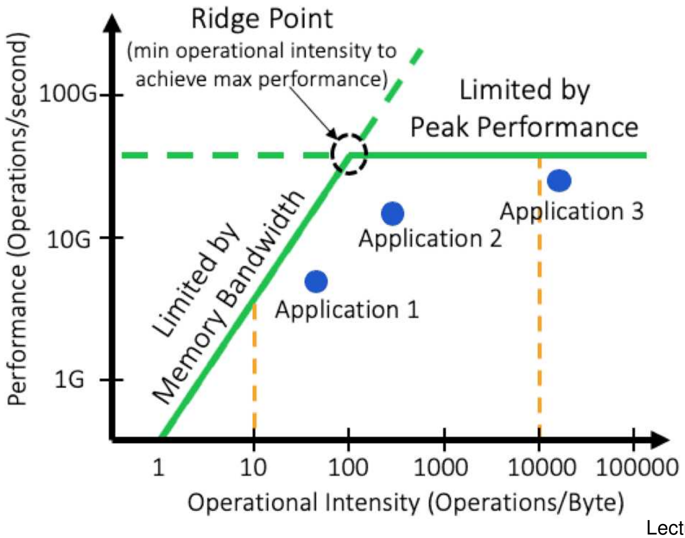
Hardware trends

Compute / bandwidth ratio



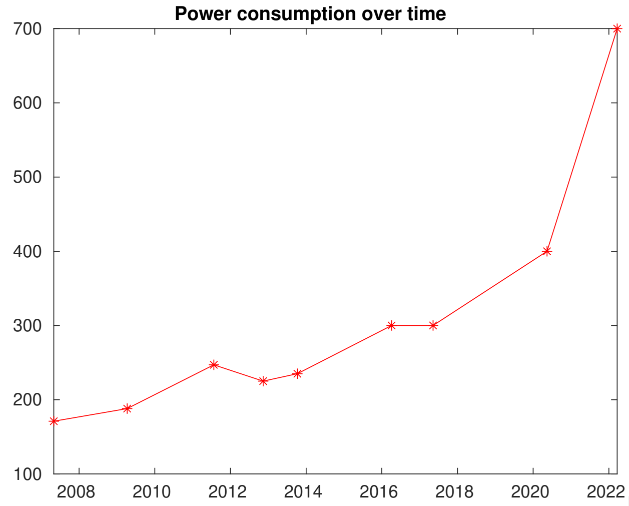
Hardware trends

Roofline model (image copyright Rambus Inc.)

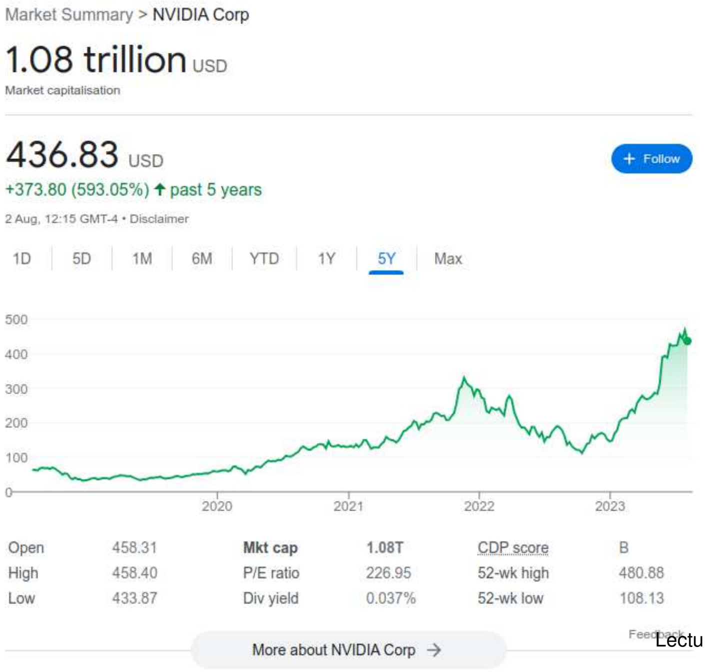


Hardware trends

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



NVIDIA



Market Summary > NVIDIA Corp

1.08 trillion USD

Market capitalisation

436.83 USD

+373.80 (593.05%) ↑ past 5 years

2 Aug, 12:15 GMT-4 • Disclaimer

Open	458.31	Mkt cap	1.08T	CDP score	B
High	458.40	P/E ratio	226.95	52-wk high	480.88
Low	433.87	Div yield	0.037%	52-wk low	108.13

NVIDIA

- Volta came out in 2017/18:
  - V100 for HPC
  - 80 SMs
  - 32GB HBM2 memory
  - special “tensor cores” for machine learning – much faster for TensorFlow & PyTorch

