

Prof. Simon McIntosh-Smith

Head of the HPC research group

University of Bristol, UK

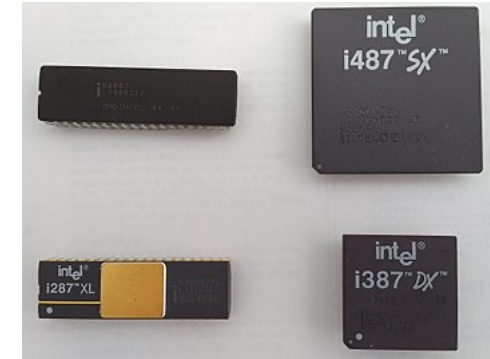
Twitter: [@simonmcs](https://twitter.com/simonmcs)

Email: simonm@cs.bris.ac.uk

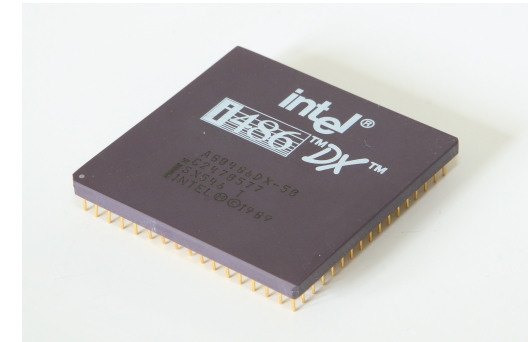


Heterogeneous Computing: past, present and future

Heterogeneity has been around since the dawn of computing

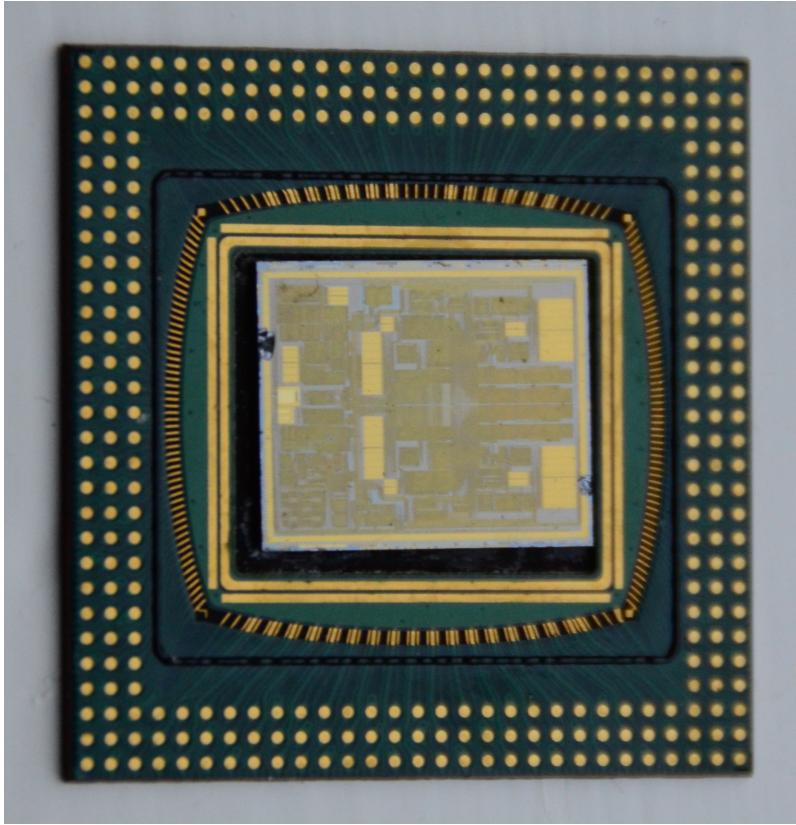


Intel floating point co-processors, 1980s

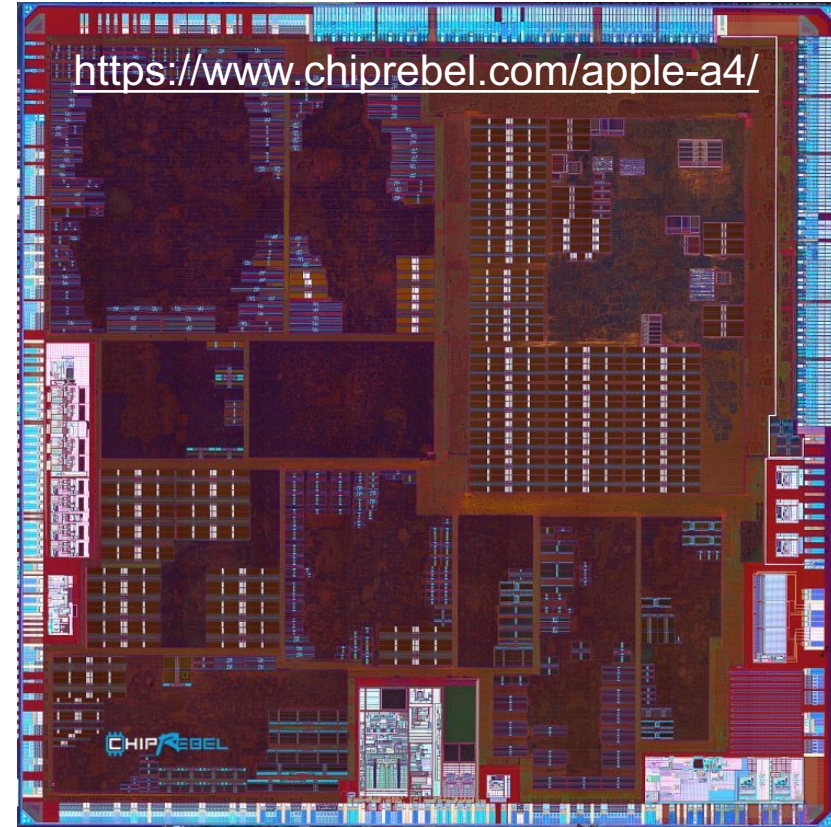


Intel 486 with integrated f.p., c.1989

Other areas of computing had been heterogeneous for years...



Inmos/STMicro "Chameleon" ST40
Dual 64-bit cores, dual issue SIMD,
accelerators for video, audio c.1996.



Apple A4 (first in-house design), c.2010.
iPhone 4, integrates CPU, GPU and other
accelerators.

The modern era of accelerators dawned 15 years ago...

