Italian Gaia data management and processings infrastructure for the AVU systems at DPCT

A joint collaboration between INAF and ALTEC

Enrico Licata
on behalf of INAF & ALTEC teams

INAF
ISTITUTO NAZIONALE DI ASTROFISICA

ALTEC

ASI
Agenzia Spaziale Italiana

esa
gaia

INAF USC VIII - General Assembly
October 14 – 18, 2024
Galziniano Resort Terme & Golf - Viale delle Terme, 84 Galzignano Terme (PD)

The Team

ALTEC

Lorenzo Bramante
Ruben De March
Daniele Gontero
Rosario Messineo
Luigi Squillante
Leonardo Tolomei

Former Members

Fabio Filippi
Andrea Fonti
Enrico Pigozzi
Marco Vaschetto

INAF
ISTITUTO NAZIONALE DI ASTROFISICA

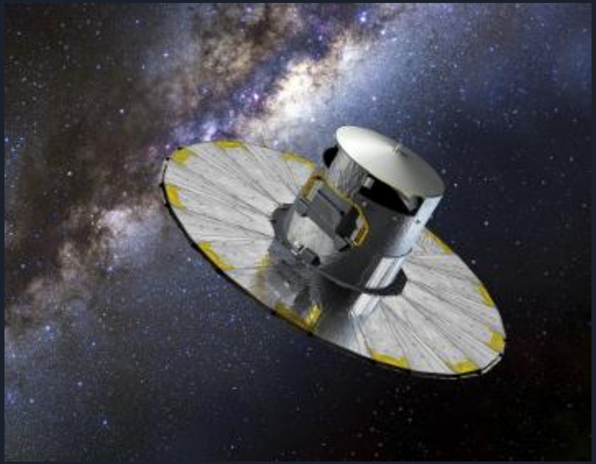
Ummi Abbas
Beatrice Bucciarelli
Deborah Busonero
Raffaella Buzzi
Mario G. Lattanzi
Enrico Licata
Roberto Morbidelli
Alberto Vecchiato

Former Members

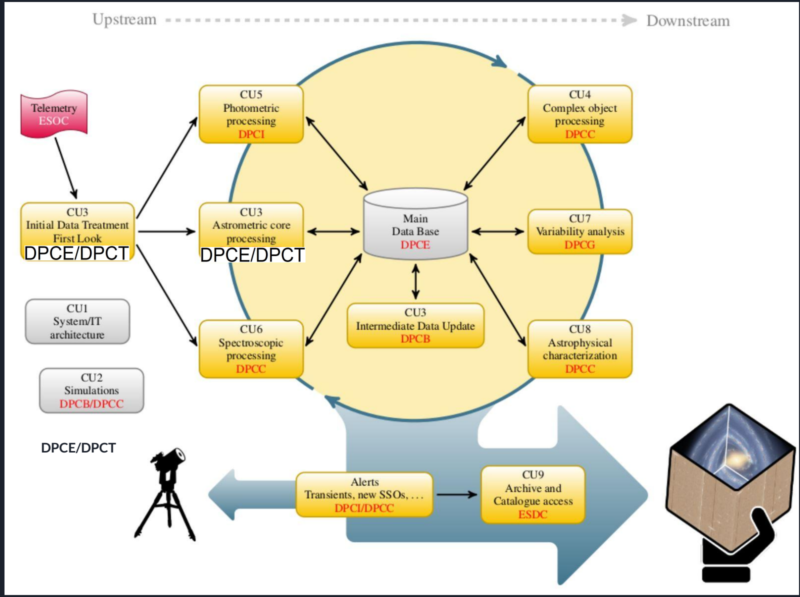
Luca Bianchi
Marco Pecoraro

A few info on Gaia and its Data

- Launched on 19/12/2013
- Expected duration ~5 years
- Actual duration 10+ years 😱
- Expected end of Gaia scientific data collection: Jan 2025
- Up to now, observed and measured 2+ billion sources
- Each source has been observed on average ~200 times
- Each source transit is comprised by
  - telemetry data
  - actual raw data (6.5k ~ 35k windows, 18x12 px, every 4.41s for every one of the 62 AF CCDs)
  - data and metadata generated by multiple processing pipelines across europe (9 CUs)
  - and much more…
- Estimated size of the full dataset @DPCT (including outputs) ~ 3PB



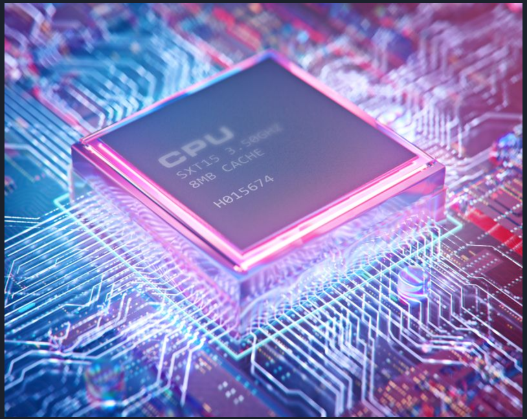
Raw data are NOT images



DPCT: one of the 6 data processing centers (DPC) of the Gaia SGS, hosted in ALTEC in Turin. Under dedicated ASI industrial and scientific contracts, its construction and operation is the result of the work of an integrated INAF-OATo / ALTEC team.

