UNIVERSITY OF CAMBRIDGE

AIRRFED
DSIT UKRI AI Research Resource (AIRR) Federation Demonstrator Project

Creating a joined up AIRR service
&
A look at the art of the possible for a re-imagined UKDRI

Dr Paul Calleja : Director Research Computing Service

Today's discussion

- Cambridge Overview, service & innovation lab
- AIRR, introduction to Dawn
- AIRRFED

Cambridge RCS - overview

- Formed 19 years ago pioneering the use of large scale HPC clusters as leadership class HPC systems, back in 2006 the fastest HPC system in the UK at #20 in top500

- Today RCS is significant UKRI and DSIT national AI & HPC centre excellence
  - Multiple national / International HPC / AI activities - AIRR, EPSRC Tier2, DiRAC, IRIS, UKAEA, SKA-SRC, ExCALIBUR H&S, Schmidt Futures Climate Centre
  - Serving a cross domain HPC user community of ~4000 users
  - Providing ~48% of UK Top500 performance over last 7 years
  - Well developed TRE strong focus on AI & HPC within medicine

- Successful Industry partnership model, over £80M industry investment over last 4 years,

- Strong in-house technology team - system design, implementation and operations

- Long standing technology partnership across Africa

UK Research and Innovation

Department for Science, Innovation & Technology

UK Atomic Energy Authority

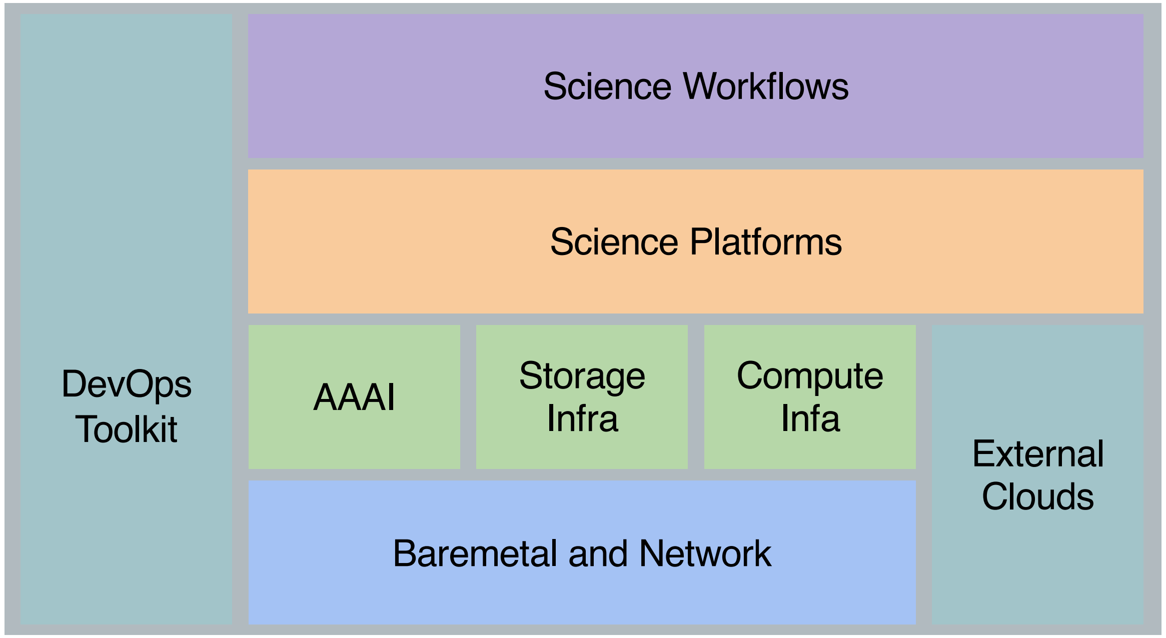
UNIVERSITY OF CAMBRIDGE

Cambridge RCS - infrastructure & people

- £100M of HPC / AI equipment in operation (<4years old)
- 50 staff across Platforms, DevOpps, RSE, outreach business development
- 1.8 MW Leading UKRI water cooled AI HPC Data center 100 Racks - £30M University investment in UK national service provision infrastructure
- 30 PF heterogeneous HPC/AI system
- 3000 Dell servers X86/GPU Intel & NVIDIA
- 45 PB storage (disk/NVMe/tape)
- ISO27001 compliant secure computing for commercial & medical users
- Private cloud – system stack using OpenStack in collaboration with StackHPC leading UK SME

Scientific OpenStack – On premise science cloud

Full-Stack Science platforms



- Deliver HPC & AI via cloud APIs
- Under strong active development in partnership with StackHPC with funding from UKRI, SKA, Industry & AIRR
- Creating UK "Community Cloud Middleware Stack" UKRI project
- Revolutionises flexibility and end user functionality of HPC systems
- Controlling all Cambridge infrastructure
- Gaining traction within large HPC centres in UK and Europe

CAMBRIDGE OPEN ZETTASCALE LAB

- $10M 6-year program - Academic / industrial partnership for the co-design, development and testing of leading-edge HPC, AI and HPDA solutions
- Democratising HPC and AI technologies - increasing functionality, flexibility, efficiently, accessibility, lowering cost, widening user base & impact

UNIVERSITY OF CAMBRIDGE

DELL Technologies

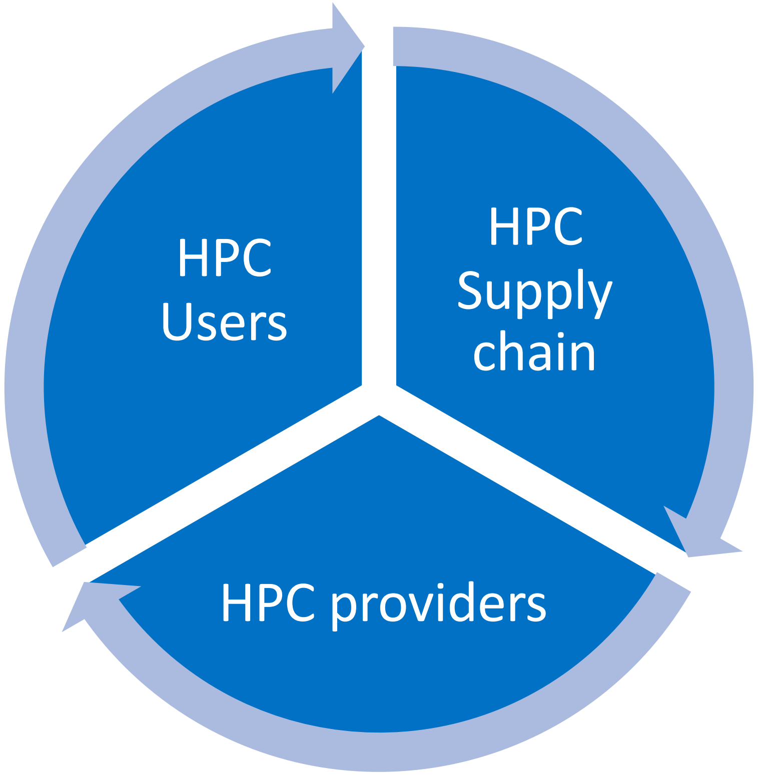
intel

DiRAC

UK Research and Innovation

UK Atomic Energy Authority

The co-design virtuous circle



- Fusion of science use-case inputs, service provider knowledge and technology vendor capability
- Critical mass of requirements, experience, skills and infrastructure
- Creates a science led technology development process
- Driving the innovation cycle of requirements capture, development, deployment, evaluation, iterate

ZettaScale Lab technology development themes

- Energy efficiency
- oneAPI Centre of Excellence
- Research Computing middleware, accessibility & tools
- Large scale tiered storage solutions, how to use lage scale NVMe file systems
- AI workflows and tools merging AI cloud into HPC infrastructure, converged AI & HPC systems
- HPC networking technologies
- Extreme scale visualisation
- Health informatics (TRE’s) HPC within clinical medicine setting

The UK AI Research Resource - AIRR

AIRR represents £300M investment in two of the UK's largest supercomputers designed and configured for AI workloads. Kick-starting the UK's next generation AI infrastructure. Providing large scale AI capability for UK research, industry and government

