CNES Gaia Data Processing Centre, a complex operation plan

Veronique Valette and Kader Amsif
CNES, Centre National d'Etudes Spatiales, Toulouse, France

CNES is in charge of a large part of the Gaia data processing. The Gaia astronomy mission aims at mapping one billion stars and objects. The Gaia project, the Data Processing and Analysis Consortium and the role of CNES are presented. The satellite is planned to be launched in 2013.This paper focus on the complexity of the operation plan in terms of data volume, of processing systems and of organisation as scientific algorithm development will continue during operations. Gaia scientific data processing will be daily and cyclic, cycles planned to last from 4 months to one year. Two daily and 5 cyclic large scientific chains will run in the CNES Data Processing Centre (DPCC). Due to the iterative data reduction, the number of systems in operation and the data volumes will increase at each cycle. The operation plan required to obtain the final stars catalogue will be presented. The data volumes will go from some TeraBytes at the beginning of the mission to PetaBytes at the end. To deal with this we have chosen an Hadoop solution that is highly scalable in terms of data volume and processing. We will present the operation framework that will enable to run the operations. Some scientific systems will come into operations more than two years after launch so they will not be ready at launch time and development will go on. We will present the organisation planned to deal with this.

I. Introduction

To explain the CNES Gaia Data Processing Center complex operation plan, we first quickly present the Gaia mission and the Data Processing and Analysis Consortium DPAC organization. Inside this consortium, CNES plays an important role during development and operation phases.
An overview of the Gaia data processing operation phase is then presented and finally we focus on operations from CNES point of view.

II. The Gaia Mission

A. The Gaia project

Gaia is an ambitious space astronomy mission of ESA with a main objective to map the sky in astrometry (the measurement of stellar position, parallax, and proper motion), photometry (the measurement of photometric magnitudes) and spectroscopy (for the acquisition of radial velocities and astrophysical parameters). It will measure the positions, distances and physical characteristics of more than one billion stars in our galaxy and beyond.
The management model adopted for Gaia assumes that the entire satellite, including payload and operations is under ESA responsibility.
The Gaia satellite is developed by Astrium and the control and mission centres are under ESOC responsibility.

Satellite launch is planned in September 2013 by a Soyouz-Fregat launcher in French Guyana Space Centre. The 2 tons satellite will be positioned at Lagrange L2 point, 1.5 millions kilometres from Earth. The mission is foreseen to last 5 years.
Raw data will be received in ESA Cebreros station near Madrid and will be first processed in ESAC (European Space Astronomy Centre near Madrid) before being sent to the other processing centres for further processing.

B. Data processing and Analysis consortium

The scientific activities of the Gaia mission are conducted by members of the astronomy community, nationally funded.
Following an ESA Announcement of Opportunity, the Data Processing and Analysis Consortium (DPAC) has been created in 2006 and represents now about 400 people, engineers and scientists, from 21 countries all over Europe. The annual workload of the Gaia DPAC is about 250 FTE (Full Time Equivalent).

The Gaia DPAC has been divided in 9 Coordination Units (CU) and 6 Data processing Centres with an executive committee the DPACE as presented in figure 1. Each of these CU are based upon Data Processing Centers (DPC) which ensure the technical coordination of the scientific softwares developments and their operational deployment.

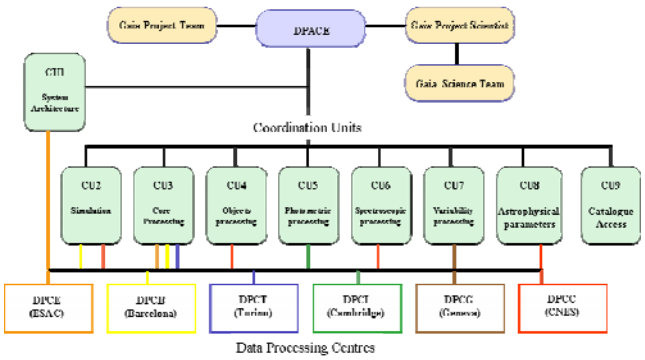

Figure 1. The Gaia DPAC organization, CUs and DPCs

The data processing centres are located all over Europe, in charge of processing one or more CU scientific algorithms:

-DPCB at BSC/CESCA, Barcelona, Spain, CU3
-DPCC at CNES, Toulouse, France, CU4, CU6, CU8
-DPCE at ESAC, Villanueva, Spain, CU1, CU3
-DPCG at ISDC, Geneva, Switzerland, CU7
-DPCI at IoA, Cambridge, England, CU5
-DPCT at Altec, Turin, Italy, CU3

The scientific codes are developed in java language by the CU members, then delivered to the corresponding DPCs for integration, system testing and operation.

