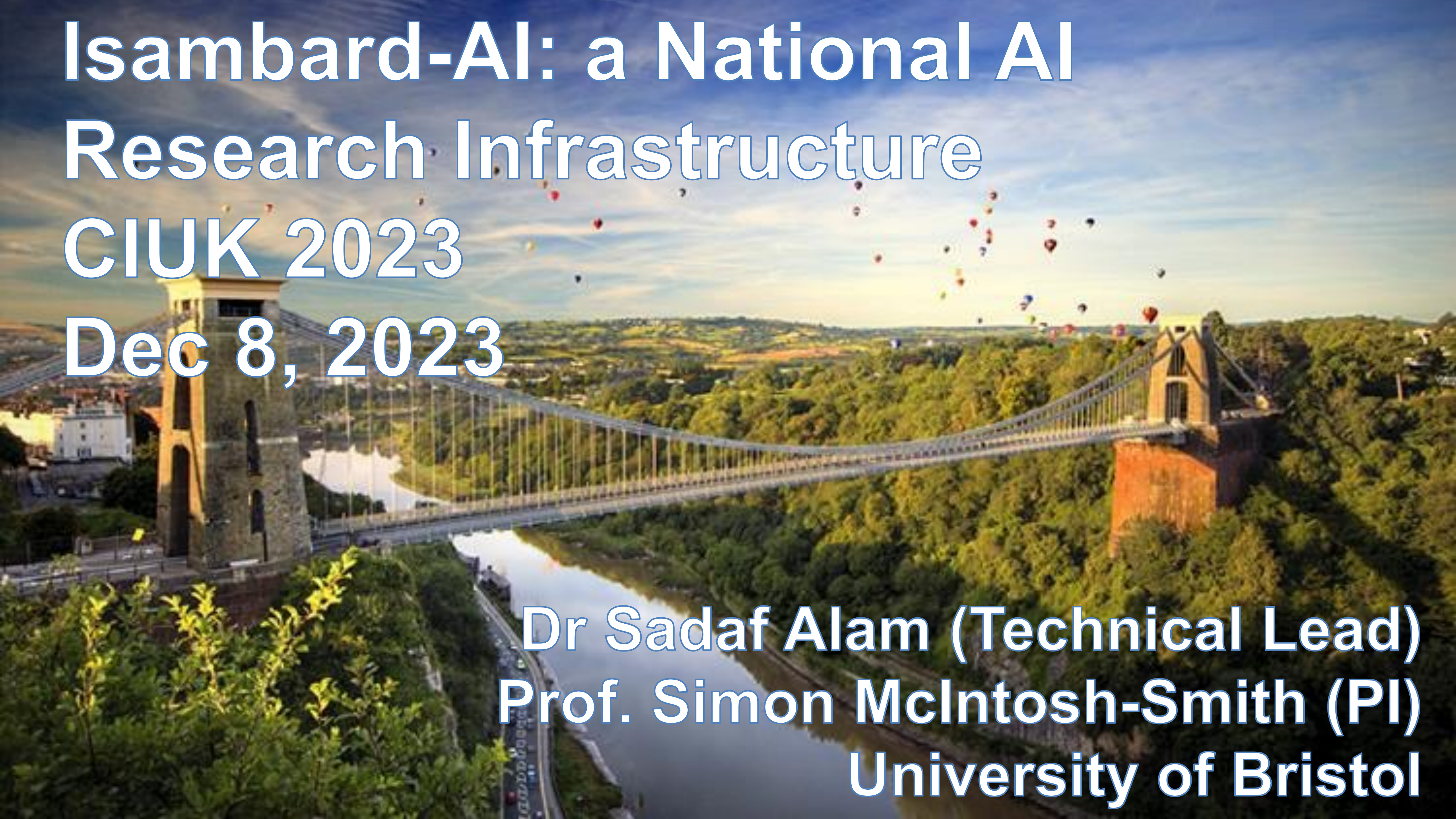


Isambard-AI: a National AI Research Infrastructure

CIUK 2023

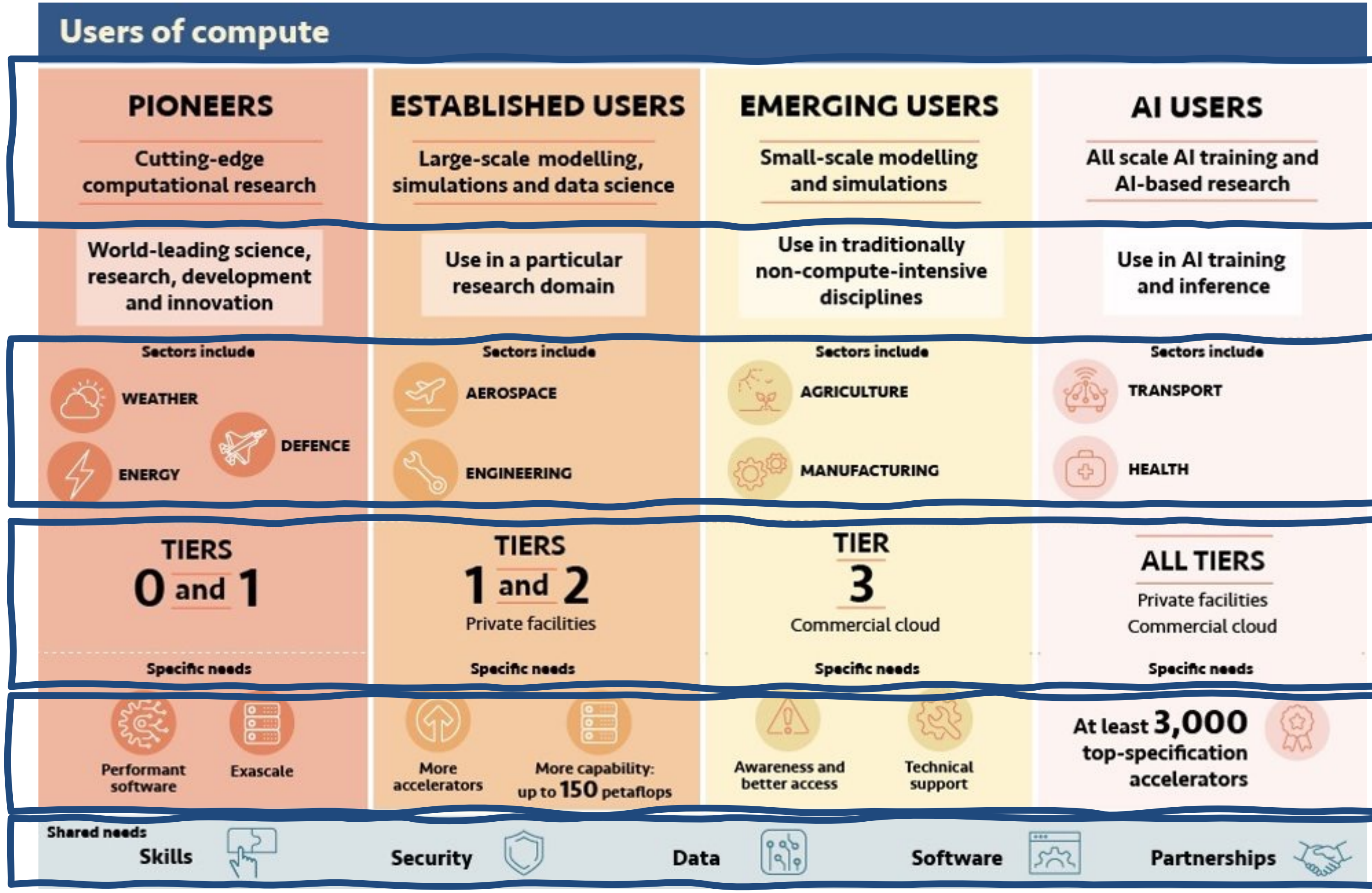
Dec 8, 2023

Dr Sadaf Alam (Technical Lead)

Prof. Simon McIntosh-Smith (PI)

University of Bristol

Design Specifications for AI Research Resource (RR)