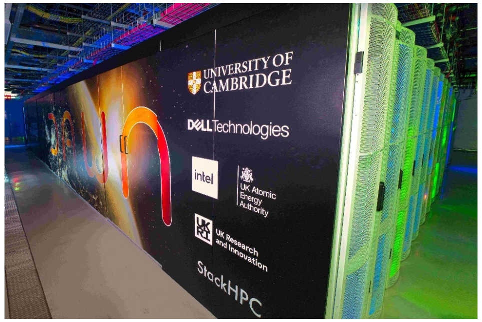
Isambard AI at University of Bristol

Phase 1 168 Grace-Hopper GPUs June 2024

Phase 2 5000 Grace-Hopper GPUs H2 2025

Dawn at University of Cambridge

1000 Intel data centre max 1550 GPUs Nov 2023

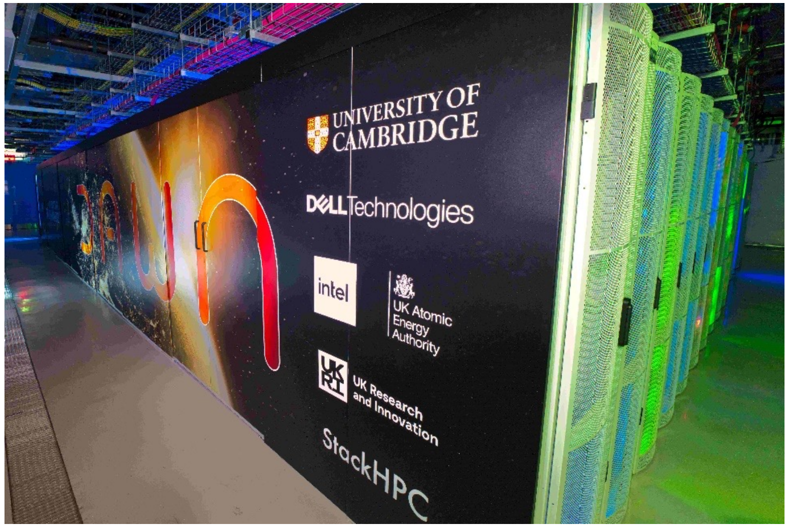


University of BRISTOL

UK AI RESEARCH RESOURCE

UNIVERSITY OF CAMBRIDGE

DAWN



- UK’s Fastest AI supercomputer for the past year 19.45 PF, #40 Top500 at launch (11/2023)
- Highly innovative co-design, co-investment partnership - Dell, Intel, Cambridge, UKAEA, UKRI, DSIT, StackHPC Worked for two years on co-design of the server, cooling, network and OpenStack system software
- Run as a private cloud infrastructure using Scientific OpenStack via Collaboration with StackHPC
- Good collaboration with US Argon National Labs and the Aurora Intel PVC Exascale system

UNIVERSITY OF CAMBRIDGE

StackHPC

intel

DELL Technologies

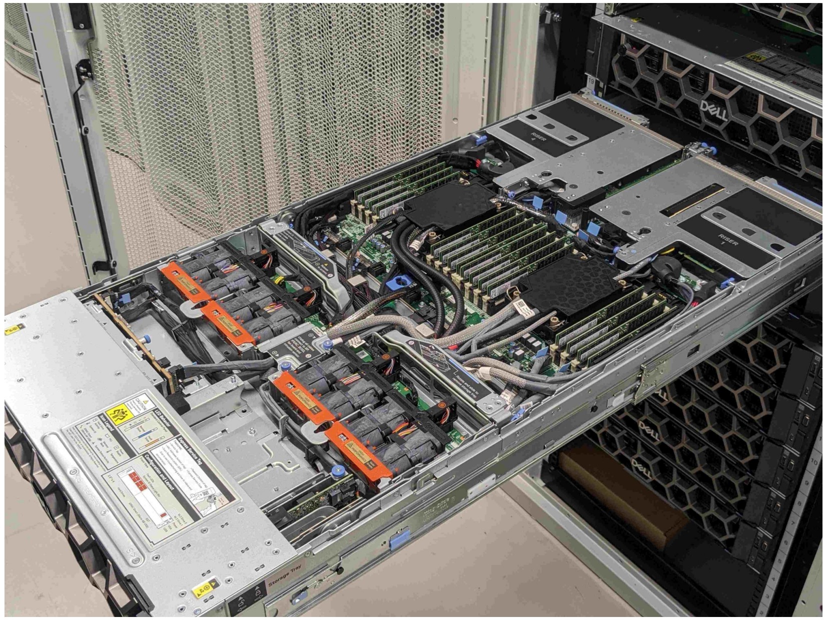
UK Research and Innovation

Department for Science, Innovation & Technology

UK Atomic Energy Authority

DAWN

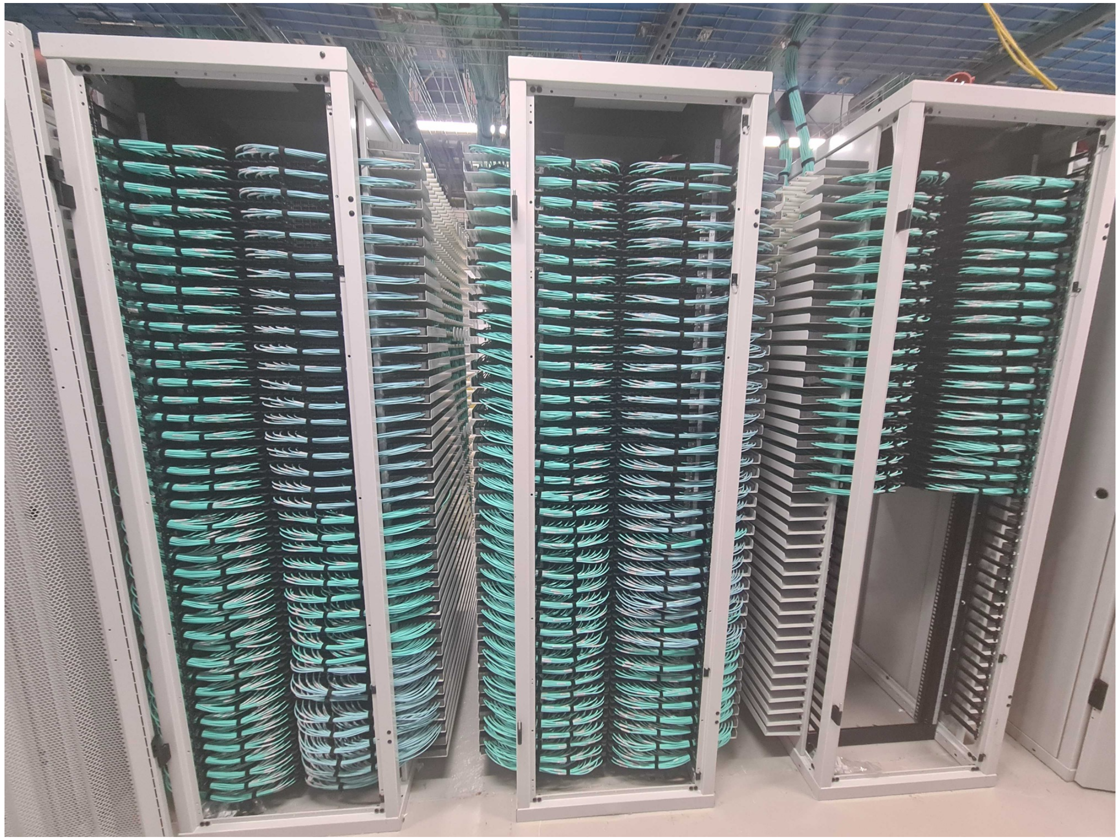
- 256 Dell XE9640 2U DLC cooled GPU servers
- Each with :-
  - 2 * 4th Gen Intel Xeons
  - 4 way SMP Intel Data Centre Max GPU
  - 1TB RAM,
  - 4 * 3.6 TB local NVMe,
  - 4 * HDR200 (fully non blocking)
- 1024 Data Centre Max GPU – 19.45 PF HPL
- 5PB spinning disk Luster + additional 3PB NVMe storage with over 3000 GB/s R/W bandwidths



