National Supercomputing in the UK for Modelling and Simulation

Professor Mark Parsons

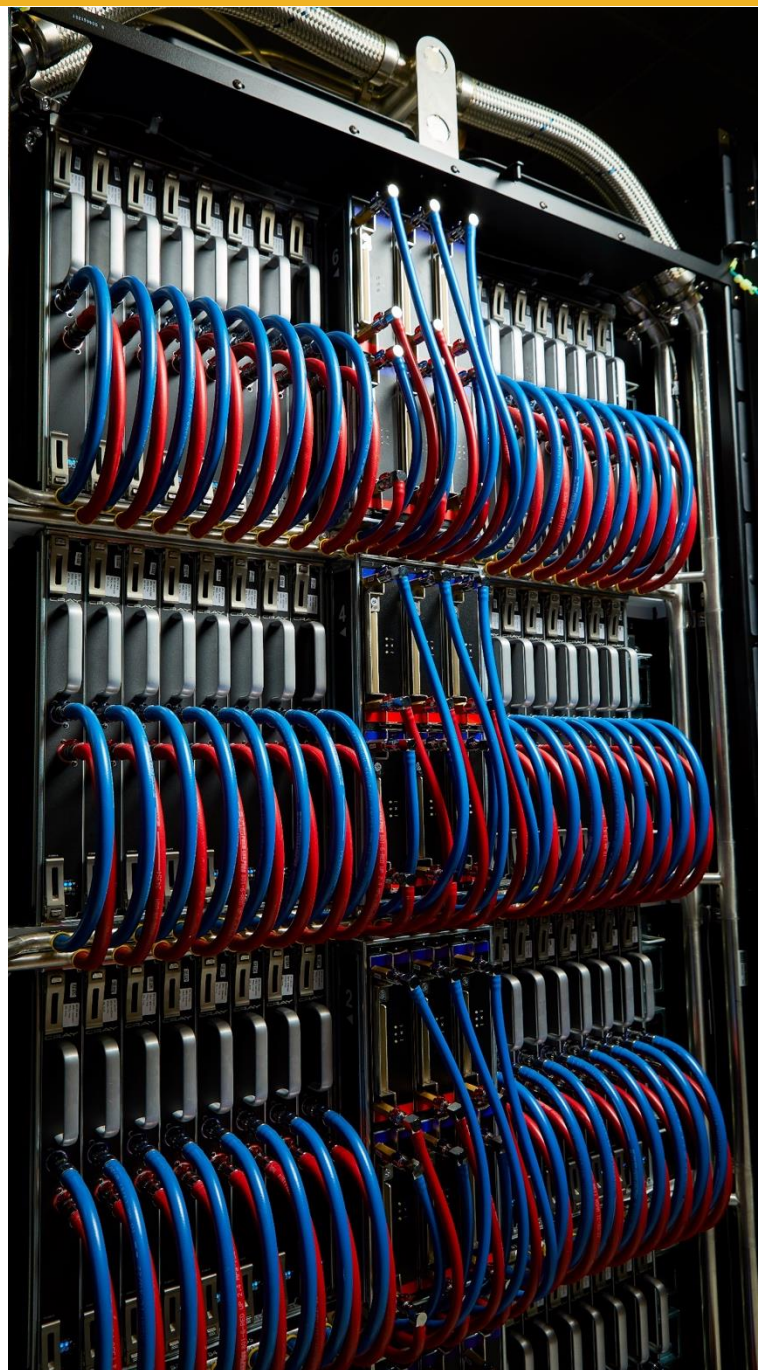
EPCC Director
Dean of Research Computing

