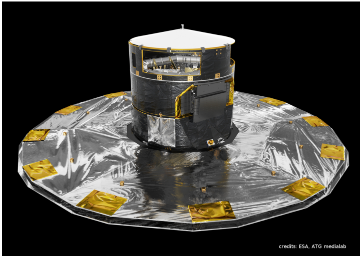
Data processing challenges in Gaia

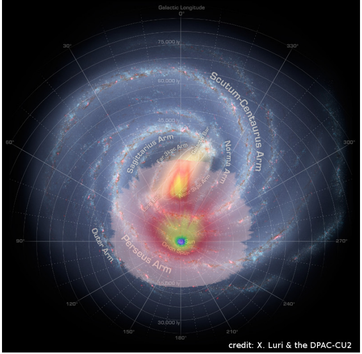
Pasquale Panuzzo

GEPI, Observatoire de Paris, PSL Research University, CNRS,
Univ. Paris Diderot, Sorbonne Paris Cité

Data Processing and Analysis Consortium - CU6

The Gaia mission





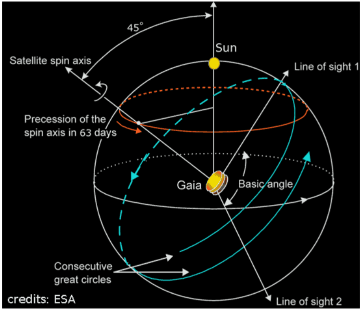
- 6th ESA Corner Stone mission
- Launched December 19 2013, orbiting around L2
- Aim: Produce the most accurate 3D map (+3D velocities) of the Milky Way
- μarcsec Astrometry $G < 20$ (10^9 sources, 1% of the Galaxy)
- Radial Velocities $G < 16$ (10^8 sources)
- Photometry millimag $G < 20$
- 5 years of nominal mission + extension
  - On average 70 visits for each source
- Huge impact on stellar physics, galaxy formation and many other fields
- Catalogues will be public without guaranteed time

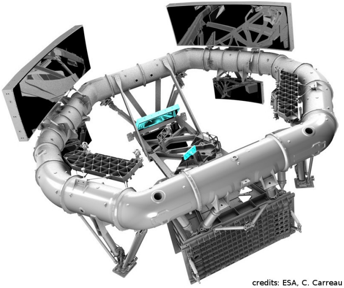
Gaia DPAC

How it works

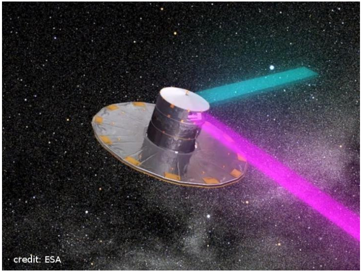
gaia

- The satellite rotates at a constant speed of 1 rotation in 6 hours
  - Spin axis at 45° of the Sun direction
  - Spin axis precesses in 63 days
- Two rectangular telescopes looking in 2 directions at 106.5°
  - Objects transiting in the first field of view (FoV) are seen in the second after 106.5 minutes.



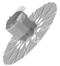


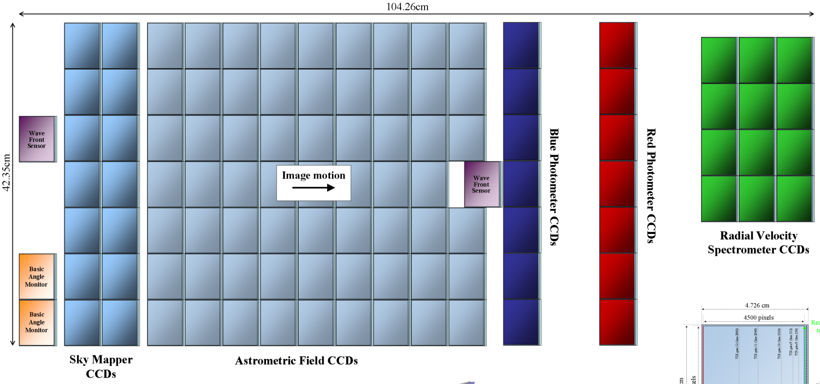
credits: ESA, C. Carreau

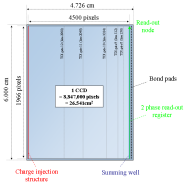


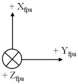
Gaia Focal Plane

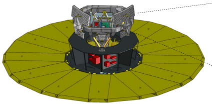
106 CCDs ≈ 938 million pixels ≈ 2800 cm^2



