Astronomical Data Analysis Software and Systems XIII
ASP Conference Series, Vol. 314, 2004
F. Ochsenbein, M. Allen, and D. Egret, eds.

Gaia: understanding our galaxy

X. Luri

Dept. Astronomia, Universitat de Barcelona, Diagonal 647, 08028 Barcelona (Spain), Email: xluri@am.ub.es

S.G. Ansari

European Space Agency / ESTEC, 2200AG Noordwijk, The Netherlands

J. Torra, F. Figueras, C. Jordi, P. Llimona and E. Masana

Dept. Astronomia, Universitat de Barcelona, Diagonal 647, 08028 Barcelona (Spain)

Abstract.

Gaia is an ambitious mission to chart a three-dimensional map of our Galaxy, the Milky Way, in the process revealing its composition, formation and evolution. Gaia will observe about 1 billion objects in the Galaxy, about 100 times each one, during its 5-year lifetime, providing astrometric data of unprecedented accuracy (about 10μas at 15th magnitude) as well as radial velocities and photometric measurements in 16 broad and medium band filters.

Gaia will produce about 20 Terabytes of raw telemetry data that, after treatment and reduction, will generate a database of the order of 1 Peta byte. Contrary to other PI-based missions, Gaia data is required to reside in a database in its entirety, due to the complex interaction of the algorithms that will operate on the data to derive distances, proper motions astrophysical properties and create the final three-dimensional model of the Galaxy.To estimate the processing power and complexity required to build and manage such a database the European Space Agency issued the Gaia Data Access and Analysis Study (GDAAS) contract. This ongoing study is focusing on several important development issues. On the one hand, it aims to collect a relatively complete set of algorithms required to process Gaia Data and obtain from them estimates on CPU power, memory and archive size for the whole mission. On the other hand, it also aims to identify the most appropriate database management system technology that will not only be reliable, but will last beyond the 5-year lifetime of the mission.

An overview of GDAAS and its current results is presented.

1. Introduction

The Gaia astrometric mission was approved by the European Space Agency (ESA) in 2000 as a *Cornerstone* mission, to be launched around 2010-2012. Gaia will observe more than one billion stars, several millions of galaxies, hundreds of thousands of solar system objects and many types of other exotic objects (ESA (2000), Perryman et al., 2001). Gaia will provide astrometry (10μas precision at V=15), multiband multi-epoch photometry and radial velocities for all of them.

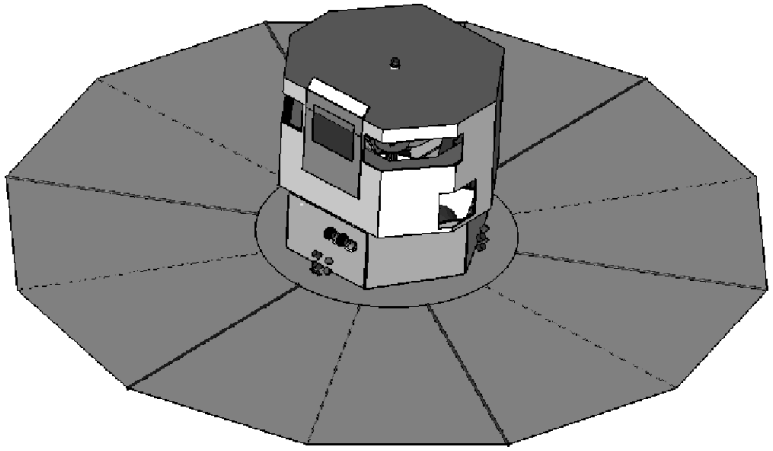

Figure 1. Top-view of the Gaia satellite

According to present estimates Gaia will downlink about 120TB of raw telemetry that will require about 10^{19} flops of processing to produce the final Gaia catalog. In addition, the data presents complex relationships between different data sets and scientific, instrument and satellite data.Therefore, the design and implementation of the Gaia Data Management System is non-trivial.

In the spring of 2000, ESA issued a Call for Proposals for the development of the Gaia Data Base and Access Study (GDAAS), having as its main goal "To define an efficient, scalable, maintainable and usable system for populating the Gaia database (DB) from the satellite data stream allowing not only the data storage but also the processing of scan data". The challenge was to establish the technical baseline concepts for the system on a realistic basis and to prove the feasibility of the approach chosen for the reduction of the mission data. The contract was awarded to a Consortium constituted by GMV (Software Company, Madrid), CESCA (Supercomputing Center of Catalonia) and the group at the University of Barcelona as the scientific partner.

We present here an overview of the mission and the development of GDAAS, describing the approach we have followed and some results of the tests performed for critical algorithms.