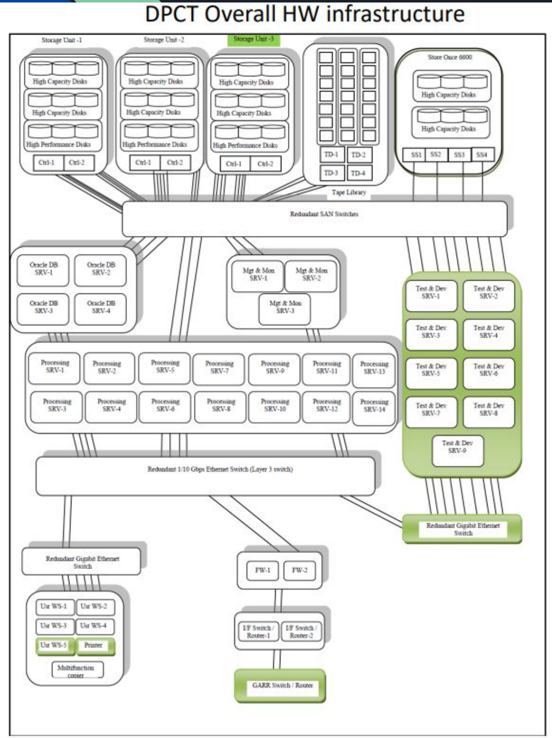
DPCT provides the infrastructure support (in terms of HW, DB and software framework) to run the CU3 sw systems that are part of the AVU (Astrometric Verification Unit): AIM, BAM and GSR from raw data to catalogue data.

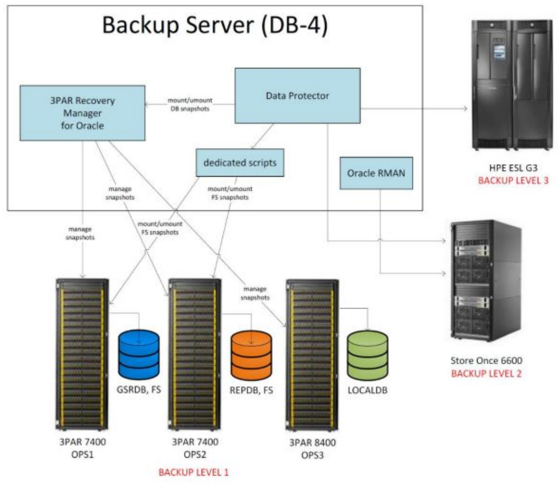
Processing Requirements @ DPCT



1. Handle a massive dataset: 10^6 ~ 10^7 transits / day
2. Pipeline integrated DBMS capable of supporting OLTP (online transactional processing) and OLAP (online analytical processing) operations
3. Max processing time for Daily pipeline < 24h
4. Heterogeneous workloads
   - Computation
   - Monitoring
   - Visualization
5. Store all mission data and all produced outputs of both daily and cyclic processing

HW infrastructure 1/2





Credits: ALTEC

INTERNET LINK: 10Gbps (300 Mbps guaranteed) via GARR

STORAGE CAPACITY: about 2.6 PB overall rawdisk space distributed among 2 HP P7400 and 1 HPE P8400, HPE Nimble HF40, Oracle ZFS and Oracle ODA

COMPUTING: 14 servers HP DL580 G7/G9 with a total of about 600 CPU cores and 4.5TB RAM.

DEV & TEST: 7 servers HP

DB SERVERS: 3 HP DL580 G10 dedicated to the database cluster based on Oracle RAC technology

NETWORK CONNECTION: LAN network up to 10 Gbps. SAN network redundant at 8/16 Gbps.

SECURITY SERVICE: redundant firewall on 2 HPE DL380 G10 servers for replacement. based on pfSense, enabling secure remote access via VPN.

INFRA MONITORING AND MANAGEMENT: services based on VMWare virtual environment configured with 2 HP DL580 G7 servers clustered and managed by vCenter Server.