epcc

ARCHER2 National Service

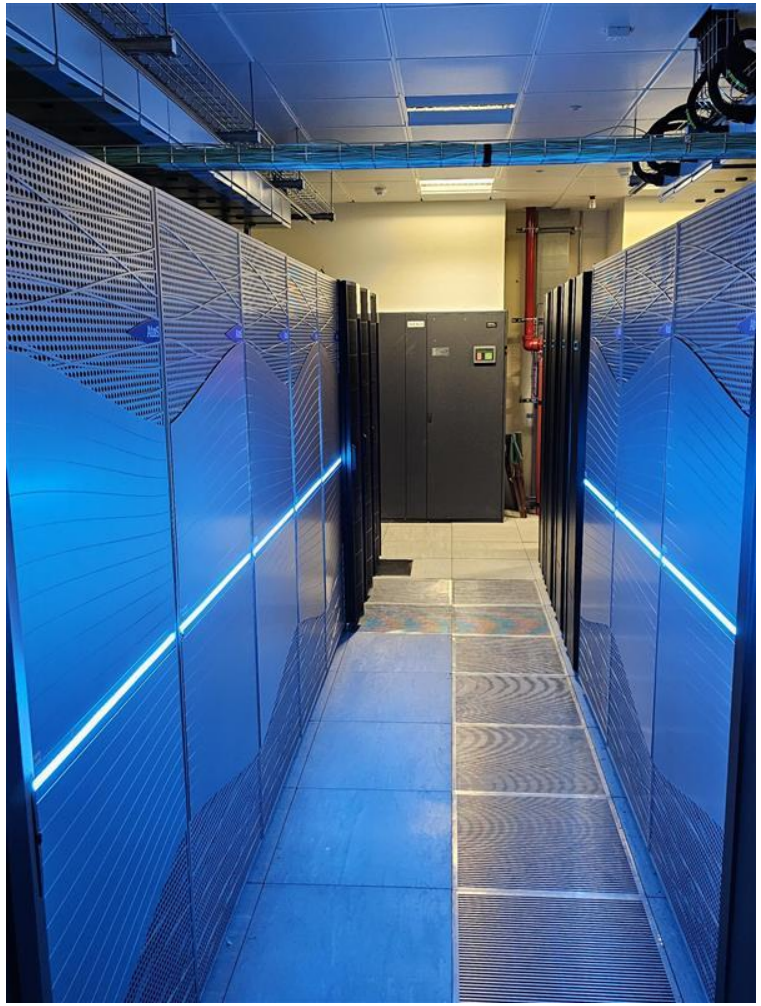
Still the 5th largest cores-only machine in the world

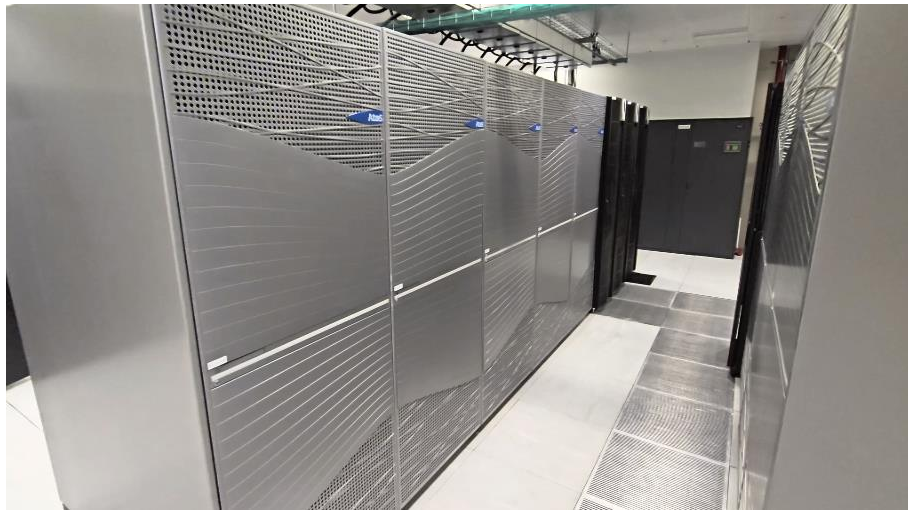
- HPE Cray EX Supercomputer – Serial no. 1!
- 5,860 compute nodes (750,080 CPU compute cores)
- HPE Slingshot 10 interconnect
- Compute nodes:
  - Dual AMD EPYC™ 7742 Processors, 64 cores, 2.25 GHz
  - 256 GiB / 512 GiB memory per node
  - Two 100 Gbps HPE Slingshot interfaces per node
- 4 x ClusterStor L300 Lustre file systems, each 3.6 PB
- 1 PB ClusterStor E1000F solid state storage
  - Available to users via Slurm BB/DW directives
- 4 x NetApp FAS8200A file systems, 1 PB total

