**STFC IRIS**
Science Director:
Jon Hays
Technical Director:
Andrew Samsun

IRIS Resource Request
v10/22

# IRIS Resource Request

## 1 Administrative details

**Project Name: GAIA**

**Lead Contact: Nicholas Walton naw@ast.cam.ac.uk**

**Further Contacts:**

- **Core Processing: Patrick Burgess: pwb@ast.cam.ac.uk**
- **Data Mining Platform: Nigel Hambly: nch@roe.ac.uk / Dave Morris: dmr@roe.ac.uk**

*Version: v2: 20230104*

## 2 Glossary

- BP/RP: Red and Blue Photometer – Gaia's low resolution spectrophotometry
- DMP: Data Mining Platform
- DPAC: The Gaia Data Processing and Analysis Consortium - https://www.cosmos.esa.int/web/gaia/dpac/consortium
- DPCI: Data Processing Centre, IoA, Cambridge – one of the core Gaia DPAC processing centres. https://www.gaia.ac.uk/gaia-uk/ioa-cambridge/dpci
- ESA: European Space Agency
- ML: Machine Learning
- RSE: Research Software Engineer

## 3 Usage made of IRIS resources in the previous year

We provide usage information for the two aspects of the Gaia allocation, namely the Core Processing Activity, and the Data Mining Platform work.

### 3.1 IRIS resources allocated to your project

#### 3.1.1 Gaia Core Processing

Openstack cloud resources are located at Cambridge. The allocation previously has been used to prototype replacement system for non-IRIS hardware that is to be decommissioned.

The previous allocation fulfilled requirements for:

- 5824 cores for distributed processing cluster.
- 216 cores for development and operations systems.
- 384 cores for Science Alerts systems.

Allocated 6424 cores, 4272TB disk on the Cambridge Arcus Openstack cloud.

| Resource description (machine view) | | | | | Summary (please see notes) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Count | Cores/ total RAM | GPU cards/ | Attached fast storage | Location | CPU cores | CPU mem/core | GPU cards | GPU mem/ card | Storage/ core |

|  |  | total onboard memory |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| N/A | N/A | N/A | N/A | Cambridge | 6424 | 10G | 0 | N/A | N/A |

Table 3.1.1A: Current allocation for Gaia Core Processing

**Storage**

| Amount | Location | Disk/Tape |
|---|---|---|
| 4272TB | Cambridge | Disk |
| 600TB | N/A | Tape |

Table 3.1.1B: Storage allocated to Gaia Core Processing

### 3.1.2 Gaia Data Mining Platform

The resources allocated to date have been used to maintain and develop a live end-user science exploitation platform known as the UK Gaia Data Mining Platform. This involved a large investment of effort in familiarisation with relevant technologies (OpenStack; Apache Spark and associated "big data" handling components) and working with the providers (StackHPC and DiRAC operations personnel) as well as routine operations such as data transfer and loading. Much experimentation was needed to determine optimal data organisation and in the development of example workflows. The resulting system presents a web notebook (Apache Zeppelin) user interface with Python as the main interpreter. PySpark/SQL provides a convenient and user-friendly API to distributed data objects that can be operated on using functionality provided via Spark including common machine learning (ML) algorithms and a high degree of end-user programmability. Hence the system provides a platform on which users can run code next to the data, leveraging distributed processing on a (Spark) cluster to provide scalability to workflows processing very high data volumes. At the time of writing, around 30 users (including post-graduate students and early-stage post-doctoral researchers) have been on-boarded to exercise the system and exploit the science data. Science results generated from use of the system are beginning to appear (e.g. Crake et al. 2023). In parallel with running a live end-user platform we continue to develop the system for scaling to the next major data release (Gaia DR4, end 2025) which will see a huge step change in data volume from 9T (DR3) to ~500T.

**CPU resources provided as VMs by Cambridge Arcus (OpenStack Cloud) via the "CCLake" pool**

| Resource description (machine view) |  |  |  |  | Summary (please see notes) |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| Count | Cores/ total RAM | GPU cards/ total onboard memory | Attached fast storage | Location | CPU cores | CPU mem /core | GPU cards | GPU mem / card | Storage/ core |
| 5 | 55/188 GiB | 0 | 0 | Camb. | 550 (virtual) | 1.7 GiB | 0 | 0 | 0 |
| 2 | 55/512 GiB | 0 | 0 | Camb. | 220 (virtual) | 10 GiB | 0 | 0 | 0 |

Table 3.1.2A: Current (October 1st 2023) allocation for the Gaia DMP. The two rows distinguish between 'standard' and 'high' memory nodes. A total of 770 vCPU was allocated and provided; attached fast storage was applied for in the last round but is not available at the time of writing.

