Gaia, Data Processing and Challenges

William O'Mullane

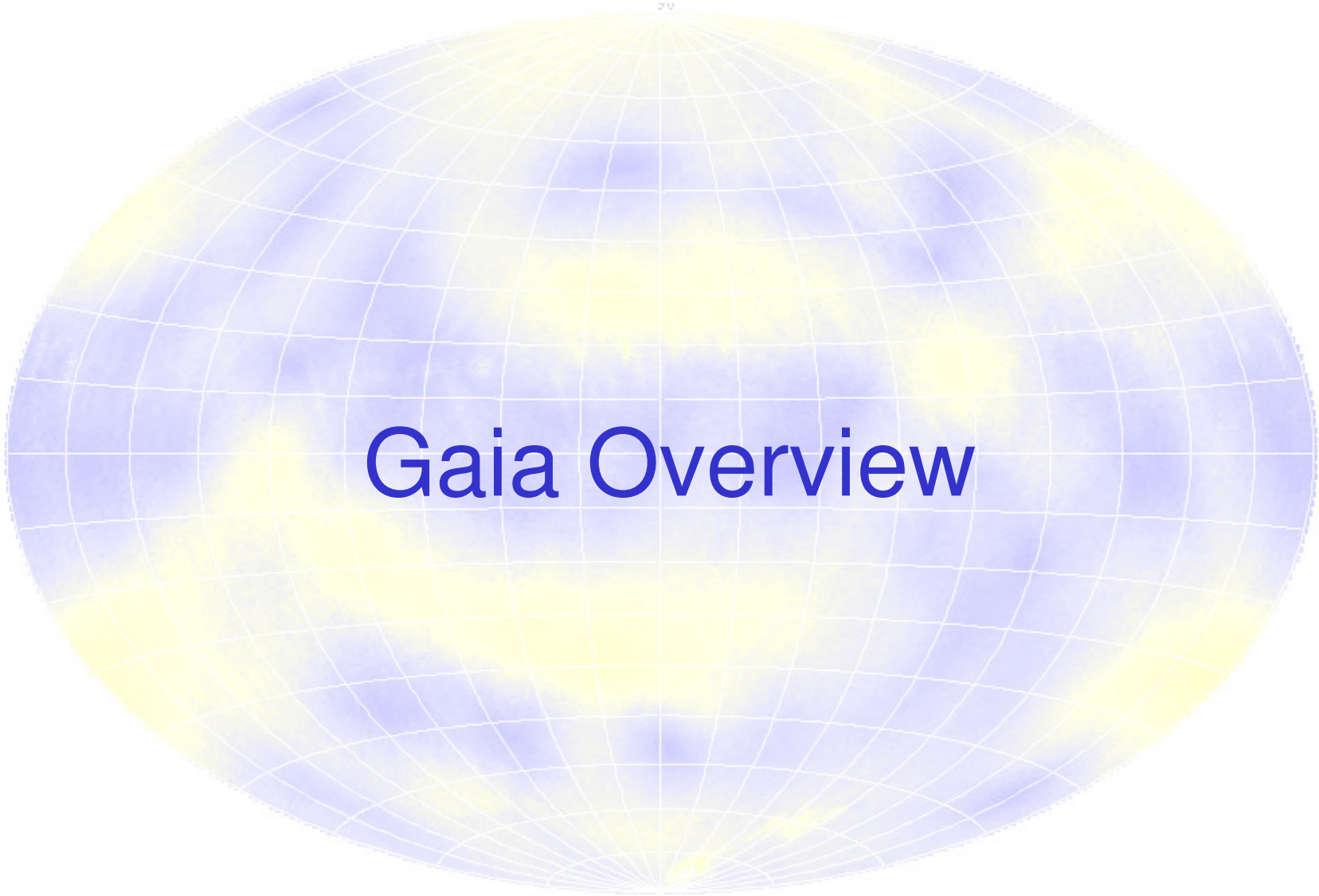
Gaia Science Operations Development Manager

European Space Astronomy Centre (ESAC)

