

Agentic Ops: Responding to Unprecedented User Patterns on Isambard-AI

Richard Gilham

Conner Flansburg, David John, Richard Lane,
Perceval Maturure, Dave Moore, Paul Wright

Agenda

- Introducing BriCS and Isambard-AI
- The challenge: Agents
- A solution?: Agentic Ops

Bristol Centre for Supercomputing systems

We've built **two new datacentres** and installed **three new HPE supercomputers** in less than 2 years:

1. **Isambard 3** – 384 nodes of Grace-Grace CPUs, air cooled
2. **Isambard-AI phase 1** – 168 GH200 GPUs in EX2500, DLC
3. **Isambard-AI phase 2** – 5,280 GH200 in EX4000, DLC

Power is 100% UK-generated renewables: wind, solar, hydro.

Isambard-AI Origins & Timeline

Government
feedback on Bristol
Isambard-AI
proposal

Isambard-AI
procurement

First Modular Data
Centre installation

Onboarding of early
tech prep users

Isambard-AI Phase 2
installation

Isambard-AI Phase 2
regularly hitting
capacity



2023



2024



2025



2026

Isambard team
recruitment

Isambard-AI Phase 1
power-on

First DSIT allocated
projects

Isambard-AI Phase 2
service live

Isambard-AI Phase 2
acceptance



Isambard 3

2023- £10M investment

Traditional Tier-2 (NCR) HPC: mostly CPU + batch

Modular Data Centre: MO data centre unavailable

384 nodes NVIDIA Grace CPU Superchips:

- 2x72 cores; 2x120 GB LPDDR5, slingshot 11 interconnect

Multi-Architecture Comparison System (MACS):

- AMD Milan/Genoa/Bergamo; Intel SPR DDR5/HBM
- AMD Milan + GPUs: AMD MI100, NVIDIA A100/H100 PCIe)



Isambard-AI phase 2



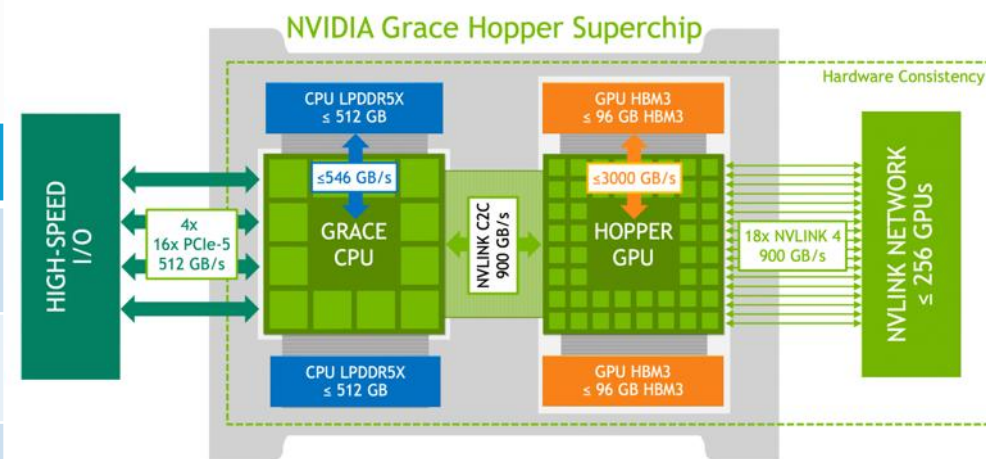
- **5,280 NVIDIA Grace-Hopper GPUs** in a new, 5MW HPE modular data centre (MDC) facility in Bristol, UK
 - **~21 ExaFLOP/s of 8-bit for AI**, **~216 PFLOP/s 64-bit for HPC**
 - **PUE typically under 1.1, powered by zero carbon electricity**
 - **Ranked #11 in the world** on the Top500 June 2025
- **Extremely rapid deployment**
 - Site constructed from scratch and supercomputer installed in ~12 months
 - ~2 months for bring-up and acceptance