Machine view numbers correspond to physical hardware to which our project is “pinned” while the summary provides numbers factoring in x2 hyperthreading.

**Storage**

| Amount | Location | Disk/Tape |
|---|---|---|
| 72 TiB | Cambridge | CephFS share |
| 72 TiB | Cambridge | Ceph volume |
| 20 TiB | RAL | Echo S3 |

Table 3.1.2B: Storage allocated to the Gaia DMP at 1st October 2023. An additional 32TiB of direct attached storage was allocated in the previous round yielding a total of 196TiB for the Gaia Data Mining Platform, but at the time of writing this has not been provided.

## 3.2 Current usage of IRIS resources

### 3.2.1 Gaia Core Processing

**CPU/GPU Usage for 01/10/2022 to 31/09/2023 (or closest reporting period)**

The latest allocation (Oct 23 to Oct 24) brings the scale up to the level required to replace the previous system. Currently approximately 10,252 vCPUs in use to provide 155 VM instances. The majority are used to form the data processing cluster and the remaining allocation are being brought online to replace the services noted in section 4.1.2.

Storage allocation configuration is still under discussion with the provider. Final experiments are being carried out to determine configuration and performance trade-offs to reduce overall resource usage by allocating a section of 4PB Ceph storage to erasure coding instead of replication, which requires less storage capacity to provision. Plans are in place to provide high bandwidth network connection between the current non-IRIS storage and IRIS new allocation of IRIS storage to allow rapid transfer of remaining current data.

**Storage usage (October 2023 or closest reporting period)**

| Amount | Location | Type | Usage (Oct 2023) | 85% expected |
|---|---|---|---|---|
| 4272TB | Cambridge | Disk | 473TB | February 2024 |

Table 3.2.1: Storage use Gaia Core Processing

### 3.2.2 Gaia Data Mining Platform

**CPU/GPU Usage for 01/10/2022 to 31/09/2023**

All standard memory CPU is in use, either permanently allocated to our end-user live system, or split between development and testing platforms.

**Storage usage (October 2023)**

| Amount | Location | Type | Usage (Oct 2022) | 85% expected |
|---|---|---|---|---|
| 72 TiB | Cambridge | CephFS share | 100% | - |
| 72 TiB | Cambridge | Ceph volume | 85% | - |
| 20 TiB | RAL | Echo S3 | 100% | - |

Table 3.2.2: Storage use currently for the Gaia DMP. RAL Echo S3 storage acts as backup and publicly accessible repository for Spark Parquet formatted file sets.

# 4 Computing Model and Computing environment

## 4.1 Computing Model: Gaia Core Processing

The core photometric processing makes use of computing resources in three main areas:

- Core distributed processing.
- Development and operations.
- Science Alerts.

### 4.1.1 Core distributed processing

The core distributed processing is the cyclic processing of Gaia data to produce calibrated data products for use in generation of Gaia data releases.

The main challenges presented by the processing of the Gaia data are due to the large data volume and to the intrinsic complexity of the data. Gaia is a self-calibrating instrument which implies the need for many iterations over the entire dataset to capture and calibrate the behavior of the instrument over the entire focal plane and in different observing configurations. The calibration process is further complicated by the rapid variability both in space and time of many instrumental and sky-related effects.

Current data processing, for data release 4, deals with approximately 120 billion instrument observations - composed of over 1 trillion individual sensor observations - matched to 2.8 billion sources. Inputs and outputs of data processing are a few hundred TB with more data volume generated in intermediate results.

During the request period (Q4 2024-Q4 2025), we will be carrying out final preparation for the processing of data release 5. This is the final data release of the Gaia mission and is approximately twice the scale of data release 4. This preparation will involve large scale test executions of processes to verify performance and data quality. In addition, analysis of data release 4 outputs will be ongoing to support the publication of data release 4 which is currently planned after the end of 2025.

The nature of the core photometric processing requires that all storage and compute should be located in a single installation (excluding tape backup which could be more isolated if sufficient bandwidth is available). Large input data sets are used to produce large intermediate and final data sets and data must be combined in a variety of ways during various elements of processing.