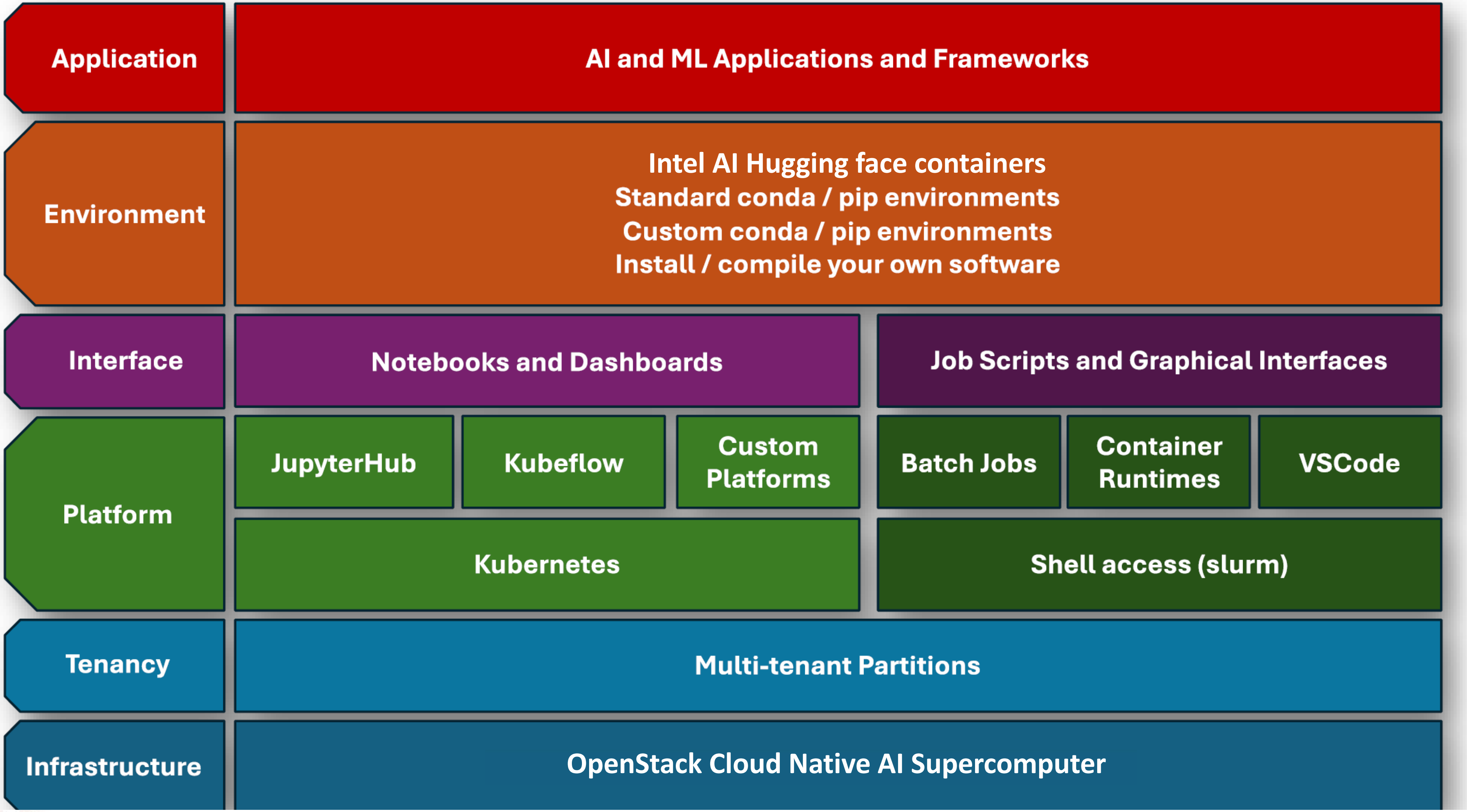
DAWN

- Consumes 1 MW of power
- 2 MW DLC water cooling retrofitted to enable dawn
- 240 L/M of flow !!
- First generation Dell water cooled GPU server
- First generation Intel GPU IB cluster
- 8 weeks P/O to delivery - 3 weeks delivery to top500 !!!

DAWN



- 32 top of rack L1 switches, 100 L2, L3 core switches
- 34 Km of cables
- 1024 * 200 Gb/s ports fully non blocking
- 100Tb/s cross sectional bandwidth in the core



Application: AI and ML Applications and Frameworks

Environment:
Intel AI Hugging face containers
Standard conda / pip environments
Custom conda / pip environments
Install / compile your own software

Interface: Notebooks and Dashboards | Job Scripts and Graphical Interfaces

Platform:
JupyterHub | Kubeflow | Custom Platforms | Batch Jobs | Container Runtimes | VSCode
Kubernetes | Shell access (slurm)

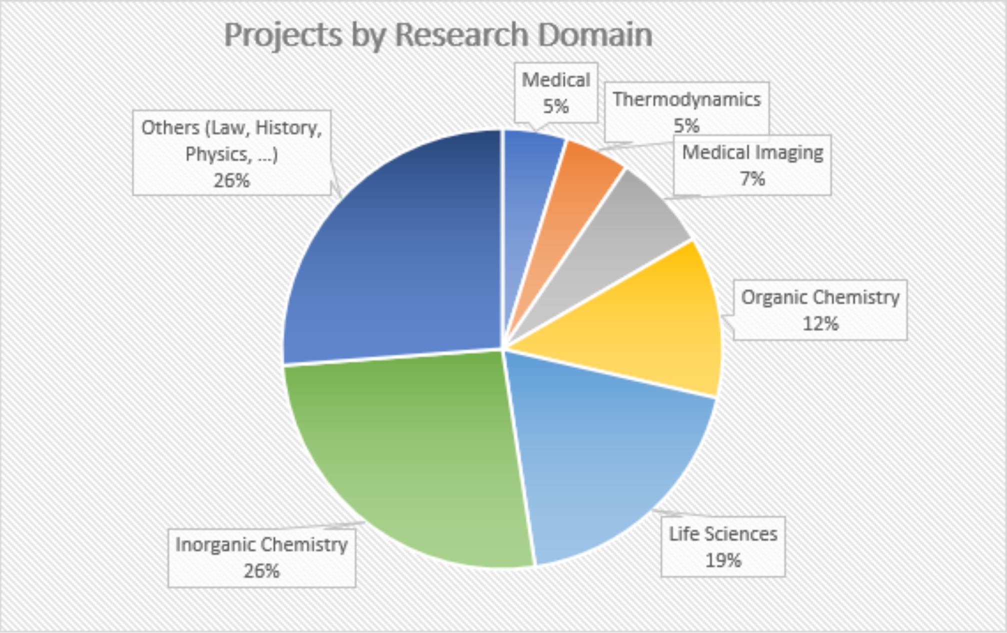
Tenancy: Multi-tenant Partitions

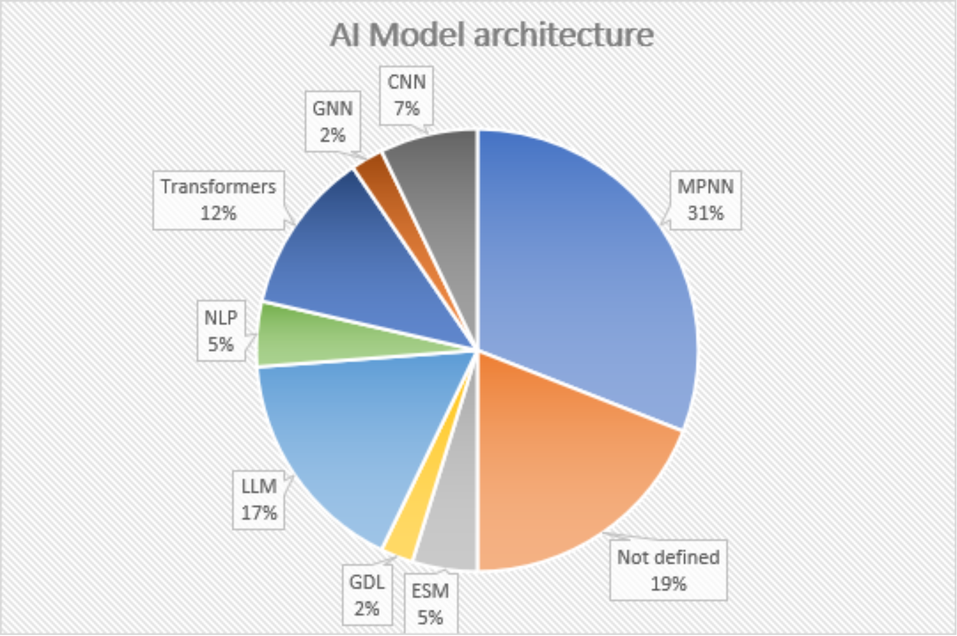
Infrastructure: OpenStack Cloud Native AI Supercomputer