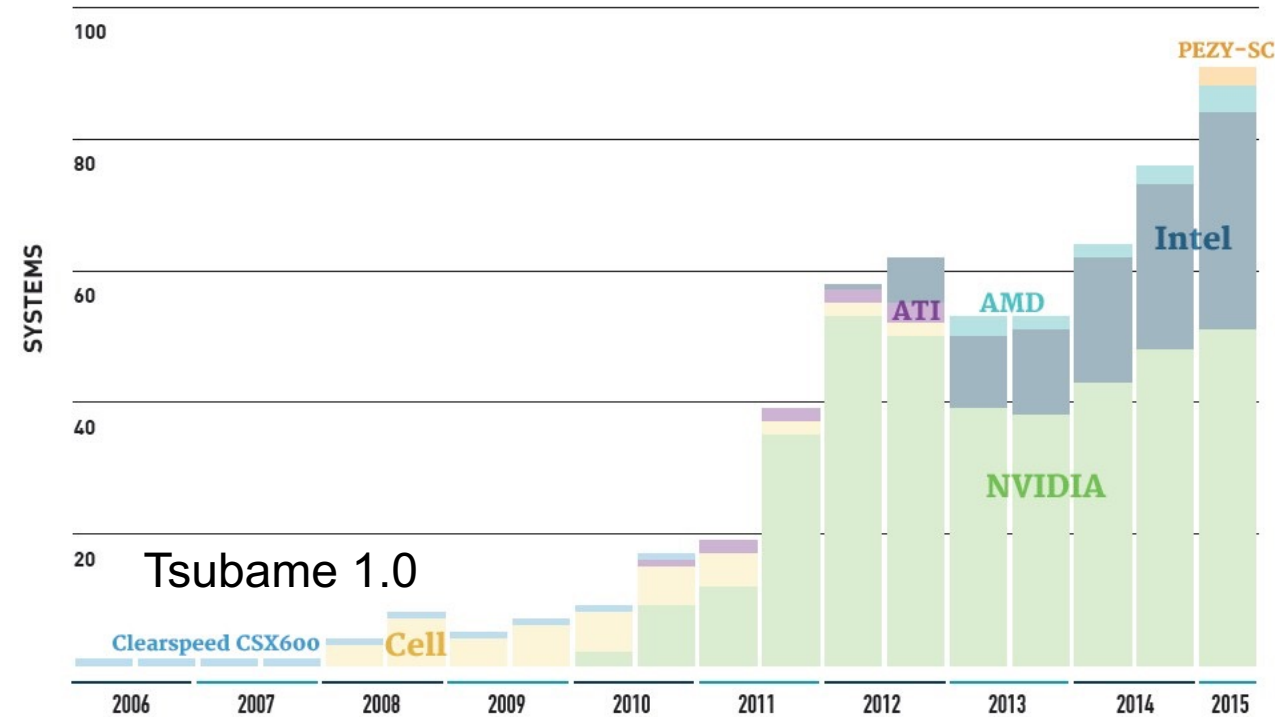


ClearSpeed e620 (2006):

- 80 GFLOP/s 64-bit
- 1GB DRAM with ECC
- 35W

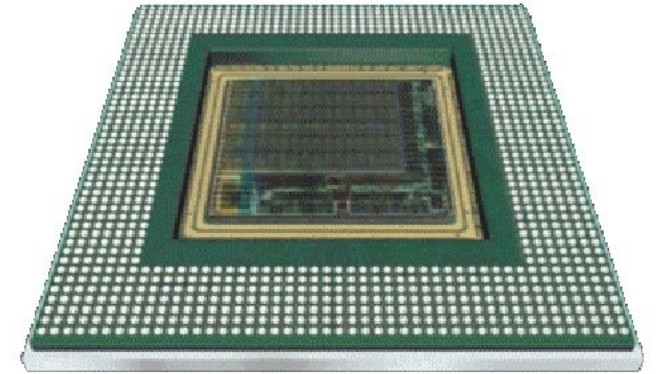
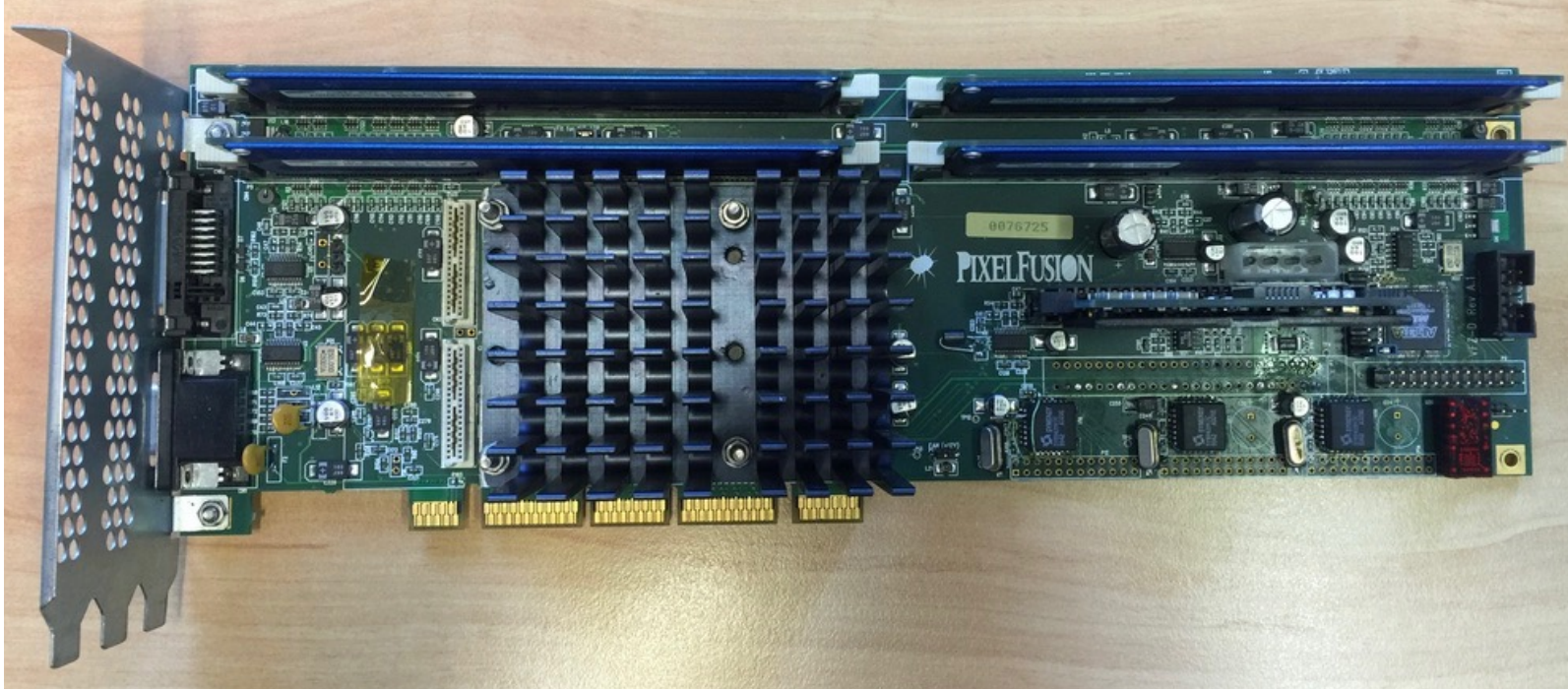


ACCELERATORS/CO-PROCESSORS



<https://www.top500.org>

But the seeds had had been sown at the turn of the millennium...