The increase in required storage since the previous request is to cover:

| Data | Increase from previous (TB) |
|---|---|
| Cyclic data deliveries in preparation for data release 5 processing. | 150 |
| 6 months of operational daily data to May 2025 | 15 |

Estimates of required storage volume in TB are:

| Data | Volume TB – previous request in () where changed |
|---|---|
| Input data archive | 765 (600) |
| Cycle 3 retained intermediate data products | 400 |
| Cycle 3 final data products | 120 |
| Pre-processed input data | 300 |
| Cycle 4 retained intermediate data products | 800 |

| | |
|---|---|
| Cycle 4 final data products | 400 |
| Cycle 4 processing workspace | 800 |
| Operational storage - Delivery database and validation database storage, scratch space for off-cluster analysis, etc. | 100 |
| TOTAL | 3685 |

Total disk data storage required is estimated as 3685TB.

### 4.1.2 Development and Operations

Various systems that are not part of the main processing cluster are required to support development and operations. These are hosted via OpenStack provisioned virtual infrastructure.

Services:

- Database server
- Web server
- Data transfer (Aspera)
  - transfers between DPCI and the rest of the consortium via the internet. Transfers often happen automatically. Requires constant availability (24 hours, 365 days). Transfers range from very small (few MB or GB) or, less frequently, very large (50TB or more).
- Software repository and data provision
  - Hosting internally developed software packages and also used for supplying validation data to DPAC members.
- Build and continuous integration
  - Jenkins system providing continuous integration and release of internally developed software.
- System monitoring and management
  - Prometheus/Grafana monitoring.
  - Syslog analysis to alert on system log errors.
  - Puppet configuration management.
  - Internal email integration with IoA mail system to allow system email notification.
  - DNS provision.
- Data analysis
  - systems provided for internal user analysis of data products that is not feasible on desktop/laptops due to volume.
- Validation database
  - provides local hosting of small (a few TB) datasets for internal user validation.

Allocation of the resources for these services are:

| **Service** | **CPU** | **RAM** |
|---|---|---|
| Database server | 12 | 24 |
| Web server | 8 | 12 |
| Data transfer (Aspera) | 8 | 12 |
| Software repository and data provision | 12 | 32 |
| Build and continuous integration | 12 | 48 |

| | | |
|---|---|---|
| System monitoring and management | 12 | 24 |
| Data analysis | 128 | 128 |
| Validation database | 24 | 48 |
| TOTAL | 216 | 328 |

### 4.1.3 Gaia Core Processing: Science Alerts

Science Alerts operation accumulates and processes data continuously, typically ingesting one batch of new observations per day and emitting alerts on those data within a few hours. Alert generation combines new data with historic data from the whole mission, so the total data-set used each day is very large.

The group has negotiated continued hosting and support for the current Science Alerts hardware systems through to May 2025. This covers the expected operational lifetime of the satellite and therefore the operational lifetime of the Science Alerts system.

IRIS resources are requested for the period Q4 2024 to Q4 2025 to support:

- storage of the Science Alerts data beyond the lifetime of the current hardware systems.
- analysis and post-processing of the accumulated data to produce required outputs for the final cyclic data release.
- continued community web access to the final Science Alerts database.

Estimated resources to support Core Processing: Science Alerts functions are given below.

| Service | CPU cores | RAM/core | Block storage (TB) |
|---|---|---|---|
| Compute | 128 | 6 | 24 |
| Back-up | N/A | N/A | 312 |

Total allocation for Science alerts is 128 CPUs and 336TB storage.

## 4.2 Computing Environment: Gaia Core Processing

### 4.2.1 Basic Compute Information

10GB RAM per core was chosen for the current cluster as a compromise. Different processes benefit from different core/memory balance. This level works well for most moderately memory hungry processes allowing most processes to utilise close to the full core count.

### 4.2.2 Access requirements

Access to resources will be mainly via ssh as it is with our current system. Sysadmins use ssh to VMs in order to deploy, manage and maintain the system. Users have limited  ssh access to specific VMs to submit processing jobs, access cluster file resources, run ad-hoc local processing or carry out remote data visualization.

### 4.2.3 Storage

Disk storage will be configured and provided though the OpenStack system. A mixture of block device provision for persistent attached volumes and Ceph network shares. The Ceph network shares will be provisioned using a mixture of erasure coding and replication. The replication provides better performance but the erasure coding requires less capacity per unit usable storage.