Madrid

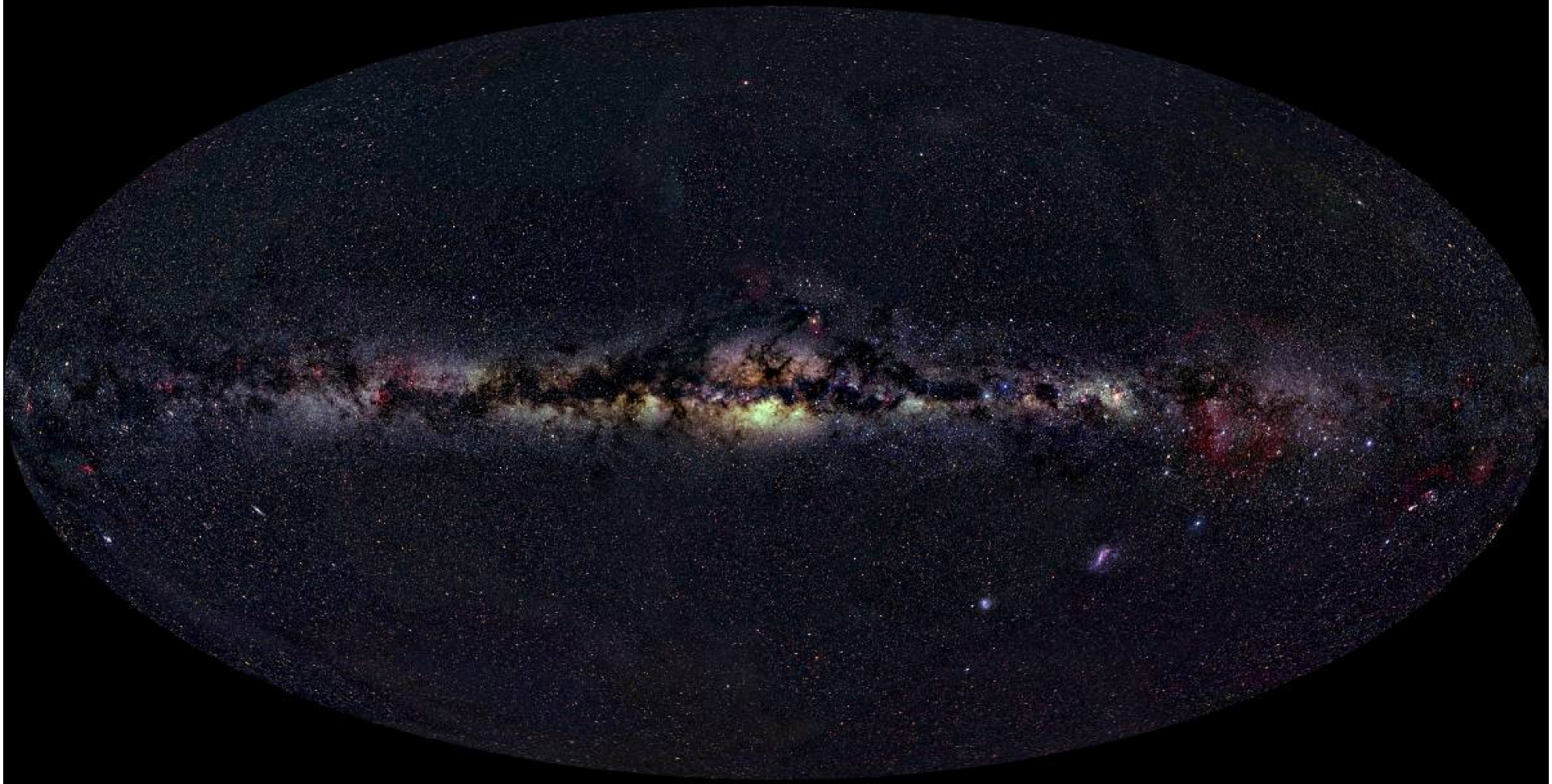
Spain

http://www.rssd.esa.int/Gaia



Gaia Overview

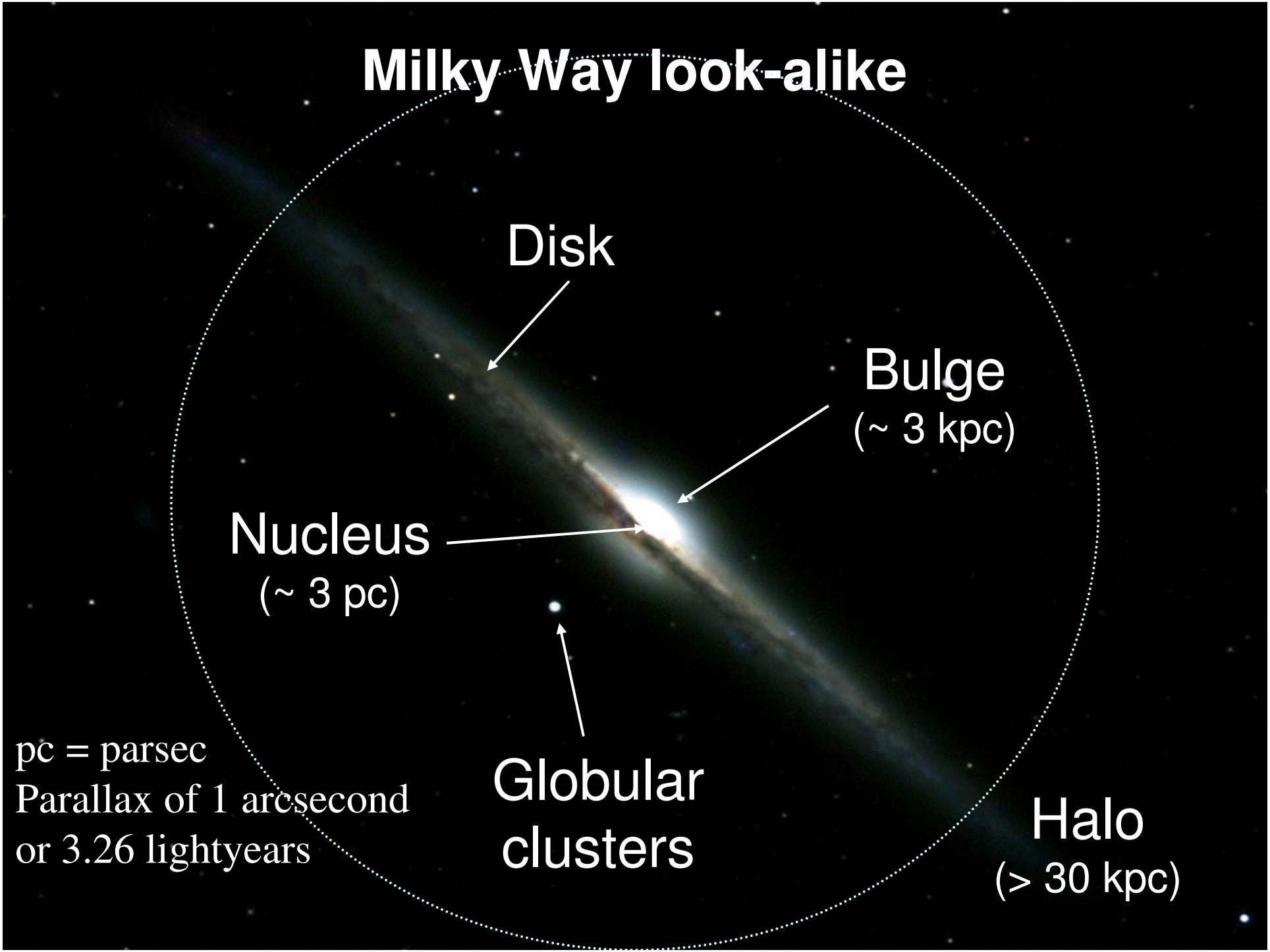
The Milky Way



Our view is severely obstructed by the dust in the disk and relatively little is known about the origin, history, and structure of our own Galaxy

(De Bruijn)


The origin of the Milky Way
Jupiter
Hercules
Juno
(De Bruijn)
Tintoretto (1575, National Gallery, London)


Milky Way look-alike
Disk
Bulge
(~ 3 kpc)
Nucleus
(~ 3 pc)
Globular
clusters
Halo
(> 30 kpc)
pc = parsec
Parallax of 1 arcsecond
or 3.26 lightyears

The need for Gaia

- Studies of the Galaxy require:
  - representative, complete sample of stars down to faint magnitudes
  - accurate distances and motions, combined with temperatures, metallicities, surface gravities, reddenings, … (masses, ages, …)
- Astrometry:
  - 'Hipparcos', but with improved aperture and detector (efficiency, bandpass, multiplexing), plus full control of all error sources (aberrations, attitude, basic-angle stability, …)
- Photometry:
  - astrophysical diagnostics
  - astrometric chromaticity correction
- Spectroscopy:
  - third component of space motion (perspective acceleration for nearby stars, binaries, chemistry, stellar rotation, …)

Why Gaia needs to fly …

- Astrometry from the ground suffers from:
  - Atmospheric refraction
  - Atmospheric scintillation
  - Mechanical telescope flexures
  - Thermal telescope flexures
  - Earth rotation, nutation, polar motion
  - Horizon

Gaia's key science driver

Stellar motions can be predicted into the future and calculated for times in the past when all 6 phase-space coordinates (3 positions, 3 velocities) are known for each star

The evolutionary history of the Galaxy is recorded mainly in the halo, where incoming galaxies got stripped by our Galaxy and incorporated

In such processes, stars got spread over the whole sky but their energy and (angular) momenta were conserved. Thus, it is possible to work out, even now, which stars belong to which merger and to reconstruct the accretion history of the halo

(De Bruijn)

Exo-Planets: Expected Discoveries

- Astrometric survey:
  - monitoring of hundreds of thousands of FGK stars to ~200 pc
  - detection limit Period < 10 years and about Jupiter size
  - complete census of all stellar types, P = 2–9 years
  - masses,
  - multiple systems measurable, giving relative inclinations

- Results expected:
  - 10–20,000 exo-planets (~10 per day)
  - displacement for 47 UMa = 360 μas
  - orbits for ~5000 systems
  - masses down to 10 M_{Earth} to 10 pc

- Photometric transits: ~5000?