C. CNES participation

CNES is responsible for the technical coordination, quality assurance and integration of the scientific developments of spectroscopic processing (CU6), Object processing (CU4: Solar System Objects, Multiple Stars, Extended Objects), Astrophysical parameters processing (CU8).
CNES participates also actively to the CU1 which is the system architecture unit, provides technical coordination and computing facilities for simulators (CU2) and also contributes to the preliminary definition of the final Gaia catalogue (CU9).
CNES is in charge of the development, validation and operation of the DPCC, CNES Data Processing Centre. The operations are foreseen for 7 years : 5 years of analysis and data collection and 2 years for the final reprocessing catalogue.

The DPCC architecture is based on three main functions, embedded in a framework named SAGA :

- The GaiaProc function that allows to run all the scientific algorithms and to supervise and monitor their processing. The algorithms are executed as lots of jobs on a large dedicated computing cluster in the CNES computing centre. This function is based on a Job Scheduler adapted from the Pleiades project and called Phoebus. An important part of the GaiaProc function is also the DataManager whose role is to prepare the data in the most efficient way to optimise the time spent to access the data. The large amount of data, database volume of PetaByte order, and the associated complex processing have led us to choose a NoSql technology based on Hadoop/HDFS technology for the data manager.

- The GaiaDex function that allows to exchange the data with ESAC and other partners via network in a Gaia specific binary format called Gbin.

- The GaiaWeb portal that will be used during testing and operations to exchange data with the scientists, essentially in order to validate the scientific results.

The SAGA framework is developed by a sub-contractor THALES.
The scientific codes, developed in Java by the scientists themselves (about 80 developers) are integrated and tested in CNES before delivery to Thales. About 20 huge scientific processing chains are planned to run in CNES.

III. Overview of Gaia DPAC Operations

Gaia DPAC operations will be very complex compared to other scientific space missions. The ground segment is composed of six Data Processing centers, including the Science Operation Center, located in ESAC, near Madrid. The data will be exchanged between SOC and the five other DPCs in a hub-and-spokes manner, the hub being DPCE (SOC).
The data reduction task will be cyclic over the five years of the mission and for final re-processing until the release of the final Gaia catalogue. This basic principle is dictated by the self-calibrating nature of the Gaia mission. It starts from a very rough (as viewed by the standards of Gaia's scientific goals) initial knowledge of the spacecraft, instruments, attitude and sky. The actual measurements will be used to cyclically get improved estimates for all these parameters, until a convergence of the procedure is achieved.

Each cycle is planned to last from about six months to one year. The raw and processed data are stored in the so-called Main Data Base (MDB) located in ESAC. At the beginning of each cycle, a version N of the MDB is sent to all the DPCs. During the cycle each DPC process the data, coming from the last cycle results from its processing but also from all the other processing chains of the DPAC. At the end of the cycle all the DPCs sent back the results to ESAC. ESAC will collect all the results, ingest them in the MDB and create the MDB cycle N+1 that will be the entry of the next cycle.

In addition to this cyclic processing, some daily processing has to be done, to manage the Science Operations Centre (SOC) and the Mission Operations Centre (MOC) interface, to verify the spacecraft instruments status, to calibrate the instruments and raise alarms. Daily processing will be done in DPCE, DPCC, DPCI and DPCT.

In this framework, the DPAC operations plan must be set up to satisfy the following goals:

-Perform a continuous monitoring of the spacecraft and payload status and of the data quality, in near-real time and during the entire operational space mission. This is necessary to assure that at any time the best-possible raw data are produced on board.