DAWN

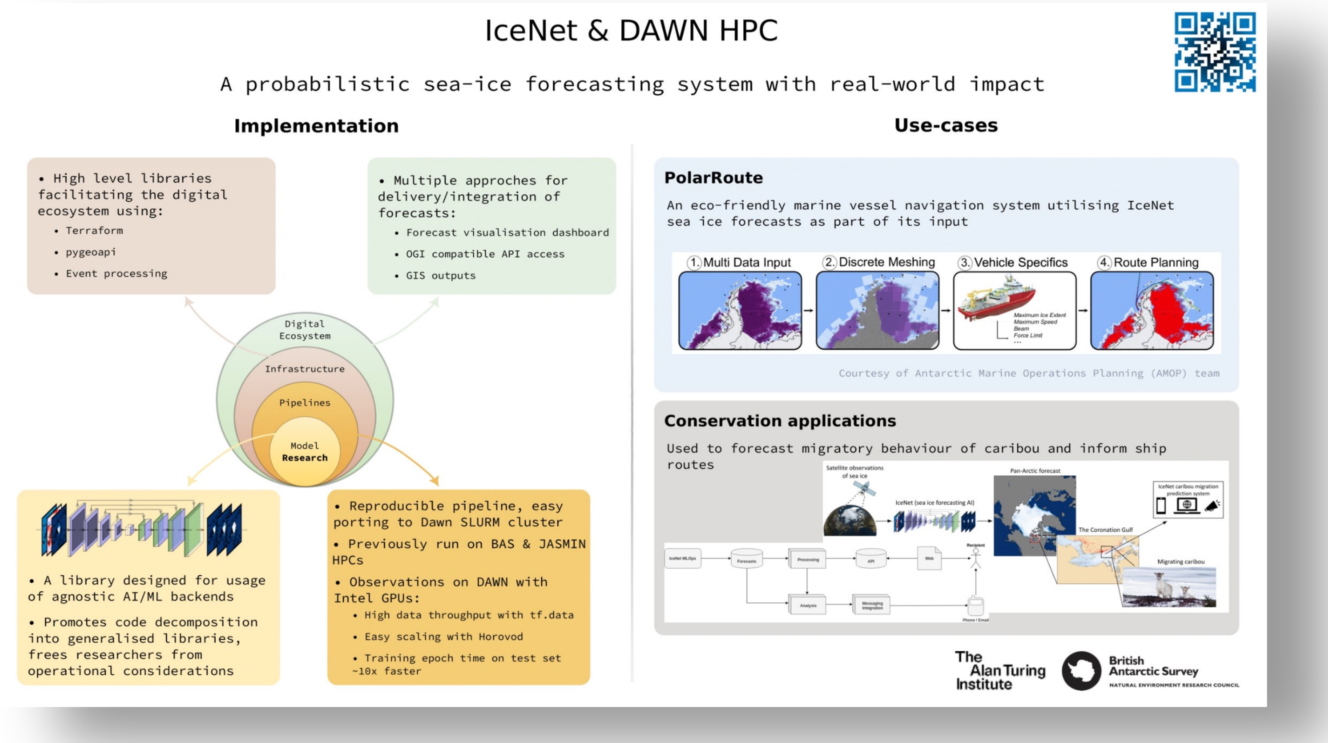
- Dawn entered early science mode in start of last year
- We currently have 140 users across 65 different AI and simulation project
- Most of the focus has been AI projects and we are surprised how easy it is to get NVIDIA PyTorch codes up and running on the Dawn with out of the box performance being acceptable
- AIRR national rolling EOI call now open with new nationally allocated projects being onboarded

DAWN


Projects by Research Domain
Medical 5%
Thermodynamics 5%
Medical Imaging 7%
Organic Chemistry 12%
Life Sciences 19%
Inorganic Chemistry 26%
Others (Law, History, Physics, ...) 26%


AI Model architecture
CNN 7%
GNN 2%
Transformers 12%
MPNN 31%
NLP 5%
LLM 17%
Not defined 19%
GDL 2%
ESM 5%

DAWN



IceNet & DAWN HPC

A probabilistic sea-ice forecasting system with real-world impact

Implementation

- High level libraries facilitating the digital ecosystem using:
  - Terraform
  - pygeoapi
  - Event processing

- Multiple approches for delivery/integration of forecasts:
  - Forecast visualisation dashboard
  - OGI compatible API access
  - GIS outputs

- A library designed for usage of agnostic AI/ML backends
- Promotes code decomposition into generalised libraries, frees researchers from operational considerations

- Reproducible pipeline, easy porting to Dawn SLURM cluster
- Previously run on BAS & JASMIN HPCs
- Observations on DAWN with Intel GPUs:
  - High data throughput with tf.data
  - Easy scaling with Horovod
  - Training epoch time on test set ~10x faster

Use-cases

PolarRoute

An eco-friendly marine vessel navigation system utilising IceNet sea ice forecasts as part of its input

Courtesy of Antarctic Marine Operations Planning (AMOP) team

Conservation applications

Used to forecast migratory behaviour of caribou and inform ship routes

The Alan Turing Institute

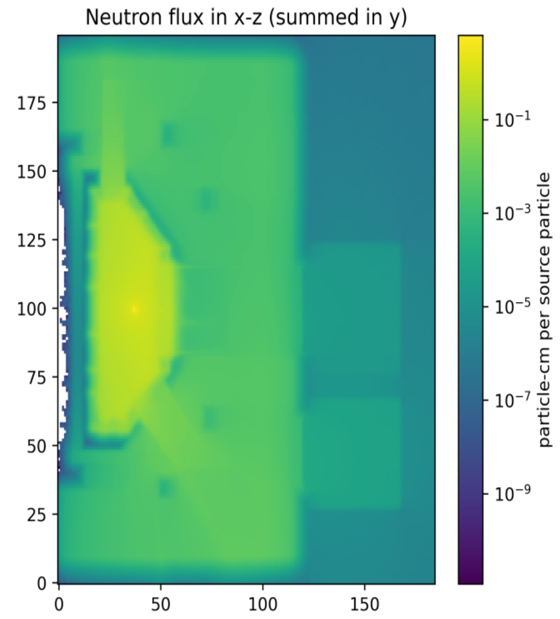
British Antarctic Survey
NATURAL ENVIRONMENT RESEARCH COUNCIL

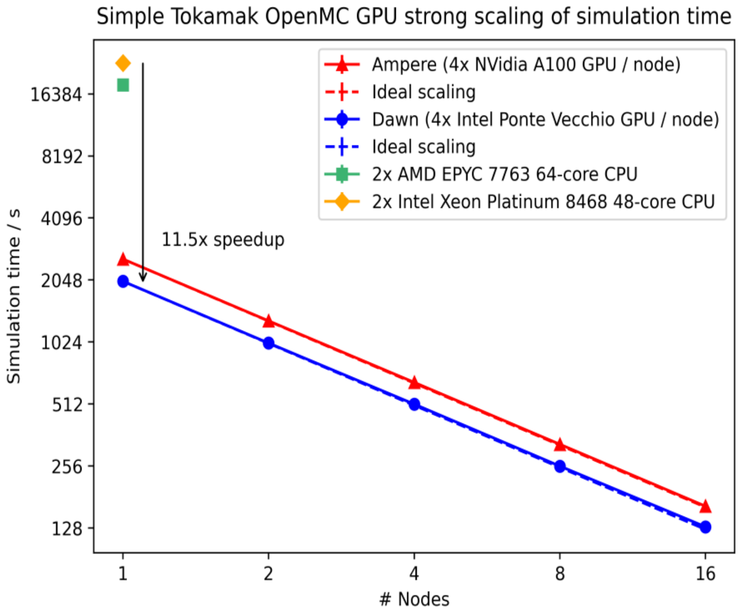
British Antarctic Survey and **The Alan Turing Institute**

Use Dawn to generate AI-powered sea ice forecasts.