### 4.2.4 Networking

#### *4.2.4.1 Internal Network*

The processing cluster needs to be able to read and write large amounts of data with reasonable performance. Input and output data volumes approaching 100TB are not unusual during processing.

Additionally, some processes require iteration with exchange of several TB intermediate data products between processing nodes at each step.

#### 4.2.4.2 External IP addresses

We require a small number of externally visible IP addresses permanently assigned to the project and routed to project VMs.

- SSH access for development and operations.
- Target and source of data transfers with the other DPAC data centres.
- External web services (some public but also proxy for continuous integration, data provision etc).
- Integration with ESA authentication system to provide common access to hosted services.

5 IP addresses, assignable within Openstack is sufficient. Our current Openstack project provides for this.

#### 4.2.4.3 External Bandwidth

The system needs to be able to reliably send/receive large amounts of data via the internet at close to 1Gbps.

Data transfers are carried out automatically throughout the year (24 hours a day, every day). The external bandwidth and public IP address for the data transfers must be always available – except for planned maintenance periods during which transfers can be suspended.

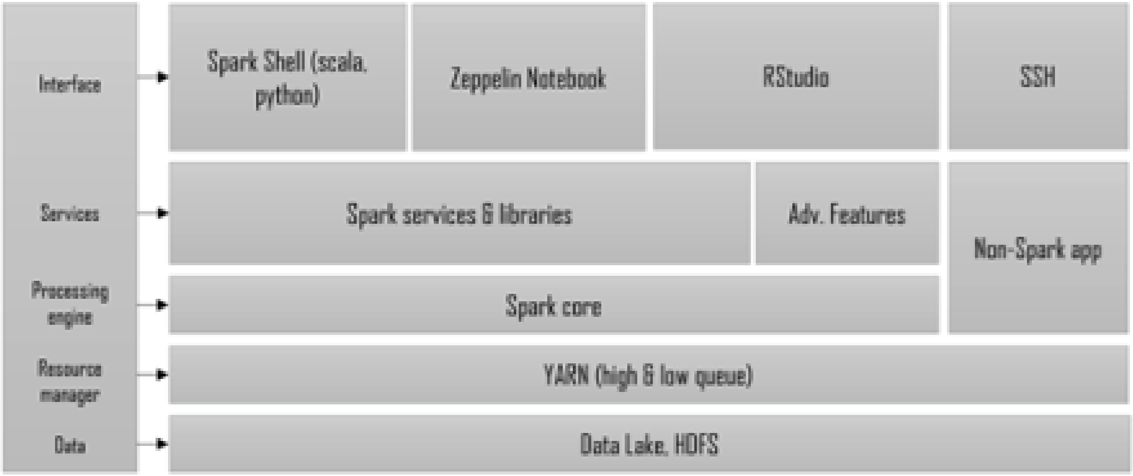
#### 4.2.4.4 Software

Software including OS will be provided, installed and maintained by DPCI team as is done with the current installation.

**Acknowledgement of limitations of IRIS computing support:** There is a team of 3 people based at Cambridge funded through the data processing grant who are responsible for management of the current hardware. This team will be responsible for utilization of allocated IRIS resources.

### 4.3 Computing Model: Gaia Data Mining Platform

The high-level architecture of the Gaia DMP is as follows:



The system is based on the Gaia Data Analytics Framework developed by our Gaia DPAC colleagues at the University of Barcelona under the auspices of the FP7 project Gaia European Network for Improved User Services (GENIUS; EC FP7 606740). The architecture is built upon the Apache "big data" processing stack. At the core of the system are the Hadoop Distributed File System and the

Parquet partitioned data format providing high throughput along with horizontal CPU scaling (clustering) to provide a high-performance data analysis platform via the Spark computing framework. Use of the platform includes:

- A range of compute jobs, from single threaded CPU (using the parallelism inherent to Spark to provide the scale-up) to expert users employing parallelism with multi-threaded tasks
- End-user usage is data analysis, (forward-) modelling
- Any substantial simulations will be done by other means - user compute jobs will be mainly I/O bound
- Batch scheduling via a bespoke resource reservation system is planned for development in 2024.