2. Overview of the Gaia mission

Gaia is an ambitious space observatory, adopted within the scientific program of ESA and building on the success of the *Hipparcos* astrometric mission.

Gaia's main goal is to obtain very precise astrometric data (positions, parallaxes and proper motions) of an extremely large number of stars and other astronomical objects. For this, Gaia will carry on a *complete and unbiased* all-sky survey up to 20^{th} magnitude that will create a catalog of about one billion objects. This catalog will contain mainly stars, but will also include many other types of objects, namely:

- 10^6 - 10^7 galaxies
- $\sim 5 \times 10^5$ quasars
- $\sim 10^5$ extragalactic supernovae
- 10^5 - 10^6 (new) asteroids
- $\sim 50,000$ extrasolar planetary systems
- many others

Gaia's angular measurements will have a precision of about 10 microarseconds (μas) at 15^{th} magnitude. In order to achieve this, Gaia will use two telescopes combined onto a single focal plane composed of 180 state-of-the-art Charge Coupled Devices (CCDs). The satellite will continuously scan the sky, allowing for about 50 measurements for each star during the 5 years of duration of the mission. Full sky coverage will be possible because of the spin motion of the satellite over its own axis, combined with a precession motion.

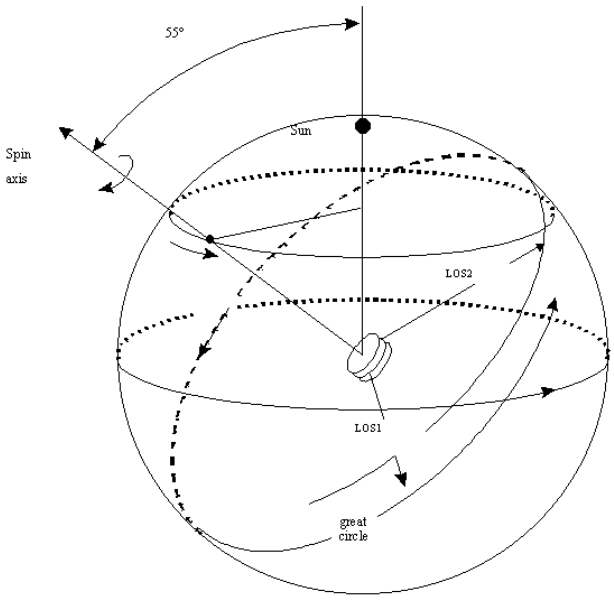

Figure 2. The Gaia scanning principle

Continuous measurement of stellar sources using CCDs implies a special operation, different than the typical shutter based imaging. Time Delayed Integration (TDI) is the best option for this case, which is based on a continuous charge shift from one pixel row to the next synchronized with the satellite spin motion. It will be done in each CCD, thus accumulating the charge during its corresponding integration period allowing long exposure times without distortion or blur.

The main focal plane will also provide Broad Band Photometry (BBP) for the observed objects while a third telescope, projected over another CCD focal plane, will measure radial velocities and Medium Band Photometry (MBP). This will provide the third component of the spatial motion as well as complementary physical information of the objects, allowing the determination of astrophysical parameters like the metallicity or the effective temperature.

The combination of all these types of data and its high accuracy allows a very ambitious scientific program for Gaia. The main science driver of the mission is the study of the origin, formation and evolution of our Galaxy, including:

1. **Structure and kinematics of our Galaxy**
   - Shape and rotation of the bulge, disk and halo
   - Internal motions of star forming regions, clusters, etc
   - Nature of spiral arms and the stellar warp
   - Space motions of all Galactic satellite systems
2. **Stellar populations**
   - Physical characteristics of all Galactic components
   - Initial mass function, binaries, chemical evolution
   - Star formation histories
3. **Tests of galaxy formation**
   - Dynamical determination of dark matter distribution
   - Reconstruction of merger and accretion history

However, the Gaia scientific case is much wider, because the amount and quality of the observations will have a strong impact in many other areas, namely:

- Stellar astrophysics
- Solar System studies
- Extra-solar planetary science
- Cosmology
- Fundamental physics

3. GDAAS

In the spring of 2000, ESA issued a Call for Proposals for the development of the Gaia Data Base and Access Study (GDAAS), having as main goal to

> *Define an efficient, scalable and maintainable system for populating the GAIA mission database from the satellite data stream*
> *The system should not only be designed for data storage and retrieval but also to allow the reduction processes to be easily implemented and optimally run*

The amount and complexity of the observations produced by Gaia makes the data reduction a very challenging problem. The Gaia data processing can be decomposed in 4 main categories of processes, as described in Figure 3:

1. **Data ingestion:** Gaia will downlink a total of about 120TB of (uncompressed) data. The amount of daily data will heavily fluctuate depending on the sky region being scanned by the satellite; GDAAS should be able to cope with these variations, maintaining an ingestion rate compatible with the downlink.