PixelFusion F150, developed in Bristol c.1999
Predecessor to ClearSpeed
The world's first fully programmable GPU, 1,536-way parallel SIMD.

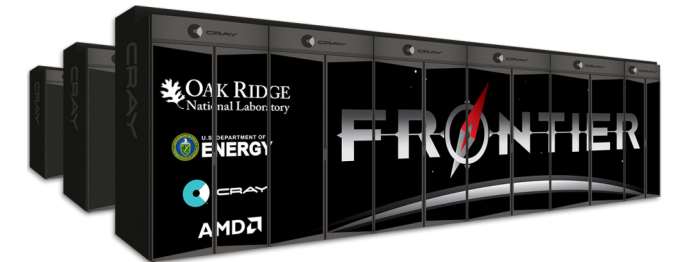
Heterogeneous considerations

- **Motivation:** *energy efficiency*
 - Wider, slower → more energy efficient
 - Counteract dark silicon problem
- **Drawback:** *radically different programming models*
 - From 1 thing to program to 2 or more
 - We were used to this for graphics
 - No universal agreement on the programming model though...

Next-generation supercomputers largely heterogeneous

The coming generation of Exascale systems will include a diverse range of architectures at massive scale, most (but not all) heterogeneous:

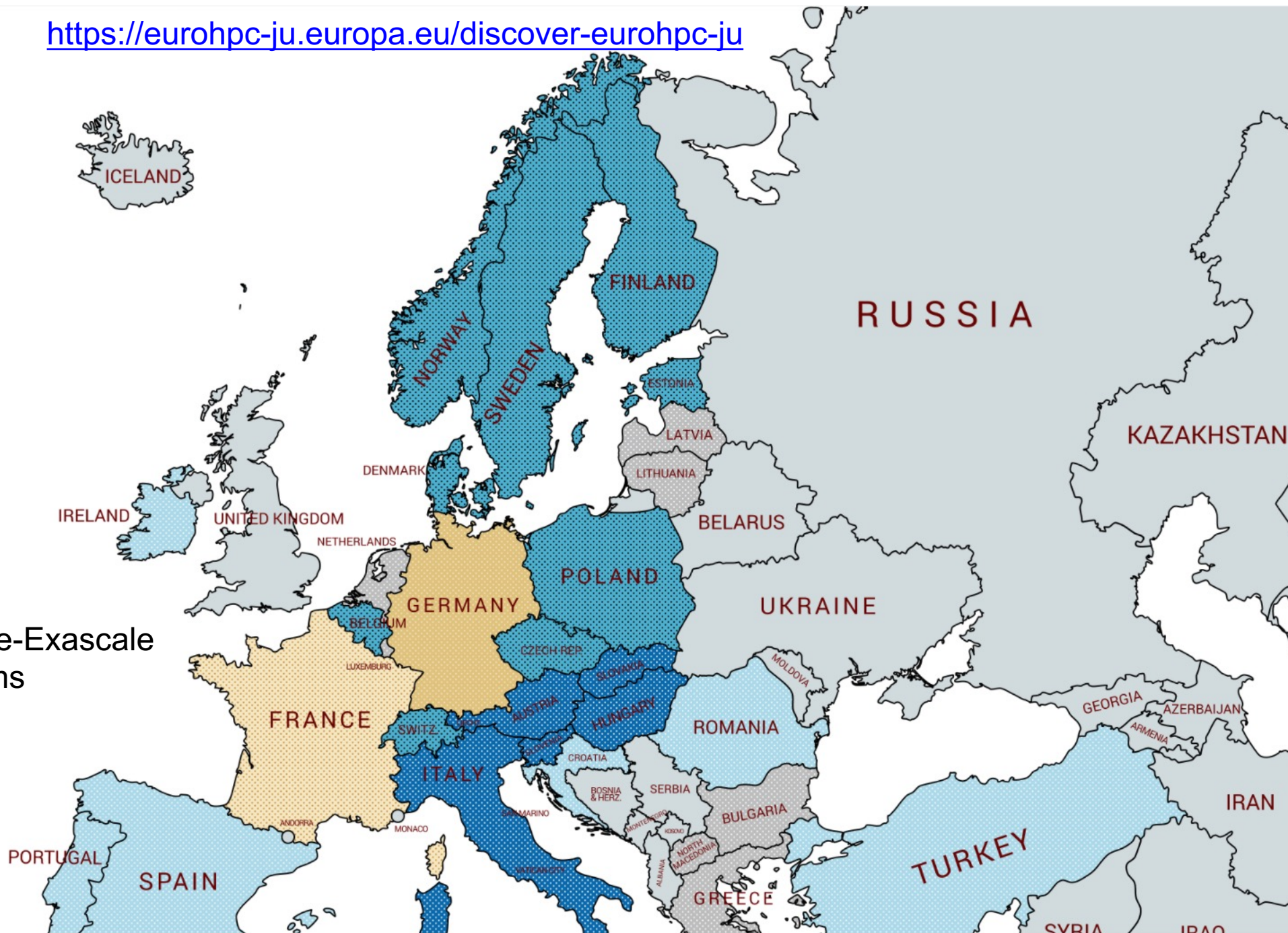
- **Fugaku:** Fujitsu A64FX Arm CPUs
- **Perlmutter:** AMD EYPC CPUs and NVIDIA GPUs
- **Frontier:** AMD EPYC CPUs and Radeon GPUs
- **Aurora:** Intel Xeon CPUs and Xe GPUs
- **El Capitan:** AMD EPYC CPUs and Radeon GPUs