-Use as few processing cycles (overall iterations) as possible. The total number of processing cycles is the prime factor determining the total computing effort needed.

-Perform a complete calibration, including a CCD charge damage model, as early as possible. This is important to keep the total number of processing cycles small.

-Produce scientifically interesting outputs as early as possible. This is to satisfy the clearly understandable desires of the scientific community.

-To start the processing of all major DPAC SW systems provided as early as possible.

-Derive the optimum possible end results within the specified time (three years after the end of the operational space mission).

IV. CNES Gaia data processing operations

A. Baseline and constraints

The DPCC have the responsibility to run the CU4, CU6 and CU8 processing all along the Gaia mission and for the post-launch re-processing. It includes both daily and cycle processing.

The scientific chains to be run are the following ones:

- -Daily processing
  - -CU6 Daily processing, RVS instrument daily calibration
  - -CU4 SSO (Solar System Objects) -Short Term, alerts on Solar System Objects
- -Cycle processing
  - -CU6 Cycle Spectroscopic processing
  - -CU4 SSO-Long Term, Solar System Objects processing
  - -CU4 SSO-Update, update of the SSO Ephemeris data base (weekly chain)
  - -CU4 NSS, Non Single Stars
  - -CU4 EO, Extended Objects
  - -CU4 EO-Training, Training of Extended Objects
  - -CU8 APSIS, Astrophysical classes and astrophysical parameters determination

The daily processing will enter operations as soon as the first science telemetry is received by DPAC, 3 months after launch, cycle 0 processing.

The cyclic processing chains will not enter operations at the same time due to cyclic data reduction approach between CUs and need of some chains to have a certain amount of data or to have results of previous cycles computed by other CUs.

So along time the DPCC will have more and more scientific chains to process.

In the same time, from a cycle to another the data volume to process is increasing: the raw data is always re-processed from the beginning of the mission, the volume of data processed by other CUs increases too at each cycle.

In summary the main difficult aspects of DPCC Gaia operations are:

- -increase of the number of scientific chains at each cycle,

-increase of the data volumes at each cycle

-run daily and cyclic processing on the same hardware cluster, giving a priority to the daily processing

-scientific codes entering operations more than 2 years after launch will go being developed and integrated as operations of the first cycle have already begun.

B. The DPCC Complex Operation Plan

The figure 2 presents the macro-planning of the Gaia cycles and the scientific processing that has to be operated.

It shows the increase of complexity as operations go on and the fact that some scientific processing chains are launched two years after launch.