BACKUP SERVERS: HP DL580 G7 dedicated to database and file system backups from data volume snaphotsand HPE DL580 G10 servers for replacement.

3 LEVELS BACKUP: L1 on primary storage array, L2 on disks (StoreOnce 6600) and L3 on tape libraries (HP ESL G3).

HPC INTERCONNECTION: access to HPC supercomputer at CINECA for dedicated processing

HW infrastructure 2/2

The DPCT hardware infrastructure is based on a distributed environment, including a computational grid, a database grid and a storage area network.

The DBMS choice is Oracle that provides advanced availability and scalability features. Oracle allows multiple computers to run the Oracle DBMS software simultaneously while accessing a single database, thus providing a clustered database.

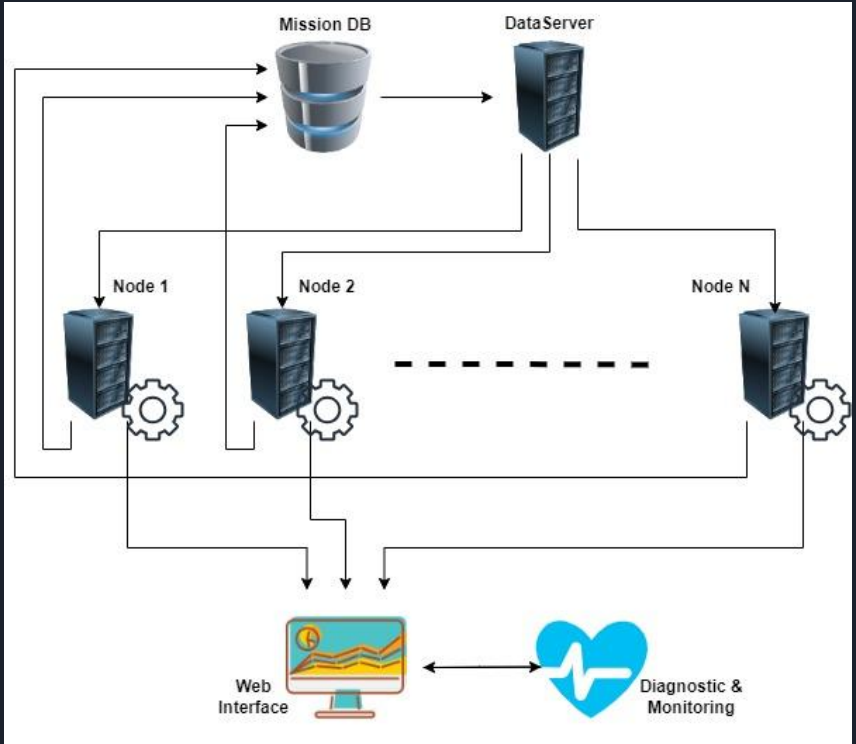
The database grid will use the following Oracle products:

- Oracle Server
- Oracle RAC
- Oracle Partitioning
- Oracle ASM to manage storage used by database

Simplified Functional Schema

For the execution of each production run:

- Data is retrieved from Mission DB and sent to the DataServer (DS)
- The DS performs checks, selects data, creates data packages to be sent to the processing nodes
- Each node runs the scientific SW.
  - Statistical data collection is performed at this level
- Outputs produced by the nodes is used to populate the Web Interface, and to perform Monitoring and Diagnostic Operations, while being stored back to the Mission DB



Production & Test Environments

To be able to develop your project you also need an appropriate dev and testing environment, together with a solid SW life-cycle

PROD ENV: 14 servers with a total of about 600 CPU cores and 4.5TB RAM.

DEV & TEST: half as much (is it enough?)

- Ensure consistency with sw baseline and configuration
- Unit Tests → Use Cases → System Test → Integration Test → Prod
- SRN → STR → Release
- ALL documentation adhere to ECSS standards (European cooperation for space standardization)
- Test automation via Jenkins + Ant Tasks
- 3 SW systems competing for HW resources

Considerations

- Is it reasonable / necessary to have a full scale test env?
- Plan for a procedure to keep Prod and Test envs configuration aligned

Infrastructure Software

- DPCT_SoftwareSystem
  - manages workflow
  - node deployment
  - db interaction
- DPCT_SoftwareConfiguration
  - manages the conf of daily and DRC workflows
  - separate conf for all project modules
  - plot descriptors
- DPCT_ConnectorLibrary
  - Integration between the Infra SWs and the scientific SWs

The integration between Infra and Scientific code managed by the ConnectorLibrary required:

- A Shared effort between INAF and ALTEC on specific parts of the source code
- Required tight cooperation and coordination between research institution and industry for the developent and maintenance

