

Isambard-AI's Early Years and Future Roadmap

Driving open and sovereign supercomputing

Sadaf Alam · Richard Gilham · Ashish Tom · Jake Watson

Bristol Centre for Supercomputing (BriCS), University of Bristol

Oct 2023

Oct 2024

Oct 2025

Oct 2026

Oct 2027 · projection

Bristol Centre for Supercomputing

BriCS was formed in Q3 2024 within the University of Bristol's Faculty of Engineering to manage the Isambard Digital Research Infrastructure (DRI) projects.

It builds on the GW4 Isambard project of Bath, Bristol, Cardiff and Exeter, which set out in 2016 to prove that Arm-based technology was ready for supercomputing.

Services

Isambard-AI

~5,500 NVIDIA GH200 Superchips in a modular data centre

Isambard 3

~400 nodes of NVIDIA Grace–Grace Superchips

Digital Labs

software-defined, OpenStack-based cluster for the University

AIRRPortal

federated user and project management for AI and HPC services

National AI Data Infrastructure Network–*Coming*

Complementing AIRR compute with storage services for open & sensitive data

Inference Technology Watch–*Coming*

Platform for exploring novel technologies and software stacks for inference

Isambard-AI at a glance



5,448 GH200

HPE Cray EX, Slingshot 11, 4 NICs per node

>21 EF FP8

>250 PF FP64 within under 5 MW

All-flash storage

20 PiB ClusterStor + 3.5 PiB VAST

~5,500 users

academia, industry and government

The journey: three stories, five Octobers

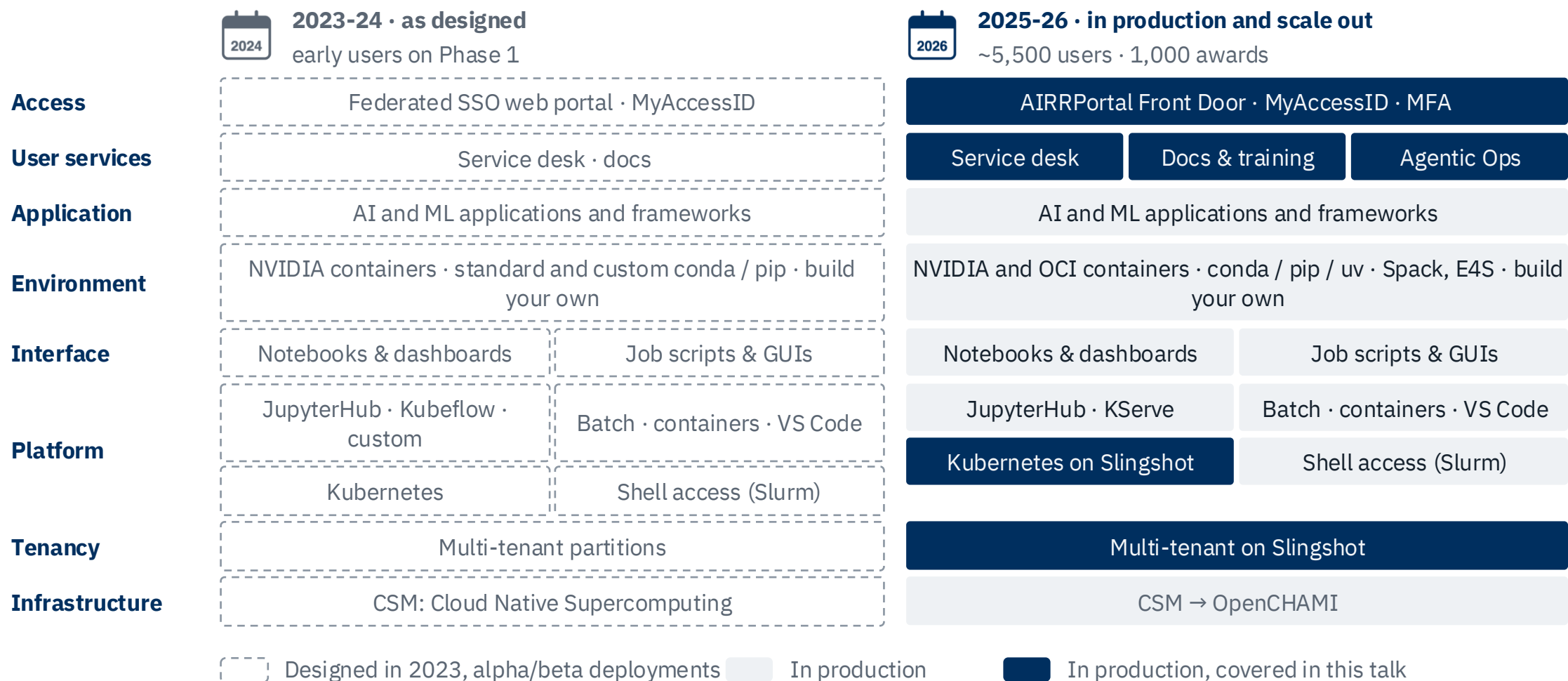
	Oct 2023	Oct 2024	Oct 2025	Oct 2026	Oct 2027 plans
AIRRPortal	No team, facility or computer Isambard-AI first announced in September 2023 – procurement completed in October 2023	Designing federated portals and identities	Live: awards flow to Bristol automatically	Rolling calls at scale; AI Open Access; AIRR Cloud	One identity, routed support.
Kubernetes on Slingshot		Kubernetes planned as part of the core stack	FRIDGE TRE on RKE2	CXI DRA driver; vLLM served via KServe	Kubernetes service for early users
Rise of the Agents		Isambard-AI Early Access	Ramping Isambard-AI	Agents are here	Differentiated AI services