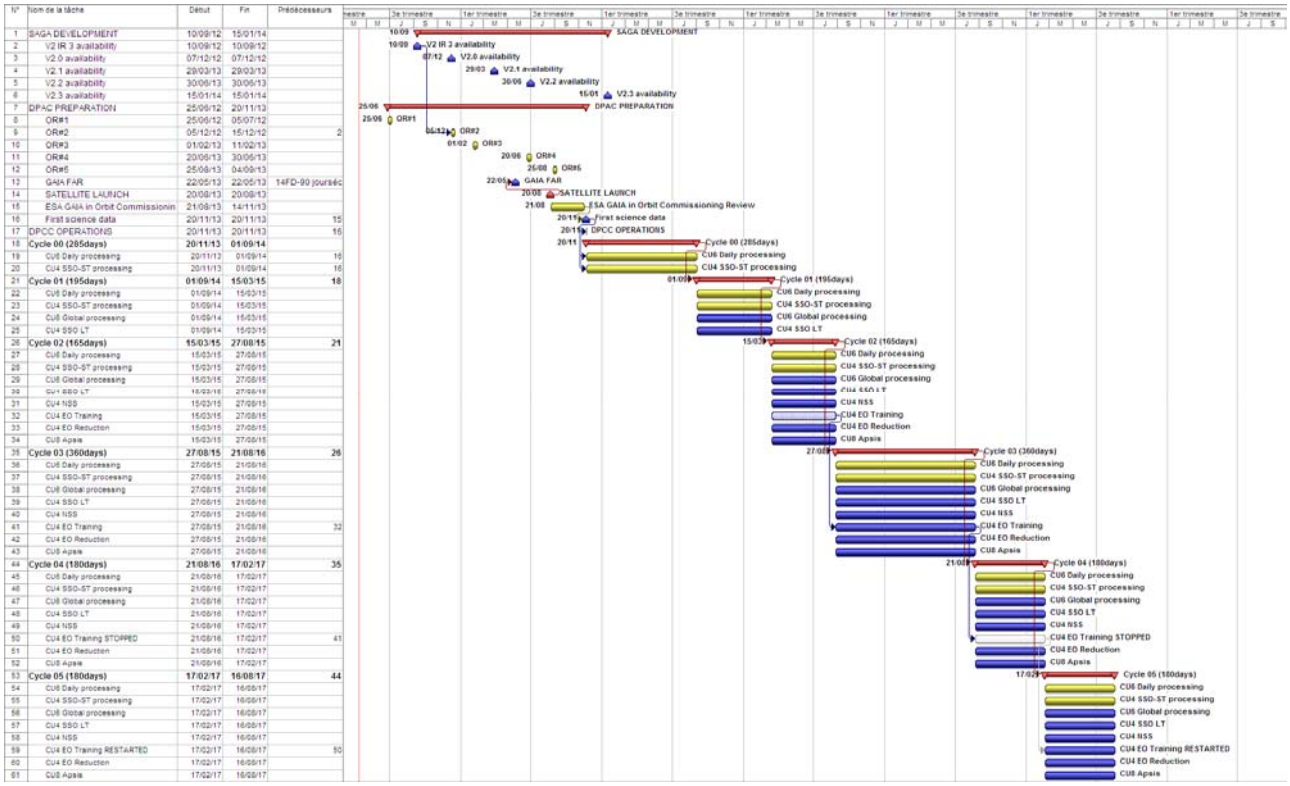

Figure 2. DPCC operations chronology, scientific chains and cycles

If we try to model the operation plan complexity all over a day, assuming that we will run every day all the chains that have to be exercised at this time we got the figure 3 schema. It shows the complexity in planning and conducting the operations.