DAWN

A New Dawn for Fusion Neutronics


Neutron flux in x-z (summed in y)


Simple Tokamak OpenMC GPU strong scaling of simulation time

Obtaining strong scaling results on Dawn for Monte Carlo Simulation in fusion.

Helen Brooks[1], John Tramms[2], Paul Romano[2], Alex Valentine[1] from [1]United Kingdom Atomic Energy Authority and [2]Argonne National Laboratory

Artificial Intelligence (AI) for Automated Early Detection of Renal Cancer

Bill McGough[1] , Dr. Mireia Crispin-Ortuzar[1]

[1]Department of Oncology, University of Cambridge

ACED

INTERNATIONAL ALLIANCE FOR CANCER EARLY DETECTION

CANCER RESEARCH UK

CANARY CENTER AT STANFORD

UNIVERSITY OF CAMBRIDGE

OHSU KNIGHT CANCER Institute

UCL

MANCHESTER 1824 The University of Manchester

What is AIRRFED

ExCALIBUR 10

University of BRISTOL

StackHPC

OpenNode

UNIVERSITY OF CAMBRIDGE

intel

AIRRFED principles

- One year demonstrator project, started March 2024. Develop fast, don't be afraid to change out later
- Small group – coalition of the willing, with strong agreement on core design elements
- Firm aim to spend MAXIMUM TIME DOING NOT TALKING driven by the need to have a product not to do research on how we can build a product.
- Learning by doing, demonstrate what could be done by modern API driven software infrastructure within a DevOps development and deployment environment
- Use existing access portal technology from Europe, yielding cloud native user environment, re-use, add, contribute back, not re-invent the wheel and spend a year talking about it.
- Strong industry partnership model to drive fast development, with open source commercially supported software components, that already exist.

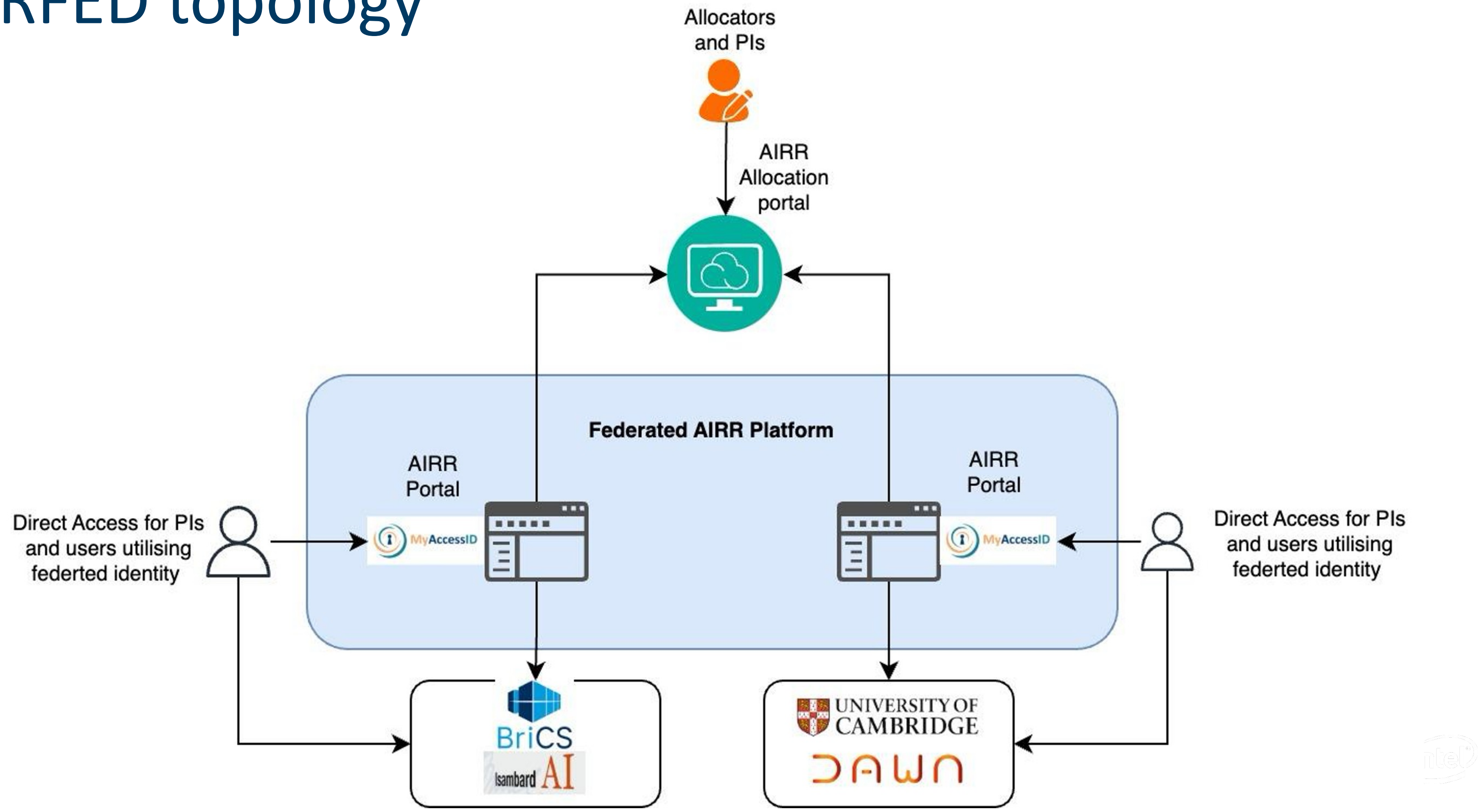
What is AIRRFED

- The AIRRFED project provides a single pain of glass, graphical interface for secure federated discovery, allocation, access and usage to UK national AI and HPC resources for both allocators and users (users being from UKRI, government and industry).

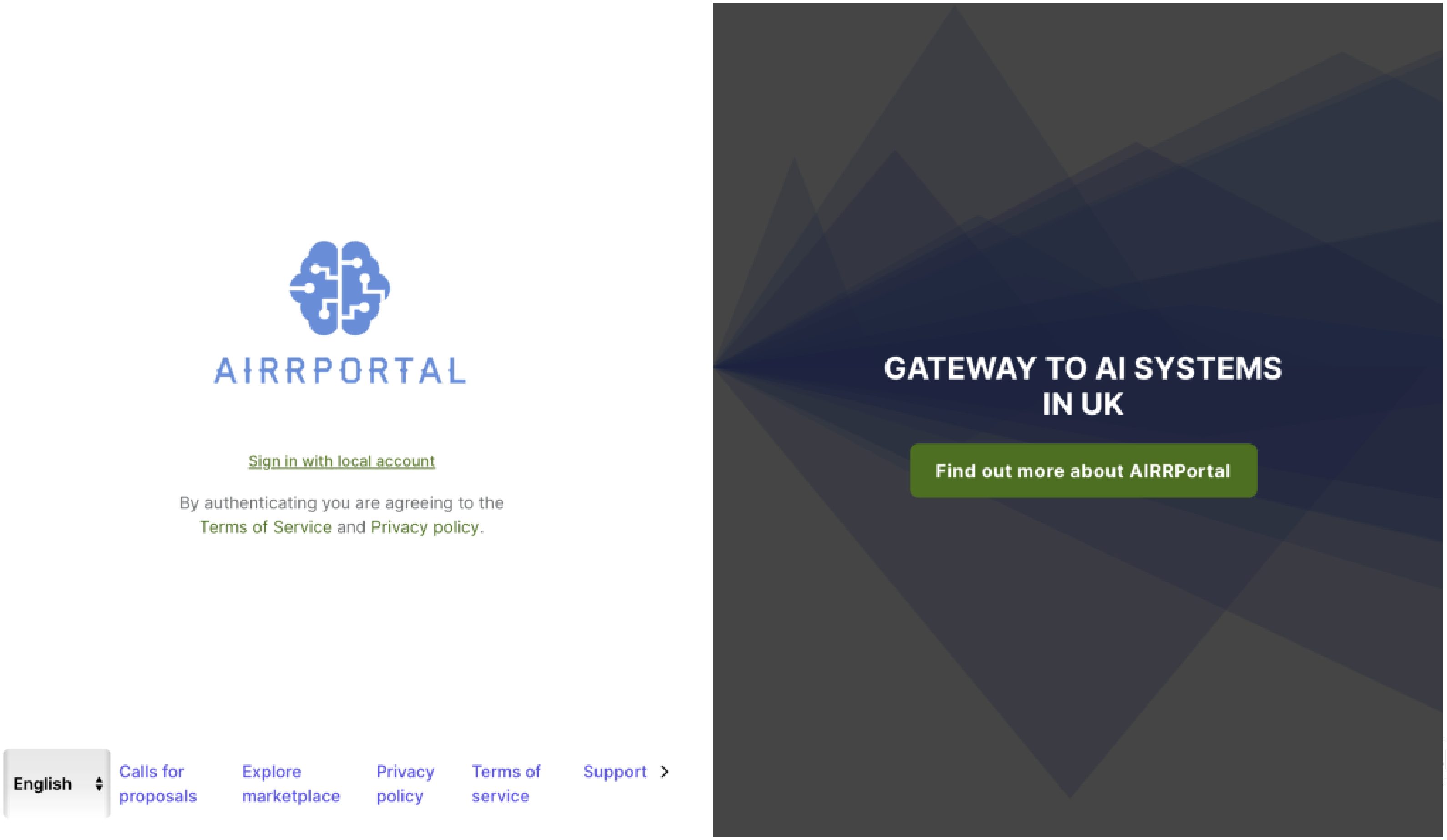
- It is designed to lower the barrier of access to national HPC and AI resources by creating single a top level holistic national gateway for running and responding to calls for resources, allocating projects, award credits, redirect requests, view statistics, monitor usage directly, without waiting for quarterly reports or emailing site team.

