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 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.

Allocated 1200 cores, 480TB disk on the Cambridge Arcus Openstack cloud.

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
11	56(112)/192 GiB	0	1TB SSD	Cambridge	1200 (virtual)	1.7GiB	0	0	9GiB

Table 3.1.1A: Current allocation for Gaia Core Processing

Storage

Amount	Location	Disk/Tape
480TB	Cambridge	Disk

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 deploy 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, loading of the Gaia DR3 bulk data products is complete and an initial set of ~10 test users (including post-graduate students and early-stage post-doctoral researchers) have been on-boarded to exercise the system and exploit the science data.

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
6	55/188 GiB	0	~ 512 GiB (available)	Camb.	550 (virtual)	1.7 GiB	0	0	9 GiB
2	55/512 GiB	0	~ 512 GiB (available)	Camb.	220 (virtual)	10 GiB	0	0	9 GiB

Table 3.1.2A: Current allocation for the Gaia DMP. The two rows distinguish between 'standard' and 'high' memory nodes.

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

3.2 Current usage of IRIS resources

3.2.1 Gaia Core Processing

CPU/GPU Usage for 01/10/2021 to 31/09/2022 (or closest reporting period)

The 1200 core allocation was assigned at the end of August 2022. Due to current high workload on the Gaia team we have not yet deployed the processing infrastructure. This is planned by the end of this year. Various test systems have been deployed during the year taking up to 100% available cores of the previous allocation of 300.

Storage usage (October 2022 or closest reporting period)

Amount	Location	Type	Usage (Oct 2022)	85% expected
480TB	Cambridge	Disk	120TB (25%)	End of January 2023

Table 3.2.1: Storage use Gaia Core Processing

300TB of this storage was made available at the end of August 2022 and had not been fully utilised in October 2022. (Update on 5th Jan 2023: 80% storage resources in use).

3.2.2 Gaia Data Mining Platform

CPU/GPU Usage for 01/10/2021 to 31/09/2022 (or closest reporting period)

All standard memory CPU is in use, either permanently allocated to our end-user live system, or split between development and testing platforms. We have only recently (Q3 2022) been allocated the high memory systems and have not yet fully utilised those resources.

Storage usage (October 2022 or closest reporting period)

Amount	Location	Type	Usage (Oct 2022)	85% expected
72 TiB	Cambridge	CephFS share	100%	
72 TiB	Cambridge	Ceph volume	50%	
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 filesets.

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 overall workflow starts from intermediate parameter determination outputs for the G-band photometric data and raw BP and RP low-resolution spectral data. The final output of the processing is a catalogue of mean and epoch photometry in the G, integrated BP and RP bands and mean and epoch spectra in the BP and RP wavelength ranges. These quantities will be accompanied by a number of additional parameters reporting on the statistics and quality of the output data.

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.

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.

Estimates of required storage volume in TB are:

Data	Volume (TB)
Input data archive	600
Cycle 3 retained intermediate data products	400
Cycle 3 final data products	120
Cycle 4 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	3520

Total disk data storage required is estimated as 3520TB rounded up to 3600TB usable storage.

4.1.2 Development and Operations

Various systems that are not part of the main processing cluster are required to support development and operations. The intention is to host all these via OpenStack provisioned virtual infrastructure unless issues are identified that prevent this.

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.

Estimates 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 retained data-set grows at roughly 20 TB per year (with eight years of data already stored) and up to 24 TB of raw data may be held pending processing. In addition to daily processing, occasional large-scale operations are needed on the retained data-set and these need free space of ~30% of its permanent storage.

The main database, holding the retained data-set, is a PostgreSQL installation. Currently, science alerts runs a replica pair of 32-core servers with data ingestion on one and read-only queries on the other. This sub-system is currently saturated in both CPU and I/O. In the transition to IRIS, we propose to combine the two DB servers into one and to give up the replication, saving storage and processing capacity. This is feasible if the back-up arrangements are robust.

Back-ups of the main database are taken weekly, and two such are available at any time.

Reduction of incoming data and generation of alerts requires a separate server that runs a multi-threaded application in Java. This work cannot be distributed across computers due to the structure of the application.

A second compute-server is used for development of the system and for tests on unexpected features in the data.

Science alerts also operate a number of low-traffic web-applications that can be served from nodes of the core processing

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

Service	CPU cores	RAM/core	Block storage (TB)
Database	128	16	312
Compute	128	6	24
Development	128	6	24
Back-up	N/A	N/A	624

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. Integration of the Lustre filesystem with OpenStack is required for necessary IO performance.

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. Therefore network between cluster compute nodes and storage and also connecting cluster compute nodes should be higher bandwidth.

Current cluster uses Mellanox FDR Infiniband interconnect.

Specifically during 2022 and 2023 we will need to move large amounts of data between our current installation and the new IRIS disk storage. It would be very useful if we could configure relatively high bandwidth between these installations. Assuming the hardware is hosted in the West Cambridge Data Centre this may be possible since our current system is also there. Hosting the IRIS resources elsewhere would be problematic for this reason.

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.

10 IP addresses, assignable within Openstack would be sufficient. Our current Openstack project provides for more than this in any case.

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. The current system will achieve an average of approx. 800Mbps for sustained periods. Sometimes these transfers will be 10s of TB and take several days to transfer.

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.