One chapter per row, each told in the same five snapshots.

Looking back for lessons learned, co-design successes.

Looking forward for continuous improvement of services.

The stack: as designed in 2023, in production since 2024



CUG'24: doi.org/10.1145/3725789.3725794 · ICPP'26: doi.org/10.1145/3816891.3834897

01

AIRRPortal Front Door

How do thousands of new users get from application to compute in days?

Ashish Tom



Oct 2023

Oct 2024

Oct 2025

Oct 2026

Oct 2027 · projection

AIRRPortal: five Octobers



Oct 2023

**Isambard-AI announced;
no way to apply yet**

Isambard-AI announced
Sept 2023; part of a planned
30× boost to UK public AI
compute; no route to apply
yet;



Oct 2024

**Designing federated
identity and portals**

AIRRFed: MyAccessID logins,
IdP of last resort for non-
academics; OpenPortal
evaluated; awards sent as
spreadsheets



Oct 2025

**Live: awards flow to
Bristol automatically**

Demoed Mar 2025; first call
23 Jun, first award 17 Jul;
Bristol connected in August
and Cambridge connected in
November



Oct 2026

**One front door, many
paths, at scale**

~2,300 proposals, ~1,000
awards, ~5,500 users across
6 calls and 10 rounds; AIRR
Cloud awarding since August;
same open model for UKRI's
NCRs