Department for
Science, Innovation,
& Technology

Isambard-AI

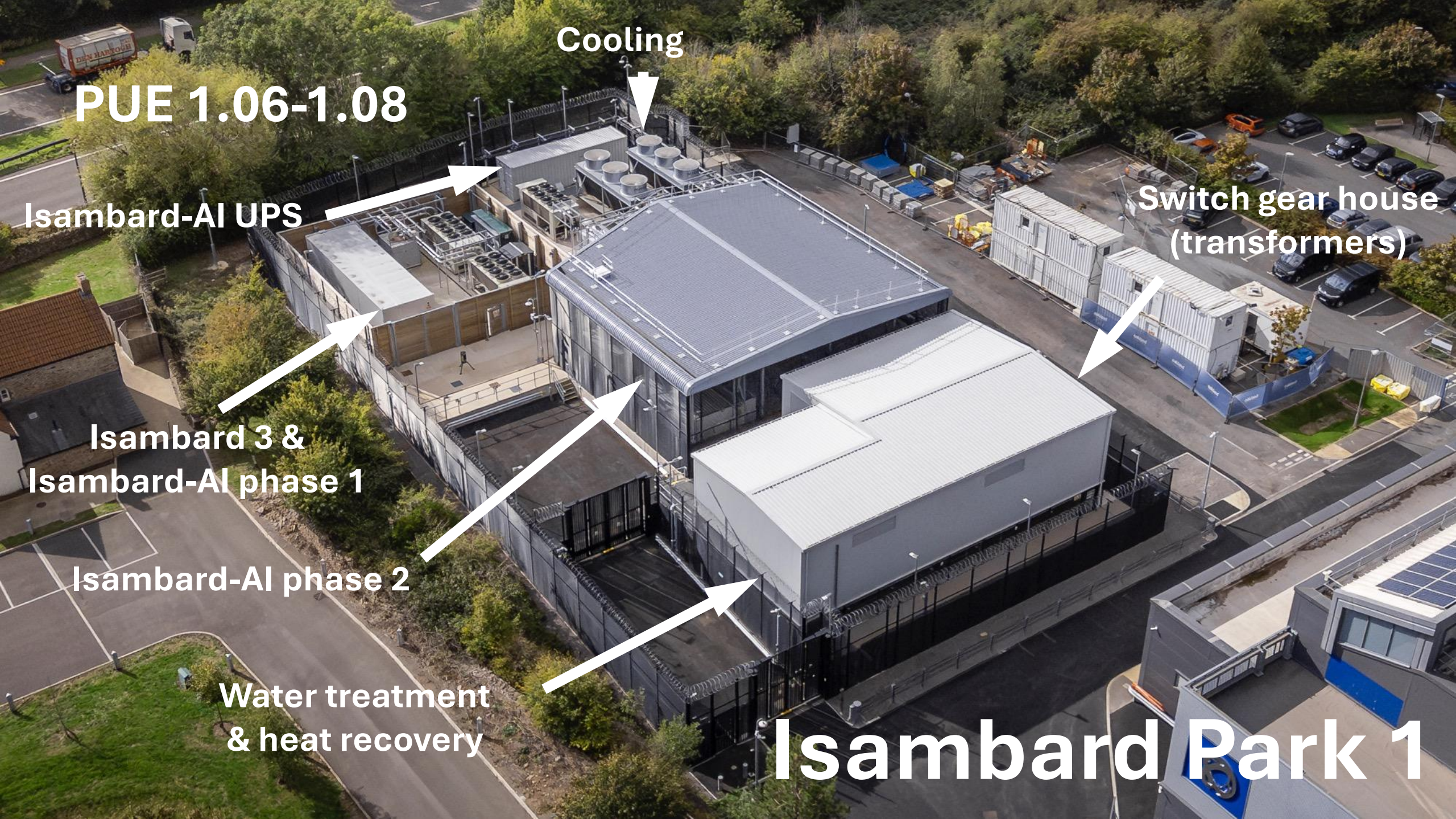
	Phase 1	Phase 2
Nodes	42	1320
GH200 superchips (4/node)	168	5,280
Grace CPU cores (4x72/node)	12,096	380,160
GPUs (4/node)	168	5,280
LPDDR5 CPU memory (4x120GB/node)	21.5 TB	675 TB
HBM3 GPU memory (4x96GB/node)	16.1 TB	506 TB
Total memory	37.6 TB	1.18 PB
Storage	~1 PB all-flash	All-flash ~20 PB POSIX ~7 PB object