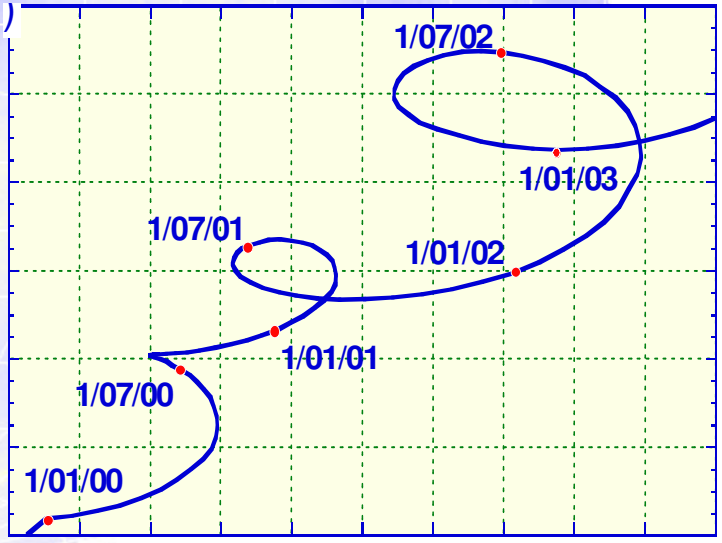
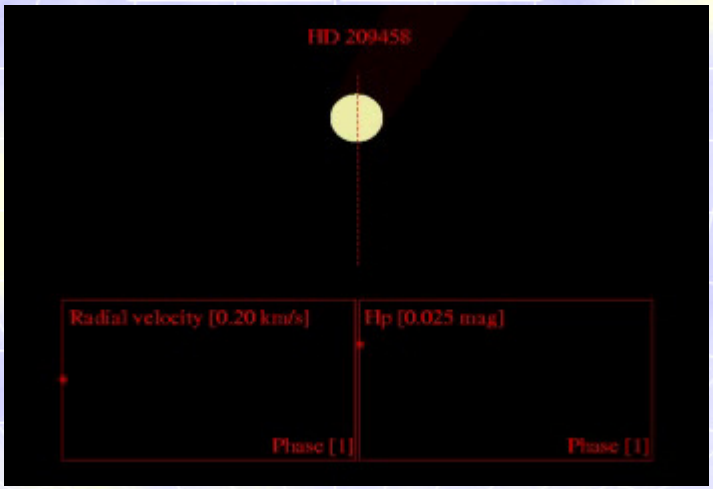


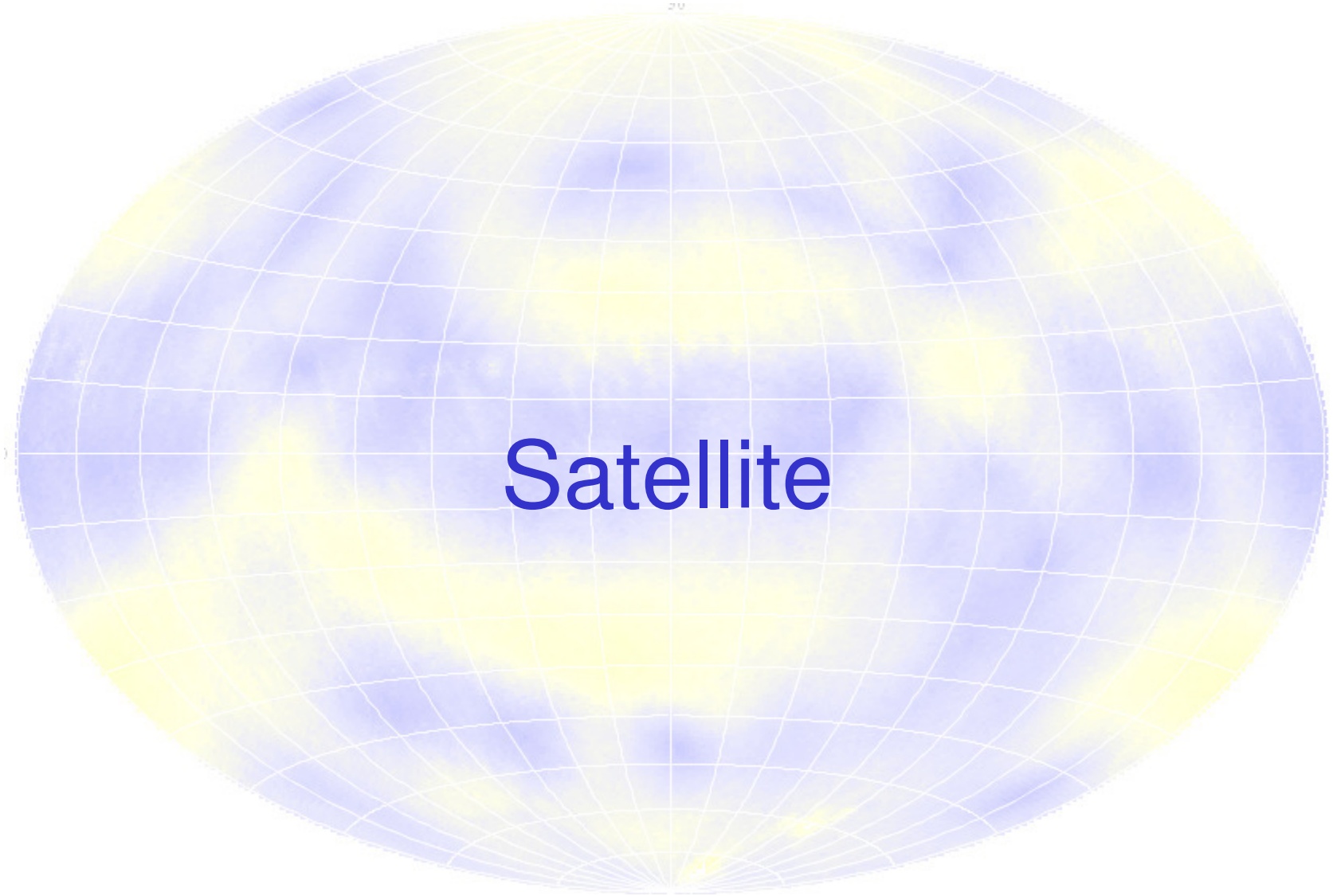
Figure courtesy François Mignard

Detecting planets …

Using :

- Astrometry
- Photometry
- Radial Velocity



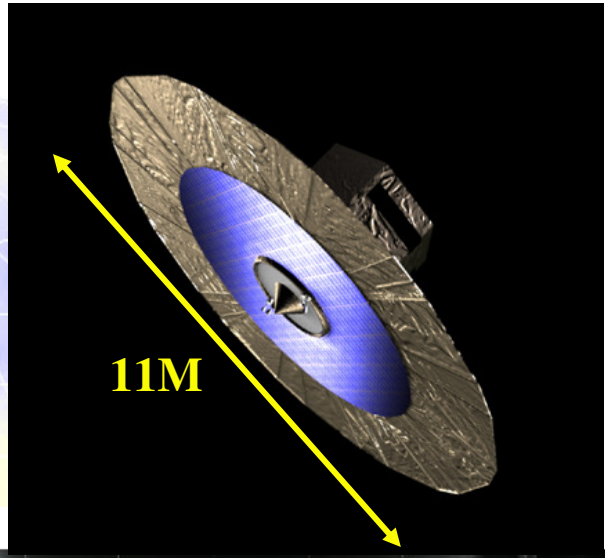


Satellite

Satellite and Mission

- Mission:
  - Stereoscopic Census of Galaxy
  - μarcsec Astrometry G<20 (10^9 sources)
  - Radial Velocities G<16
  - Photometry millimag G < 20
- Catalogue due 2020
- Status:
  - ESA Corner Stone 6
    - ESA provide the hardware and launch