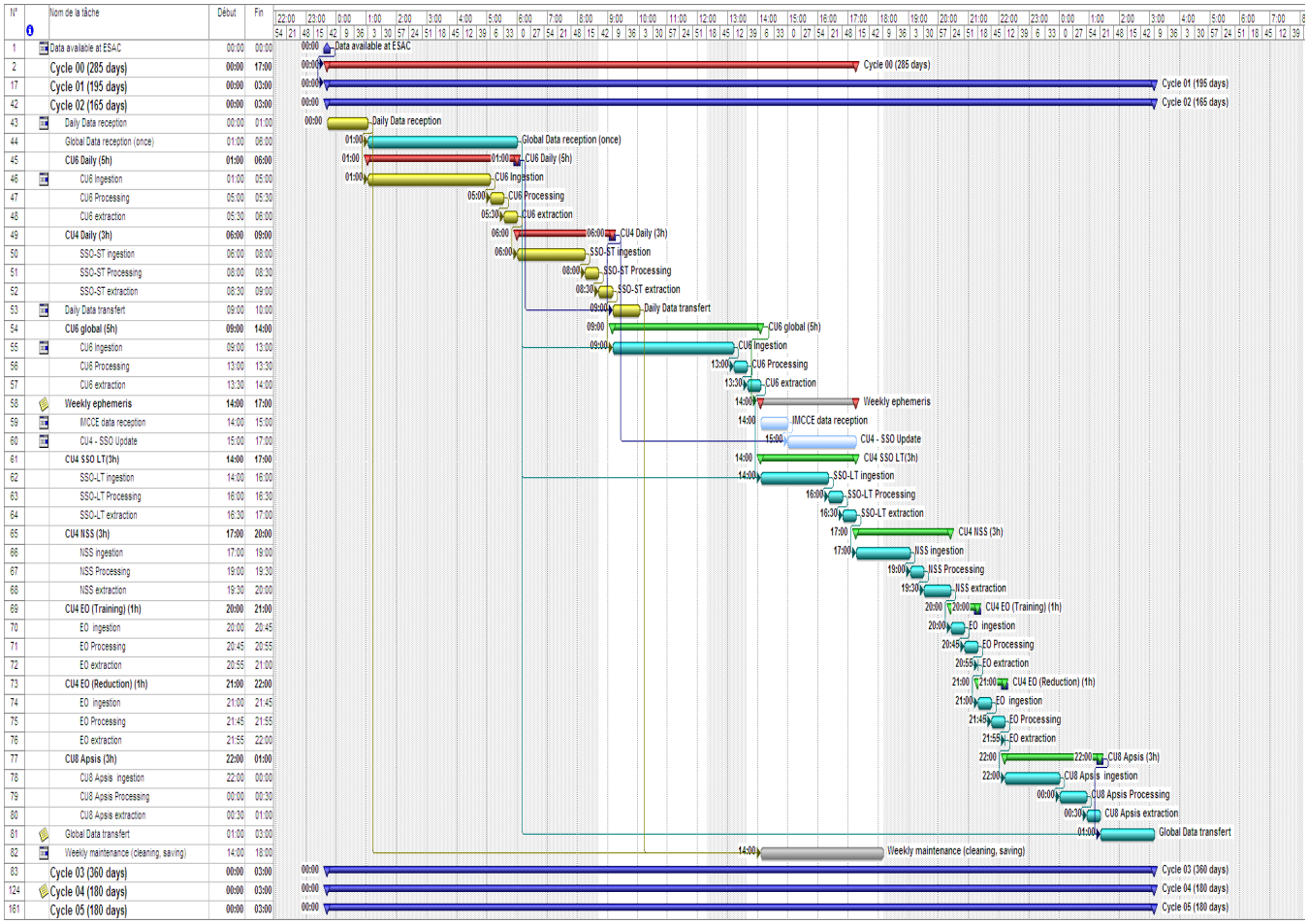

Figure 3. DPCC cycle 2 operations chronology, one day simulated

C. The Tools to manage the operations

We have shown the complexity of the operation plan in terms of number of chains to run and chronology to be respected. The management of the execution of the different scientific chains on the number of objects to be processed (more than one billion sources) is done with a job scheduler (THALES Phoebus product).

For each scientific chain, the dataset to be processed is splinted in small datasets, leading to about one hour of processing, and the associated jobs are spread on the computing cluster. The estimated size of the cluster at the end of the mission is more than 6000 cores on more than 300 nodes for a data volume of PetaByte order.

Cyclic processing for one scientific chain can take weeks on the cluster. In order to follow the evolution of the operation a tool called “operation plan” has been developed.

We show in figure 4 an example of the HMI showing the status of the different on-going scientific processing. It shows the progress of the different chains in terms of percentage of processed objects, the processing speed in terms of number of objects per second. It is a high level view.

The tool permits to configure the chains execution, the timing of the launch, the number of objects per job, the priorities.