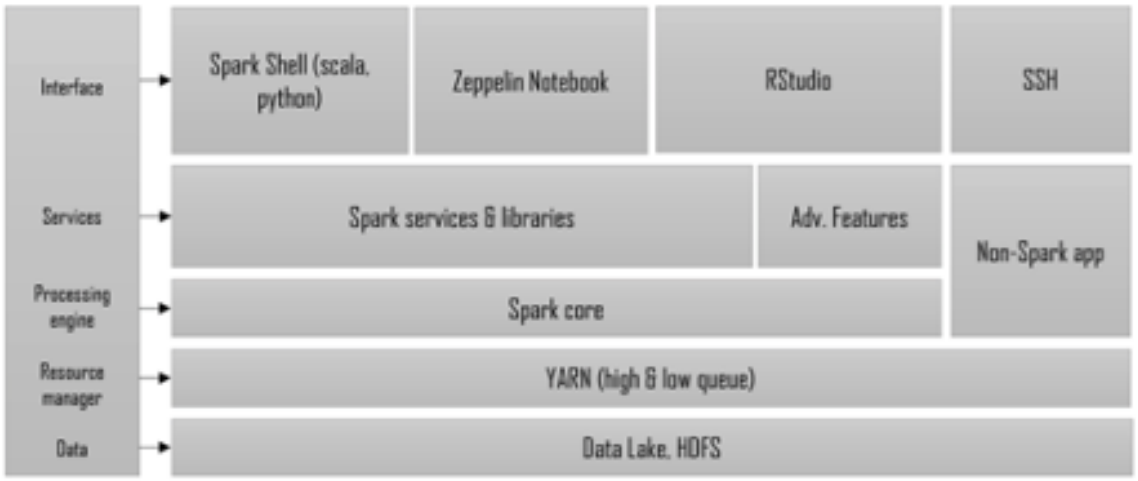
4.2.4.4 Software

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

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



The system is known as the Gaia Data Analytics Framework and has been 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 2023.
- Authentication/authorization via delegation to IRIS-IAM or similar.

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 would be clearly advantageous. 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 will be useful. Overheads on user staging and scratch space are likely to be 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).

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:

We are consulting with Cambridge and StackHPC to advise on an appropriate allocation to provide an experimental deployment to evaluate our future requirements for 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 (~300TB 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 project has RSE staff resources from a special STFC Gaia “CU9” (exploitation-enabling) grant.

5 Resource request for September 2023 –August 2024

Requesting additional resources to enable complete replacement of current Gaia hardware systems and transfer of all operations/processing to the IRIS infrastructure.

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

5.1 Additional Request 23/24: Gaia Core Processing

An additional 5224 cloud cores are requested in addition to the 1200 currently allocated to provide a total of 6424 cores. Total of 6424 is derived from:

- 5808 cluster processing cores to match the scale of the existing processing cluster (5280 cores) with a 10% increase in core count to mitigate continuously increasing data volume in processing jobs.
- 144 cores for cluster infrastructure and services and user login nodes for cluster operations. Based on the resources in the existing processing cluster.
- 256 cores for Science alerts cloud resources (Compute and Development from table in 4.1.3).
- 216 cores for Development and Operations resources (Total from table in 4.1.2).

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	5224	10GB	N/A	N/A	N/A
1	128/2048 GB	N/A	312 TB	N/A	N/A	N/A	N/A	N/A

Table 5.1A Additional resources requested for Gaia core processing to be added to current cloud resources. The first line is for the 'normal' Gaia core processing (including the bulk of resources for the science alerts processing) and the second line is a specialized server (with specific hardware requirements) for the science alerts database.

Storage

Amount	Preferred Location	Type (Disk/Tape)
4272TB	Cambridge	Disk
600TB	Cambridge	Tape

Table 5.1B Total amount of storage requested (includes current allocation)

5.2 Additional Request 23/24: Gaia Data Mining Platform

CPU/GPU

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
6	55/188 GiB	N/A	4TB/node	550	3.5GiB	N/A	N/A	72GB
2	55/512 GiB	N/A	4TB/node	220	10.0 GiiB	N/A	N/A	72GB

Table 5.2A Maintain resource level previously requested / already allocated but with modest additional allocation of direct attached fast storage

We request that 25% of compute nodes are high-memory (e.g. 512G) nodes in order that correspondingly high-capacity VMs can be provisioned for some memory-intensive, non-distributed workflows. CPU numbers in the machine view are physical, not virtual. We expect a x2 hyperthreading factor over these numbers when allocating virtual cores to vms in OpenStack. Memory is given in GiBytes per physical core and is split into two separate lines of standard and high memory. The only addition here over our current allocation is the availability of fast local storage.

Storage

Amount	Preferred Location	Type (Disk/Tape)
72 TiB	Cambridge	CephFS share
72 TiB	Cambridge	Ceph Volume
20 TiB	RAL	Echo S3
32 TB	Cambridge	Direct attached SSD

Table 5.2B Total amount of storage requested: maintain existing levels previously requested and already allocated but modest additional directly attached storage (last row in the table).

6 Long term forecast

6.1 Long term Forecast: Gaia Core Processing

Year	GPU (note type)	CPU (cores)	Storage/Disk (TB)	Storage/Tape (TB)	Notes
2024-2025	N/A	6000	6000	700	
2025-2026	N/A	9000	8000	800	
2026-2027	N/A	9000	8000	800	

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
2024-2025	N/A	660	196 TiB	0	Maintain DR3 levels
2025-2026	N/A	1200	496 TiB	0	+300T Gaia DR4
2026-2027	N/A	1200	496 TiB	0	Maintain DR4 levels

Table 6.2 Long term forecast totals

Here we maintain allocation levels apart from 25/26 where we anticipate the arrival of a new Gaia data release with 300TB data volume. 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

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