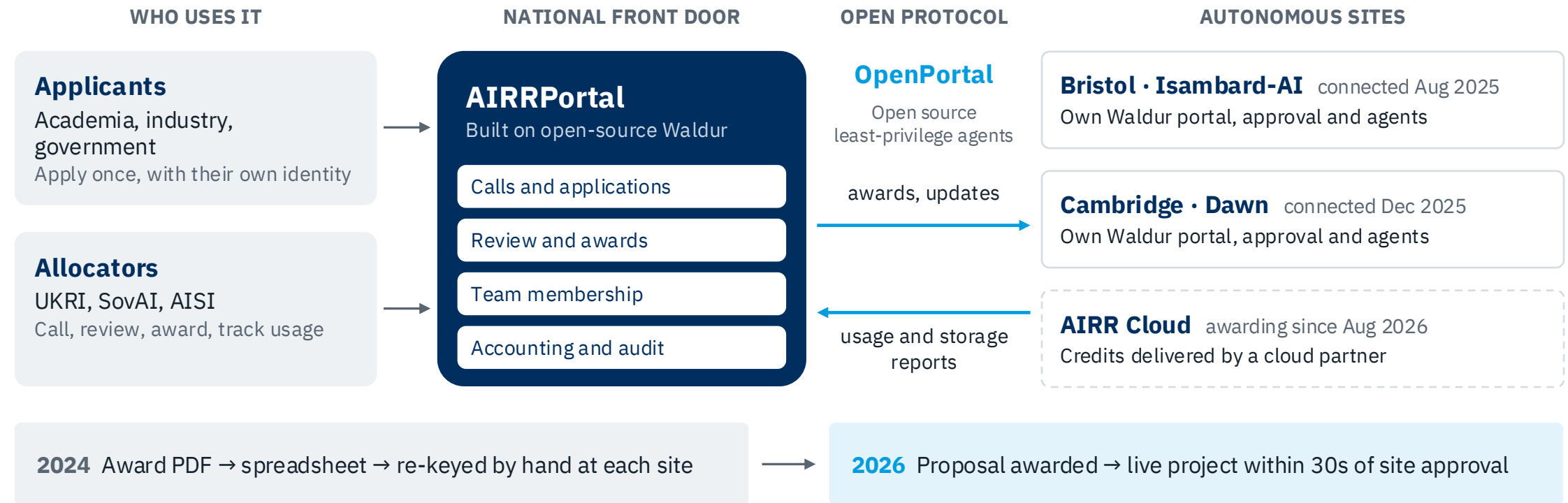
Oct 2027 · projection

**One identity, routed
support, new services**

ConnectID brokering for all
sites; routed AIRR Support;
AIRR Data awards

From spreadsheets and email to one open national front door that routes awards to autonomous sites.

AIRRPortal: open protocol, autonomous sites



The portal sends the award, not the business logic: each site keeps control of its own systems.

AIRRPortal: what worked, what surprised us, what's next

What worked

- One self-service front door for users and allocators: proposals, awards, usage
- Proposals for light-touch calls reviewed in weeks, awards onboarded in 1–2 days;
- Distributed peer review piloted on AI4Science and Innovator calls
- Open Waldur and OpenPortal let each site keep its own stack

What surprised us

- Human review and compliance checks, not software, set the pace
- Self-managed identities: 13% of accounts, 70% of identity support
- Once onboarded, users and allocators run awards without waiting on site staff

What's next

- ConnectID brokering for Isambard-AI and Dawn sites, Q4 2026
- AIRR Support and AIRR Data over the same open protocol
- UKRI reusing the same collaborative open model for the NCR

02

Kubernetes on Slingshot

How do we run multi-tenant, cloud-style services on a supercomputer at bare-metal speed?

Jake Watson



Oct 2023

Oct 2024

Oct 2025

Oct 2026

Oct 2027 · projection

Kubernetes on Slingshot: five Octobers



Oct 2023

Funding and procurement

no data centre, no team, just an architectural and design objective



Oct 2024

Architecture or stack proposed

Kubernetes was laid out as a key part of the under lying service



Oct 2025

FRIDGE use case

Initial for running Kubernetes as well as supporting TREs Using RKE2



Oct 2026

Kubernetes Work matures

CUG'26 paper on initial slingshot work, Built a CXI DRA Driver results in SC26 workshop paper, VLLM running using KServe, FRIDGE work continued under AIRR-BRIDGE



Oct 2027 · projection

Kubernetes Service

Early users training, internal and external service provisioning

| From a design objective to Slingshot RDMA as a native, multi-tenant Kubernetes resource at near bare-metal speed.