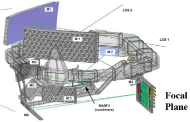








Gaia (MLI removed)

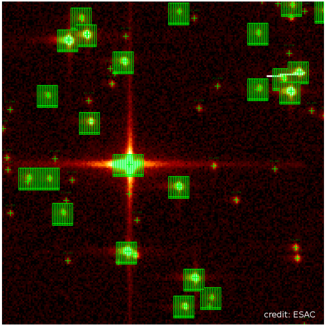


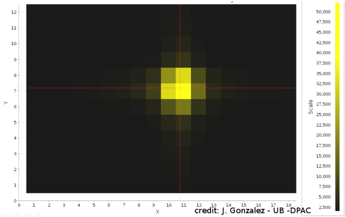
Payload Module (telescopes)

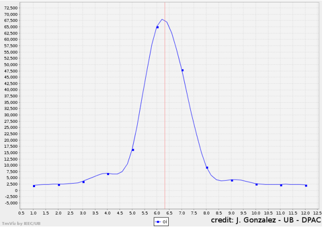
A.D. Short – July, 2009

On-board image processing

- Both FoV are projected on same CCDs
- Charges on the CCDs are moved in synchrony with the image motion
- Not all the CCD images are read and sent to ground. The onboard software:
  - identifies sources coming from each FoV
  - reads only the CCD areas containing the star images/spectra
  - rebins data into 1D (2D for bright stars)

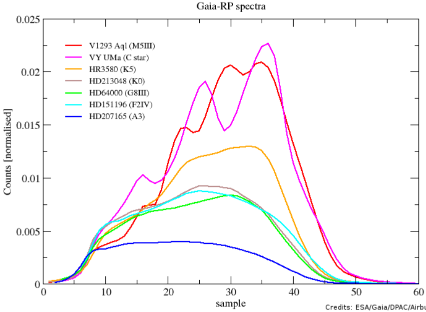




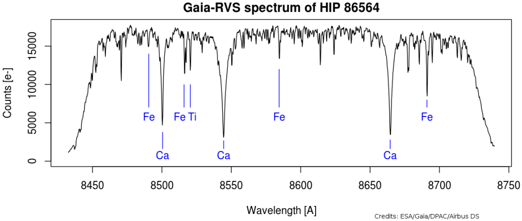


Photometry & Spectroscopy

- Gaia has also a photometer (BP/RP)
  - which is in reality a low resolution spectrometer with 2 bands
  - high precision SED measurements 320-1000 nm
- And a spectrometer (RVS)
  - R $\sim$ 11500, 847-874 nm
  - For radial velocity measurements (1 km/s)