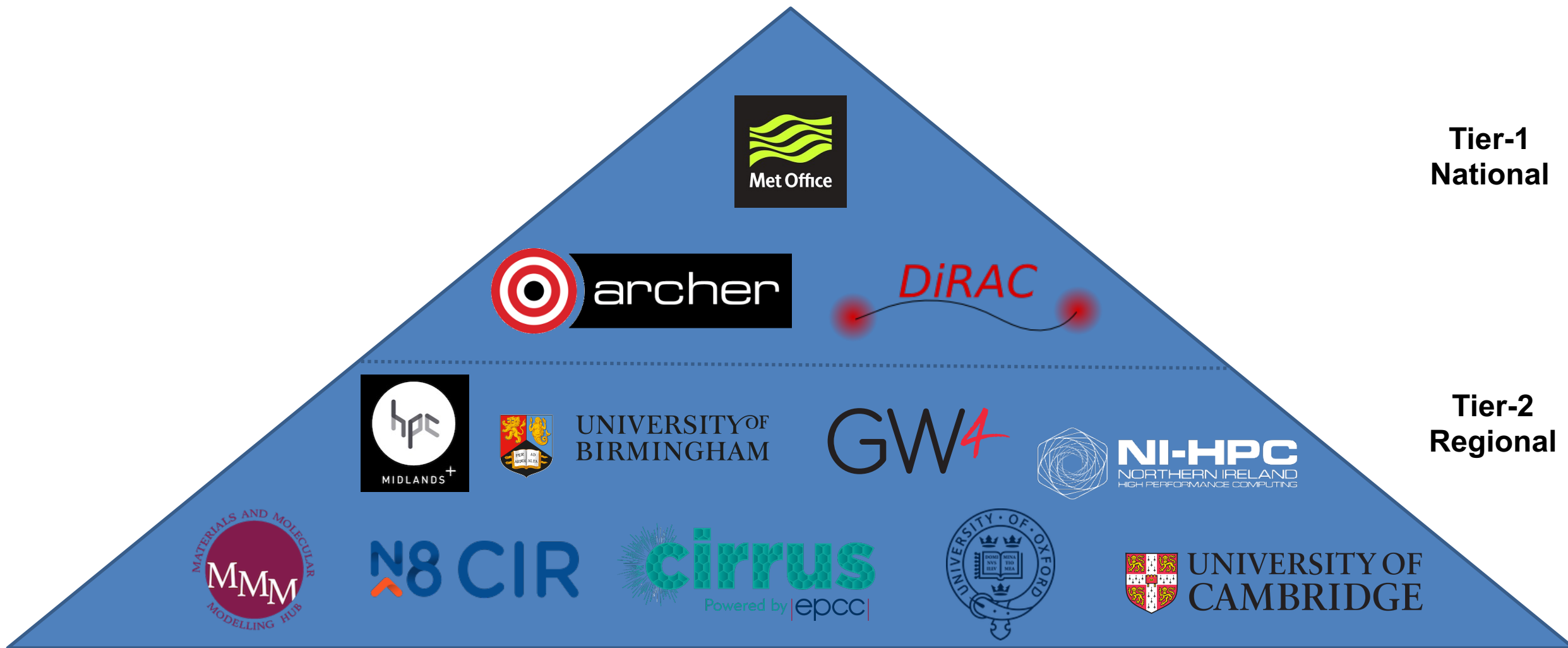
- Pre-exascale – Finland led consortium
- Pre-exascale – Italy led consortium
- Pre-exascale – Spain led consortium
- Exascale – Germany
- Exascale – France
- Other EuroHPC countries

<https://eurohpc-ju.europa.eu/discover-eurohpc-ju>

EuroHPC includes 3 pre-Exascale
and 5 Petascale systems



The UK's HPC service ecosystem is intentionally diverse



Isambard 2

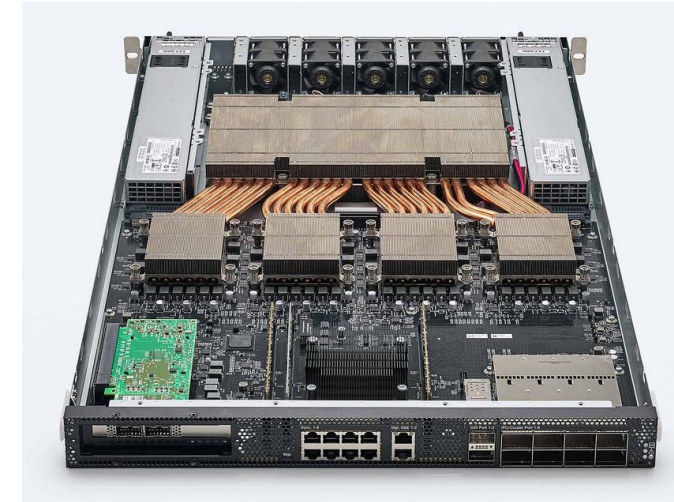
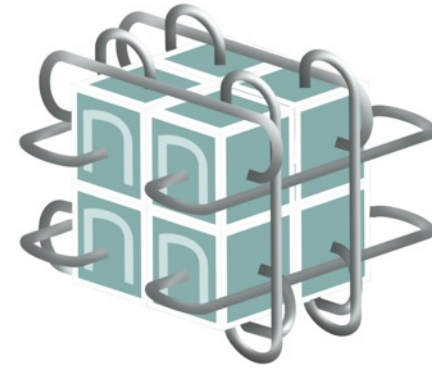
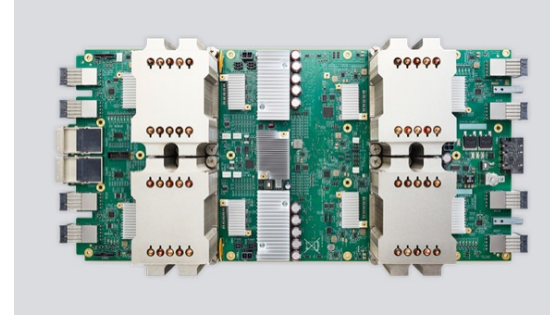
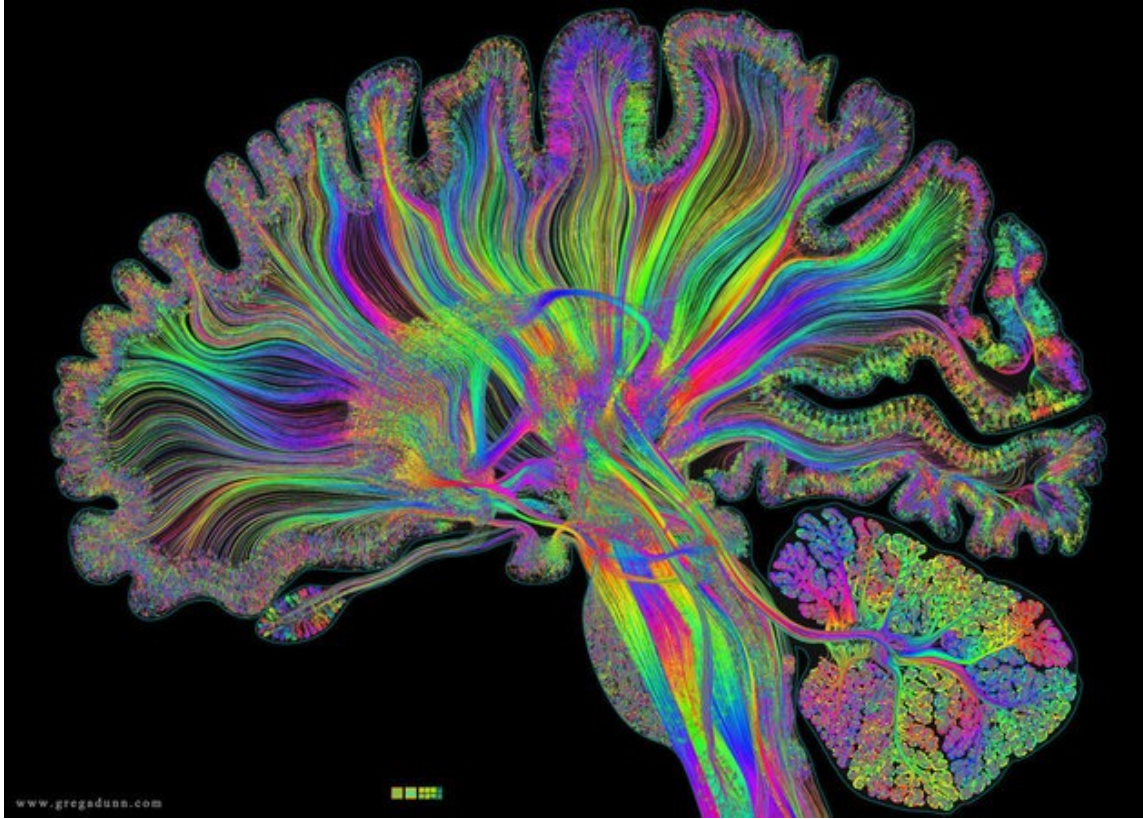
- **21,504** Armv8 cores (336n x 2s x 32c)
 - **Marvell ThunderX2 32 core @2.5GHz**
- 3,456 core Fujitsu A64fx system
- >600 registered users
- Includes a “Multi Architecture Comparison System (MACS)”
 - Includes interesting CPUs and GPUs:
 - AMD Rome, Intel Cascade Lakes, IBM POWER9
 - NVIDIA V100 and P100 GPUs
 - Currently adding AMD Milan, Intel Icelake, NVIDIA A100 GPUs and AMD Mi100 GPUs