Kubernetes on Slingshot: DRA at near bare-metal speed

```
metadata:
  annotations:
    modules.isambard.ac.uk/load: aws-ofi-nccl-1.18.0
  name: vllm
spec:
  containers:
    - image: nvcr.io/nvidia/vllm:26.05-py3
      name: vllm
  resources:
    claims:
      - name: gpu
  resourceClaims:
    - name: gpu
      resourceClaimTemplateName: gh200-x1-cxi-x1
```

NCCL All Reduce (K8s vs. Bare Metal)

Between two nodes, single GPU+NIC to single GPU +NIC

Size	Bare Metal (GB/s)	K8s (GB/s)	Delta
128 KiB	1.55	0.44	-72%
512 KiB	3.58	1.67	-53%
1 MiB	6.64	6.12	-8%
4 MiB	17.80	17.03	-4%
16 MiB	20.97	20.73	-1%
64 MiB	22.75	22.69	-0.3%
256 MiB	23.08	23.10	+0.1%
1 GiB	23.21	22.35	-4%
4 GiB	23.01	22.41	-3%
8 GiB	23.13	22.56	-2%

K8s within 4% of bare metal from 4 MiB up; peak 23.2 GB/s, ~93% of line rate.

Matters for multi-node LLM inference, with its large tensor collectives.

Stock containers get a NUMA-matched GPU and Slingshot NIC, with libraries injected through CDI: no privileged pods, and 2.43 μ s point-to-point latency against 2.23 μ s on bare metal.

Watson et al., CUG'26 (to appear) · Watson et al., SC26 Workshops, Chicago (to appear)

Kubernetes on Slingshot: what worked, what surprised us, what's next

What worked

- RKE2 for deploying k8s, simple install and easy to configure, joining nodes is also straightforward
- Dynamic Resource Allocator, comprehensive CXI device scheduling, NUMA Node matching with GPU
- Admission Policies, allow us control over what user can submit, simplify UX

What surprised us

- Networking challenges when running K8s Control Plane outside of the cluster
- Injecting Libraries into containers is hard, evaluation order, dependencies, GLibc
- User Resource Usage tracking
- AI users really wants K8s not Slurm

What's next

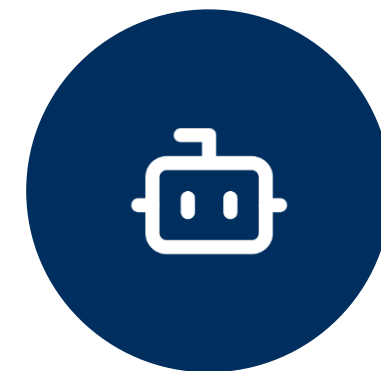
- Improve Resilience
- Observe Users Behaviour and Tune
- Dedicated clusters

03

Rise of the Agents

In 2026, the paradigm for consuming supercomputing services changed.
How do we support users?

Richard Gilham



Oct 2023

Oct 2024

Oct 2025

Oct 2026

Oct 2027 · projection

Rise of the Agents

				
Oct 2023	Oct 2024	Oct 2025	Oct 2026	Oct 2027 · projection
Isambard 2 legacy	Isambard-AI Early Access	Ramping Isambard-AI	Agents are here	Differentiated AI services
ARM is niche in the datacentre	GH200 becoming available. Patchy support for libraries & frameworks	GH200 widely available. ARM support for libraries and frameworks improving	GH200 support on parity with x86. Software 'just works'	Looking beyond GH200 as one-size-fits all.
Small community of academic/gov HPC experts with compiled-language modsim workload	Academic/gov experts and well-supported generalists start bringing Python-based AI workloads. Some use notebooks & IDEs	Academia/Gov but broader experience/expertise spread. More without traditional HPC experience. Some hybrid AI-modsim workloads	SMEs spearhead a step-change. Widespread use of Agents. Experience/expertise spread huge. Demand for inference service	Inference/other specialized services enable access to even broader audiences

Isambard-AI is at the centre of a paradigm shift of the why, what, who and how people use supercomputers

