IRIS resources used at Cambridge for Gaia and PLATO data processing

NAM Durham 2025

Patrick Burgess

Gaia Photometric Processing

Gaia Mission

- European Space Agency mission to form the most precise 3D map of the galaxy.
- December 2013 : Spacecraft launched. Scientific observations started in July 2014.
- June 2022: Latest published Data Release DR3.
  - 34 months of mission data.
  - 1,811,709,771 sources.
- Next release is DR4: 66 months of mission data.
- Cambridge processing for DR4 recently completed:
  - Generated 236TB, 433 billion records

- The entire mission is processed from scratch for each Data Release.
- Each Data Release introduces more complex processing to improve results in addition to more observations.
- Computational requirements continuously increase.

https://www.cosmos.esa.int/web/gaia

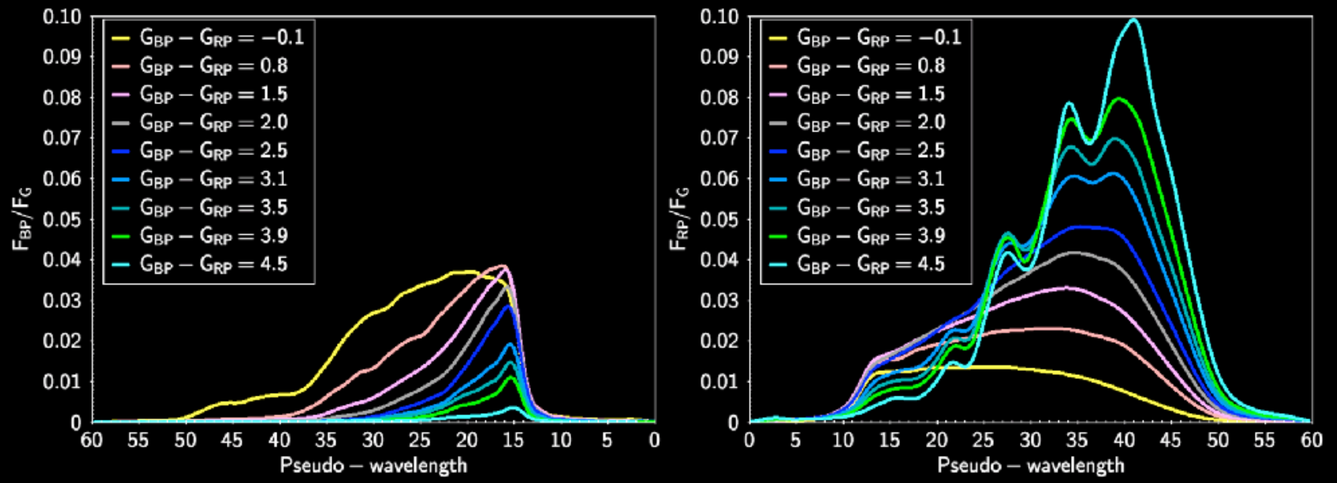
https://www.cosmos.esa.int/web/gaia/data-release-3

Gaia Photometric Processing

Cambridge

- Photometric Processing carried out at Cambridge.
- One of six Data Processing Centres (with Barcelona, Geneva, Madrid, Toulouse, Turin).
- Java software based on Apache Spark, “Apache Spark is a unified analytics engine for large-scale data processing.”
- During processing for DR4 (May 2024), migrated from dedicated hardware to virtual IRIS resources.

https://www.cosmos.esa.int/web/gaia/iow_20200812



- Cambridge DR4 data products approximately 236TB:
  - Calibrated mean and epoch integrated photometry.
  - Calibrated mean and epoch BP/RP spectra.
  - Passbands, absolute calibrations.
  - Source environment.

Gaia Photometric Processing

Data Release 5 (DR5)

- Development and testing for DR5 processing in progress. Photometric processing at Cambridge starts mid 2026.
  - Processing for DR5 already underway at DPCB.
- DR5, “not before the end of 2030”, will cover full mission duration.
- Three trillion observations taken over ~126 months.

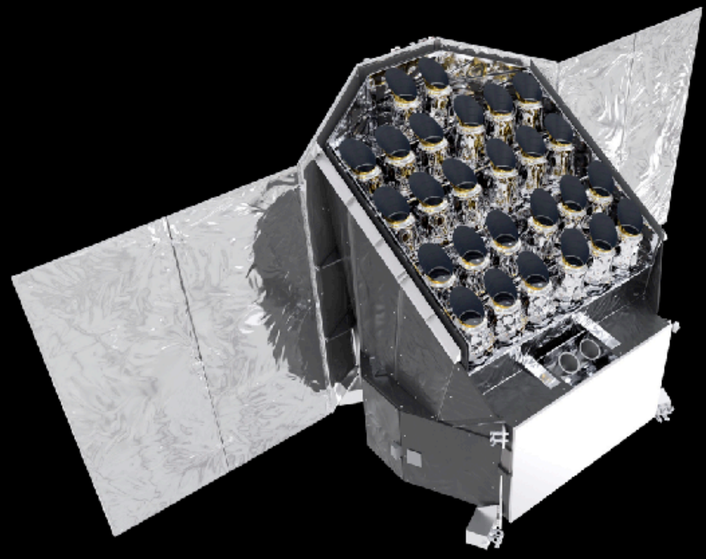
- Gaia Spacecraft “passivated” in March this year. Final science observations taken in January this year.

PLATO

Overview

- Space based observatory primarily tasked with detection of terrestrial exoplanets around bright, solar type stars.
- PLATO Spacecraft has 26 cameras covering 2132 square degrees in an overlapping field of view.