- Ampere came out in 2020:
  - A100 for HPC
  - 108 SMs
  - 40-80 GB HBM2 memory
  - wider range of “tensor core” capabilities

NVIDIA

- NVIDIA DGX Station A100
  `https://www.nvidia.com/en-us/data-center/dgx-station-a100/`
  - 4 NVIDIA A100 GPUs, each with 80GB HBM2
  - 64-core AMD CPU
  - 512 GB DDR4 memory, 10 TB SSD
  - 600GB/s NVlink interconnect between the GPUs
- 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 has come out in 2023:
  - H100 for HPC
  - 228-264 SMs
  - 80GB HBM2 memory
  - 40MB L2 cache
  - NVlink improvements – up to 50% faster, 900GB/s
  - PCIe v5.0 – $2\times$ improvement
- Grace CPU has also arrived in 2023:
  - Arm-based
  - up to 72 cores
  - 550GB/s bandwidth to LPDDR5X memory
  - 900GB/s NVlink connection to Hopper GPU

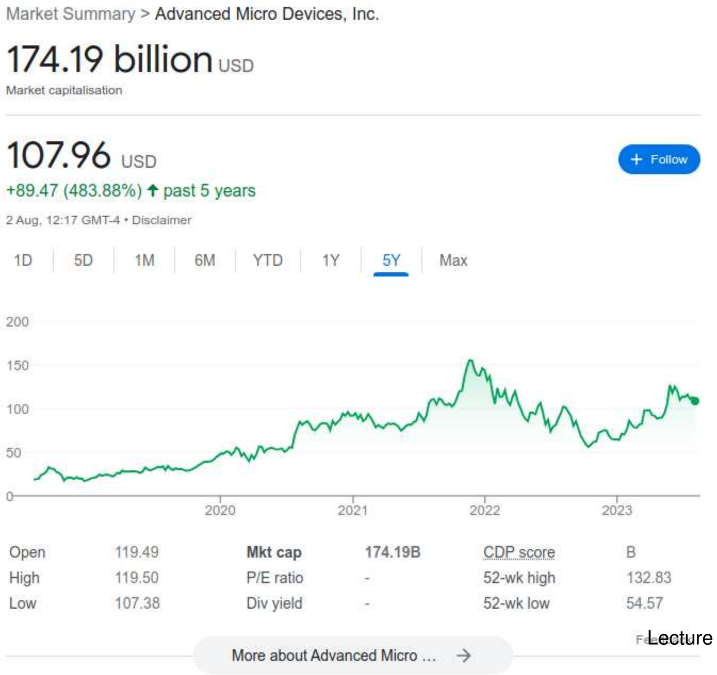
NVIDIA

Current status:

- big AI companies are competing to buy huge numbes (10,000+) of Hopper H100 GPUs – some 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, like AMD

  (maybe also signifies ARM breakthrough into the server market?)

AMD



Market Summary > Advanced Micro Devices, Inc.

174.19 billion USD

Market capitalisation

107.96 USD

+89.47 (483.88%) ↑ past 5 years

2 Aug, 12:17 GMT-4 • Disclaimer

Open	119.49	**Mkt cap**	**174.19B**	CDP score	B
High	119.50	P/E ratio	-	52-wk high	132.83
Low	107.38	Div yield	-	52-wk low	54.57

Top500

Top 5 on Top500 list, June 2023:

- #1 Frontier (DoE/ORNL, USA)
  - HPE: 40,000 AMD MI250X GPUs
- #2 Fugaku (RIKEN, Japan)
  - Fujitsu: 160,000 Fujitsu/ARM CPUs with vector units
- #3 Lumi (EuroHPC/CSC, Finland)
  - HPE: 10,000 AMD MI250X GPUs
- #4 Leonardo (EuroHPC/CINECA, Italy)
  - Atos: 14,000 NVIDIA A100 GPUs
- #5 Summit (DoE/ORNL, USA)
  - IBM: 28,000 NVIDIA V100 GPUs

AMD

Frontier: #1 supercomputer based on Linpack performance

- sited at Oak Ridge National Laboratory (DoE)
- 1.7 Exaflops, 21 MW
- system from HPE; CPUs and GPUs from AMD
- 9,472 compute nodes, each with one EPYC CPU, four MI250X GPUs and 4TB of flash memory