Rise of the Agents: The challenge



HPC and cloud are complex

Making them useful and productive is not free: some barrier to entry is inevitable



Service providers can lower the bar

Abstract the complexity; documentation, training and support



Users can jump higher

Use documentation, training and support, or use an AI agent

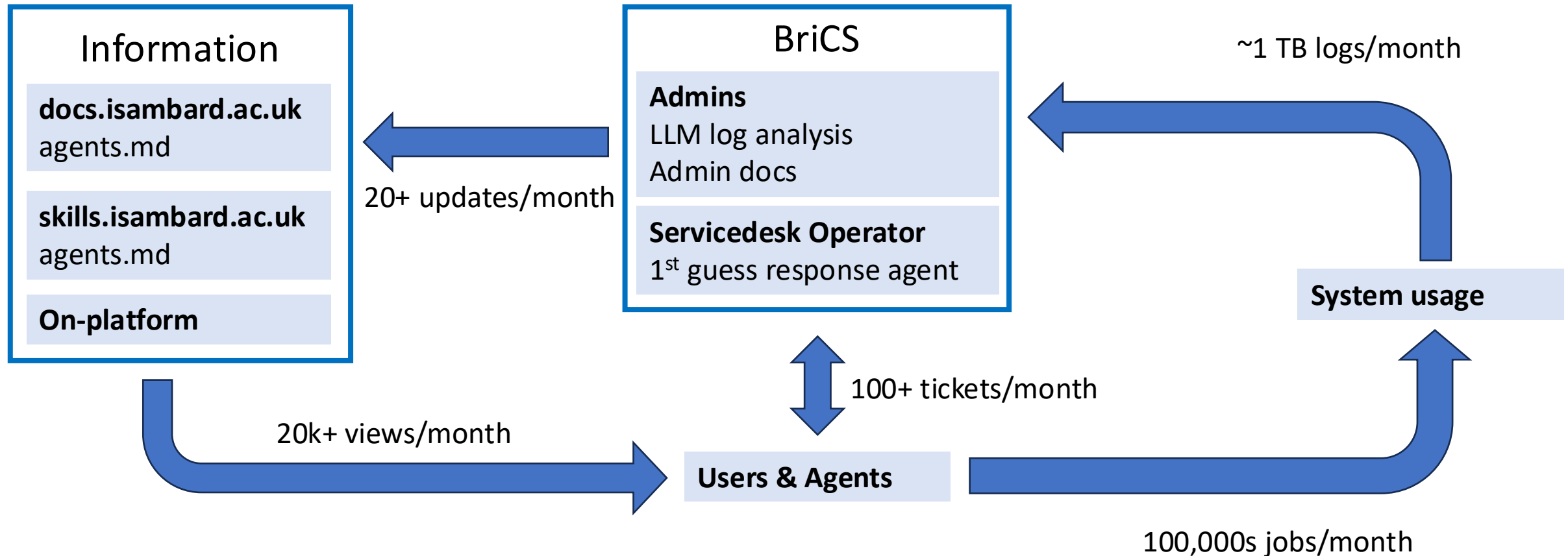


Agents let users jump 10x higher, 100x faster

What if they jump in the wrong direction?

AI Agents could turn all users into superusers, but only if they do the right thing

Rise of the Agents: **Agentic Ops**



Agentic Ops- influencing user agents at scale

User services & Agentic Ops: what worked, what surprised us, what's next

What worked

Making documentation central

What surprised us

Speed of agent uptake and impact on how people work

Low volume of support tickets

What's next

Getting ~~users~~ agents to read the ~~docs~~ skills



Outlook to 2027: evolving AI, HPC, Data DRI services at pace

Oct 2023

Oct 2024

Oct 2025

Oct 2026

Oct 2027 · projection



What we will evolve

One identity and routed support: ConnectID, AIRR Support and AIRR Data

Kubernetes as a service: multi-tenant, bare-metal RDMA for AI workloads

Differentiated AI services beyond GH200, including inference, for broader audiences