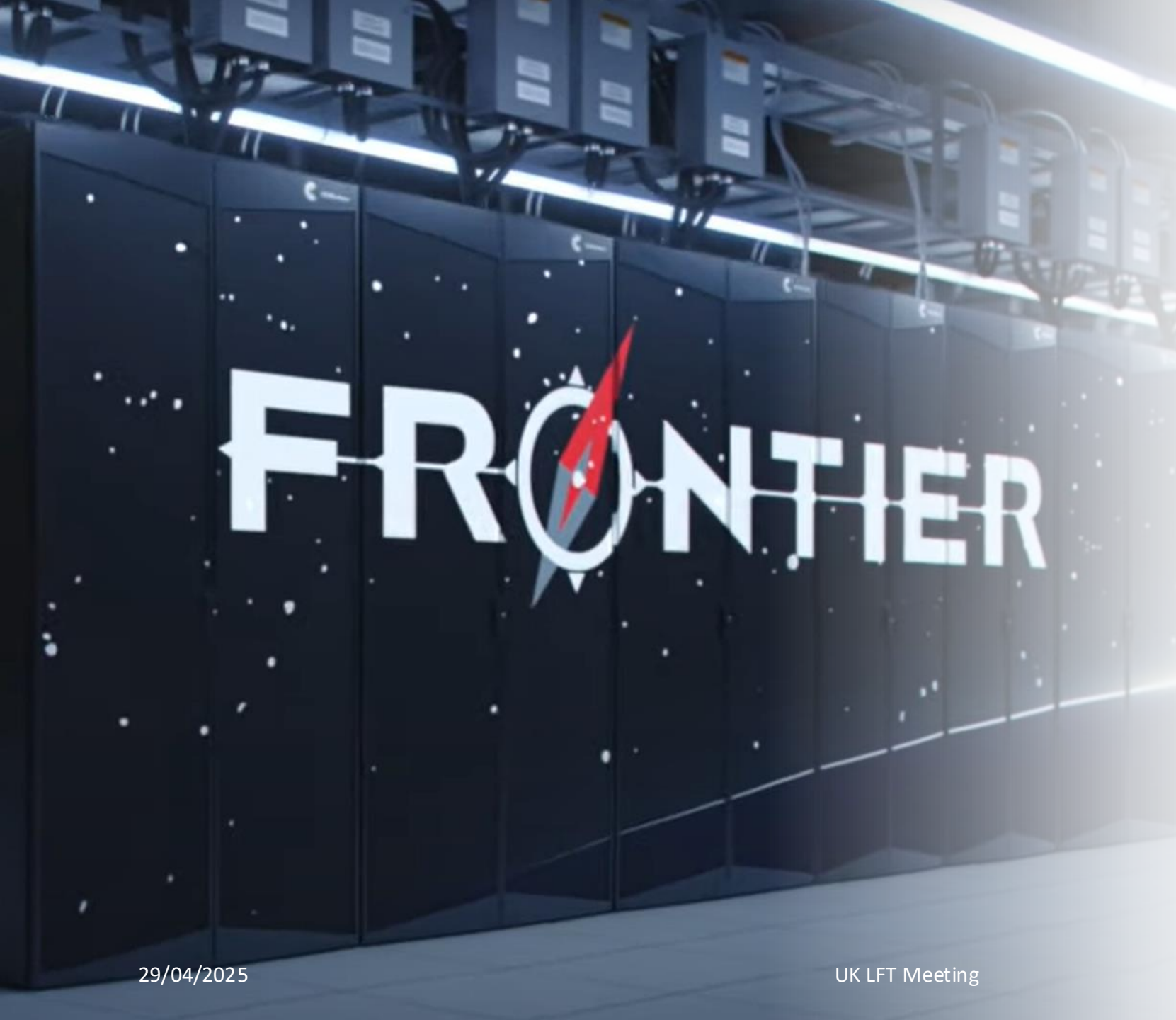


DiRAC Tursa Service

- **Bull Sequana XH2000** system from Eviden
- 712 Nvidia GPUs
  - A100 40/80GB
- Each node has
  - 4 x A100 GPUs
  - 2 x AMD CPUs
    - Rome and Milan
  - 4 x 200Gbps Infiniband NICs