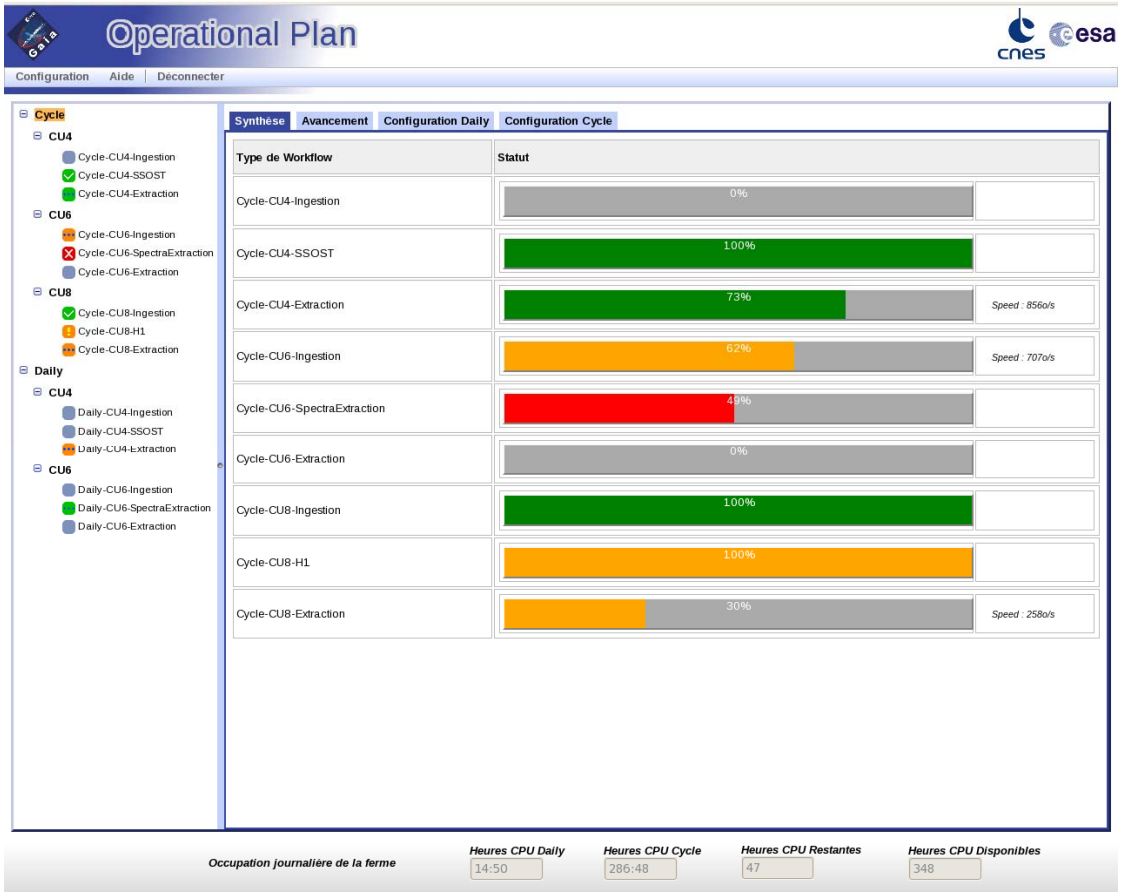

Figure 4. An example of the DPCC Operation plan

D. The organization

The difficulties in organizing the DPCC project in CNES after the beginning of operations is that the processing software will not all be at the same level of development at launch date.

The cycle 0 software will be ready, validated and tested in CNES and in DPAC due to system end to end tests and to operation rehearsals. Five operation rehearsal to exercise cycle 0 software are planned in DPAC from June 2012 to September 2013.

Concerning the cycle data reduction processing chains , some have been validated during “end to end” DPAC testing in 2011 and 2012. Due to the lack of realistic data it has been impossible to validate some chains (CU4 Non Single Stars and Extended Objects) at DPAC “end to end” level. The chains that have been exercised during system tests were in preliminary versions and are still subject to huge evolutions before they will enter operations.

The development will go on as operations have begun and CNES integration teams will have to remain working some years after launch. They will have to take into account the corrections and evolutions on the scientific codes being already in operation in the early cycles and in parallel go on working on the development and testing of the chains entering mater in operations.

One of the important point to take into account is that the scientific part of the DPCC codes are developed and planned to be maintained by the CU scientists. It implies that the whole DPAC organization that has been installed for the development has to remain in place some years after launch. In CNES people who ensure the link between CUs and DPCC will be necessary all operations long in order to manage the scientific codes corrections and evolutions. They will also be the link between DPCC and scientists in their responsibility to control the scientific quality of the results.

V. Conclusion

We have presented in this paper the complexity of the CNES Gaia operations in terms of:

- volume of data to process (PetaByte order at the end of the mission),
- very complex scientific code developed by the scientists and that are integrated and tested in CNES,
- complex planning of the operations due to daily and cyclic chains and to the numbers of chains to process,
- scientific chains entering in operations more than 2 years after launch that implies development going on during operations.

The development is still going on and systems tests going to be passed at the end of 2012. Operation Rehearsals will take place in 2012 and 2013 to prepare cycle 0 operations. Development and testing of cyclic chains will continue more than 2 years after launch. The CNES Gaia project organization for development and operations has to take into account all these constraints and this will be a big challenge.