With whom

National · UKRI, DSIT, AIRR sites, and FRIDGE partners (Turing, UCL, Cambridge), ... everyone welcome

International collocation and communities · Linux HPSF, CNCF, OpenCHAMI, OpenPortal, open source and open standards

Industry · with several entities

Services

Isambard-AI

Isambard 3

Digital Labs

AIRRPortal

National AI Data Infrastructure Network–Coming

Complementing AIRR compute with storage services for open & sensitive data

Inference Technology Watch–Coming

Platform for exploring novel technologies and software stacks for inference

Across the three stories: an open, cloud-style stack kept pace with a fast-moving AI ecosystem.

Selected papers and more at SC26

Papers

Publications

CUG24 doi.org/10.1145/3725789.3725794. Isambard-AI: a leadership-class supercomputer optimised specifically for Artificial Intelligence

S-HPC'24 (SC24) doi.org/10.48550/arXiv.2410.18411. Federated Single Sign-On and Zero Trust Co-design for AI and HPC Digital Research Infrastructures

ICPP'26 NVSUG doi.org/10.1145/3816891.3834897. Portable and Resilient by Design: A Cloud-Style Stack for HPC and AI across NVIDIA Grace and Grace Hopper Superchips

At SC26 · workshop papers

HUST Supporting a Fast-Paced National AI Resource: Portal, Identity, and Federated Access at Scale

Fabrics Slingshot as a First-Class Kubernetes Resource: Bare-Metal RDMA for Multi-Tenant AI

HPCSYSPROS Security and Day-2 Resilience as First Class Control Plane Primitives for Heterogeneous Supercomputers

S-HPC A Threat Model for Sensitive-Data Workloads on Multi-Tenant AI Supercomputers

ECHO Placing the Control Plane: Cloud-Native Federated Learning Across Batch-Scheduled HPC

PMBS NVIDIA Vera: A Microarchitecture-Level Comparison with Grace, Turin, and Granite Rapids

P3HPC The Price of a Single Source: Performance Portability of Chapel and Julia across CPUs and GPUs

Meet us at SC26 in Chicago: University of Bristol, Booth 6804.

Sessions at SC26 · Chicago, 15–20 Nov

Mon · Workshop Sovereign AI Supercomputing Cloud (organiser)

Tue · BoF OpenCHAMI Community

Tue · BoF Confidential Supercomputing: Essential Evolution or Elegant Illusion?

Wed · BoF A Single Pane of Glass for AI Supercomputing: Interoperable Access Portals Worldwide

Wed · BoF Open Source and Sovereign Infrastructure for HPC and AI

Thu · BoF Trusted Research Environments for AI and Integrated Science: Launching the Genesis Mission

Fri · Panel Brain Dump: HPC Experts Volunteer to Be Replaced by AI

Thank you

-  **The BriCS team**, who design, deliver and run Isambard-AI every day.
-  **Our users**, whose workloads and feedback shaped everything shown here.
-  **Stakeholders and funders**: DSIT, UKRI, STFC and DARE UK.
-  **Collaborators**: HPE, NVIDIA, VAST Data, our FRIDGE partners, AIRR collaborators, ...

 Web · bristol.ac.uk/research/centres/bristol-supercomputing

 Docs · docs.isambard.ac.uk

 BriCS training · docs.isambard.ac.uk/training

 AIRR access calls · apply.isambard.ac.uk/ukri

 YouTube · [@BriCS-UoB](https://www.youtube.com/@BriCS-UoB)

 LinkedIn · [brics-uob](https://www.linkedin.com/company/brics-uob)

 Bluesky · [@brics-uob.bsky.social](https://bsky.app/profile/@brics-uob.bsky.social)

 Instagram · [@bristolcentreforsupercomputing](https://www.instagram.com/@bristolcentreforsupercomputing)

These slides were prepared with the assistance of Claude (Anthropic). Content and conclusions are the authors' own.