Credits: ESA/Gaia/DPAC/Airbus DS



Credits: ESA/Gaia/DPAC/Airbus DS

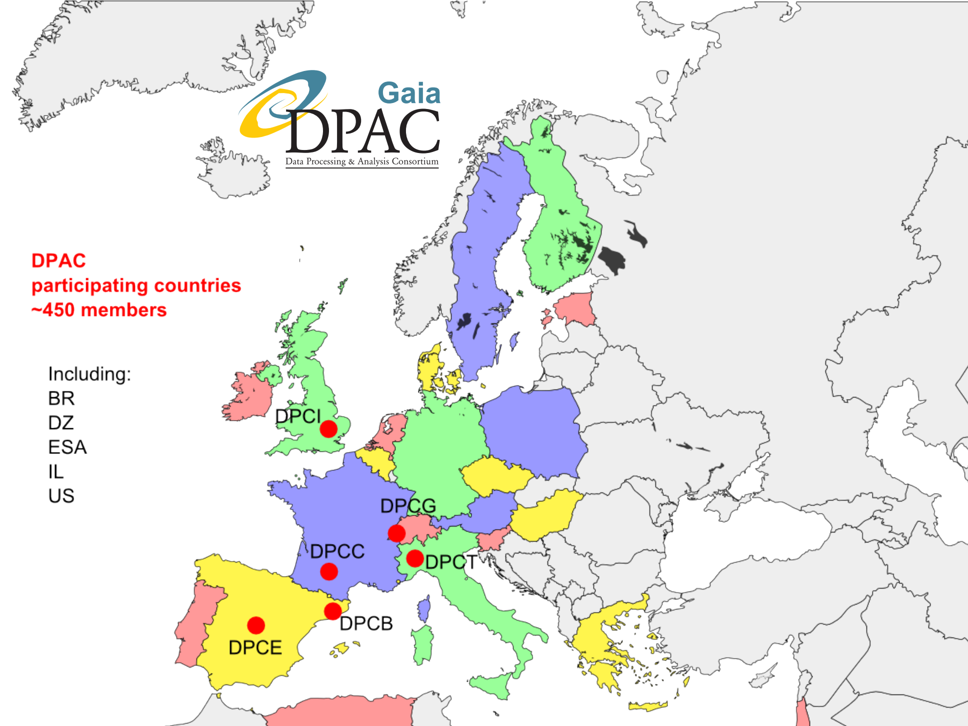
Data Processing and Analysis Consortium

- Formed to answer the Announcement of Opportunity for Gaia data processing (2006)
- Involves large number of European institutes and observatories ($>$ 450 people, $>$ 20 institutes)
- **The DPAC's role is to process the Gaia data and produce the catalogs for the community**
  - The DPAC doesn't have any proprietary time on the data!
  - The DPAC members can do science only with published data.
- It is composed by 6 Data Processing Centers (DPCs) and 9 Coordination Units (CUs)
  - The DPCs are charged to process the data
  - The CUs are charged to develop the algorithms and the software that will be used in the DPCs
- Each specific processing pipeline is implemented in a single Data Processing Center
  - The infrastructure is different in each DPC

Gaia DPAC
Data Processing & Analysis Consortium

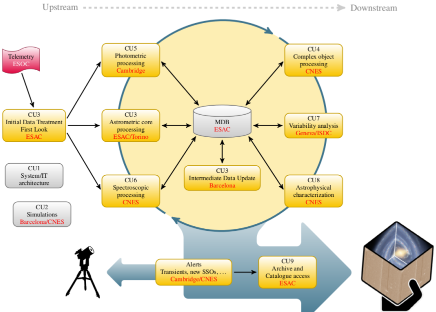
DPAC participating countries ~450 members

Including:
BR
DZ
ESA
IL
US



Data processing responsibilities

- ESAC, Madrid (DPCE):
  - Initial Data Treatment
  - First Look
  - Astrometric Global Iterative Solution
  - Archive and catalogue access
- UB, Barcelona (DPCB):
  - Intermediate Data Update
  - Simulations
- CNES, Toulouse (DPCC):
  - Spectrometry
  - Object Processing
  - Astrophysical Parameters
- IoA, Cambridge (DPCI):
  - Photometry
  - Science alerts