ExCALIBUR programme in the UK tackling the challenge

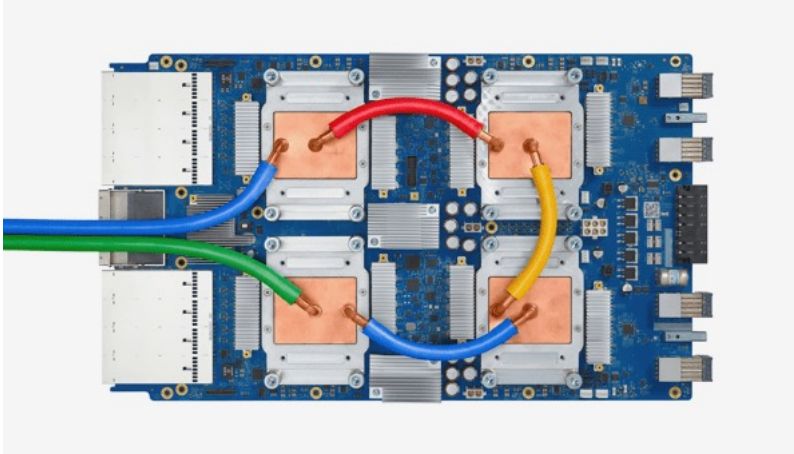
- See talks earlier today from Dr Elizabeth Bent (UKRI), Dr Rob Akers (UKAEA) and Dr Nigel Wood (Met Office)
- £46M over 5 years, focused on getting the UK's science codes Exascale ready
- Addressing heterogeneity by multiple routes:
 - Separation of concerns
 - Domain Specific Languages (DSLs)
 - Performance portable programming languages, etc.

Emerging architectures for AI / Machine Learning



Google's Tensorflow Processing Unit (TPU), GraphCore, Intel's Nervana

Google's Tensor Processing Units



Cloud TPU v3:

420 TFLOP/s

128 GB HBM

\$2.40 / TPU hour

V4 supposedly improves
performance by 2.7x

Cloud TPU v3 Pod:

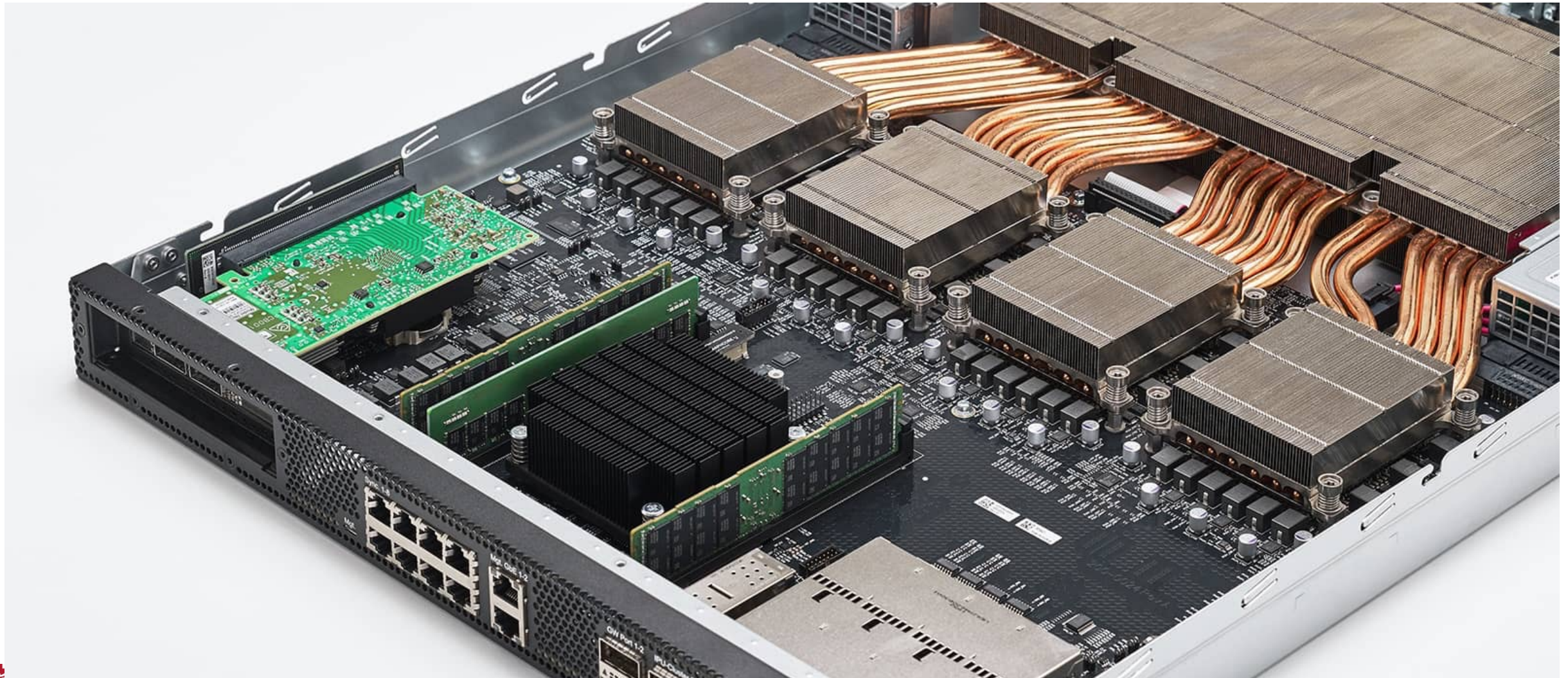
100+ PFLOP/s

32 TB HBM

2-D toroidal
mesh network



Graphcore is already onto their 2nd generation “IPU”