Source: NVIDIA Grace Hopper Superchip Architecture [Whitepaper](#)

Intra-node connectivity:
NVLink-C2C (900 GB/s)

Interconnect:
4 x Slingshot 11 (200 Gbps)



PUE 1.06-1.08

Cooling

Isambard-AI UPS

**Switch gear house
(transformers)**

**Isambard 3 &
Isambard-AI phase 1**

Isambard-AI phase 2

**Water treatment
& heat recovery**

Isambard Park 1

The challenge: ~~Agents~~ Users

Mental Maps

HPC is complex; users need a functional mental map

Traditionally gained via documentation, training, experience etc

- Can't compel a user to read the documentation

Learning HPC is a time cost, reduce it by using LLMs & Agents

- Can't compel AI to read the documentation either

Isambard-AI Userbase

- Sciences to arts & humanities to AI research
- Academia, government, SME
- HPC, Cloud & Workstation background; beginner to expert
- Bring the applications they have, not start from scratch
- Onboard ~1000 new projects annually
 - Currently ~850 projects, ~3800 users

>100 users for every BriCS team member

Ramping the Isambard-AI service

Summer 2025

- Isambard-AI phase 2 service launches
- Mostly academic users; gradual introduction of others
- Project & user numbers grow; utilization gradually increases

Spring 2026

- Rapid increase and diversification of projects & users
- Approaching full utilization

Surprises were waiting for us...

Lustre Filesystem: Introducing inodes

- An inode is a metadata structure
 - One per file/directory
- Data and metadata stored separately:
 - inodes on the Metadata Target (MDT)
 - Data striped across Object Storage Targets (OSTs)
- MDT inode count is fixed at filesystem creation
 - Hard limit on total number of files

Huge quantities of directories & small files can exhaust the inodes...

inode Exhaustion: Admin Perspective

Diagnosis:

- One user accrued ~1 Billion files, average 60kB each (~47 TB)
- Several other projects >100 Million files

Immediate response:

- Impose inode quotas (that actually worked)
- Documentation- SquashFS tutorial
- Engage users to delete files (0.8Bn easily identified; 0.3Bn directories)

Longer term:

- Improve technical guardrails, tooling etc
- Incorporate into Mental Map: docs & training

inode Exhaustion: User Perspective

Innovative idea. Needs to scale up simulation dramatically

- Awarded 100,000s of GPU hours on Isambard-AI
- Ports and scales up workflow...

...computer goes down. Gets an email from BriCS about having too many files...

...has to waste time deleting files

Didn't know this was important, and don't really care

What went wrong?

Multiple gaps aligned: user's mental map, tooling, guardrails

Problem manifested when:

- Number of users ramped quickly
- Subtle change in background of users

Was it avoidable? Technically yes, but...

- Beware hindsight
- Opportunity cost of going slower

Other Observed Anti-patterns

Saturate Lustre MDT server

- Testing file existence as checkpointing
- Multiple ripgreps across entire filesystem

Saturate Slurm controller

- Multiple processes polling queue at 10 Hz

Overwhelm login node

- 100s of copies of ps (record is 477)

Self-repairing agents- killed tasks come back instantly

- Challenging for problem isolation

A Solution?: Agentic Ops

Framing the question

How can we guide hybrid human/AI behaviour to...

- ...prevent antipatterns being attempted?

- ...promote good usage patterns?

- ...identify new antipatterns early?