credit: A. Brown - DPAC

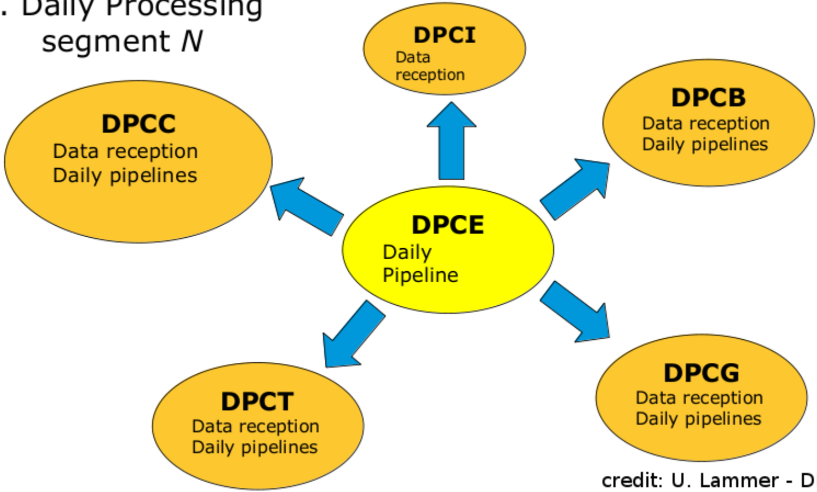
- ISDC, Geneva (DPCG):
  - Variability
- OATO, Torino (DPCT):
  - Astrometric Verification

Cyclic processing

- The DPAC works in Data Release Cycles
  - We do a release every 6 months - 1 year (first release summer 2016)
  - The final catalogue will be in $\approx$ 2022
- At each cycle we reprocess all the data since the start of the mission (Global processing)
  - The astrometric solution improves adding new observations
- At each cycle we add new algorithms and we treat fainter/more complex sources
- In parallel to Global processing, we have Daily pipelines used to:
  - produce raw data (will not be reprocessed in following cycles)
  - monitor the health of the payload
  - produce boot-strap calibrations for the Global processing
  - produce science alert (Supernovae, detection of asteroids, etc)