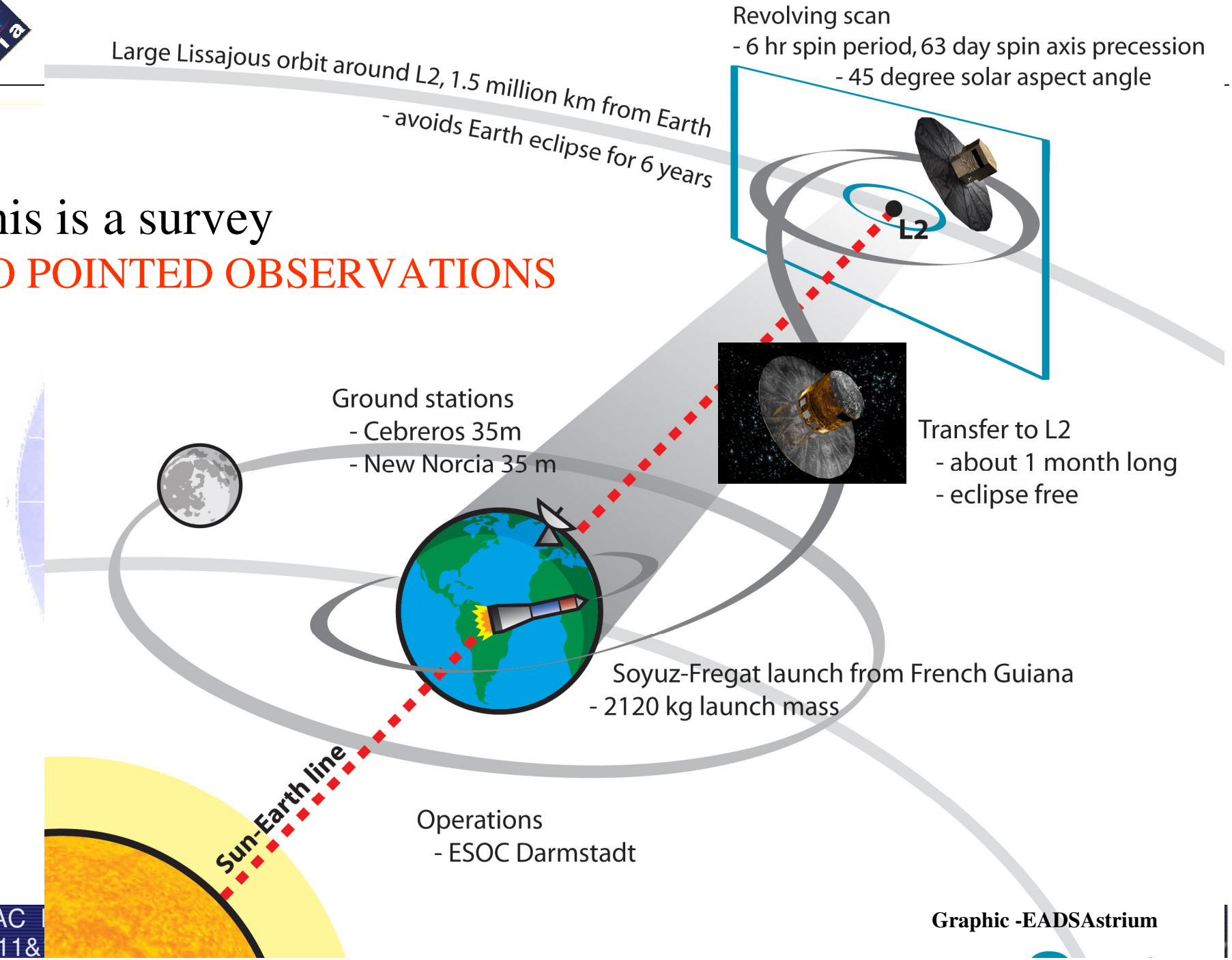
    -Mass: 2120 kg (payload 743 kg)
    Power: 1631 W (payload 815 W)
  - Launch: 2012 to L2 (1.5 Million Kms)
  - Satellite In development /test (EADS/Astrium)
    - Toros complete
    - Many mirrors, most CCDs



Launch August 2012 Soyuz from Kourou

This is a survey

NO POINTED OBSERVATIONS

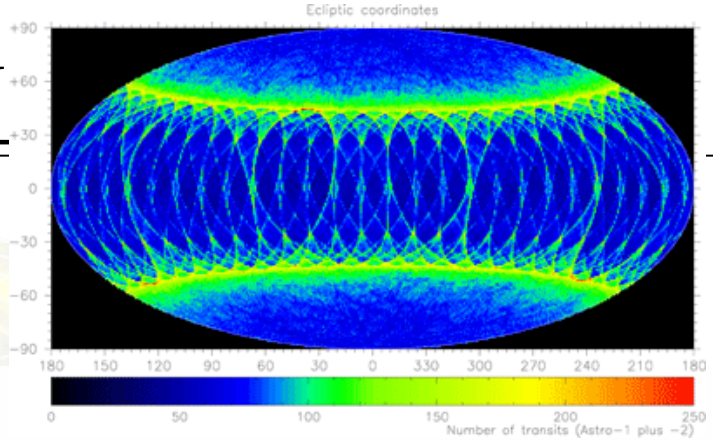


Graphic -EADSAstrium

Lissajous Scanning I

Full sky 3 fold every six months

5 year coverage →

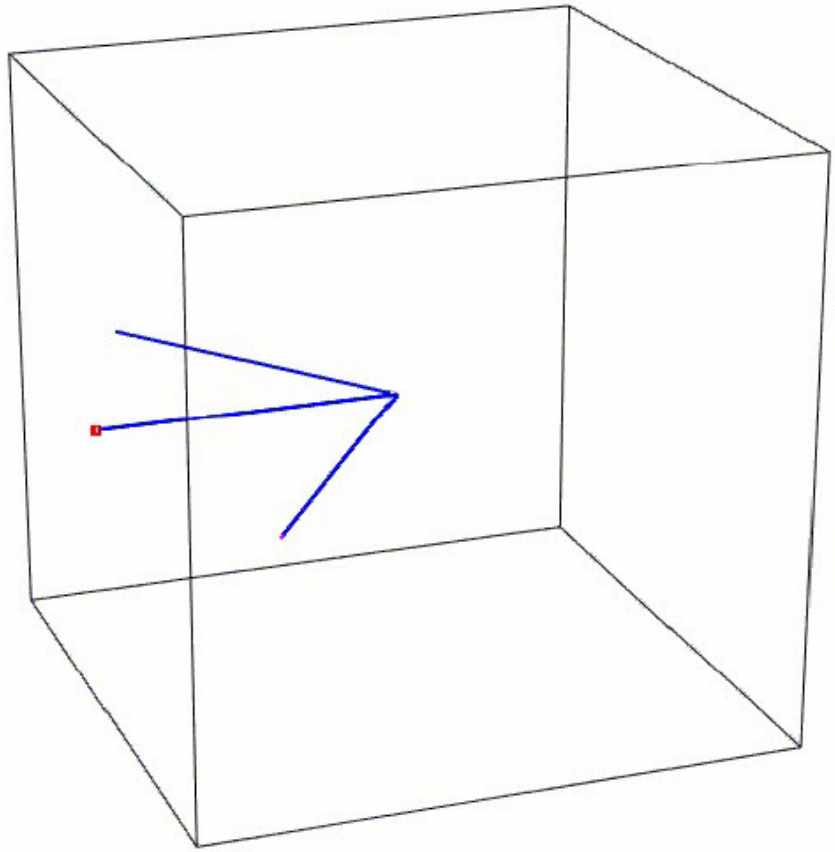


2 month animation 1 FOV

Astro (green) 4 complete circles per day

Sun (red)

Spin Axis (Purple) 50 Degrees away from sun – looping.