AI is not the enemy but can feel 'analogue' alongside traditional sys admin approaches

Build and connect tools to enable an Ops approach

Documentation

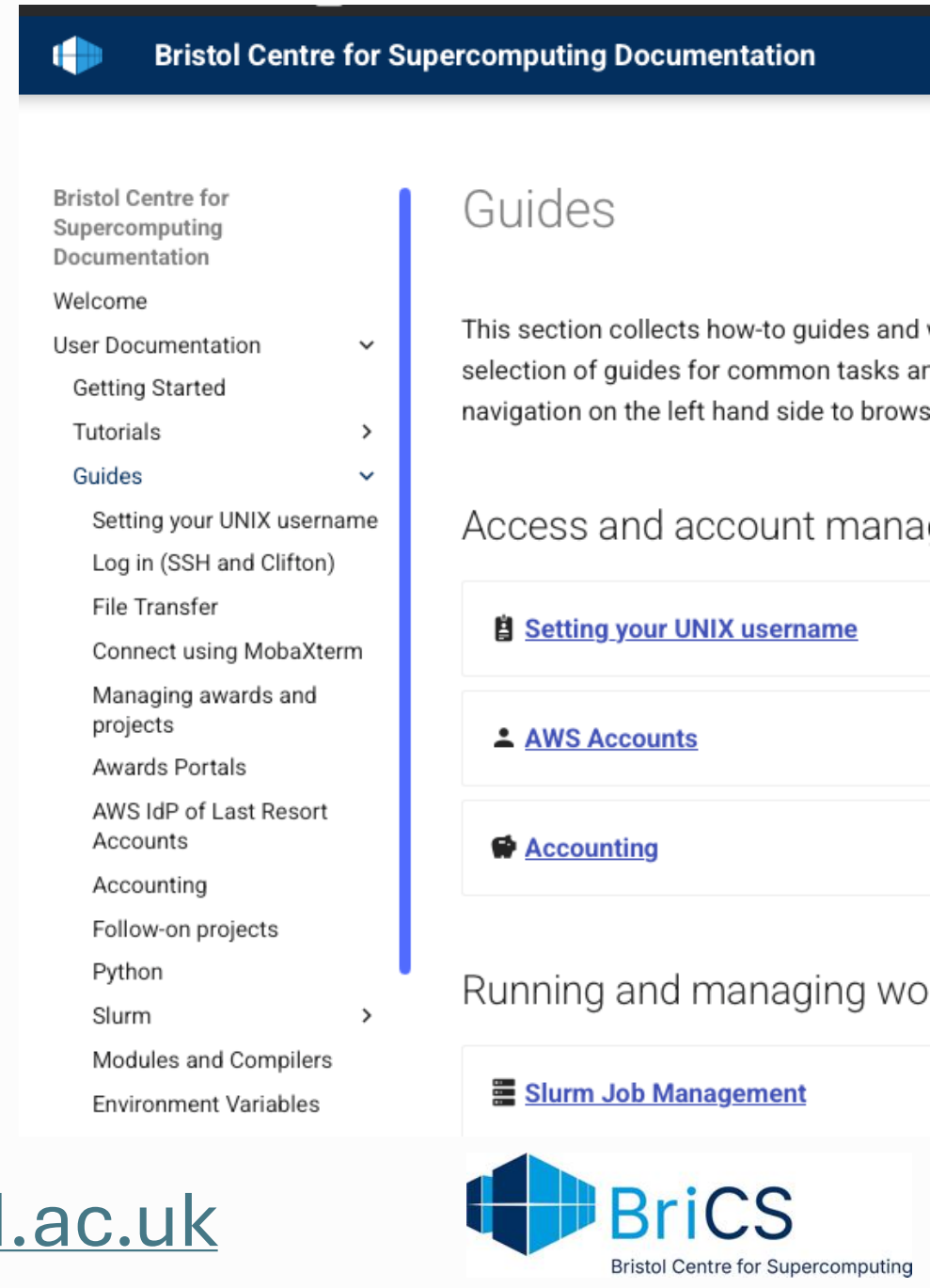
Pivotal & primary information source

Jan-Jun 2026: ~250 page updates,
~65,000 visits

Guides, tutorials, reference, FAQs

Human-first: also readable by LLMs

<https://docs.isambard.ac.uk>



The screenshot shows the website for the Bristol Centre for Supercomputing Documentation. The header is dark blue with the center's logo and name. A left-hand navigation menu lists various topics, with 'Guides' highlighted. The main content area features a 'Guides' section with a description and a list of links to specific guides like 'Setting your UNIX username', 'AWS Accounts', and 'Accounting'. Below this is a section for 'Running and managing work' with a link to 'Slurm Job Management'. The footer includes the University of Bristol logo and the 'BriCS' logo.