- Also, each HPC or AI site then has its own site specific customisable AIRRFED instance and graphical portal allowing federated access to each site, again lowering barrier to access and increasing interoperability between sites, enriching user experience, making it easier for users to access and use a sites HPC and AI systems with rich graphical portal, and highly flexible Science Platform as a Service model. Creating a marketplace for cloud like research platforms across a National DRI

AIRRFED topology



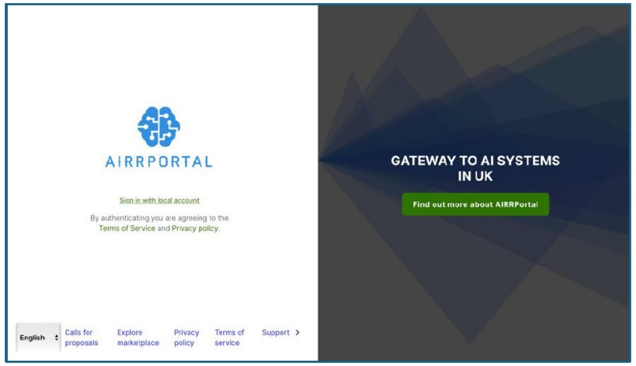
What is AIRRFED



By authenticating you are agreeing to the Terms of Service and Privacy policy.

AIRR Allocator Dashboard

Allocators use AIRRPortal as a single dashboard to allocate projects to the Dawn and Isambard AIRR Systems Can query / graph / ask LLM to understand how allocations are being used



AIRRPORTAL

Sign in with local account

By authenticating you are agreeing to the Terms of Service and Privacy policy.

English | Calls for proposals | Explore marketplace | Privacy policy | Terms of service | Support >

GATEWAY TO AI SYSTEMS IN UK

Find out more about AIRRPortal

AIRRPortal communicates the allocator's commands to the Isambard and Dawn sites. Projects are created. Users onboarded. Resource consumed. All reported back to the allocator's AIRRPortal Dashboard

BriCS

Sign in with local account

By authenticating you are agreeing to the Terms of Service and Privacy policy.

English | Explore marketplace | Privacy policy | Terms of service | Support >

AIRRPORTAL BRISTOL

University of Bristol

UNIVERSITY OF CAMBRIDGE

Sign in with local account

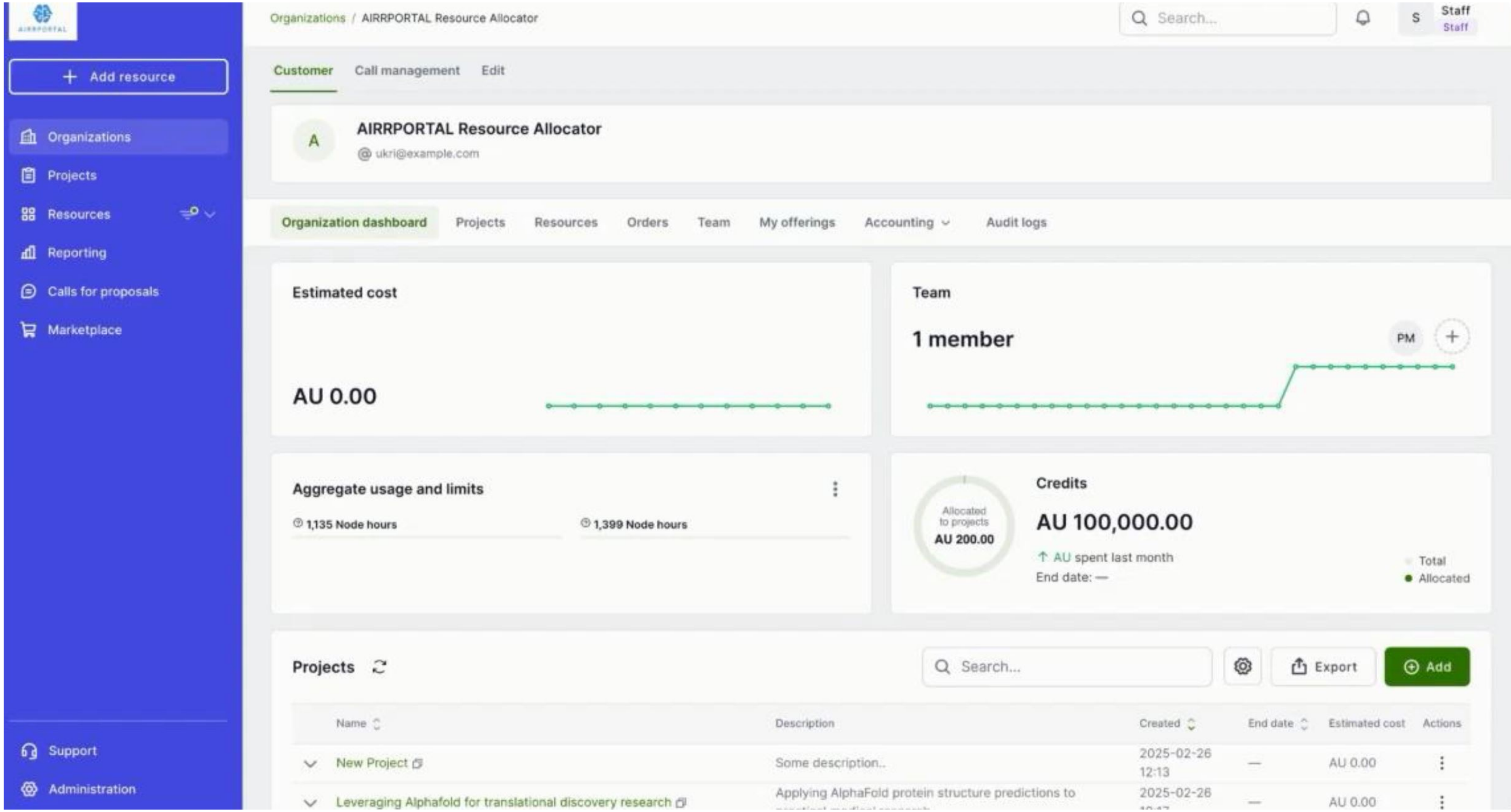
By authenticating you are agreeing to the Terms of Service and Privacy policy.