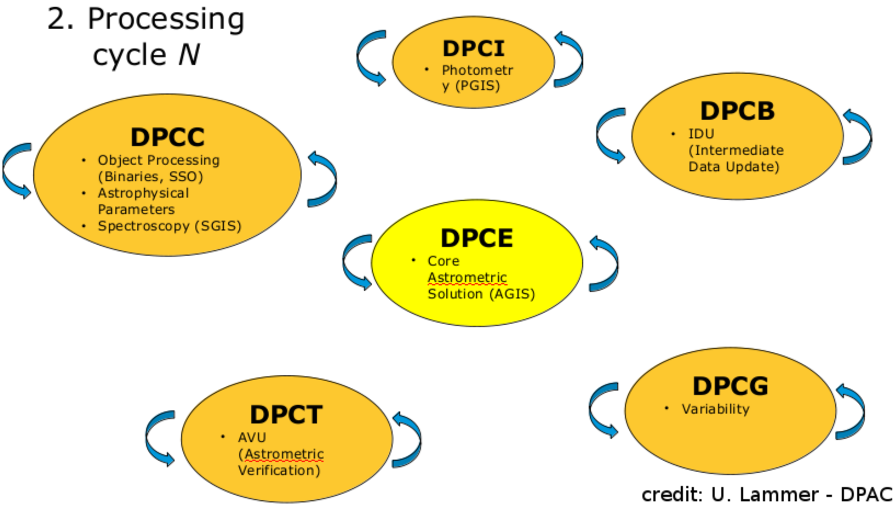
Cyclic processing

1. Daily Processing segment *N*



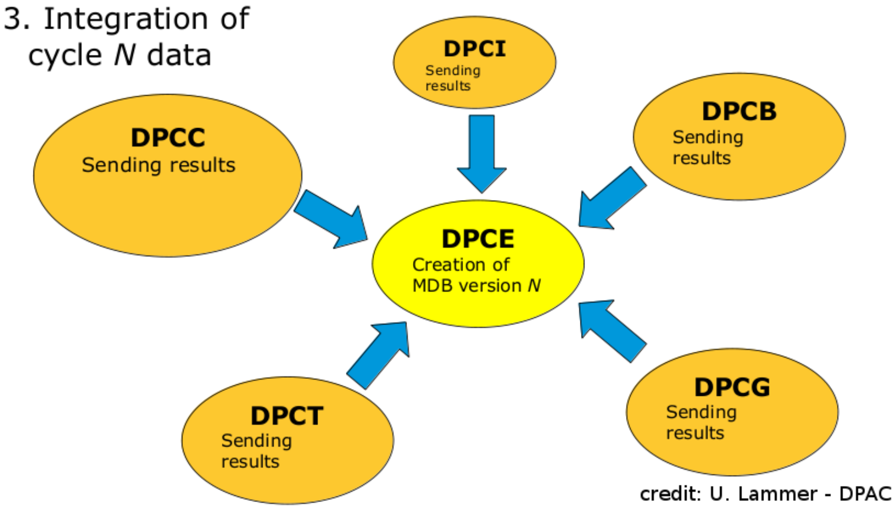
credit: U. Lammer - DPAC

Cyclic processing



- Daily processing of segment N+1 running at the same time

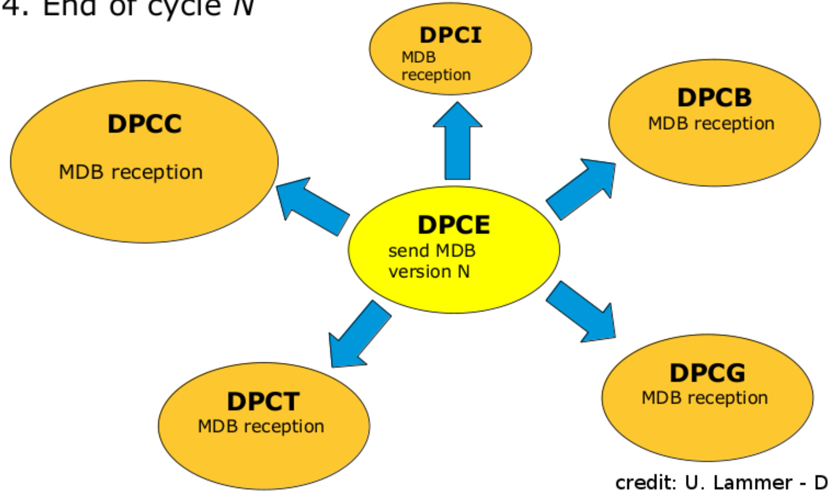
Cyclic processing



- Consolidate the results in a single database (the Mission DB)

Cyclic processing

4. End of cycle *N*



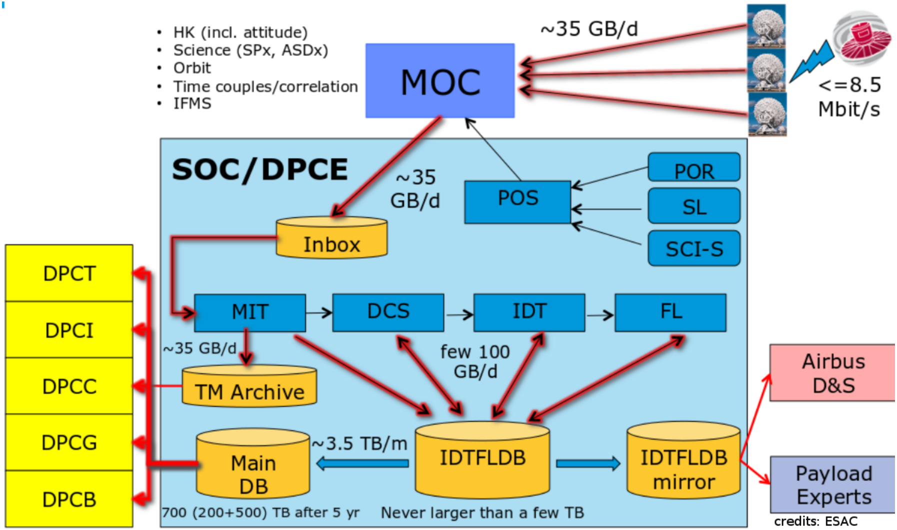
credit: U. Lammer - DPAC

- The MDB of cycle N is sent back to be used as basis for cycle N+1

Lessons learned #1

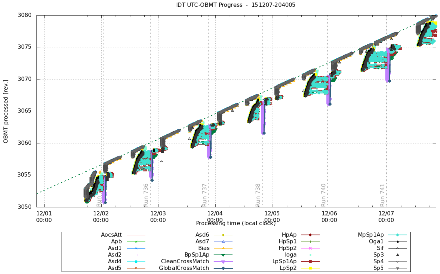
- Data transfer between DPCs is a bottleneck
- 60 days to transfer the data at next release at a rate of 1Gbps
- The MDB is expected to be $\sim$400 TB at the end of the mission (MDB $\sim$80 TB for the first release)
- ESAC is connected to the internet via a 10 Gbps (to be shared with other missions)
- Transfer is done with Aspera (a proprietary data transfer software)
- Data are serialized in a Gaia-specific binary file format ("gbin"), not FITS.