## 4.4 Computing Environment: Gaia Data Mining Platform

### 4.4.1 Basic compute information:

Benchmarking of a bare-metal deployment of the prototype analysis system built on the Apache Spark software stack used a 6-node, 16 cores/node, 64 GB per node system which we would consider to be the minimum useful configuration. Higher memory nodes (512 GB) and a larger number of nodes have been allocated previously via IRIS. The current resource request is based on our experience of deploying and maintaining an end-user service for the 9TB Gaia Data Release 3, and the likely requirements for scaling up to the 500TB DR4. The combination of IO-bound and CPU-bound usage scenarios envisaged implies that both distribution over more nodes, and greater cores-per-node and/or hyper-threading within each core are beneficial. Overheads on user staging and scratch space no more than 10% of the bulk data release volume.

We have exercised many workflows on the OpenStack deployment of Spark via the IRIS allocations to date, and find that small-scale virtualized clusters with tens of nodes easily service distributed use cases, while CPU-heavy, non-parallelized workflows require configuration of a single worker node with higher memory. The orchestrated environment provisioned via OpenStack Magnum/Kubernetes is a good fit to our requirements as it provides the flexibility to create these kinds of different user environments on-demand.

Some level of provision of fast local persistent storage would be advantageous for the most commonly queried catalogue subsets. For network attached storage we observe CPU wait times of up to 50% in the systems provided to date (comprising OpenStack VMs with storage provisioned through CephFS). Hence we again request a modest amount of directly attached fast disk with which to address this limitation.

### 4.4.2 Access requirements:

Requested enabling infrastructure:

- Cloud middleware: OpenStack with Magnum, Manila and Blazar
- Authentication and Accounting Infrastructure: IRIS-IAM

### 4.4.3 GPUs:

No GPU resources are requested in the current allocation.

### 4.4.4 Storage:

Some further considerations on storage and data handling:

- Gaia DRs are a few TBs in size (DRs 1, 2 & EDR3) to 10 TB (DR3) to a few tens of times larger (~500TB to 1 PB, full DR4 to DR5, projected). A typical scale-out usage mode might require additionally a few percent of this volume (anywhere between 10s of GB to 10 TB) in temporary analysis space. Any derived data products requiring longer-term preservation are likely to be insignificant in comparison to these requirements

- There is no requirement for bulk data archiving on tape - back-up of public released data is provided by copies hosted elsewhere
- Data releases are generated periodically (every few years) at the European Space Astronomy Centre (ESAC) near Madrid, Spain. A commercial Content Delivery Network on the WAN is being used to distribute bulk data.
- Compute jobs are strongly I/O limited

Data management (following https://stfc.ukri.org/funding/research-grants/data-management-plan/)

- Type of data generated: published data generated from Gaia public data releases.
- Published data preservation: as per individual research project data management plans
- SW and metadata: anticipate end-users will employ external repositories (e.g. GitHub) with further description in associated publications in learned journals
- Long term preservation: as per individual user research project DMPs
- Shared data: anticipate research groups will share data via shared space allocated within IRIS prior to publication of results, and then ideally in a publicly accessible area available once their analyses are published
- Proprietary periods: as per standard research practice, derived data will be proprietary to any research group during their analyses and prior to publication of their results. Once published, anticipate that groups will want to make their derived data publicly available according to their own individual plans as part of their grant awards.
- How data will be shared: see above.
- Specific resources required for preservation and sharing: no staff resources, but see above for computing resources.

### 4.4.5 Networking

No special requirements

### 4.4.6 Software

Software we need:  OpenStack with Magnum, Manila and Blazar

Software we use

- Apache Hadoop, Spark and YARN
- Apache Zeppelin
- Apache Parquet
- Kubernetes
- Python (including various third-party libraries) and Java
- ML libraries in Spark and Python

**Acknowledgement of limitations of IRIS computing support:**
We acknowledge the statement of limitations of support available directly from IRIS. Our Gaia DMP project has RSE staff resources from a special STFC Gaia “CU9” (exploitation-enabling) grant.

# 5 Resource request for October 2024 – October 2025

**These resources are in addition to the current resources indicated in section 3.1 IRIS resources allocated to your project**

## 5.1 Additional Request 24/25: Gaia Core Processing

For the Gaia Core Processing, due to the renegotiation of Science Alerts maintenance and therefore reduction of the resources required for that section, the resource request is a slight reduction from the previous allocation.