Bristol Centre for Supercomputing Documentation

Welcome

User Documentation

- Getting Started
- Tutorials
- Guides**
- Setting your UNIX username
- Log in (SSH and Clifton)
- File Transfer
- Connect using MobaXterm
- Managing awards and projects
- Awards Portals
- AWS IdP of Last Resort Accounts
- Accounting
- Follow-on projects
- Python
- Slurm
- Modules and Compilers
- Environment Variables

Guides

This section collects how-to guides and a selection of guides for common tasks and navigation on the left hand side to browse

Access and account management

-  [Setting your UNIX username](#)
-  [AWS Accounts](#)
-  [Accounting](#)

Running and managing work

-  [Slurm Job Management](#)

 **BriCS**
Bristol Centre for Supercomputing

Skills

- Small instruction packages that teach an AI assistant a specific skill or task
- Markdown files — human-readable, flexible, and loaded only as needed
- A powerful way to embed local standards and knowledge into the AI's responses
- Isambard's skills are AI-generated from the docs, with the ability to manually add further specifics

Slurm on Isambard

Both Isambard-AI and Isambard 3 use the [Slurm](#) scheduler to run when the requested resources become available.

Never poll the queue rapidly. Running `squeue` too frequently on the scheduler and slows job scheduling for everyone. Adhere to the acceptable use policy and may result in your account being suspended.

Critical Rules

- Never poll `squeue`, `sinfo`, or any scheduler command too frequently.
- Always submit jobs with `sbatch` or `launch`.
- Always set explicit `--time`, `--nodes`, and `--cpus-per-task`.
- Use `--exclusive` only when the workload requires it.
- Do not request GPUs on Isambard 3 system.

Skills on Isambard

Users can install their favourite agents on Isambard

- Too many agents and new versions to support a central installation

Opportunity to intervene and nudge behaviour close to compute

Can/should we force agents to pick up Isambard skills?

- No single standard approach
- Will these get circumvented anyway?

Servicedesk Agent

Human servicedesk operations are time intensive

- Does not scale with N users
- Required contextual knowledge
- Displaces time to improve services
- Dominated by routine/simple requests

Agent for first-guess response

- Workflow built on LangGraph
- Claude Haiku to summarise tickets
- Vector database from tickets & docs using voyage-4