Frontier was the first Exascale system

- Frontier became first Exascale system on Top500 at ISC 2022 conference in Germany
- HPE Cray EX system with AMD Trento CPUs and MI250X GPUs
- Slingshot-11 network
- 1.194 Exaflop/s sustained on HPL
- 37,888 GPUs
- 606,208 cores
- 22.7 MW

Exascale supercomputing worldwide

41 million cores!

Country or Region	Timescale	Detail
Japan 	2020	Fugaku – based on Fujitsu A64FX Arm processor- [illegible]F
China 	2021	Two systems in operation - next generation Sunway and Tianhe 3 system. Probably more systems by now.
USA 	2022 2024 2024	Frontier – Oak Ridge National Lab – 1.35EF Aurora – Argonne National Lab – 1.01EF El Capitan – Lawrence Livermore National Lab – 1.74EF
Europe 	2022/3 2025 2026	Pre-Exascale systems in Finland, Italy and Spain – ~0.3EF Jupiter system in Germany Summer 2025 – 1.0EF Alice Recoque system in France – 1.0EF + 13 AI Factories

EuroHPC will have invested **€10 billion by 2027**. **Finland**'s new project has just received **€612m** – UK GDP is 11 times that of Finland …

epcc

Exascale in the UK has taken far too long

2011

CRESTA Project

2011 - 2014

EPCC's first EU-funded Exascale project

2017

EuroHPC established

March 2017

Subsequently UK decides not to participate

2019