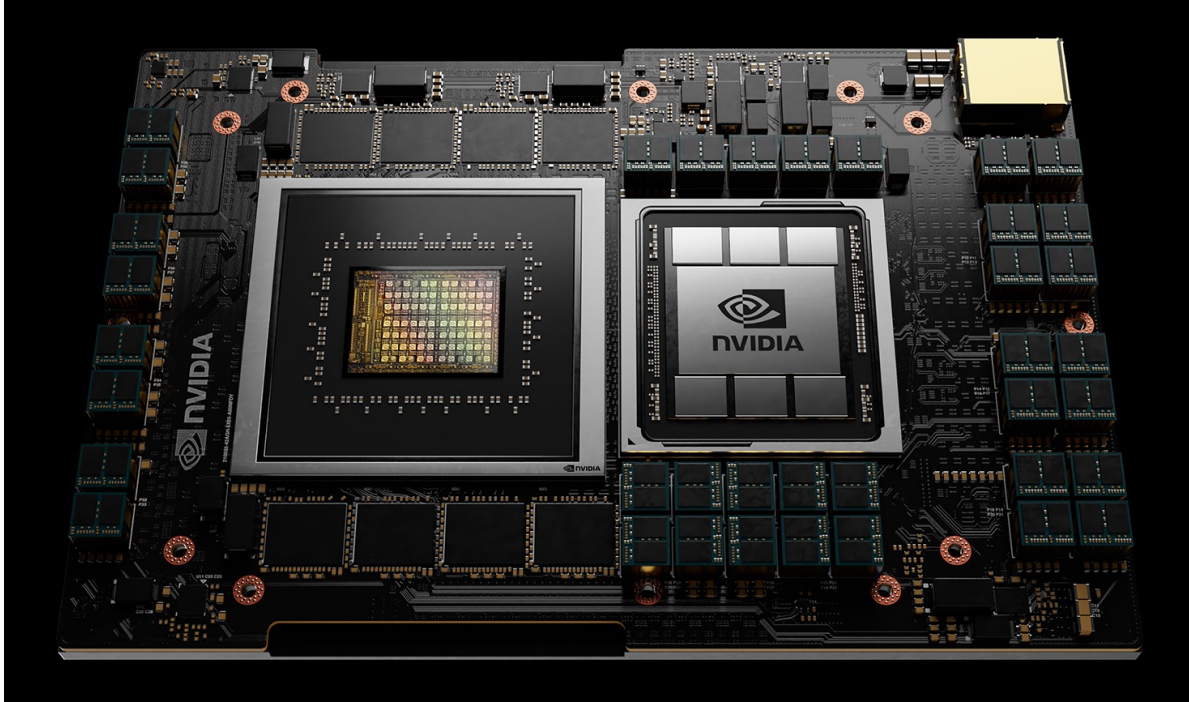


Graphcore IPU-M2000

- 4 x Colossus MK2 GC200 IPUUs in a 1U box
 - 1 PetaFLOP “AI compute” (**16-bit FP**)
 - 5,888 processor cores, 35,328 independent threads
 - Up to 450 GB of exchange memory (off-chip DRAM)
-
- 59.4B 7nm transistors in 823mm²
 - 900MB of on-chip fast SRAM per IPU (3x first gen.)
 - 250 TFLOP/s AI compute per chip, 62.5 TFLOP/s single-precision



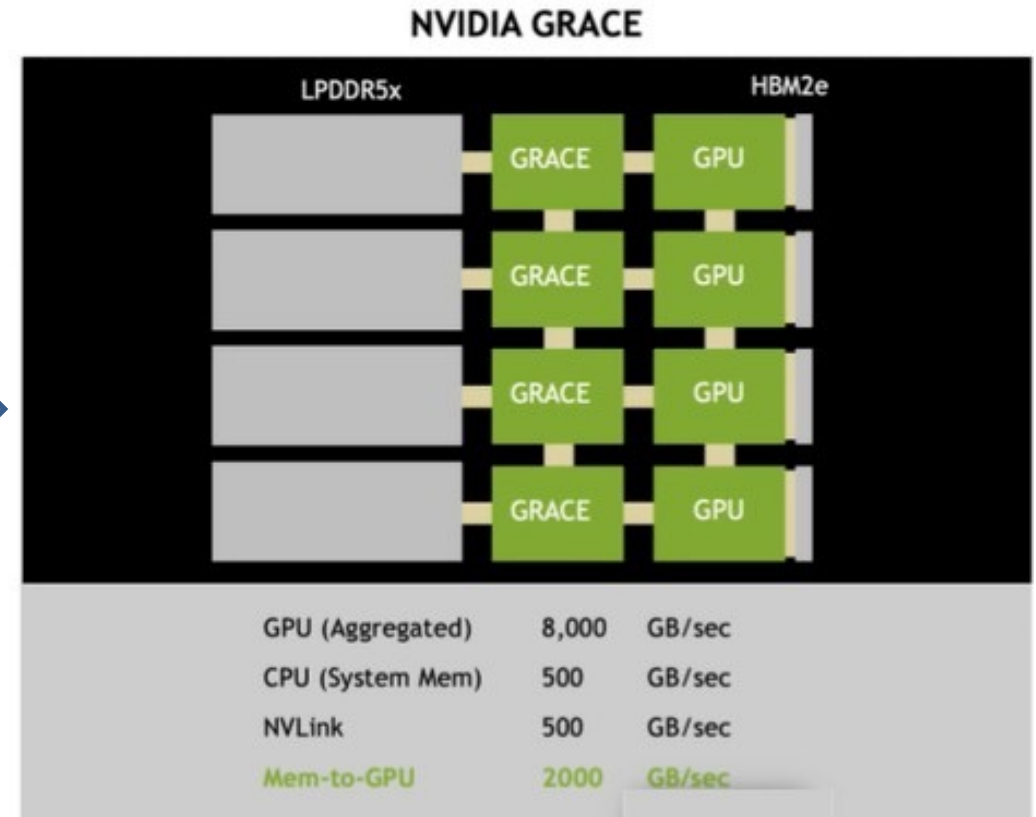
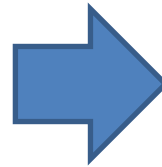
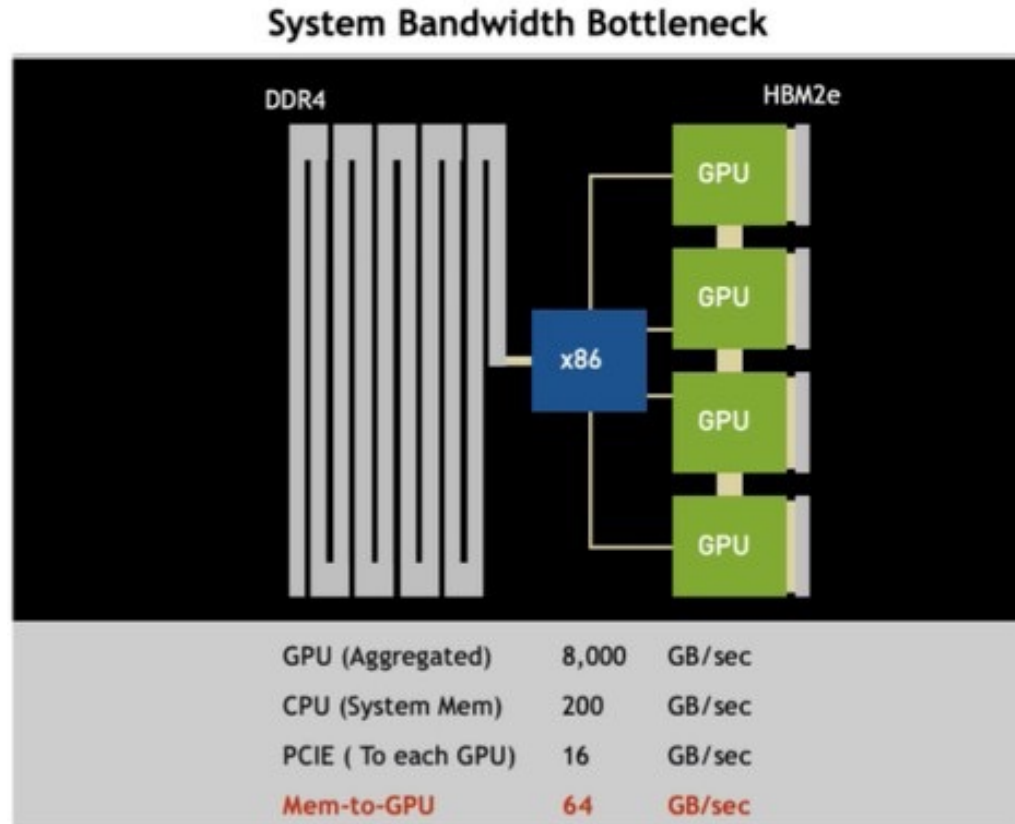
CPUs and GPUs becoming more tightly coupled