Users of compute

PIONEERS	ESTABLISHED USERS	EMERGING USERS	AI USERS
Cutting-edge computational research	Large-scale modelling, simulations and data science	Small-scale modelling and simulations	All scale AI training and AI-based research
World-leading science, research, development and innovation	Use in a particular research domain	Use in traditionally non-compute-intensive disciplines	Use in AI training and inference
Sectors include WEATHER DEFENCE ENERGY	Sectors include AEROSPACE ENGINEERING	Sectors include AGRICULTURE MANUFACTURING	Sectors include TRANSPORT HEALTH
TIERS 0 and 1	TIERS 1 and 2 Private facilities	TIER 3 Commercial cloud	ALL TIERS Private facilities Commercial cloud
Specific needs Performant software Exascale	Specific needs More accelerators More capability: up to 150 petaflops	Specific needs Awareness and better access Technical support	Specific needs At least 3,000 top-specification accelerators

Shared needs: Skills · Security · Data · Software · Partnerships

AI users base cover all levels
familiar access interfaces (browser)
framework abstractions (Jupyter)

Opportunities for multiple sectors
non-traditional & non-digital
enabling AI and HPC workflows

Usage is interactive & batch
X-as-a-Service where X is software
platform and infrastructure

5000+ Nvidia GH acc (Isambard AI)
1000+ Intel PVC acc (DAWN)

AI RR collaboration areas

University of BRISTOL

References: Independent Review of The Future of Compute: Final report and recommendations, March 2023; National AI Strategy - AI Action Plan, July 2022; £300 million to launch first phase of new AI Research Resource

AI RR Collaboration Workgroups (Isambard AI and DAWN)

Cybersecurity for Digital Research Infrastructure

Federated Identity and Access Management

AI and ML Environments and Frameworks

Data motion and management

User support, training and outreach

Nov 27-28 kick-off meeting between Bristol and Cambridge
Established collaborative space
Excalibur proposal for federation and linking of two AI RRs

Anatomy of a national AI Research Resource (AIRR)

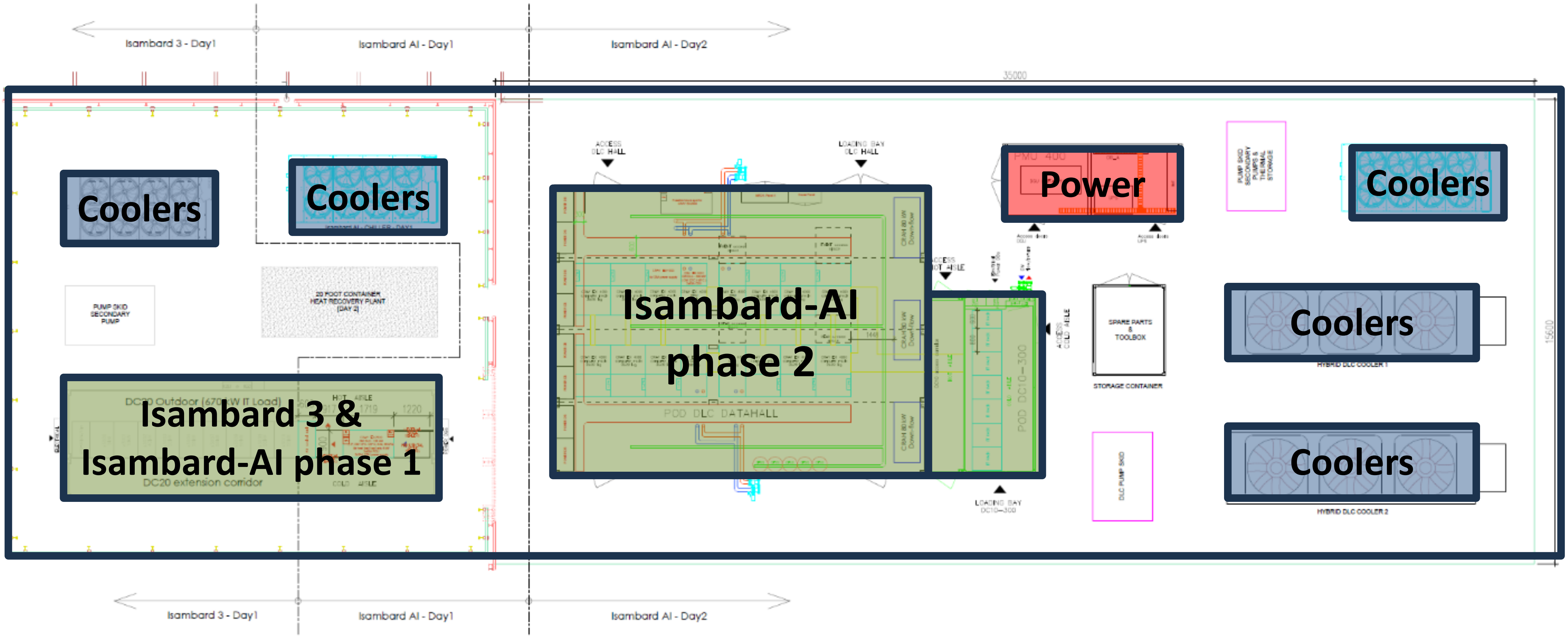
- Access methods as similar as possible to existing resources
  - E.g. cloud-style, Jupyter notebooks, as well as HPC style, e.g. ssh & batch
- Fully **federated**, with support for true **multi-tenancy**
- Highly scalable resource, providing from 1 to 1,000s of GPUs
- Flexible, fast storage solution, optimized both for high IOPs and small file accesses, as well as bulk parallel file access for BW
- System architecture optimized for a wide-range of AI workloads:
  - Training, including next-generation VLLMs
  - Inference
  - Hybrid workflows, including AI+HPC