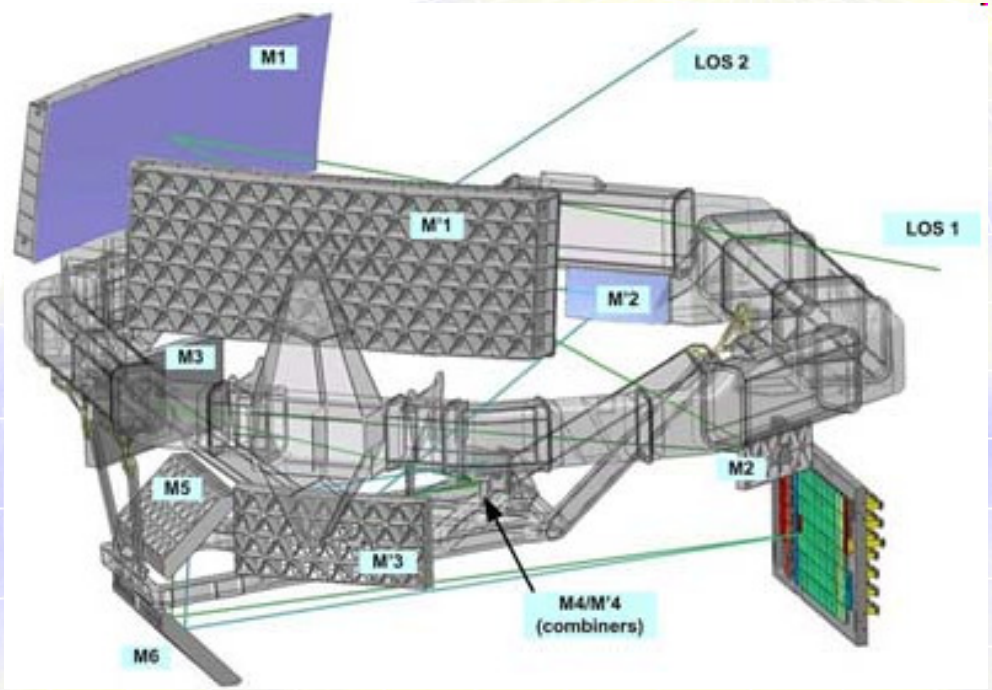


Attitude, velocity, spin rate

Determination of micorarcsecond astrometric positions requires accurate satellite attitude

- Instantaneous rotation of Gaia needed to about 20 microarcseconds
  - Spacecraft and optics diameter of 3 meters
  - relative location of any two parts on its outer periphery to about 1.5 nanometers with respect to inertial outer space
- Also require Gaia's three dimensional velocity to ~ 1 millimeter per second
- Spin rate < mas/s
  - Focal plane scale ~5 arcsec /mm

Mirrors



2 LOS - 6 reflectors - 1 Focal Plane

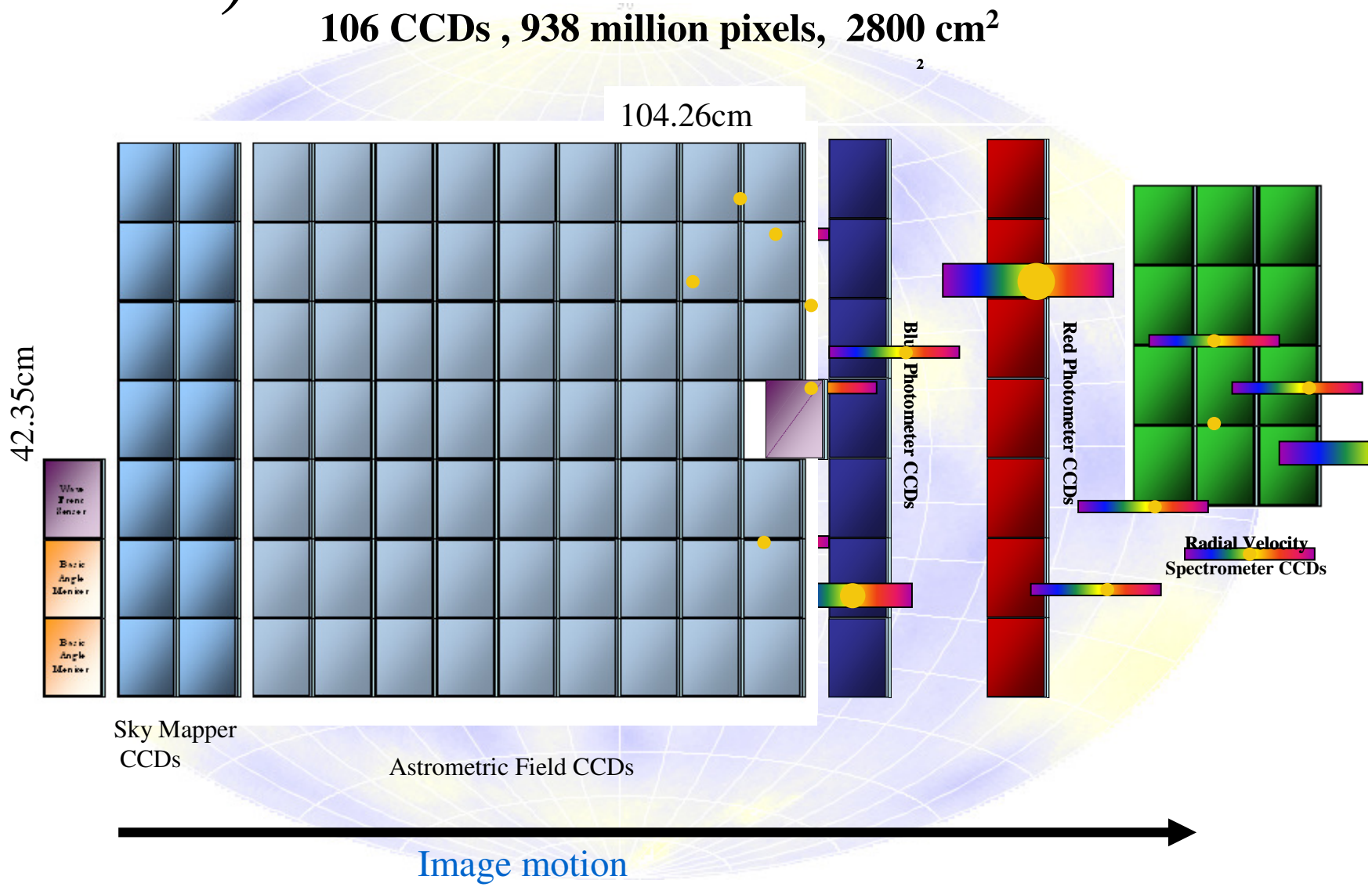
M5,6 common

Focal Length 35 M

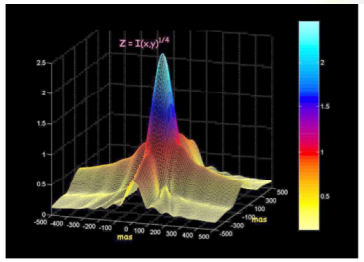
Geometric Imperfections Calibrated in AGIS