Our CPU request drops by 256 cores from the previous 6424 cores to 6168. These are computed as **physical** cores. This is made up of:

- 5824 cores for processing cluster.
- 216 cores for services.
- 128 cores for Science Alerts.

The storage request drops by 251TB from the previous 4272TB to 4021TB. Science Alerts has reduced storage requirements by 336TB but a combination of increased data requirements for future data inputs and the removal of an 80TB rounding used last time makes the actual reduction 251TB.

This is made up of:

- 3685TB for processing cluster and services.
- 336TB for Science Alerts.

**CPU/GPU**

| Resource description (machine view) | | | | Summary (please see notes) | | | | |
|---|---|---|---|---|---|---|---|---|
| Count | Cores/ total RAM | GPU cards/ total onboard memory | Attached fast storage | CPU cores | CPU mem/core | GPU cards | GPU mem/ card | Storage/ core |
| N/A | N/A | N/A | N/A | 6168 | 10G | N/A | N/A | N/A |

Table 5.1A Absolute request for Gaia core processing.

**Storage**

| Amount | Preferred Location | Type (Disk/Tape) |
|---|---|---|
| 4021TB | Cambridge | Disk |
| 600TB | Cambridge | Tape |

Table 5.1B Absolute amount of storage requested from IRIS

## 5.2 Additional Request 24/25: Gaia Data Mining Platform

In this additional resource request for the Gaia DMP we ask that the following be taken into consideration:

- Although the next major data release (DR4) is beyond the current allocation period we require to stress test the handling of DR4-like data volumes with mock data (at least) during 2024/25;
- Certain bulk datatypes in DR4 may be deliverable to us earlier than the official release date of end 2025 (exact timescale uncertain at this point);
- There is inevitably a lag between applying and receiving a new storage allocation.

For these reasons we are requesting sufficient storage to handle Gaia DR4 in the current allocation round. A more detailed justification is provided following the tables.

**CPU/GPU: Gaia DMP: Cambridge Arcus Cloud live plus dev system deployments, in absolute numbers**

| Resource description (machine view) | | | | Summary (please see notes) | | | | |
|---|---|---|---|---|---|---|---|---|
| Count | Cores/ total RAM | GPU cards/ total onboard memory | Attached fast storage | CPU cores (virtual) | CPU mem/core | GPU cards | GPU mem/ card | Storage/ core |
| | 1400/2 TiB | N/A | (32 TiB) | 1400 | 1.4GiB | N/A | N/A | (23GiB) |

Table 5.2A: overview of Gaia DMP live and dev system deployments resource requested: total vCPU, RAM and direct attached fast storage at Cambridge Arcus Cloud. At the time of writing the 2023 allocation of 32 TiB direct attached fast storage is in the process of being provided with the Cambridge Arcus cloud and will be removed in the final submission if successfully deployed.

**CPU/GPU: Gaia DMP: Edinburgh Somerville Cloud live deployment, in absolute numbers**

| Resource description (machine view) | | | | Summary (please see notes) | | | | |
|---|---|---|---|---|---|---|---|---|
| Count | Cores/ total RAM | GPU cards/ total onboard memory | Attached fast storage | CPU cores (virtual) | CPU mem/core | GPU cards | GPU mem/ card | Storage/ core |
| | 700/1 TiB | N/A | 16 TiB | 700 | 1.4GiB | N/A | N/A | 23GiB |

Table 5.2A: overview of Gaia DMP system deployment resource requested: total vCPU, RAM and direct attached fast storage at Edinburgh Somerville Cloud.

**CPU/GPU: Gaia DMP: RAL STFC Cloud system deployments, in absolute numbers**

| Resource description (machine view) | | | | Summary (please see notes) | | | | |
|---|---|---|---|---|---|---|---|---|
| Count | Cores/ total RAM | GPU cards/ total onboard memory | Attached fast storage | CPU cores (virtual) | CPU mem/core | GPU cards | GPU mem/ card | Storage/ core |
| | 700/1 TiB | N/A | 16 TiB | 700 | 1.4GiB | N/A | N/A | 23GiB |

Table 5.2A: overview of Gaia DMP system deployment resource requested: total vCPU, RAM and direct attached fast storage at RAL STFC Cloud.