Scientific Code - Overview

- 5 pipelines:
  - AIM - Astrometric Instrument Modeling - daily/cyclic
  - BAM - Basic Angle Monitoring - daily/cyclic
  - GSR - Global Sphere Reconstruction - cyclic
- Parallelization is managed by the infrastructure
  - but many parts have been parallelized to allow testing of large batches of data
- Source code is under version control using SVN (chosen by ESA and imposed to the Consortium)
- Implements established design patterns such as: Chain of Responsibility / Singletons / Abstract Factory
- The data packages send to the processing nodes are not stored “as is”: they are runtime items, built from an underlying db and following a specific DM
- the whole AIM and BAM pipelines are performed locally on the DPCT infrastructure
- Computational requirements for GSR exceeds the capability of the DPCT
  - the SOLVER is executed on LEONARDO (see Cesare talk this conference)

Workflow - CD/CI - Issue Tracking

- Centralized management utility VMware vCenter: deployment of VMs across the servers
- Ingestion, Archive, and daily Workflow, are automated using custom GAIA SW system developed in Java
- A monitoring web interface updated in real time (in house sw)
- All main project's branches on SVN (both for infrastructure and scientific SW) are monitored by a Jenkins instance that manages
    - svn checkouts
    - compilation
    - tests with apache ant tasks
- Daily pipelines programmatically generate a report each day
- Issue tracking is managed with JIRA

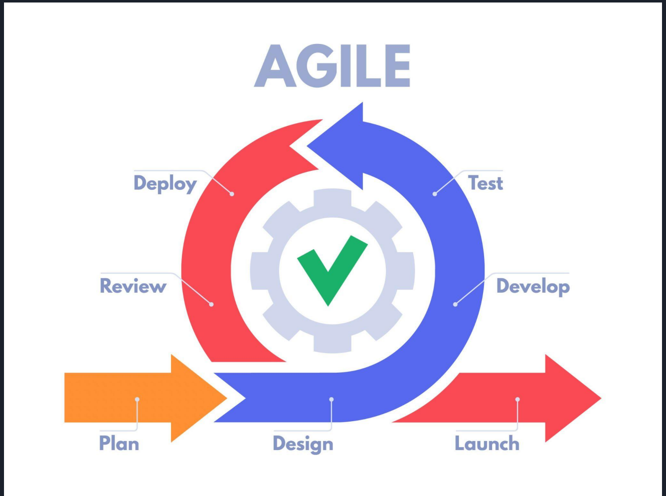
GAIA
TOOLS

Agile Project Management

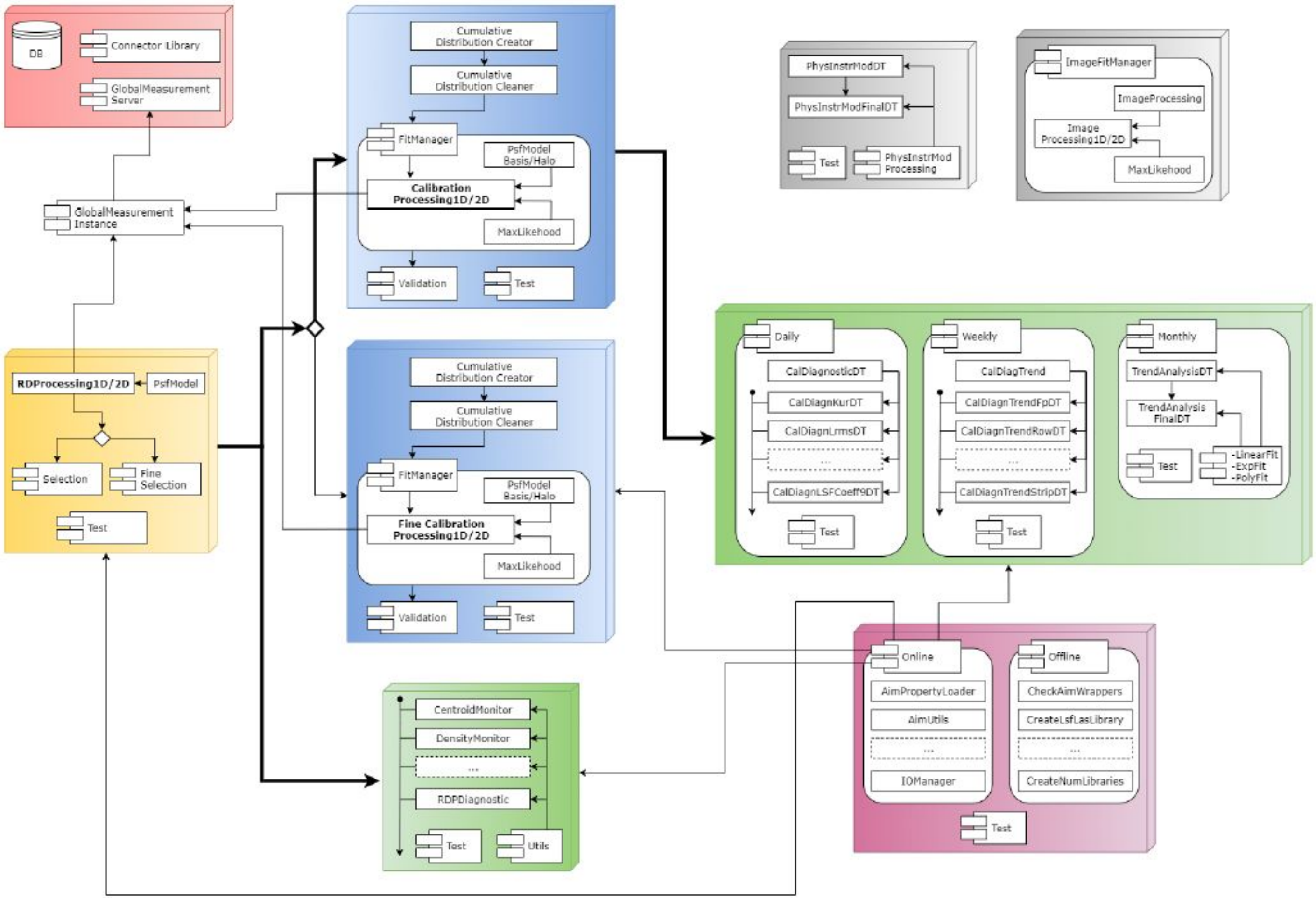
A project management approach that involves breaking the project into phases and emphasizes continuous collaboration and improvement.