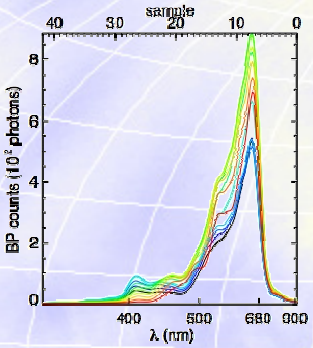
- Only M2 movable (Wave Front Sensors)
- To be adjusted in Commissioning

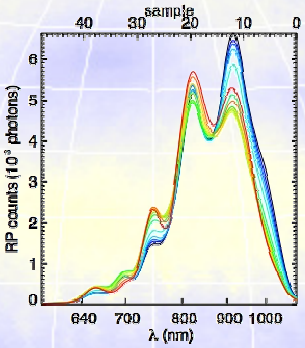
Giga Pixel Focal Plane (TDI mode)

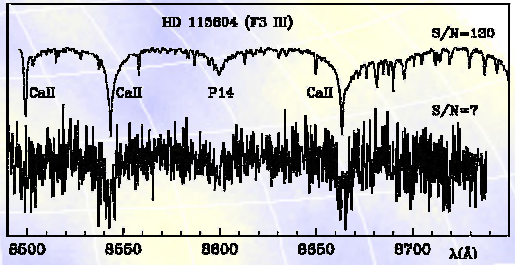


Gaia CCD principles



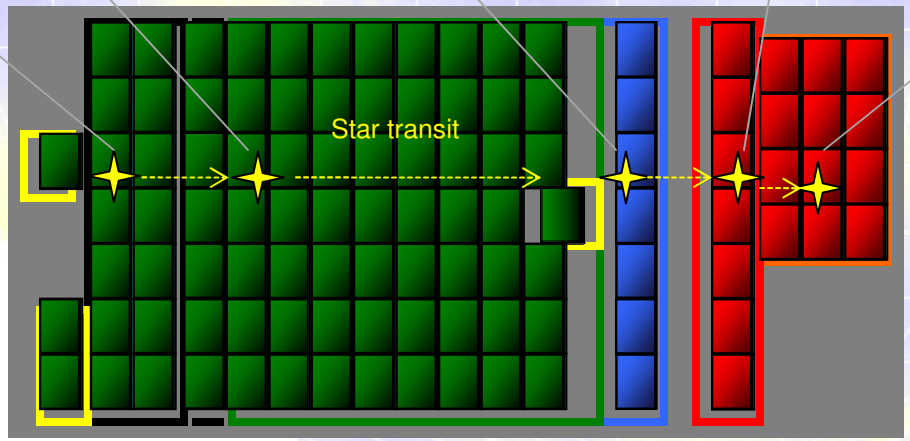






PSF PSF/LSF Low resolution spectra (R ~ 20-90) High resolution spectra (R ~ 11500)

S/N = 7 (single measurement)
S/N = 130 (summed over mission)



--------- IM + VPU
--------- IM + VPU
--------- IM + VPU
--------- IM + VPU
--------- IM + VPU
--------- IM + VPU
--------- IM + VPU

PSF = Point Spread Function
LSF = Line Spread Function
IM = Interconnection Module
VPU = Video Processing Unit

Operations

- CCDs operated in Time Delayed Integration (TDI) mode
  - Sources tracked across Focal Plane – only data around the source (windowing) read out and kept.
  - 'Virtual Objects' injected for background determination
  - Each type of data requires special processing and there are Gates and different sized windows ..
  - Never see Gaia "pictures" as such
- Orbit corrections required @L2 each ~28 days (MOC)
  - Handled by MOC
- MOC (Flight Dynamics) provide actual orbit
  - DPAC provide Gaia observations
  - MOC provide improved orbit

Physical Challenges

- Need the astrometric centroid of the CCD image determined to an accuracy of 1% of the pixel size!
  - There will be 10^12 images ~100TB downlink need to handle ~1PB
    - At 1 millisecond each that's ~ 30 years
  - Processing estimate remains ~ 10^20 FLOP
- Gaia attitude is required at the nanometre level
- Path of light through instrument needed to similar accuracy:
  - System must be extremely stable
- Require Gaia's three dimensional velocity to ~ 1 millimetre per second
- Spin rate < mas/s
  - Focal plane scale ~5 arcsec /mm
- Attitude and Geometric calibration can only by done using Gaia's own observational data. (AGIS)

Light bending

Exaggerated : but Jupiter even at 90 degrees has significant effect

General Relativity

- Gaia microarcsecond level measurements must consider bent space-time
- Require high-precision general-relativistic theory of light propagation
- Gaia truly observes ‘events’
  - Time is very important
  - On board clock must be very accurate !
    - Will also require calibration
- Relativity and Clock Calibration under Sergei Klioner in Dresden

Photometry

- G magnitudes for all stars to few millimag
  - Calibration challenge
- As shown in the cartoon prism dispersed spectra are collected using red and blue enhanced CCDs (low resolution)
  - Extracting the signal will be hard
  - Calibration
  - Flux scale (all types of objects)
    - From data, using different gates
    - Possibly specifically targeted observations
    - Ground based observations under investigation

Radial Velocity Spectra

- 250 million stars
  - ~40 times each
  - 10^10 spectra
    - At 1 millisecond each only six months !!!!
    - More data and more complex
- Challenges
  - Disentangling
    - all light in FOV dispersed on RVS
  - Calibration – like astrometry (SGIS)

Downlink

- Using Cebreros (35M)
  - 3-8Mb/s downlink
    - depends on encoding
    - which depends on weather !
  - ~ 30GB/day -> ~100TB
- occasionally New Norcia
  - during Galactic plane scans
  - data accumulated onboard
  - downlinked later
- Exact hours booked by MOC using predicted onboard memory occupancy from SOC (Science Schedule)
- MOC should also provide station delays (for time correlation)

Data Flow challenges

- Daily data flow not a problem
  - ~50GB per day over standard internet
- Using Aspera/ FASP for now
- The MainDB updates may be a problem
  - 100 Mega bit line => 1 TB in one day
    - Have done this from Marenostrum for simulation data
  - 1 Giga bit line => 10TB in one day
    - ESAC now has gigabit – have not saturated it yet
    - Ok initially but 100TB means 10 days
  - Cost effectiveness of faster lines ?
    - Should we ship ?
    - Should we put it all in the cloud ?
    - Will decide later ..