2. **First-look:** the dowlinked data will be examined "on the fly" during the ingestion process to detect transient phenomena like Supernovae or microlensing events, as well as transits of minor planets. GDAAS should allow the implementation of those time-critical processes.
3. **Core processing:** the key of the Gaia data reduction is a process called *Global Iterative Solution* (GIS), fully described in Lindegren (2001). This process runs on a subset of about 100 million "well behaved" objects and allows to simultaneously obtain the calibration of the instruments, the satellite attitude and the scientific results of the mission. Core processing involves iteratively solving a system of some hundreds of millions of equations and some hundreds of millions of unknowns.
4. **Shell processing:** once the core processing is finished, a plethora of specialized processes will run on the database to carry on specialized analysis for many types of objects and phenomena. These "shell processes" will allow the full scientific exploitation of the data

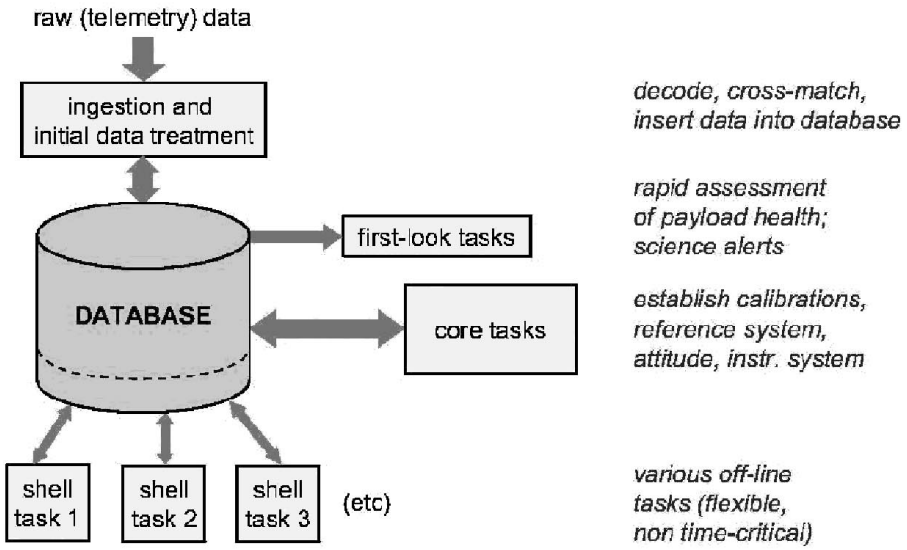

Figure 3. Overview of the Gaia data processing

Therefore, the main goal of the GDAAS study has been to establish the technical baseline concepts for the system on realistic basis and prove the feasibility of the approach chosen for the reduction of the mission (Gonzlez et al., 2002).

3.1. Technology

GDAAS has been developed from the start using the Object Oriented programming paradigm. Before any system implementation was be carried on, a UML model was defined describing, on the one hand, the data structure for the database design (Data Model) and, on the other, the structure of the system and of code to be developed.

The coding of GDAAS is done in Java. This language was chosen for its flexibility, allowing a quick development, and portability. In the future we might consider moving to faster languages like C++ to improve performance if necessary.

The system itself is designed in a classical three-tier structure, as shown in figure 4.