- Periodic “STAFF Meetings” with whole team: 1/month to keep everyone engaged and discuss procurement and contractualization between all parties involved
- Operations briefing:
  checking the status of operations, bi-weekly
- Scrum / Dev meetings:
  among INAF and ALTEC developers to tackle specific technical issues

The frequency of each of these meetings has been adjusted during the project’s life-cycle (Planning /Commissioning / Early Operations / Operations) with the intent of maintaining minimal overhead



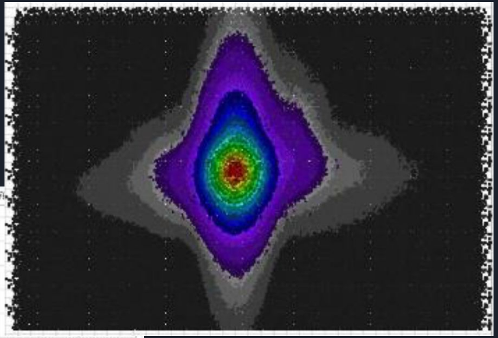
HIGH LEVEL ACTIVITY DIAGRAM of AIM PIPELINE:



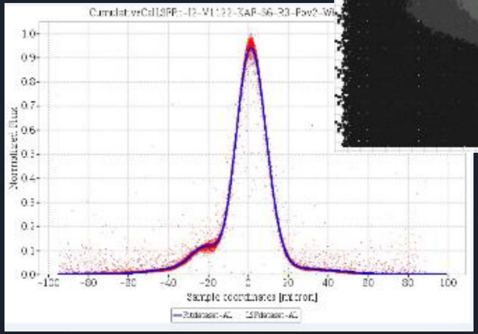
AIM - Astrometric Instrument Modeling

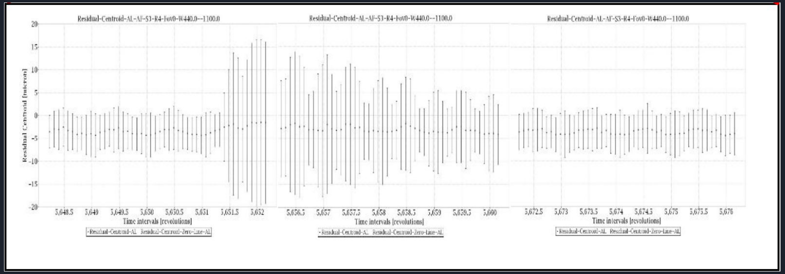
AIM daily pipeline: Raw data processing, image parameter determination, LSF/PSF modelization and calibration, astrometric instrument monitoring and diagnostics throughout the mission lifetime