Exascale Project Working Group

March 2019

EPWG starts work on business case to Government

Exascale Project Working Group

August 2019

First business case draft completed

2021

Large-scale Computing Review

September 2021

Government Office of Science publishes report

2022

Future of Compute Review

June 2022

Review announced by Government

ARCHER2 opening

February 2022

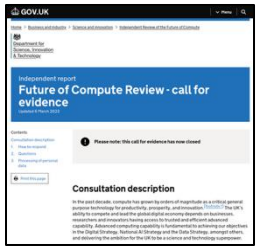
Prime Minister Johnson opens ARCHER2

2023

Future of Compute Review

March 2023

Review published recommends Exascale



Budget 2023

March 2023

Chancellor announces £900m investment in Exascale and AI

UKRI formal approvals begin

March 2023

Business Case revision and procurement preparations

Exascale hosting site

October 2023

Edinburgh confirmed as hosting site. CR4 grant application submitted

2023

BBC announces Exascale shelved

August 2024

UK Government "shelves" UK Exascale

2024

Project stalled

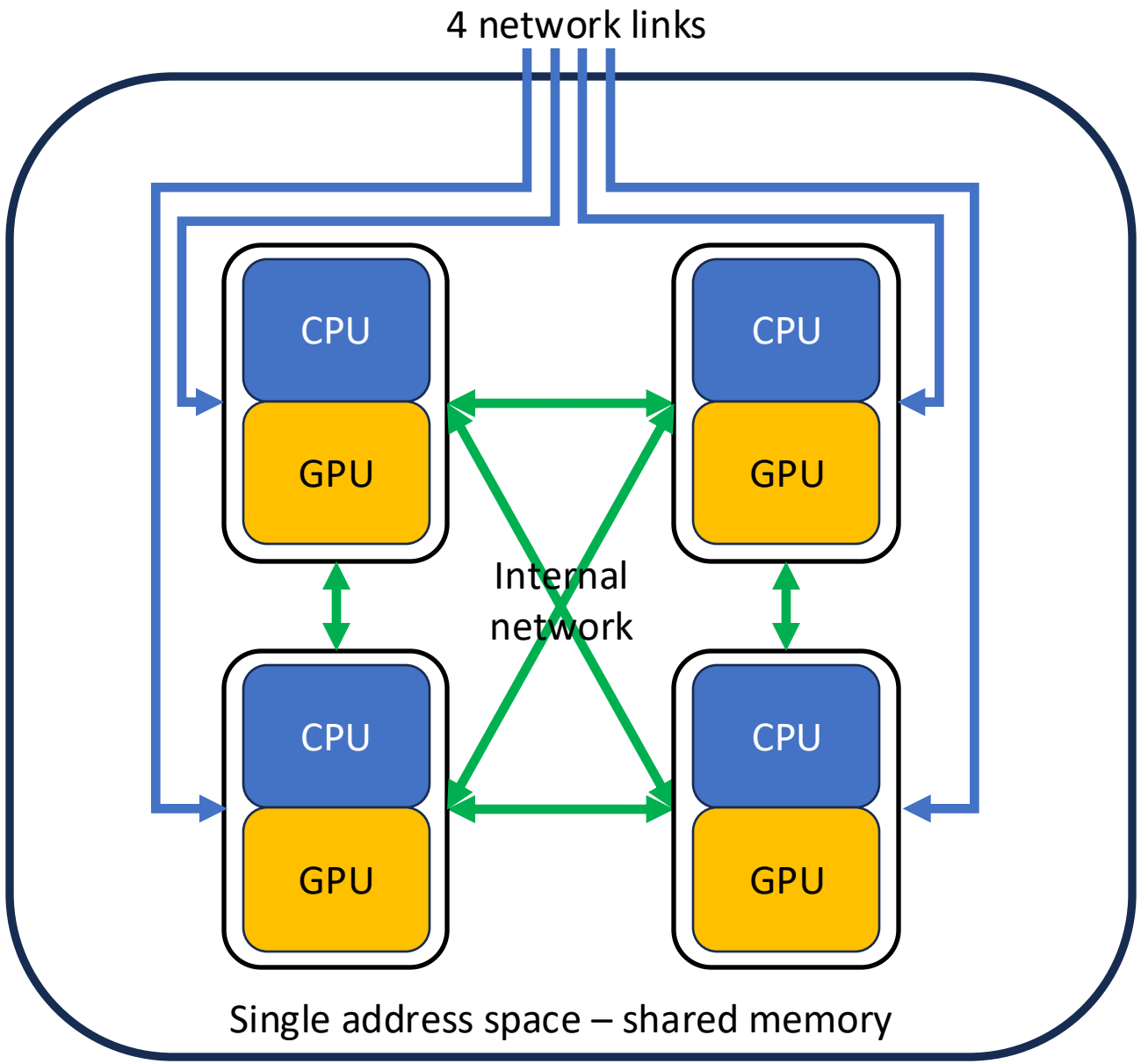
Today

High risk of no UK National Supercomputer Service after Nov 2026

2025

A "Typical" Exascale server node

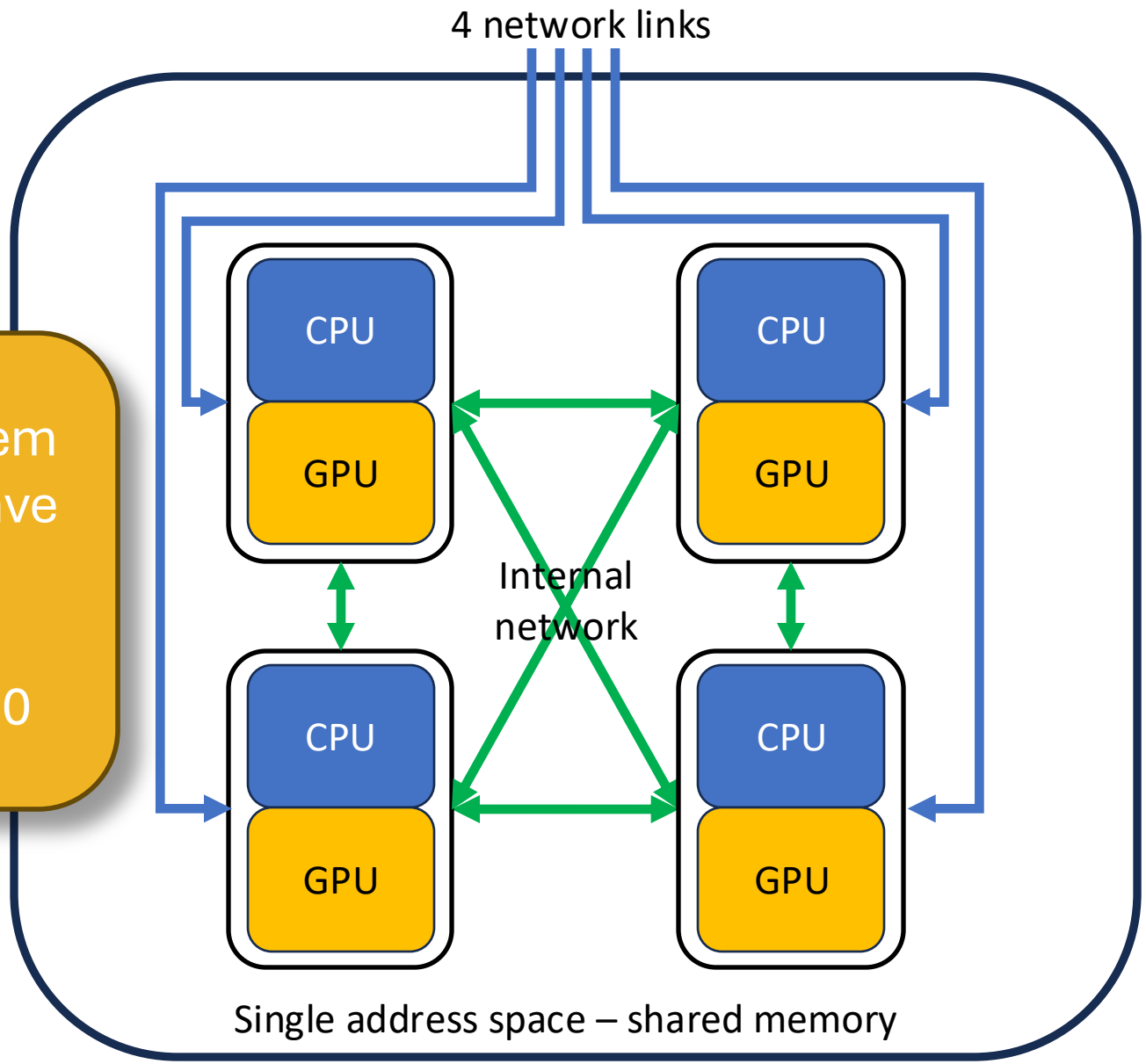
- 4 CPU/GPU modules
- Connected by very fast internal network
- 4 network links for interconnect topology
- Shared cache coherent memory
- Key innovations
  - Very high-performance GPUs
  - Power performance ratio 10X better than ARCHER2
  - Single memory space
  - Coherent shared memory
- Key challenges
  - Programming models
  - GPU utilisation optimisation



A "Typical" Exascale server node

- 4 CPU/GPU modules
- Connected by very fast internal network
- 4 network links
- Shared cache c
- Key innovations
  - Very high-pe
  - Power perfor than ARCHER
  - Single memory space
  - Coherent shared memory
- Key challenges
  - Programming models
  - GPU utilisation optimisation

An Exascale system built today will have
6,000 nodes
24,000 GPUs
MI300A or GH200



Exascale programming challenge

- All Exascale systems only possible because of use of **GPUs**
- For most performance either **NVIDIA CUDA** or **AMD ROCm** is used
  - In practical terms very little difference between them – and AMD HIP uses either
- Biggest challenge is keeping floating points units fed with data
  - All of the Supercomputing Conference prize winners focus on cache optimisation and custom data layouts to deliver performance
  - Reasonable performance through **OpenMP** or **OpenACC** frameworks
- In my opinion too much time is wasted on trying to hide complexity and deliver performance portability