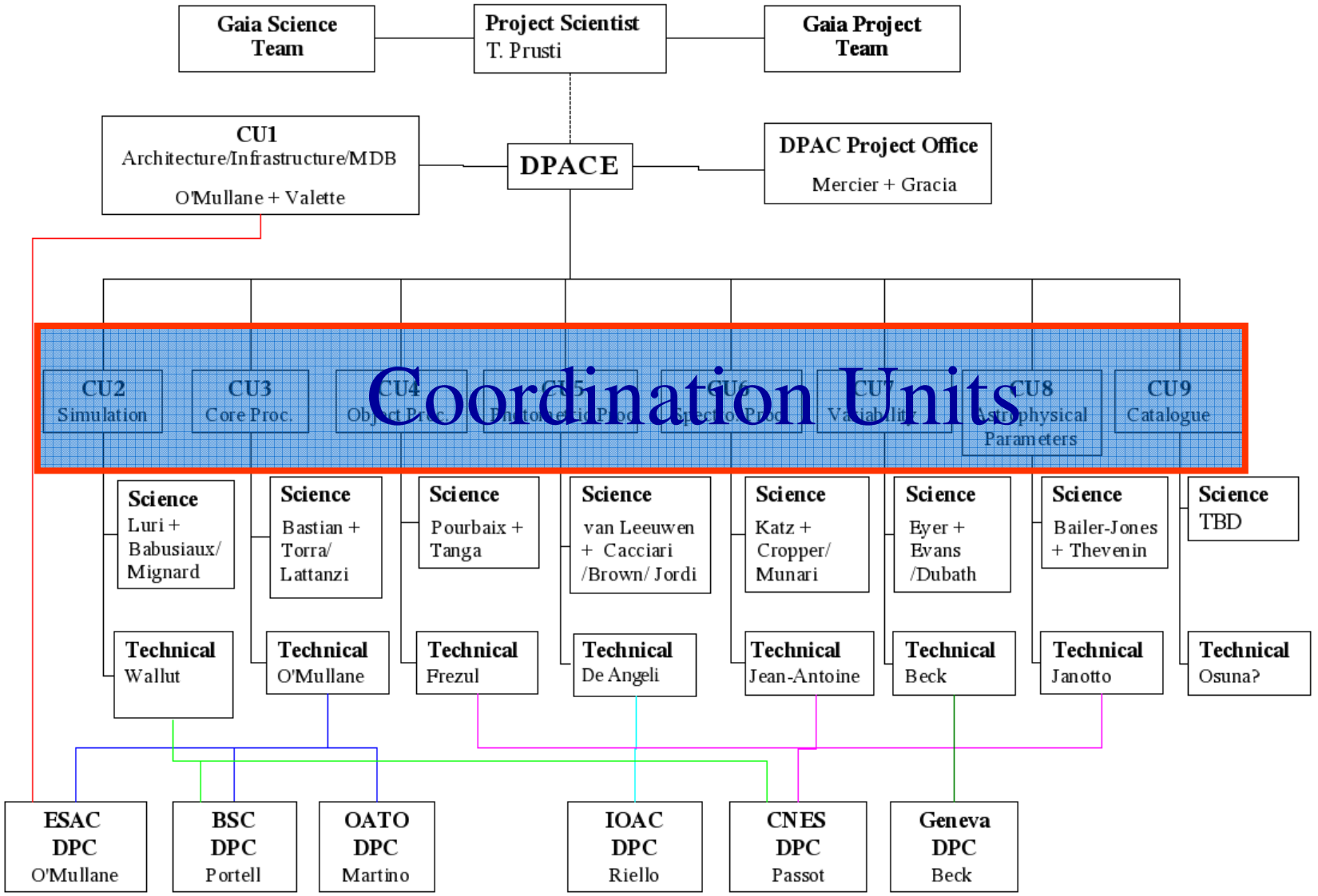
Organisation of Gaia processing and Data Processing and Analysis Consortium (DPAC)

DPAC

Data Processing and Analysis Consortium

- Formed to answer the Announcement of Opportunity (AO) for Gaia data processing
- Involves large number of European institutes and observatories (>400 people, >20 institutes)
- The science community must fund the majority of the Gaia processing
- Lead by Executive (leaders of Coordination Units) and Project office for management and planning (currently 3 people).