Isambard-AI: a national AI research infrastructure

- Isambard-AI will form the main part of the UK's national AI Research Resource (AIRR), over £300M investment in total
- New GPU system to be added alongside Isambard 3
- 5,448 NVIDIA GH200 Grace-Hopper GPUs
  - >21 ExaFLOPs for AI (8-bit), >200 PetaFLOPs for HPC
- Comfortably in top 5 fastest open AI systems, top 10 for HPC globally
- Large, fast storage system, all-flash (~25 PB)
- Software stack optimized for AI workflows and cloud-style usage
- ~5MW operating power, direct liquid cooling with heat reuse
- Modular Data Centre (MDC) technology for efficient deployment

https://www.gov.uk/government/news/bristol-set-to-host-uks-most-powerful-supercomputer-to-turbocharge-ai-innovation

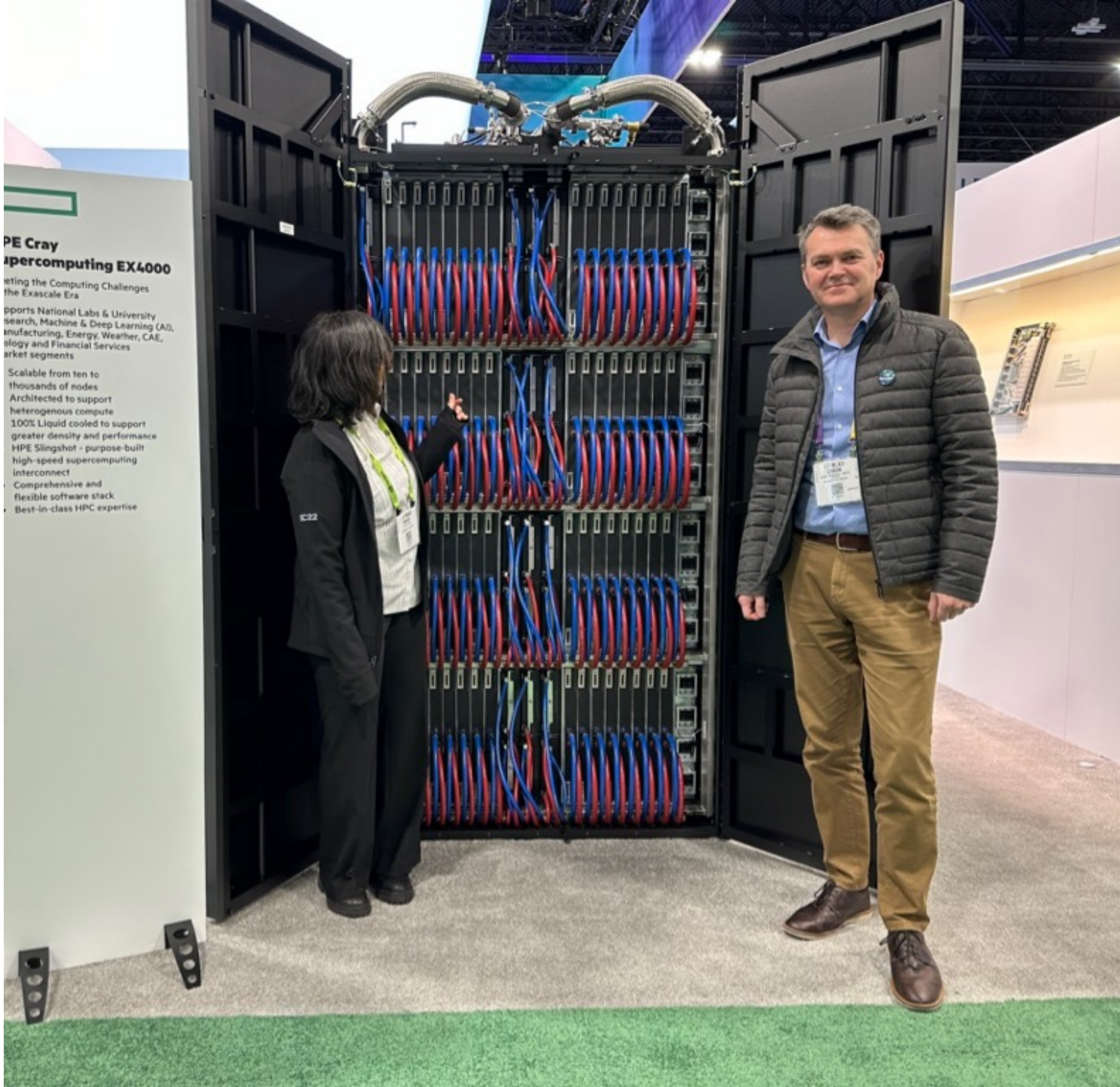
BELL
BELL CIVIL ENGINEERING
BELL PLANT
01179 533654
WC
oakland



University of
BRISTOL

- An HPE EX2500 system
- We install one of these, with 168 GPUs in it, in Isambard 3 in March next year
- Early access from May
- Equivalent to about 500 NVIDIA A100 GPUs for AI
- Picture from IEEE/ACM SuperComputing in Denver, November 2023 (SC23)

- Sadaf and Simon visited the HPE booth at SC23 last week to see the Isambard-AI hardware
- The HPE EX4000 main system
- We deploy 12 of these, each with 440 GPUs in them (5,280 in total), in the main Isambard-AI POD in Jul-Aug 2024