English | Explore marketplace | Privacy policy | Terms of service | Support >

AIRRPORTAL CAMBRIDGE

University of Cambridge

Single pane of glass empowers allocators to manage and direct AIRR Compute Traffic across national AIRR resources



A

AIRRPORTAL Resource Allocator

@ ukri@example.com

Estimated cost

AU 0.00

Team

1 member

Aggregate usage and limits

1,135 Node hours

1,399 Node hours

Allocated to projects AU 200.00

Credits

AU 100,000.00

↑ AU spent last month

End date: —

Projects

	Name	Description	Created	End date	Estimated cost	Actions
∨	New Project	Some description..	2025-02-26 12:13	—	AU 0.00	⋮
∨	Leveraging Alphafold for translational discovery research	Applying AlphaFold protein structure predictions to	2025-02-26	—	AU 0.00	⋮

Site specific AIRRFED functionality

Federated Identity and Single-Sign-On - Users use their home institution identity to authenticate with the service and get access to AIRR resources

Resource allocation management - PIs can see their allocations and manage their project resources, as well as invite users and delegate responsibilities

Resource accounting and monitoring - PIs and Allocators can monitors how resources are being used

Interactive web based access - Users can get access to remote desktop, ssh terminal, local storage file manager, jupyter notebooks and many other applications that can be executed on AIRR resources

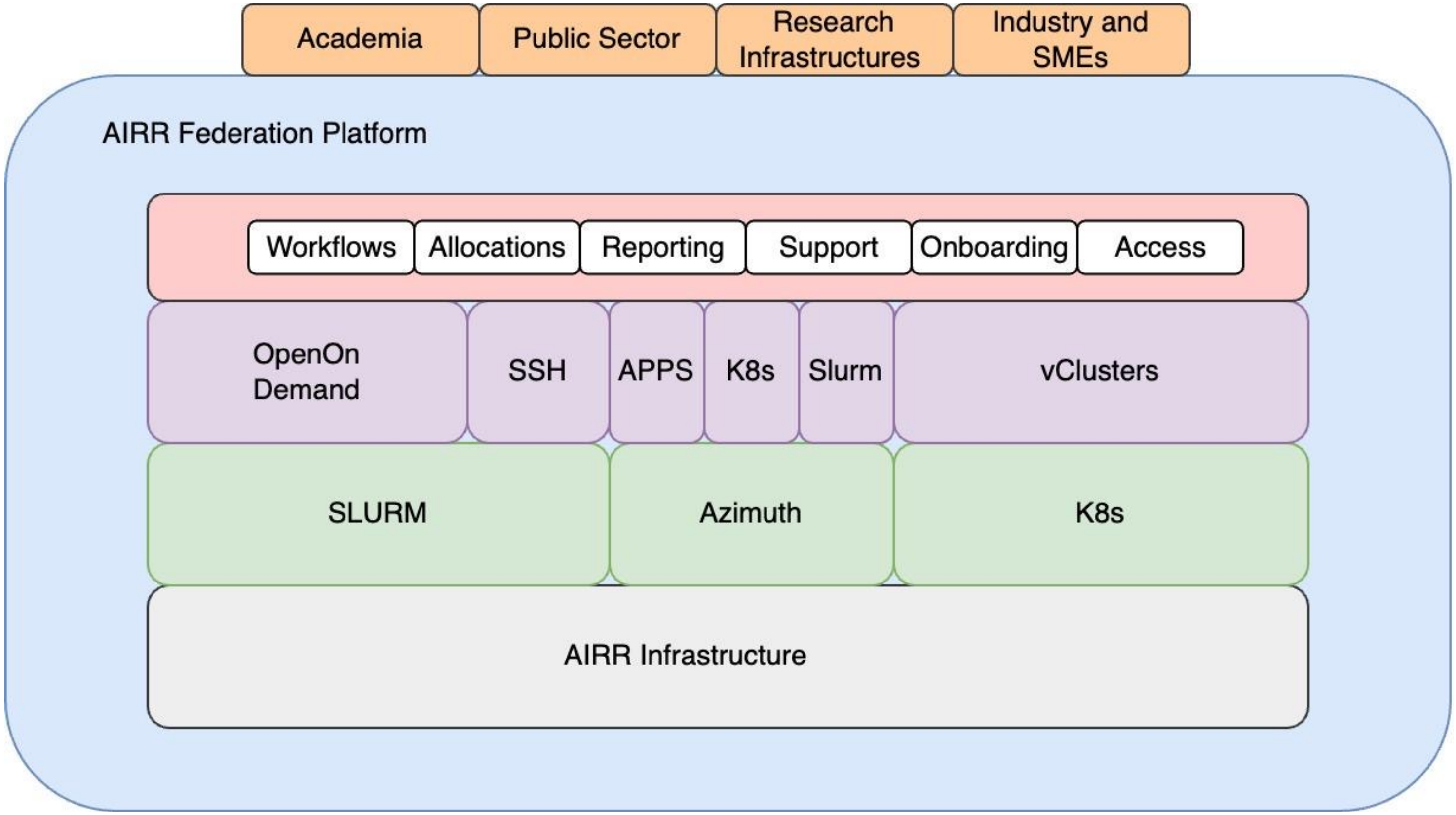
User software environment - User get access to a comprehensive library of software modules.

Direct SSH access - Users get access to the AIRR platform directly via SSH Terminal

K8s platforms - Users get access to dedicated k8s platforms and can deploy their own applications

TREs - User can get access to isolated and compliant TREs as well as integrate their own TREs to execute workloads on AIRR infrastructure

AIRRFED platform architecture



Academia | Public Sector | Research Infrastructures | Industry and SMEs

AIRR Federation Platform

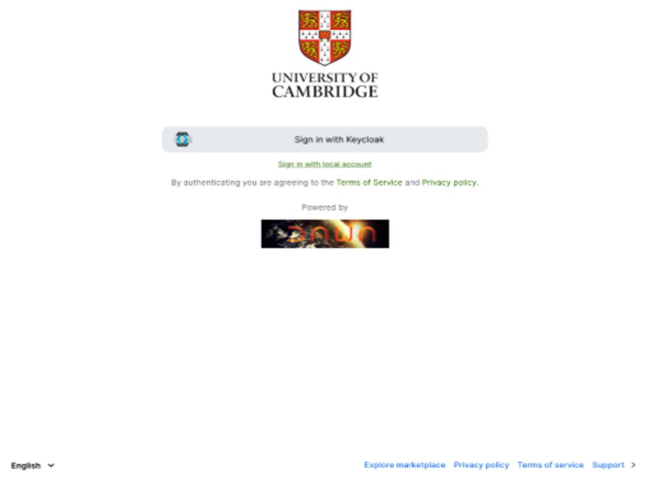
Workflows | Allocations | Reporting | Support | Onboarding | Access