DPCE Daily pipeline



DPCE Daily pipeline

- DPCE Daily Pipeline has been processing continuously the incoming TM from MOC, since launch
  - The DPCE Daily Pipeline is data-driven; processed data as they arrive
  - TM doesn't arrive in order, some data have higher priority than other



- 1000 observations per second, 35GB of TM received every day
  - reaching 85GB/day when scanning the Galactic plane
- 300GB of data produced per day by the daily pipelines

Astrometric Global Iterative Solution

- The Gaia astrometric catalogue will be computed with the Astrometric Global Iterative Solution (AGIS)
- The goal is to compute the 5 astrometric parameters α_0, δ_0, π, μ_α, μ_δ for 10^9 point sources from the 10^{12} observations
- 3 models enter in the formulation:
  - The Source model (5 parameters)
  - The Attitude model ($4 \cdot 10^7$ parameters)
  - The focal plane Calibration model (10^6 parameters)
- The problem can be expressed as a weighted least-squares problem, or a linear system
- More complex than anticipated due to basic angle variations, decontaminations, meteorites hits etc.
- For the first release of the astrometric catalogue, will also use Tycho-2 catalogue (Tycho-Gaia Astrometric Solution)

Java

- **Yes, all the Gaia software is in Java!** Why not C/C++ or FORTRAN or IDL?
- **Speed of code is an issue, but manpower is a bigger issue!**
- Easier to find skilled Java developers
  - Already many Java developers are present at ESAC
- Coding in Java is cleaner and faster, and the code is safer
  - Many algorithms are developed by scientists, not engineers
  - No memory management
  - Standard exception handling
  - Fully object-oriented allows a modularity
- Many professional software development tools are available
  - UML, Hudson/Jenkins, PMD, ant, Checkstyle, JUnit, etc.
- Automatic API documentation with Javadoc

Java

- Java is a platform independent language
  - Fundamental in a project developed by 400 people across Europe
  - Easy migration to new platforms in the future
- Easy to develop distributed infrastructures with Java EE
  - Network connections well supported by language
  - Java EE applications are very common in IT industry
  - You cannot develop an data processing infrastructure in FORTRAN...
- Negative aspects:
  - The lack of numerical libraries as good as in other languages
    - We adopted the Apache Common Math library
  - Cannot reuse legacy code
    - not much legacy code were available for Gaia
    - JNI allows Java to run code in C/C++
  - Have to migrate to new Java versions
    - Some external libraries are migrated very slowly

DPCE infrastructure

- The DPCE Daily Pipeline and the AGIS pipeline are implemented as "standard" Java EE systems based on JMS & JBoss
  - chunks of data are assigned to jobs, jobs published on a whiteboard
  - workers, installed on nodes of a cluster, grab the jobs and save the results in the database
  - a coordinator orchestrates the execution of various processing stages
  - the processing progress can be monitored from web pages
- The DPCE platform is composed by ∼100 IBM Xeon-based blades, ∼1000 cores, 32-128GB RAM per blade
  - Shared between operations, validation and reprocessing
- Database: InterSystem Caché® (SQL database)
- The DPCE Daily Pipeline and the AGIS pipeline are developed by the CU3
  - CU3 team is located in Heidelberg (ARI), Barcelona (UB) and ESAC