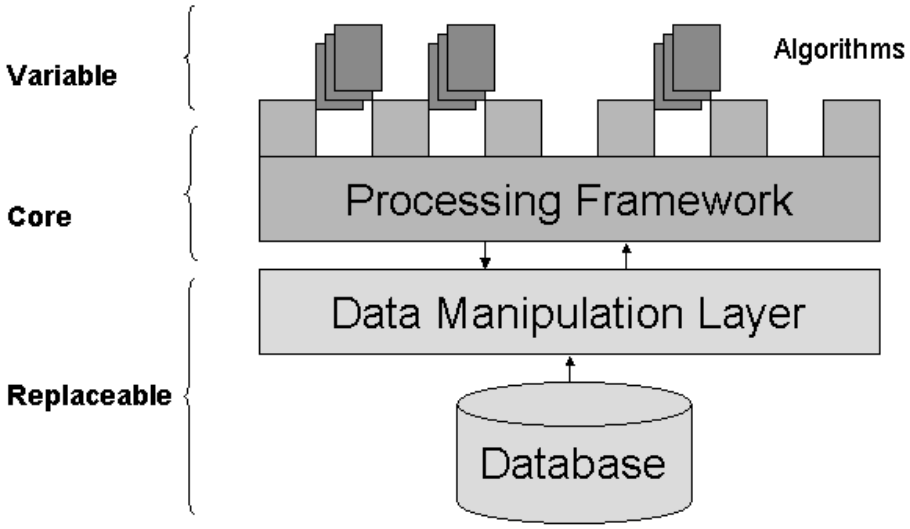

Figure 4. GDAAS architecture

The *Core layer* is the GDAAS processing framework. This framework provides the basic tools for algorithm implementation and also advanced tools for process distribution, and inter process communication. Therefore, GDAAS is designed to be a distributed processing environment.

This core layer is complemented by a *Data Manipulation Layer* providing the interface to access the database. This layer isolates the details of the database access from the implementation of algorithms in GDAAS. If the DB system is changed only this layer has to be replaced, without any other change in the GDAAS system.

Such a replacement has already been tested in the project. The database system initially used in GDAAS was *Objectivity*, a fully object-oriented database system. Due to incompatibilities with the hardware and performance problems it was replaced by Oracle 9i-RAC, an object-relational DB system, which is being currently used in the project.

The use if this layer, combined with the coding in Java, makes GDAAS a highly portable system. GDAAS has been installed and run in many different hardware and OS environments, requiring only minor adaptations.

Finally, the algorithms for data processing constitute the third layer of the system. These algorithms are provided by the scientific community and implemented using the GDAAS processing framework. Although the recommended

language for algorithm implementation is Java, we are experimenting with the integration of algorithms written in C and Fortran.

3.2. Testing

A first functional prototype of GDAAS has been implemented and is being tested using simulated telemetry. The ongoing tests cover the most critical processes expected to be run in GDAAS, namely Data Ingestion (telemetry decoding, initial data treatment and cross-matching) and GIS.

These tests are representative of the actual Gaia processing but simplified and scaled-down:

- Only Astro instrument
- Basic instrument calibration model
- Simplified telemetry model
- Simulated telemetry from a realistic galaxy model (Torra et al. (1999))
  - Observations limited to 13^{th} magnitude instead of the20^{th} magnitude limit of the actual mission.
  - Six months of observations

From these tests, several conclusions can already be drawn regarding the actual Gaia processing:

1. The Gaia database will be of the order of a few PetaBytes.
2. The system can cope with the daily data ingestion, although for the most crowded regions of the sky some extra resources should be provided occasionally.
3. The system will require some hundreds of processors working in parallel to cope with the reduction needs.

3.3. Organisation of the Work

In order to cope with this largely complex undertaking, the scientific community around Gaia has been organised into a number of Working Groups, each entrusted with a set of responsibilities towards delivering a set of algorithms that will be used to further process the data (Ansari et al., 2003). The Working Groups range from Classification to Multiple Stars Systems, Photometry and Radial Velocity (i.e. all aspects of Gaia Science.)

Each algorithm that is delivered is evaluated by the GDAAS Scientific team, then passed on for implementation to the software team and in parallel assessed for its impact on the infrastructure. Configuration Control of each module ensures that the code is versioned and documented. These aspects are not only important to ensure involvement of the Gaia community, but that long-term changes can be tracked and preserved up to the implementation of the operational system.

3.4. Future developments

It is foreseen in the near future that more than 20 algorithms covering astrometric, photometric, radial velocities and classification aspects will be introduced in GDAAS to realistically assess the overall computing power impact required to process the totality of data. The scale of the tests will also be expanded, and in the next phase up to 5TB of data will be available in the testing environment,

along with new and more poweful processing nodes. Grid-related distributed computing technology will also be taken into consideration as protocols stabilise and become more reliable.

References

Ansari, S.G., Torra, J., Luri, X., Figueras, F., Jordi, C., Masana, E., 2003. *Gaia Spectroscopy, Science and Technology*, ASP Conference Series Vol. 298 p. 97

ESA 2000, *GAIA: Composition, Formation and Evolution of the Galaxy*, Technical Report ESA-SCI(2000)4

González, L.M. Serraller, I., Torra, J., et al., 2002, *GDAAS. Final Report*, Technical Report GMV-GDAAS-RP-001

Lindegren, L., 2001, *Data Analysis Study: Core Algorithms*. Technical Report GAIA-LL-34

Perryman, M. A. C., de Boer, K. S., Gilmore, G., et al., 2000, A&A, 369, 339

Torra, J., Chen, B., Figueras, F., Jordi, C., Luri, X., 1999, Baltic Astronomy, vol. 8, p. 171