- Daily sw version 100.000+ lines of code
- Cyclic sw version about 100.000+ lines of code
- 24 hours of raw data each run : from $2x10^6$ to $15x10^6$ ~~raw images~~ transits
- Complex structure of the pipeline: 10 sw modules managed by an unsupervised coordinator, the output of one run become the input of the next one
- ~6 hours of time execution on the DPCT Operation platform for each run

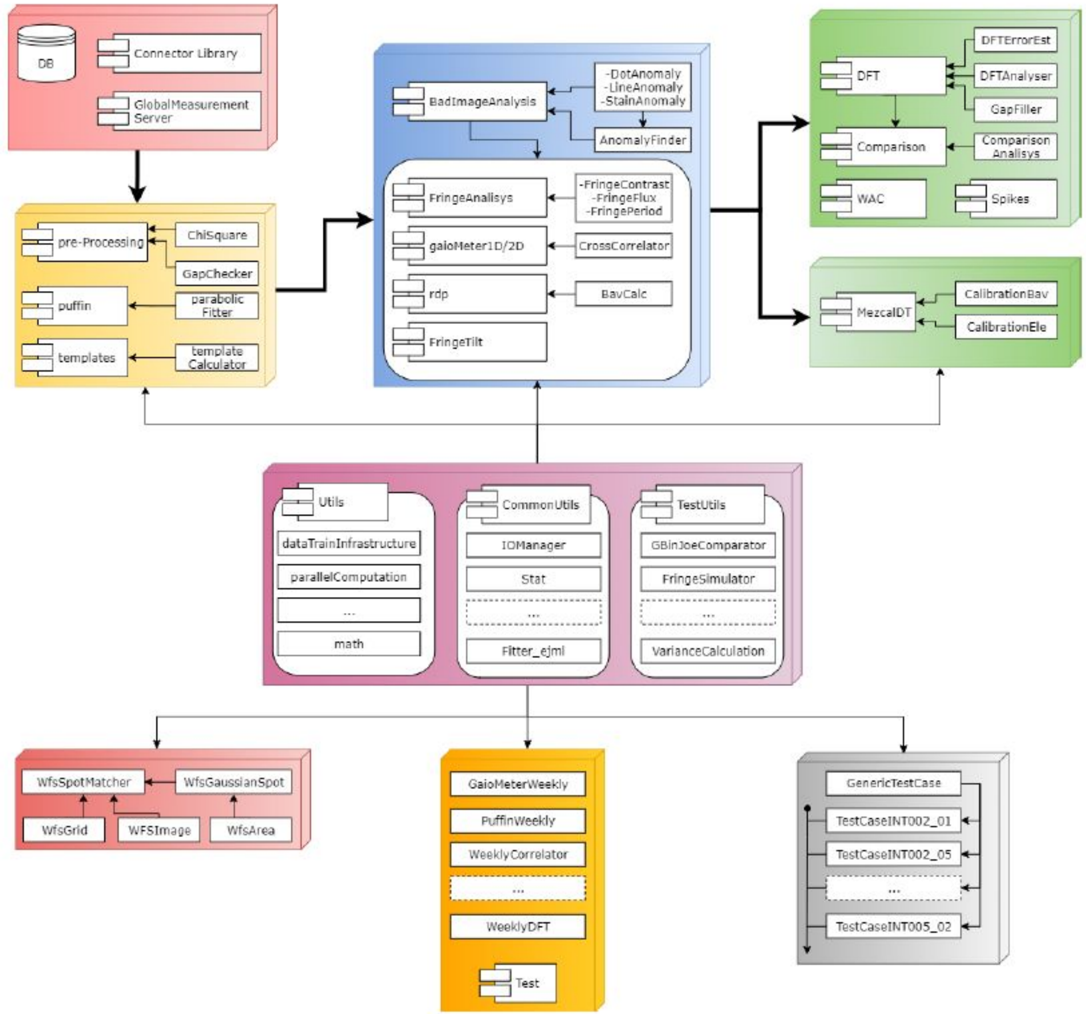


Cyclic version of AIM the software aims to reprocess images for calibration improvement and refined image parameters calculation





HIGH LEVEL ACTIVITY DIAGRAM of BAM PIPELINE:

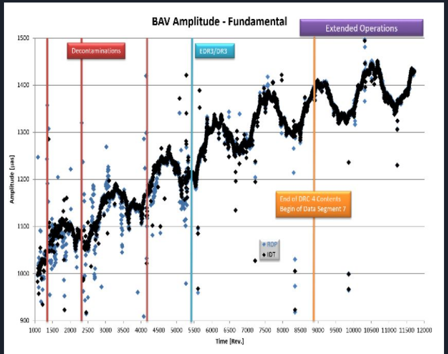


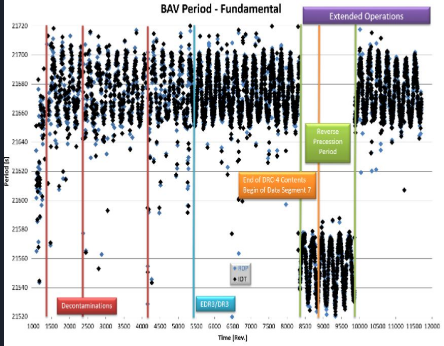
BAM - Basic Angle Monitoring

AVU/BAM daily pipeline: raw data coming from the Basic Angle Monitoring (BAM) instrument , i.e. fringes , for monitoring and analyzing the instrument behaviour throughout the mission and performing the BAV calibration

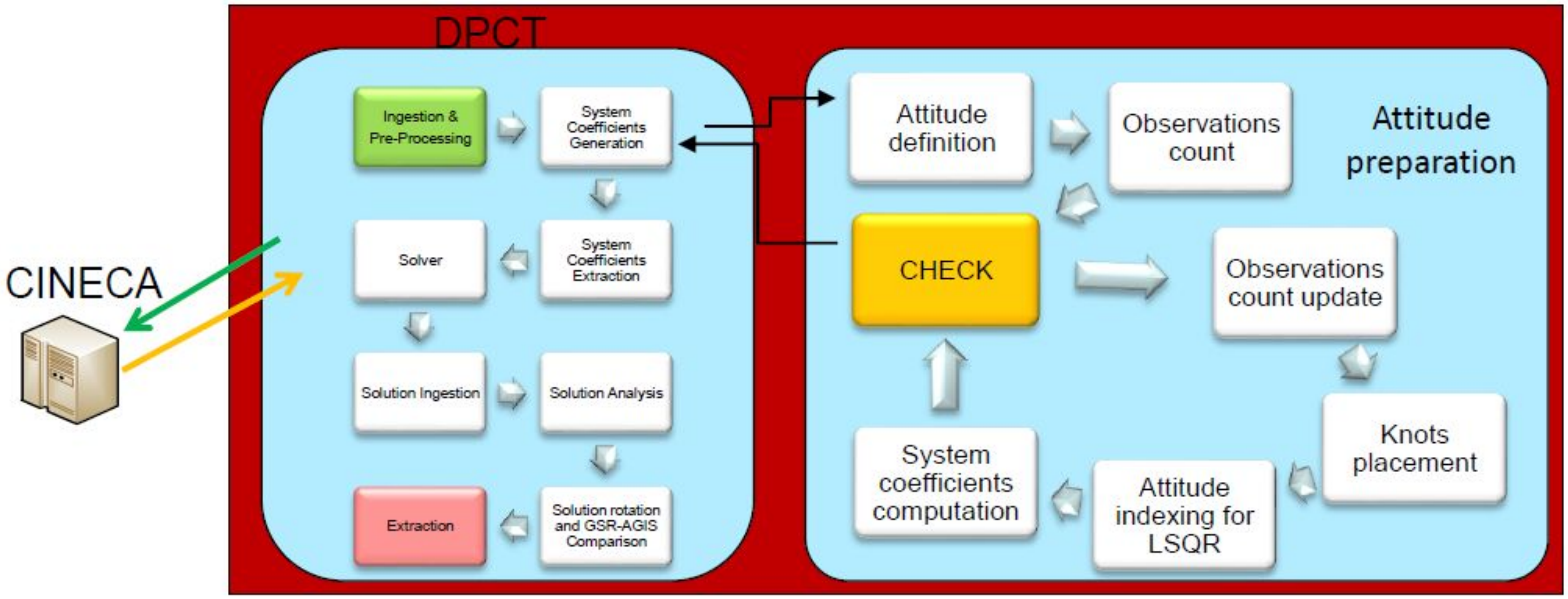
- Daily sw version 100.000+ lines of code
- Cyclic sw version circa 50.000+ lines of code
- 24 hours of raw data each run : almost $8x10^4$ images (1 pair every ~23s)
- ~2 hours for each run
- The pipeline output is sent to DPCE and ingested into the MDB

Cyclic version of the AVU/BAM software aims to reprocess fringes for calibration improvement





HIGH LEVEL DIAGRAM OF GSR PIPELINE



GSR - Global Sphere Reconstruction

The Global Sphere Reconstruction (GSR) solves a linearized system of equations who's result gives the global astrometric reference system (position, parallax, proper motions).