Lessons learned #2

- First year at ESAC was **very challenging**
  - Daily pipeline not scaling well during galactic plane scans
  - Too many "bugs"
    - Telemetry not as expected from documents
    - Many bugs not seen during pre-launch tests
  - Data not as expected
    - Had to change many algorithms
  - Too many patches installed (one every week!), with associated downtime
  - Much more manual operator control needed than anticipated
    - Many undefined contingency procedures
  - Many operations executed on-board for the commissioning
  - On-board software was changed, adding new type of telemetry
  - Data requests from payload experts
  - Unplanned reprocessing of raw data
- Now, **everything works much more smoothly**

Lessons learned #3

- The DPCE infrastructure/software proven to work
- A major bottleneck is Database cleanups
  - 300GB of data produced per day by the daily pipelines
  - need to remove old data from the database to keep a reasonable size
  - Relational databases perform very poorly when deleting data
    - Databases are optimized for insertions, updates and searches
  - Several downtimes induced by database cleanups
- Another major bottleneck is Database mirror
  - we need a database mirroring the operational one to give data access to Payload Experts
  - mirroring mechanism doesn't keep up with the rate of data produced by the pipeline
  - an alternative mechanism is under study
- The AGIS pipeline was also executed in the Cloud (Amazon EC2)
  - It worked well, but only tested on a small number of nodes (50)
  - Amazon EC2 is now used to do integration tests

DPCC (CNES)

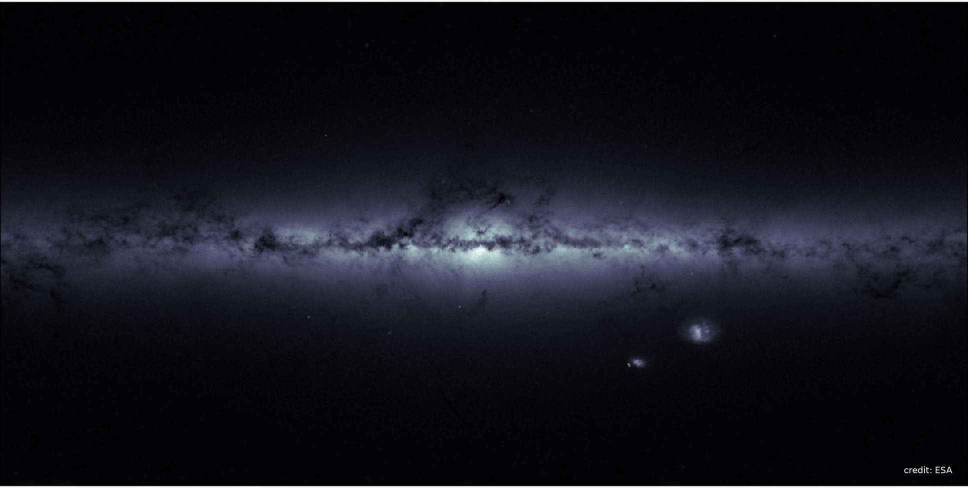
- DPCC serves 3 CUs
  - CU6 (Spectrometry), CU4 (Object Processing) and CU8 (Astrophysical Parameters)
- CNES uses a *Workflow Orchestration* software, SAGA, to run data processing pipelines on an Hadoop cluster (now with $\sim$ 1000 cores)
  - SAGA uses the Cascading framework to implement the pipelines
- SAGA is developed by Thales, which acts as sub-contractor of CNES
- Pipelines modules are developed by scientists in CUs
- Pros:
  - Should scale easily adding new hardware
- Cons:
  - Pipelines are integrated by a team that doesn't have scientific knowledge on how it should work
  - Scientists developing modules don't know how their modules are integrated
  - Cannot test the integration outside CNES
- Operations at DPCC starting only now

Final remarks

- We are now at almost 2 years after launch
- Despite some issues, the satellite is working well
- DPAC was able to handle the commissioning (much more complicated than expected), but it is now processing data in an operational way
- First data release planned for Summer 2016
  - A good fraction of the processing already done!
- Lot of work still ahead for the following releases
- The choice of Java and associated technologies was correct
- More important than good technology are:
  - **Good communication**
  - **Good Product Assurance procedures**

Questions?

Full sky from Gaia housekeeping



Thanks for your attention!