**Storage**

| Amount | Preferred Location | Type (Disk/Tape) |
|---|---|---|

| | | |
|---|---|---|
| 1000 TiB | Cambridge (Arcus) | CephFS Manilla |
| 100 TiB | Cambridge | Object Swift |
| 10 TiB | Cambridge | Block Cinder |
| (32 TiB | Cambridge | Direct attached SSD) |
| 500 TiB | Edinburgh (Somerville) | CephFS Manilla |
| 50 TiB | Edinburgh | Object Swift |
| 5 TiB | Edinburgh | Block Cinder |
| 16 TiB | Edinburgh | Direct attached SSD |
| 500 TiB | RAL (STFC Cloud) | CephFS Manilla |
| 50 TiB | RAL | Object Swift |
| 5 TiB | RAL | Block Cinder |
| 16 TiB | RAL | Direct attached SSD |

Table 5.2B Total amount of storage requested. At the time of writing the 2023 allocation of 32 TiB direct attached fast storage is in the process of being provided with the Cambridge Arcus cloud and will be removed in the final submission if successfully deployed.

**Justification:**

Peak estimates: This resource request is based on estimates for the peak load expected when Gaia DR4 is released at the end of 2025. Our current Gaia DR3 dataset is around 9TiB. Gaia DR4 is expected be of the order of 500TiB. The current timeline is to release Gaia DR4 at end Q4 2025, although we may start to receive certain high volume parts of the data during 2025. When DR4 is released at the end of Q4 2025 we are expecting to have to curate multiple copies of the 500TiB dataset, and handle a peak of interest from users in response to the publicity surrounding the event. This means that our resource requirements for this 2024/2025 RSAP round includes an estimate of the resources that will need to be deployed by the end of 2024, ready to handle the peak load leading up to the DR4 release at the end of 2025 (resources requested in the 2025/2026 RSAP round would not be in deployed, tested and ready for use in time). Once our Gaia DR4 platform has been deployed and the expected peak load has passed we may be able to release some of the requested resources.

Multiple sites: We are planning for our main live deployment to be at Cambridge, along with a second system on the same platform for development and data curation. The secondary sites at RAL and Somerville will provide scale-out capacity to handle peak load expected during the DR4 release. If secondary sites at RAL and Somerville come online with no problems, then we may be able to release some the second system at Cambridge. Our development plan for 2024 includes migrating from the current monolithic deployment to a more flexible system consisting of an initial website that handles user accounts and login, backed by a number of notebook services. This will enable us to scale out to use multiple physical sites for the notebook services while presenting the user with what appears to be a single integrated platform. Multiple sites will provide high availability of service through fail-over and allow us to test portability, autoscaling, and other aspects.

Multiple data copies: Our development plan for 2024 includes working with mock copies of the DR4 dataset to test our data ingest, indexing and partitioning processes and new deployment methods that can scale out to cope with 100s of TB. During 2025 we may also ingest subsets of high-volume the DR4 data made available prior to the formal release date. At the same time we require to run at least one development service and keep the existing live DR3 site running. Furthermore, we note that the single largest monolithic datatype to be released in Gaia DR4, the “Epoch Images” (i.e. time-resolved image sample data) will be 200TB in (compact, binary) volume. Although the exact format for delivery of such large datasets is yet to be finalised we require to stage and reformat into Spark/Parquet at a single site. Assuming a worst-case non-compact-binary scenario (DR3-like eCSV format or similar) we would require 2.5x200TB just to stage the data for ingestion. Our experience from using Spark is that reformatting into Parquet also requires temporary storage that can be up to and in addition to the final data set size. Hence we may need up to 900TB accessible in a single deployment just to ingest this one dataset, but once the ingest is finished the temporary staging space is freed up and can be returned to the IRIS providers for allocation elsewhere.

Booking system: Our development plan for 2024 includes a booking system that enables users to book sessions ahead of time, smoothing out the peaks in demand and enabling us to predict the resources needed in advance. With this in place, we may be able to release resources at some of the physical sites in response to changes in load. It is hoped that the booking system can be integrated with Blazar, the Openstack resource reservation system, if it becomes available on the Openstack platforms.

Resource pinning: Our current allocation at Cambridge is pinned to specific hardware. This is to guarantee that we can release and create sets of resources without running into problems. In the past we found that if the Openstack system is configured to make maximal use all the available resources, leaving few idle resources available for new allocations, then releasing and creating a block of 10+ VMs may fail because some of the resources get assigned to other processes during the gap between our release and create steps. Whether this will be needed at RAL or Somerville will depend on the configuration of their Openstack systems. This issue may be solved by the deployment of Blazar, the Openstack resource reservation system.