- NVIDIA announced their own Arm CPUs, “**Grace**”
- 900GB/s interconnect between the CPU and GPU
 - >10x fastest PCIe
- Very high memory bandwidth for a CPU
 - >500GB/s
- Shipping 2023 with their next-gen GPUs

<https://www.nextplatform.com/2021/04/12/nvidia-enters-the-arms-race-with-homegrown-grace-cpus/>

More balanced intra-node interconnects



Three of the big issues facing parallel programming

1. Massive parallelism

- Fugaku has over 7.63 million cores, each with 2x 512-bit wide vectors

2. Heterogeneity

- CPUs, GPUs and more, from multiple vendors
 - Intel, AMD, NVIDIA, Fujitsu, IBM, Amazon, Google, ...
- Non traditional architectures
 - Graphcore IPU, Google TPUs, Cerebras, vector engines, FPGAs, ...

3. Managing complex memory hierarchies

Are heterogeneous systems going to become more diverse?



**“Not necessarily,
mon amis!**

- Hercule Poirot

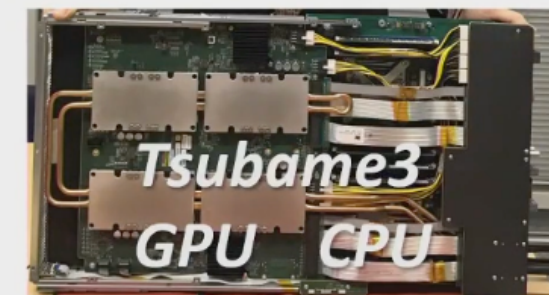


Accelerators vs. Amdahl's Law & Gustafson's Law (3)

Talking: Zoom Session 11

RCCS

- It is no accident that, every successful large-scale accelerated supercomputers (esp. GPU machines) are
 - built with a singular node configuration across the entire machine
 - tight coupling and robust interconnect (& I/O) to sustain maximum bandwidth in/out of accelerator processor
 - dominant processing on the GPU for maximum performance
 - SPMD with very good load balancing (incl. data parallel DNN training)
- Tsubame, Tianhe-2A, Titan/Summit, Piz-Daint, ABCI, Fugaku, Frontier, Lumi, Aurora, ...
- ... and this is the consequence of physical laws, so will continue to be applicable to future machines (no extreme heterogeneity, asynchrony, ...)



End

Why not?

Prof. Satoshi Matsuoka argued strongly for **limiting the diversity of heterogeneity** at SC21 in the panel session “Heterogeneity in Hardware: Opportunities and Challenges for Software and Applications”: <https://sc21.hubb.me/fe/schedule-builder/sessions/877000>. The relevant part of the panel starts 34 minutes in.

Why can't we just have lots of different accelerators for everything?

Because of **software developer productivity**, and because of **scaling and load balance**

1. Developing heterogeneous codes significantly increases the burden on software developers, and thus the cost of developing software.
2. Scaling and load balancing across a system with one type of accelerator already hard enough; with more than one, likely intractable.



Satoshi Matsuoka
@ProfMatsuoka

...

Although **diverse** heterogeneity is good for smartphones, during the Tue [#SC21](#) panel “Heterogeneity in Hardware: Opportunities and Challenges for Software and Applications” & my invited talk Thu morning plenary I will controversially present why that is a BAD idea for modern HPC.

10:30 PM · Nov 14, 2021 · Twitter

7 Retweets 1 Quote Tweet 4



Satoshi Matsuoka @ProfMatsuoka · Nov 14

...

Replying to [@ProfMatsuoka](#)

... and you will find out why machines like Tsubame, Fugaku, Summit are successful while those extrapolating that we have multitudes of heterogeneous customized accelerator components will not be.



1



3



13



Possible alternatives to heterogeneous systems

- What are the most valuable features of GPUs?
 - High FLOP/s per Watt, high peak FLOP/s, high bandwidth memory, high FLOP/s per Dollar, latency tolerance, ...?
- Could these features be integrated into, for example, CPUs?
 - Yes, mostly – see A64fx, Sapphire Rapids HBM, SiPearl Rhea, ...
- What are the main drawbacks?
 - Hard to achieve the same peak FLOP/s per Watt without heterogeneity, however, performance on real codes is a different story...