AMD

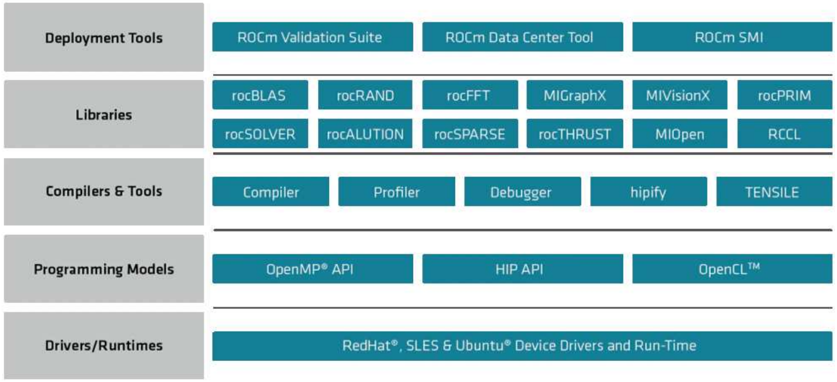
- over past decade AMD has had excellent CPUs and GPUs (and pioneered chiplet packaging) but has not invested enough in software – that is changing
- hired lots of software specialists in the past 2 years, including many of the NAG team responsible for ACML (AMD's version of Intel's MKL libraries)
- "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, each with 64 stream procs, and 128 GB HBM2e memory with up to 3.2 TB/s bandwidth: comparable to A100 GPU, including for PyTorch
  - new MI300X will be broadly competitive with H100, depending on price and availability
  - programmed using AMD's ROCm (very 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: `https://rocmdocs.amd.com/en/latest/Programming_Guides/HIP-GUIDE.html`

AMD

AMD's ROCm eco-system:



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

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.

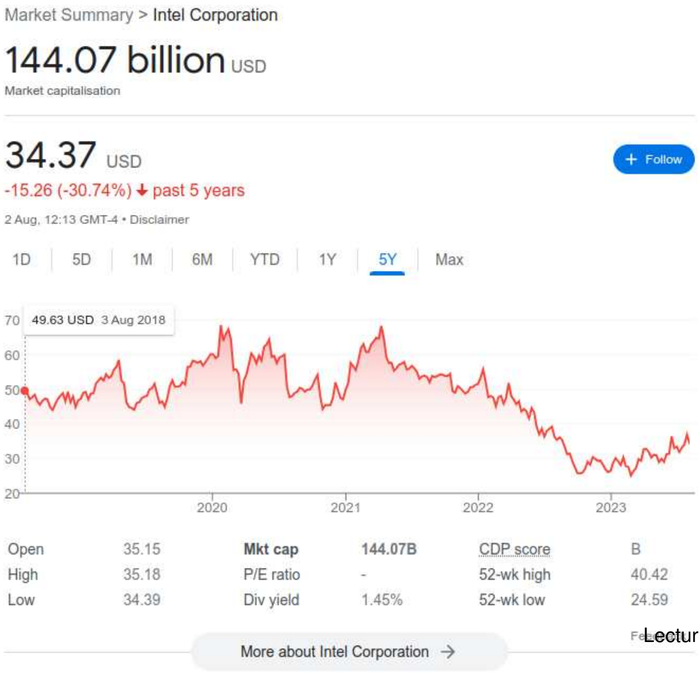
Warning: AMD GPUs have a warp size of 64, not 32, so use `warpSize` variable in your code rather than hard-coding a warp size of 32.

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



Market Summary > Intel Corporation

144.07 billion USD

Market capitalisation

34.37 USD

-15.26 (-30.74%) ↓ past 5 years

2 Aug, 12:13 GMT-4 • Disclaimer

Open	35.15	**Mkt cap**	**144.07B**	CDP score	B
High	35.18	P/E ratio	-	52-wk high	40.42
Low	34.39	Div yield	1.45%	52-wk low	24.59

Intel

- current “Sapphire Rapids” Xeon-SP CPUs:
  - up to 60 cores, each with one or two 512-bit AVX-512 vector units per core (512 bits = 16 floats)
  - up to 112.5MB L3 (shared), 2MB L2 per core
  - up to 250 GB/s memory bandwidth
  - CPU Max variants have up to 64 GB HBM2e
- “Ponte Vecchio” a.k.a. Data Center GPU Max:
  - 128 Xe cores, each with 16 $\times$ 256-bit vector units
  - 400MB L2 cache, 64GB HBM2 with 8192-bit bus
  - shipping now, but limited software support

Intel

Intel is pushing their Data Parallel C++ implementation of SYCL (an “open standard” that no-one else is adopting)

- part of Intel’s OneAPI software which aims to support all hardware platforms
- translation code (from Codeplay) enables execution on NVIDIA and AMD GPUs
- I have no experience with it, but Intel has a bad record of pushing novel hardware/software for a few years then abandoning it, so I fully expect them to axe their new Data Center GPU Max chips
- their standard C/C++ compilers and MKL libraries remain very good for multithreaded/vectorized CPU execution

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 words

- 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 of the needs of ML

  Even as the gaming market shrinks, 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 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