In the rest of the world "ModSim" and AI are converging …

- In USA, EU and Japan no distinction is made between systems for AI and for modelling and simulation
- AI's needs are now dominating supercomputing
- But AI is also bringing new ways of thinking to numerical computing … some good, some bad
- … the focus on 64bit flop/s is being de-emphasised by NVIDIA

- … but the coupling of AI with numerical computing opens completely new opportunities

Investment in Exascale computational science in UK

- Exascale programme had a large software programme planned
- Funding for 200+ FTEs of developer effort per annum
- Mixture of
  - Expanded eCSE programme – eCSE++
  - Longer software development projects for specific applications / consortia
  - “Mission” driven development programmes
- Exascale Software Programme should run at least the lifetime of the service
- Unfortunately – unclear if UKRI are still planning to fund this programme

ACF Data Centre first opened in 1970s

4 Computer Rooms
4 Plant Rooms

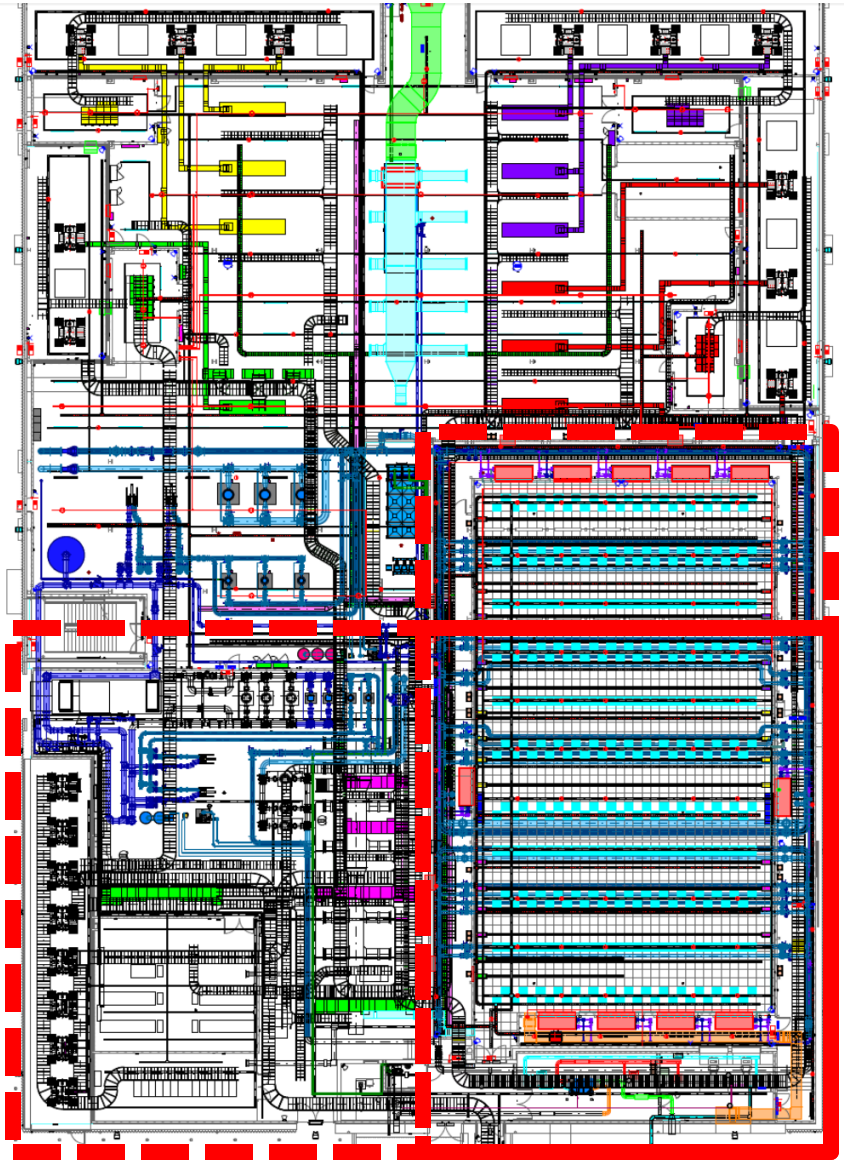
76MW power to site

100% REGO certified renewable electricity

Operationally Net Zero




PROPOSED CO-ORDINATED GROUND FLOOR



DESIGN CRITERIA

Summer	32°C dB, 50%RH
Chilled Water System	16/24°C
Computer Room Cooling	28/40°C

REPURPOSED CHILLED WATER CIRCUIT TO PROVIDE HALL AIR COOLING AND BACK UP TO COMPUTER COOLING WATER SYSTEM.

NEW PHASE 2 (18 MW) COMPUTER COOLING WATER SYSTEM TO SERVE ROW 1,4-9 CABINETS.

PHASE 1 (6 MW) COMPUTER COOLING WATER SYSTEM RECONFIGURED FROM EXISTING PLANT TO SERVE ROW 2-3 CABINETS.

Annual PUE 1.038

What next for national supercomputing?

- Government has announced ARCHER2 extension until November 2026
- Currently no news on Exascale project – now calling it “next National Supercomputer service” – **we’re post-Exascale**