- Starting from 10^7 to 10^8 objects for each run
- Very complex pipeline structure
- 130.000+ lines of code in Java for data preparation, selection and analysis, and…
- The Solver running at CINECA
  - See Valentina Cesare's presentation:" The Gaia AVU-GSR solver: a CPU + GPU parallel code in perspective of Exascale systems"
- Final GSR output sent to DPCE in the MDB
- The Solver module run at CINECA which is managed as one processing node of the DPCT
- The whole process could be iterated for Non-Linearity

Transitioning towards a new DMS 1/3

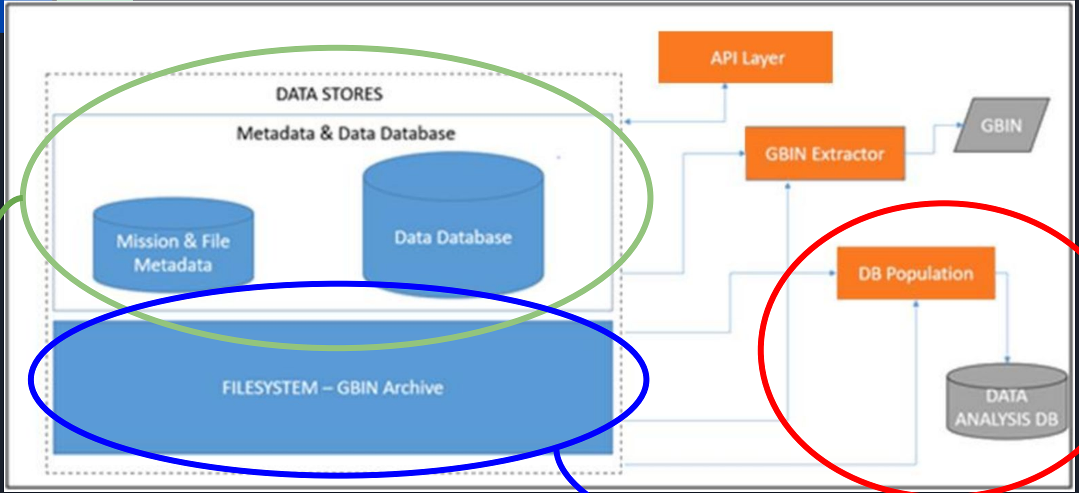
Why

1. Mission extended far beyond the expected timeframe
2. Massive databases/tablespaces (~TB) with detrimental effect on DMBS performance and maintenance
3. Daily / Cyclic operations require different DBMS paradigms
4. Heterogeneous requirements for the data model and its indexing/partitioning:
    - Hybrid DM: Relational with serialized objects (BLOB)

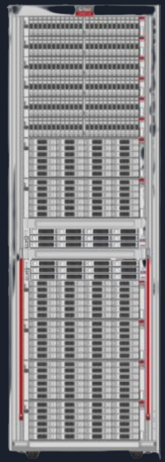
How

1. New Data Management System : Hybrid FS (ZFS) and DB solution
    - data lake with all mission data as GBINS
    - 2 different databases for medata and mission metadata
    - ODA dedicated to the scientific exploitation of the data
2. New data model:
    - Vertical cut of the original tables
    - removed / exploded every BLOB

Transitioning towards a new DMS 2/3 - HW & Logical Design



Oracle Database Appliance X8-2-HA, TAA



Oracle ZFS Storage Appliance Racked System ZS9-2

HPE Nimble HF40

Transitioning towards a new DMS 3/3

Original GAIA DM and Data format

GAIA DM it's an hybrid between relational and document based:

- Every object is represented as a table but some fields are objects as well
- Impossible to query on some of its data/metadata
- Access to the data requires specific sw tools

GAIA Data Format

- Gbin data format is a zipped serialized java object
- very limited interoperability
- very lightweight (structure not included)

New DM and Data Format

OPS4 DM is fully relational

- all fields are query-able with standard SQL
- data can be accessed directly on the DB (vertical cut) or retrieved fully from the archive

Data Format:

- Experimenting with HDF5
- greatly improved interoperability
- increased storage requirements (by a factor of 2)

Lesson Learned & Friendly advices

- Plan for the future:
  - space missions can outlive the planned lifetime and you and your infrastructure need to be ready for it
- Do not skimp on the test environments!
  - Make sure to scale them properly and plan procedures (automated if possible) to keep the envs consistent
- A broad amount of different expertise is required for the success of such projects:
  - leverage the expertise and tools of your colleagues (MIRTA :D)
- Share information and foster communication:
  - this keeps people involved and motivated, since these projects are a loooong commitment
- Be ready for the turnover:
  - Having good documentation and even better procedures to bring the new members up to speed is paramount!

The 20 years GAIA SGS experience, similarly to other mission like Euclid, Planck etc formed a community of experts, that is now an asset for INAF in perspective of future large missions and projects

If you survived this… THANKS for your attention!