- Picture from IEEE/ACM SuperComputing in Denver, November 2023 (SC23)


PE Cray
upercomputing EX4000
eting the Computing Challenges
the Exascale Era
pports National Labs & University
search, Machine & Deep Learning (AI),
anufacturing, Energy, Weather, CAE,
ology and Financial Services
arket segments
Scalable from ten to
thousands of nodes
Architected to support
heterogenous compute
100% Liquid cooled to support
greater density and performance
HPE Slingshot - purpose-built
high-speed supercomputing
interconnect
Comprehensive and
flexible software stack
Best-in-class HPC expertise

University of
BRISTOL

Summary

- The next-generation of flexible, user-oriented AI services is coming
- Optimised for LLM development, training, inference, and hybrid
- ~5,500 latest generation NVIDIA GPUs
- Fast, multi-modal storage, all solid-state
- True multi-tenancy support
- Designed to be able to evolve the software environment over time
- Ultra **energy efficient to** meet NetZero goals
- Enabling next-generation AI sciences

University of BRISTOL

Call to action

- We'll have one of the world's fastest, most advanced AI supercomputers in Bristol from next summer
- What does this make possible?
- What could you use this for which was previously unachievable?

- We'll be hiring a world-leading support team throughout 2024 to help run Isambard-AI – let us know if you're interested!

University of BRISTOL