High memory hypervisors: Two of the hypervisors at Cambridge are high memory nodes with a higher memory-to-vcpu ratio. These host the high memory VMs required by the Zeppelin head nodes to support the in-memory processing used by libraries such as HDBSCAN which are not amenable to implementation in Spark over a cluster (see e.g. Crake et al. 2023).

Project specific flavours: We have a set of project specific flavours at Cambridge, designed to optimise the packing of VMs into the pinned hypervisors. This includes a set of 'himem' flavours designed to make use of the high-memory hypervisors. We will also need to create a set of 'high storage' flavours to make use of the direct attached SSD storage when it becomes available.

Direct attached storage: This request is based on the assumption that direct attached SSD storage will be able to solve the problems with high IO wait that we have seen with our current deployment. We planned to use the SSD resources requested in our 2023 allocation to test this theory, but the resources are not available yet. We hope to be able to resolve this question during work planned for 2024.

CephFS vs Object storage: The current system uses the CephFS storage system at Cambridge, which is based on spinning disc hard drives. In theory it should be possible to serve the Gaia dataset via the S3 interface of Ceph object storage using the Openstack Swift interface. However, we have encountered some issues with the Java S3 client libraries when using this to serve the Gaia DR3 dataset. This is why this request specifies the majority of the storage, 500TiB, on CephFS via Manila and only 50 TiB of object via Swift at each site. If these issues can be solved, and if the resulting S3 interface is fast enough to provide the bandwidth needed, then we may be able to swap the bulk of the storage, 500TiB, from CephFS via Manila to object storage via Swift. We hope to be able to resolve this question during work planned for 2024.

# 6 Long term forecast

## 6.1 Long term Forecast: Gaia Core Processing

Processing for the fifth and final data release starts in Feb 2026 and continues through to June 2028. This will be followed by a long period of analytical work on the data products to inform the final set of publications to accompany data release 5. Scale is approximately double that of the current processing. Schedule is tentative at the moment. It is predicted that an increases in core count will be required to deal with higher data volume in a reasonable time. Certainly data volumes produced and analysed will be significantly larger. Detailed performance estimates from testing during 2024 will be able to refine the requirements.

| Year | GPU (note type) | CPU | Storage/Disk | Storage/Tape | Notes |
|---|---|---|---|---|---|
| 2025-2026 | N/A | 7000 | 5PB | 800 | DR5 processing starts in Feb 2026. |
| 2026-2027 | N/A | 8000 | 6PB | 800 | DR5 processing continues. |
| 2027-2028 | N/A | 9000 | 6PB | 800 | DR5 processing finishes. |

Table 6.1 Long term forecast

## 6.2 Long term forecast: Gaia Data Mining Platform

| Year | GPU (note type) | CPU | Storage/Disk | Storage/Tape | Notes |
|---|---|---|---|---|---|
| 2025-2026 | N/A | 2800 | 2200 TiB | 0 | Maintain DR3+4 levels |
| 2026-2027 | N/A | 2800 | 1700 TiB | 0 | Ditto minus returns |
| 2027-2028 | N/A | 2800 | 1700 TiB | 0 | Ditto |

Table 6.2 Long term forecast totals

Here we maintain allocation levels at the Gaia DR4 level minus the temporary staging space. CPU levels are particularly uncertain, but we anticipate roughly a factor of 2 increase to cope with increasing user demand as well as the increased data size.

# 7 References

De Angeli et al, “Gaia Data Release 3: Processing and validation of BP/RP low-resolution spectral data”, A&A, in press (2022), DOI: https://doi.org/10.1051/0004-6361/202243680

Crake, Hambly & Mann, “HEADSS: Hierarchical Data Splitting and Stitching software for non-distributed clustering algorithms”, Astronomy & Computing, 43, article id 100709 (2023)

Hodgkin et al, “Gaia Early Data Release 3: Gaia photometric science alerts”, A&A 652, A76 (2021), DOI https://doi.org/10.1051/0004-6361/202140735

Riello et al, “Gaia Data Release 2: Processing of the photometric data”, A&A 616, A3 (2018), DOI https://doi.org/10.1051/0004-6361/201832712

Riello et al, “Gaia Early Data Release 3: Photometric content and validation”, A&A 649, A3 (2021), DOI https://doi.org/10.1051/0004-6361/202039587