OpenOn Demand | SSH | APPS | K8s | Slurm | vClusters

SLURM | Azimuth | K8s

AIRR Infrastructure

AIRR Federated Platform


UNIVERSITY OF CAMBRIDGE

Sign in with Keycloak

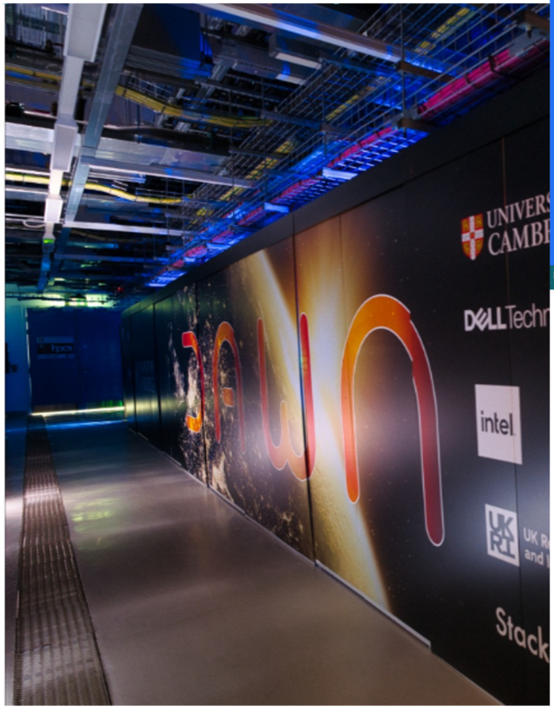
Sign in with local account

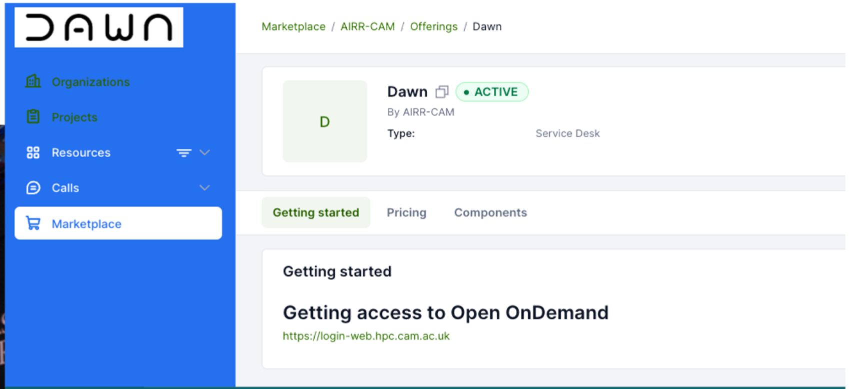
By authenticating you are agreeing to the Terms of Service and Privacy policy.

Powered by

English

Explore marketplace Privacy policy Terms of service Support




Organizations

Projects

Resources

Calls

Marketplace

Marketplace / AIRR-CAM / Offerings / Dawn

D

Dawn ● ACTIVE

By AIRR-CAM

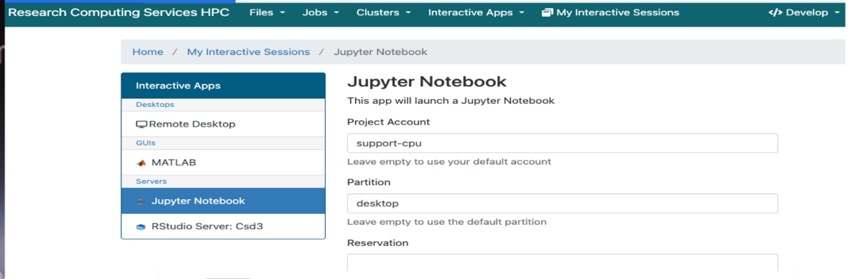
Type: Service Desk

Getting started Pricing Components

Getting started

Getting access to Open OnDemand

https://login-web.hpc.cam.ac.uk


Research Computing Services HPC Files Jobs Clusters Interactive Apps My Interactive Sessions Develop

Home / My Interactive Sessions / Jupyter Notebook

Interactive Apps

Desktops

Remote Desktop

GUIs

MATLAB

Servers

Jupyter Notebook

RStudio Server: Csd3

Jupyter Notebook

This app will launch a Jupyter Notebook

Project Account

support-cpu

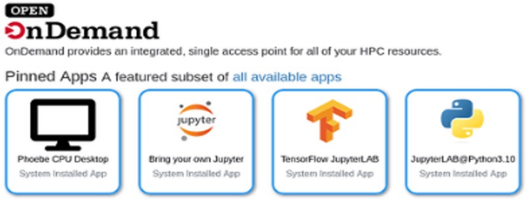
Leave empty to use your default account

Partition

desktop

Leave empty to use the default partition

Reservation


OPEN OnDemand

OnDemand provides an integrated, single access point for all of your HPC resources.

Pinned Apps A featured subset of all available apps

Phoebe CPU Desktop
System Installed App

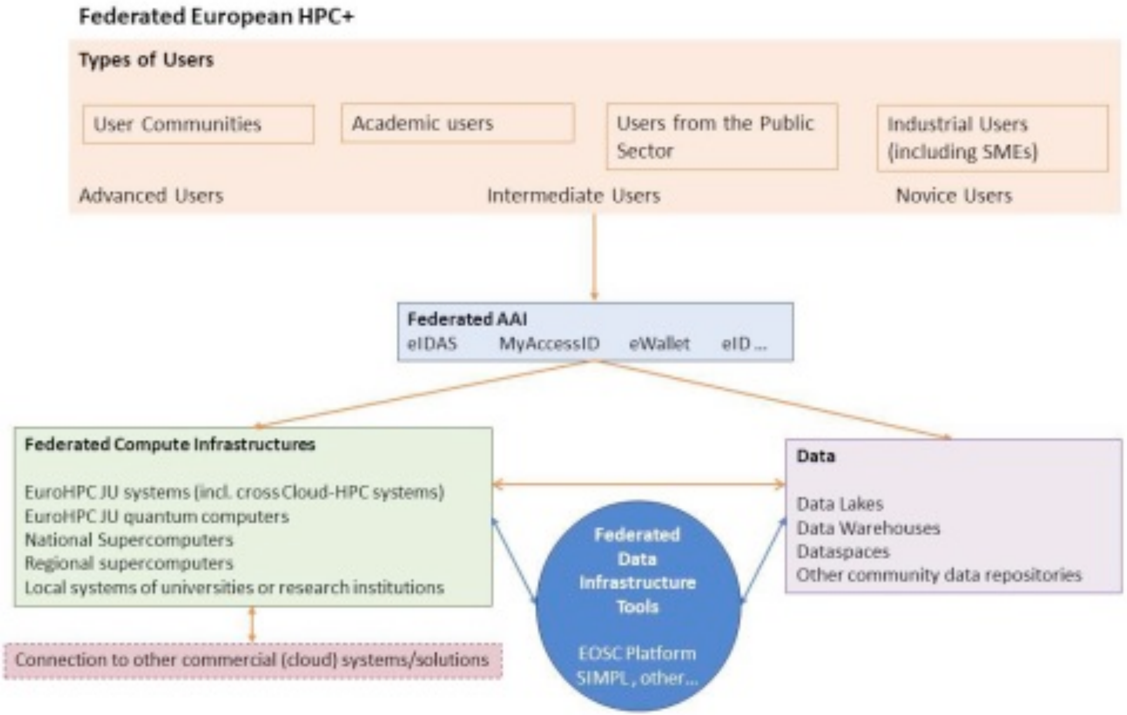
Bring your own Jupyter
System Installed App

TensorFlow JupyterLAB
System Installed App

JupyterLAB@Python3.10
System Installed App

AIRRFED Aligned with Europe's direction

Aligned with “The EuroHPC JU vision is to establish a world-leading federated and secure HPC and quantum service infrastruct ecosystem in the Union and ensure wide use of this infrastructure to many public and private users, to support the developme of key skills for European science and industry.”



Federated European HPC+

Types of Users

User Communities

Academic users

Users from the Public Sector

Industrial Users (including SMEs)

Advanced Users

Intermediate Users

Novice Users

Federated AAI

eIDAS MyAccessID eWallet eID ...

Federated Compute Infrastructures

EuroHPC JU systems (incl. cross Cloud-HPC systems)

EuroHPC JU quantum computers

National Supercomputers

Regional supercomputers

Local systems of universities or research institutions

Connection to other commercial (cloud) systems/solutions

Federated Data Infrastructure Tools

EOSC Platform

SIMPL, other...

Data

Data Lakes

Data Warehouses

Dataspaces

Other community data repositories

AIRRFED status and next steps

- AIRREFD Portals deployed and production ready at Bristol and Cambridge, Bristol already in Production usage onboarding all users to Isambard 3 Tier 2 HPC System and Isambard AI. Cambridge in production usage very shortly, all goals and stretch goals completed !
- AIRRPORTAL top level demonstrator project under development and testing, allowing UKRI and DSIT to see the art of the possible as the UK re-imagens what its future DRI ecosystem could look like
- AIRRFED project being developed activity developed with UKAEA for potential AI infrastructure deployment at Culham AIGZ – specifically looking to make it easy to use AIRRFED to scale out to public cloud providers. Also, strong TRE development underway "FRIDGE" + "PHAROS"
- AIRRFED development and testing to become a major theme within Zettascale lab unlocking investment from Dell and intel
- Need to explore future governance ,development and funding models for AIRRFED, breaking the mold of traditional academic led software projects, stronger role for industry, how do we create a sustainable model for development and operationalizing tools like this