- Stares at selected target stars for long duration - initially 2 years - rotating 90 degrees every 3 months to avoid the sun.
- Observations are imagettes (6x6 pixels) or light curves with cadences of 25s, 50s or 600s depending on the target.
- Downloading 435Gb per day at 72Mbps.
- Publishing results every 3 months during science operations.



https://www.esa.int/Science_Exploration/Space_Science/Plato

PLATO

Mission Timeline

- To be launched at the end of 2026 and start science operations in early 2027 after commissioning.
- With a nominal science mission duration of 4 years (finish early 2031) extendable to 8.5 years (finish mid 2035).
- Cambridge responsible for Exoplanet Analysis System (EAS): TransitPipe, PlanetPipe,…
    - Detection and characterisation of planet signals within calibrated input data.
- Alongside software and process development, currently carrying out a series of test campaigns leading up to launch. Execution performance, science performance, integration of data centres.

IRIS Resources

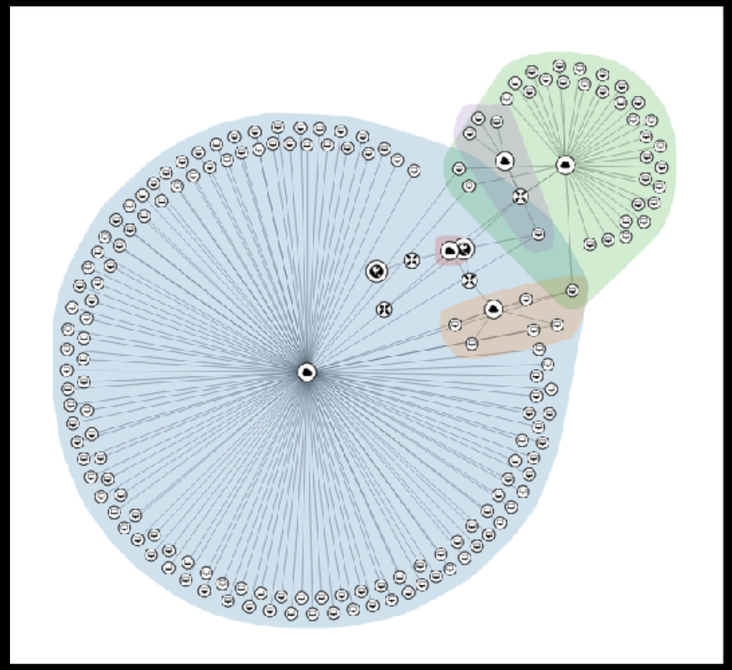
Current Usage

Project	Location	vCPU	3R Ceph	EC Ceph	S3	A100 GPU
Gaia	Cambridge	11068	1200	1741	-	-
Gaia	RAL	-	-	-	1080	-
Plato	Cambridge	3552	400	-	-	4
Plato	Imperial	288	5.4	-	-	1
Plato	RAL	200	5.6	-	-	1

- Gaia processing all Cambridge based.
- Plato has small allocations at two other sites to test operating across multiple sites.

IRIS Resources

Openstack



- Cloud resources at all sites are provided via Openstack.
- Provides a command line interface, web UI,
- Deploy, configure, operate and destroy virtual infrastructure including virtual machines, storage, networks, routers, firewalls.
- Commonality across sites makes it (fairly) straightforward to migrate infrastructure.
  - But not necessarily data.
- Some differences to sort out - e.g. software versions, different security policies - but no serious problems.

IRIS Resources

Ansible

- Infrastructure as code tools.
- Enables definition and deployment of:
  - Openstack resources - virtual machines, networks, storage.
  - Software installation and update.
  - OS Configuration - firewalls, users, networking.
- Provides repeatable architecture - rebuild the system exactly very quickly.
- Allows for fairly easy deployment of the same system at different sites.

```
- name: deploy rsyslogd configuration
  tags: logger_rsyslog
  become: yes
  template:
    src: rsyslog.conf.j2
    dest: /etc/rsyslog.conf
    backup: yes
    owner: root
    group: root
    mode: 0644
  notify: Restart rsyslog
```

IRIS Resources

Ceph Storage

- Gaia requires fairly significant disk storage (4PB).
- Most storage is configured as CephFS network shares that can attach to multiple VMs.
- Ceph storage provides resilience through:
  - Triple replication - costs 3x usable storage.
  - Erasure coding - costs 1.37x usable storage.
- Tiered storage provides significant cost savings:
  - 8+3 Erasure Coded Ceph for more than half total Ceph storage.
  - S3 for cold storage.

IRIS Resources
GPU

- PLATO using CETRA:
  - Cambridge Exoplanet Transit Recovery Algorithm
- A fast, sensitive exoplanet transit detection algorithm implemented for NVIDIA GPUs (https://arxiv.org/abs/2503.20875)
- Detect possible transits and characterise the periodicity of the signal.
- Implemented using NVIDIA CUDA framework in C by Leigh Smith.
- Used for EAS TransitPipe transit detection and characterisation.
- Open source : https://github.com/leigh2/cetra

IRIS Resources

Summary

- Both PLATO and Gaia data processing continue to rely on hardware resources provided through the IRIS collaboration and gain from the knowledge shared within the IRIS collaboration.
- Virtual provision of cloud resources at different sites allows:
  - better usage of scarce resources e.g. GPUs by moving the processing to wherever the physical resources are available.
  - some protection against site outage - recently moving test execution to Imperial to avoid maintenance downtime at Cambridge.