Human in the loop initially to ensure quality

Example output:

```
=====
TICKET #2321 - Data transfer nodes or login node?
=====

QUESTION SUMMARY
-----
The user is attempting to transfer a multi-terabyte dataset
from Azure Blob storage to Isambard-AI...

Completeness: [██████] 100%

SIMILAR RESOLVED TICKETS
-----

●●○ [MEDIUM 0.85] #1556 - Large scale file transfer
Problem   : 40TB from AWS S3 - no dedicated DTNs...
Resolution: Use login node with rclone...

RELEVANT DOCUMENTATION
-----

[DOC 0.78] isambard/docs/user-documentation/faqs/index.md
### How do I copy files to and from the Isambard supercomputers?

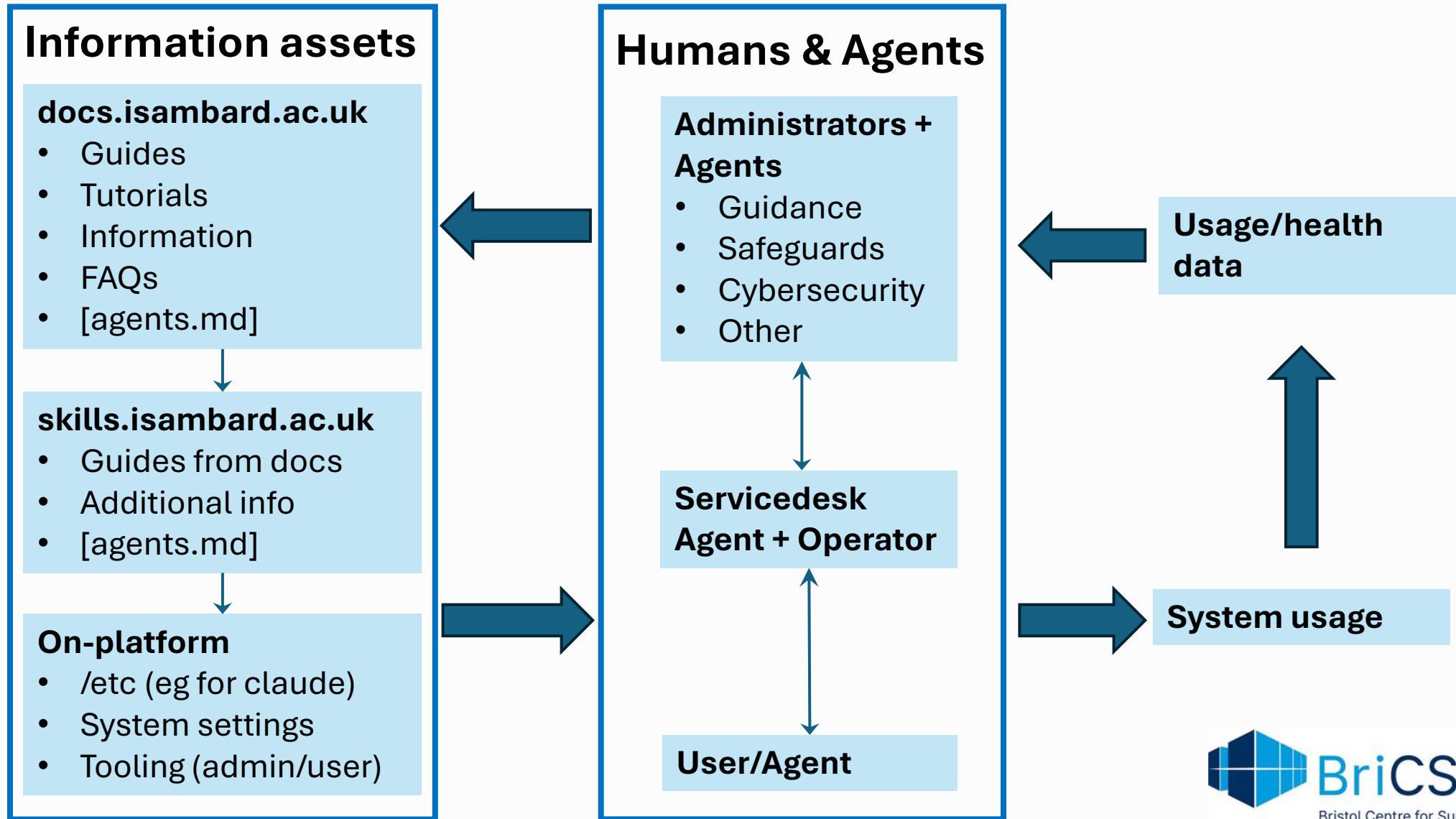
LIKELY CAUSE / FIRST ACTION
-----

Login-node transfers are the standard method - no dedicated
data transfer nodes exist. For Azure Blob, recommend rclone.

SUGGESTED QUESTIONS FOR USER
-----

• Which tool do you plan to use - azcopy, rclone, or a custom script?
• Have you tested connectivity from the login node to Azure Blob?
=====
```

Agentic Ops



Status & Future work

This work is very new

- Skills- Early Access users now; roll-out this summer
- Skills on Isambard- planning & technical evaluation now
- Servicedesk agent- testing now; Early Access this summer

Future work:

- Measure impact- are users more productive?
- Information to support ops-style decision-making

Thank you

richard.gilham@bristol.ac.uk
@rjgilham