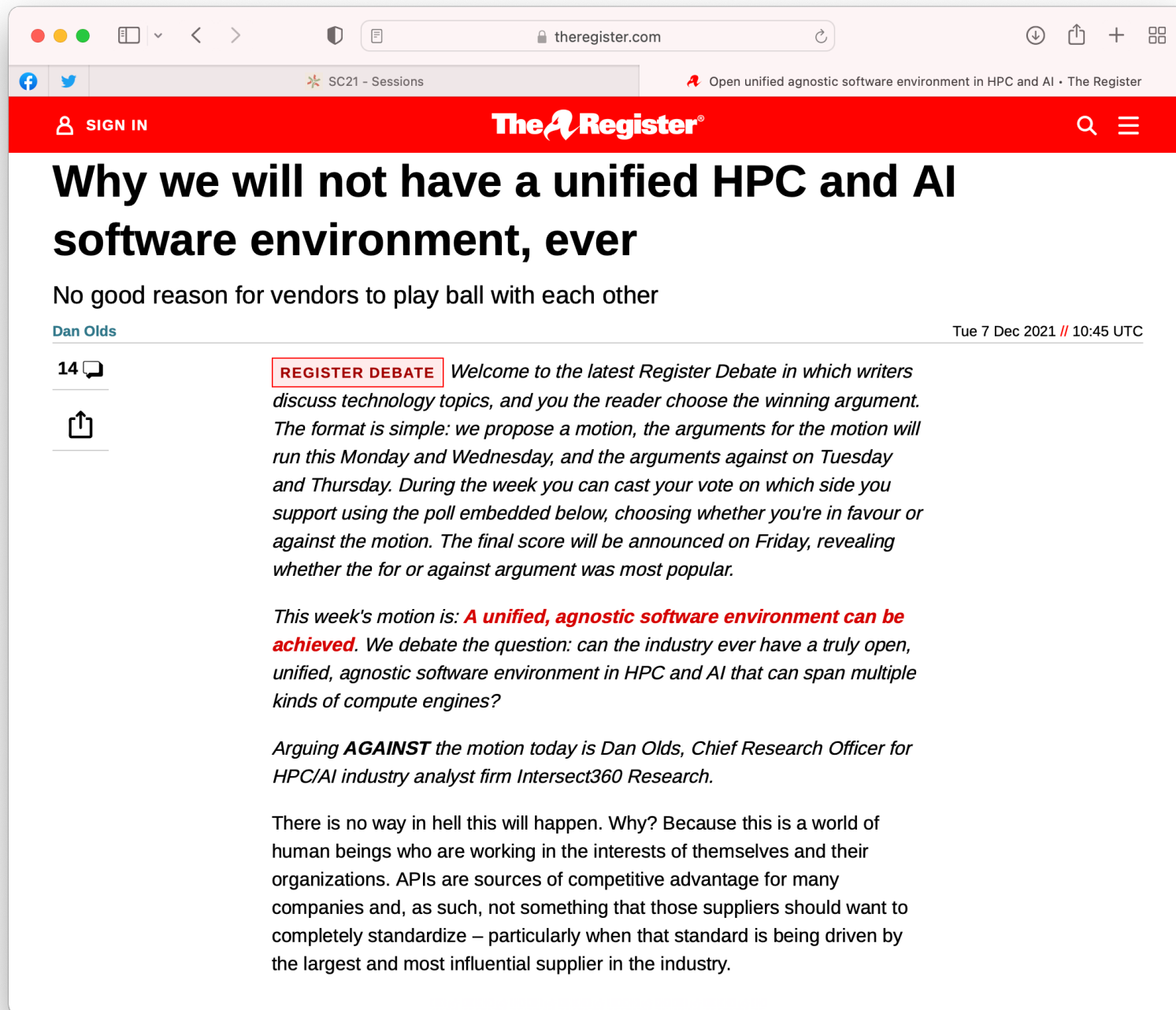
Promising heterogeneous computing developments

- Heterogeneous-aware software stacks (ECP's etc)
- Cross-platform programming standards
 - ISO C++ / Fortran, OpenMP, SYCL, Kokkos, ...
 - See talk by Jeff Hammond and Filippo Spiga on Friday 10th at 10am
- Rate of adoption of heterogeneous systems at the high end
 - All the top 10 systems are heterogeneous, except Fugaku and Sunway TaihuLight
 - 150 systems in the Nov 2021 Top500 now use accelerators (mostly NVIDIA GPUs) – that's 30% of the total list



The future of heterogeneous computing...

- Likely we'll rely on heterogeneous systems for quite some time
 - But modestly diverse, not extremely, diverse
- Closer integration between CPUs and GPUs (cache coherent, in-package etc)
- CPUs will keep integrating some of the best parts from GPUs
 - HBM, lots of cores with wide vectors, in-core accelerators (matrix etc)
- The programming situation is improving, but still has a long way to go
- The long tail of users and codes will not (and should not) go away



The screenshot shows a web browser window with the URL [theregister.com](https://www.theregister.com). The page features a red header with the "The Register" logo and a "SIGN IN" button. The main article title is "Why we will not have a unified HPC and AI software environment, ever" by Dan Olds, dated Tue 7 Dec 2021 // 10:45 UTC. The article is part of a "REGISTER DEBATE" series. The text of the article discusses the challenges of creating a unified software environment in HPC and AI, arguing that vendors have no incentive to standardize because APIs are a source of competitive advantage. The article is part of a debate where readers can vote on whether a unified environment can be achieved.

Why we will not have a unified HPC and AI software environment, ever

No good reason for vendors to play ball with each other

Dan Olds Tue 7 Dec 2021 // 10:45 UTC

REGISTER DEBATE Welcome to the latest Register Debate in which writers discuss technology topics, and you the reader choose the winning argument. The format is simple: we propose a motion, the arguments for the motion will run this Monday and Wednesday, and the arguments against on Tuesday and Thursday. During the week you can cast your vote on which side you support using the poll embedded below, choosing whether you're in favour or against the motion. The final score will be announced on Friday, revealing whether the for or against argument was most popular.

This week's motion is: **A unified, agnostic software environment can be achieved.** We debate the question: can the industry ever have a truly open, unified, agnostic software environment in HPC and AI that can span multiple kinds of compute engines?

Arguing **AGAINST** the motion today is Dan Olds, Chief Research Officer for HPC/AI industry analyst firm Intersect360 Research.

There is no way in hell this will happen. Why? Because this is a world of human beings who are working in the interests of themselves and their organizations. APIs are sources of competitive advantage for many companies and, as such, not something that those suppliers should want to completely standardize – particularly when that standard is being driven by the largest and most influential supplier in the industry.

Key takeaways

- **Increased heterogeneity** is the order of the day (esp. for Exascale)
 - + Better peak FLOP/s per Watt
 - ? Performance per Dollar on real codes
 - Harder to program, reduced developer productivity
 - Lack of vendor agreement on programming models a severe hindrance
- Some of the best features of heterogeneous systems are being integrated into homogeneous ones
- We have a very long tail of developers running thousands of different, complex, continually evolving codes, in dozens of different programming languages and parallelism models → most of this may never be ported to accelerators

For more information

Bristol HPC group: <https://uob-hpc.github.io/>

Email: S.McIntosh-Smith@bristol.ac.uk

Twitter: [@simonmcs](https://twitter.com/simonmcs)

ExCALIBUR: <https://excalibur.ac.uk>

Isambard: <https://gw4-isambard.github.io/>