- Technology is moving on at pace. By early 2027 new GPUs will be available with different characteristics – some vendors are focussing on low-precision AI rather than numerical computing
- We’d expect to be able to get 1EF with between 8,000 – 12,000 GPUs – a huge increase in performance (but a reduction in nodes 2,000 – 3,000)
- … suggests to me **we should set our target at 2EF+ to ensure scientific throughput is maintained**

MACHINE HISTORY
EPCC hardware timeline

2021 DiRAC ATOS Sequana XH2000 (4,416 cores/456 GPUs)

2021 ARCHER2 HPE Cray Ex (750,080 cores)

2019 Fulhame, HPE Apollo 70 (4,096 Arm ThunderX2 cores)

2019 NEXTGenIO, Custom Fujitsu PRIMERGY design (1,632 cores, 102TB persistent memory)

2018 DiRAC HPE SGI 8600 (35,424 cores/ 32 GPUs)

2017 Cirrus SGI ICE XA Cluster (13,248 cores/ 152 GPUs)

2014 ARCHER Phase 2: Cray XC30 (118,080 cores)

2013 ARCHER Phase 1: Cray XC30 (72,192 cores)

2012 UK-RDF (23PB)

2011 DiRAC IBM BlueGene/Q (98,304 cores)

2011 EDIM1 (240 cores, 750 TB disk)

2011 HECToR Phase 3: Cray XE6 (90,112 cores)

2010 HECToR Phase 2b: Cray XE6 (45,544 cores)

2009 HECToR Phase 2a: Cray XT4 (22,656 cores)

2007 HECToR Phase 1: Cray XT4 (11,328 cores)

2007 FHPCA Maxwell (64 FPGAs)

2006 HPCx Phase 3: IBM p5-575 (2,560 processors)

2005 HPCx Phase 2a: IBM p5-575 (1,536 processors)

2005 IBM BlueGene/L (2,048 cores)

2004 QCDOC (14,464 processors)

2004 HPCx Phase 2: IBM p690+ (1,600 processors)

2002 HPCx Phase 1: IBM p690 (1,280 processors)

2002 Sun Fire E15K (52 processors

2002 Sun Fire 6800 Cluster (66 processors

1997 Hitachi SR2201 (8 processors)

1997 Cray T3E (344 processors)

1996 Cray J90 (10 processors)

1995 Meiko CS-2 (22 processors)

1994 Cray T3D (512 processors) + CRAY Y-MP

1992 Meiko i860 CS (16 processors)

1991 TMC CM-200 (16k 1-bit processors)

1990 Meiko i860 CS (64 processors)

1988 AMT DAP608 (1,024 1-bit processors)

1986 Meiko T800 CS (400 processors)

1982 ICL DAPs (2 x 4,096 processors)

Pre-EPCC — EPCC