DPAC: Structure CU

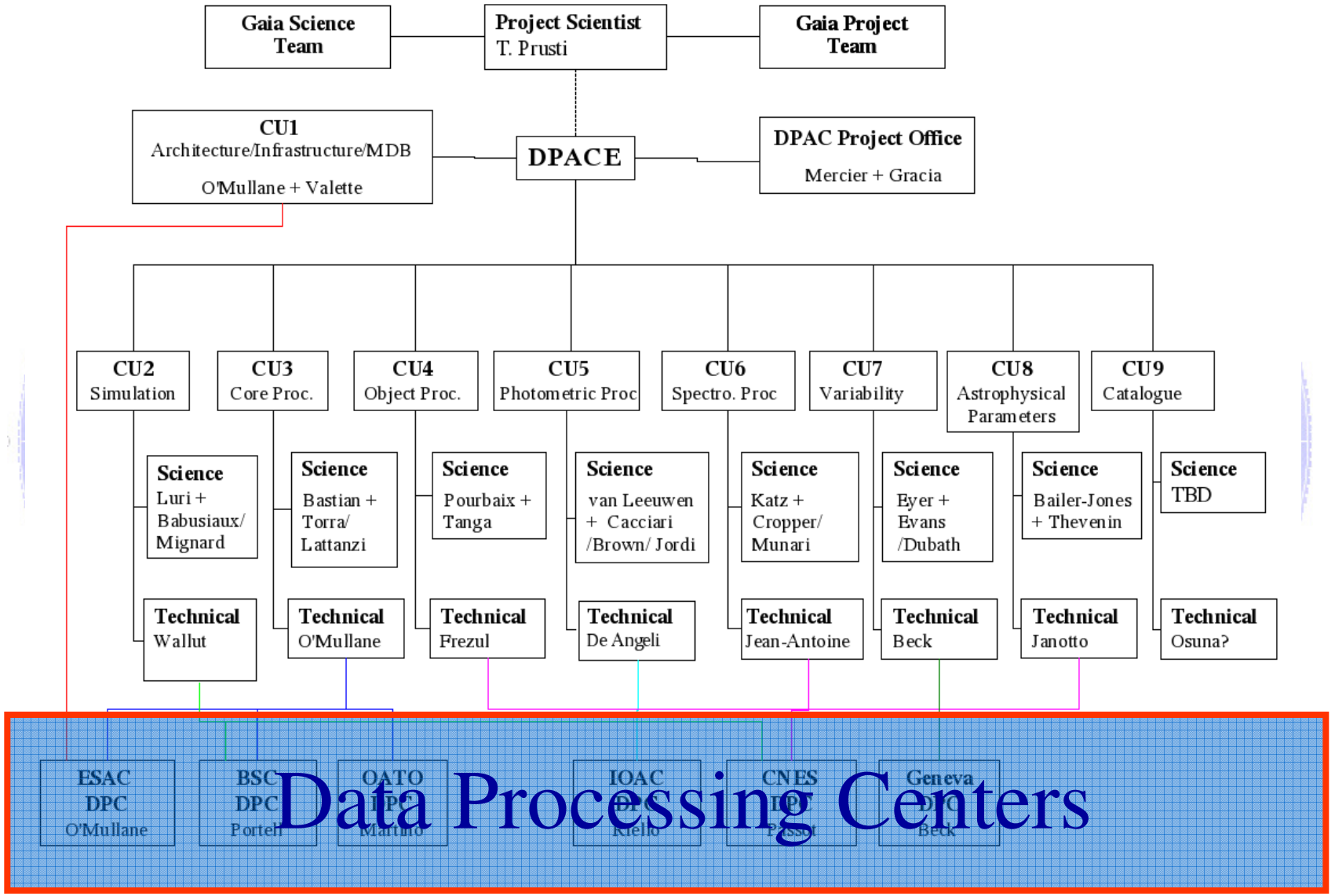


DPAC coordination units

- CU1: System Architecture
- CU2: Data Simulations
- CU3: Core Processing
- CU4: Object Processing
- CU5: Photometric Processing
- CU6: Spectroscopic Processing
- CU7: Variability Processing
- CU8: Astrophysical Parameters
- CU9: Catalogue Access

ESA areas of Contribution

DPAC : DPCs


Gaia Science Team
Project Scientist
T. Prusti
Gaia Project Team
CU1
Architecture/Infrastructure/MDB
O'Mullane + Valette
DPACE
DPAC Project Office
Mercier + Gracia
CU2
Simulation
CU3
Core Proc.
CU4
Object Proc.
CU5
Photometric Proc
CU6
Spectro. Proc
CU7
Variability
CU8
Astrophysical Parameters
CU9
Catalogue
Science
Luri + Babusiaux/ Mignard
Science
Bastian + Torra/ Lattanzi
Science
Pourbaix + Tanga
Science
van Leeuwen + Cacciari /Brown/ Jordi
Science
Katz + Cropper/ Munari
Science
Eyer + Evans /Dubath
Science
Bailer-Jones + Thevenin
Science
TBD
Technical
Wallut
Technical
O'Mullane
Technical
Frezul
Technical
De Angeli
Technical
Jean-Antoine
Technical
Beck
Technical
Janotto
Technical
Osuna?
ESAC DPC O'Mullane
BSC DPC Portell
OATO DPC Martino
IOAC DPC Riello
CNES DPC Pasot
Geneva DPC Beck
Data Processing Centers

DPAC : DPCs

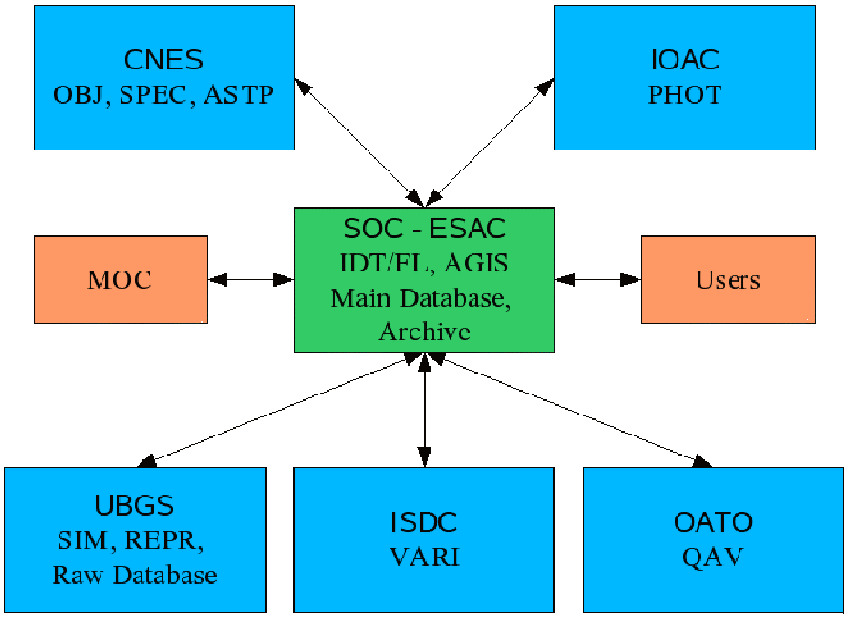
DPCs underpin and support CUs

- Software support and production
- Operation of processing system(s)

- ESAC (CU1,3) Madrid
- BPC (CU2,3) Barcelona
- CNES (CU4,6,8) Toulouse
- ISDC (CU7) Geneva
- IoA (CU5) Cambridge
- OATO (CU3) Torino

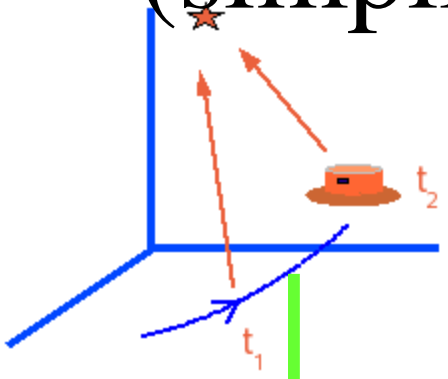
Architecture

- Highly distributed
  - Multiple independent DPCs and CUs
- Want/need decouple
  - Reduce
    - dependencies
    - risk
- Hub and spokes
  - Max flexibility for CUs and DPCs
  - Minimum ICDs



- MDB Version each 6 Months
- ~20 TB inc. each version
  - Hence ~200TB in final
- GTS push data to other DPCs
  - Governed by single ICD

Processing Overview (simplified)



Catalogue

Many iterations

Initial Data Treatment
Turn CCD transits into source observations on sky
Should be linear transform **CU3**

Astrometric Treatment
Fix geometrical calibration
Adjust Attitude
Fix source positions **CU3/ SOC**

Photometry Treatment
Calibrate flux scale
give magnitudes **CU5**

Spectral Treatment
Calibrate and disentangle
provide s spectra **CU6**

Solar System CU4

Variability CU7

Astrophysical Parameters CU8

Non Single Systems CU4

Tools

- Parameter Database provides all constants for the mission online, also as Java, Fortran, Latex (Lammers, DeBruijn, Joliet)
- MDB Dictionary provides unique distributed data modeling (Hernandez).
  - Generation of ICD, Java Classes, DB Schemas
- All code in Java (only one exception)
  - for portability have to run till 2020
  - Maintainability, testability etc.. JUnit, Hudson
  - Easier to write CORRECT code in higher level language
    - Fewer Core Dumps (Ok, ok so we get NPE)
- GaiaTools provide routines and libraries for all
  - ApacheCommon Math, Log4J etc ..
  - JBOSS for IPC
  - No special MPI or Grid libraries need

Databases and Security

- CUs and DPC relatively independent. Can see
  - Oracle,HBASE, MySql, Postgress, Intersystems Cache
- Here at ESAC Oracle since 2005
  - Not impressed with support
  - Moving some parts to Intersystems Cache
    - Excellent support and faster write times
  - May move entirely
- There are no “users” in this system.
- JDBC provides sufficient network access to data (thin DAL on top)
- So no need for
  - complex security schemes, Access control lists
  - Encryption etc …

FOCUS:
Asrtrometric Global
Iterative Solution (AGIS

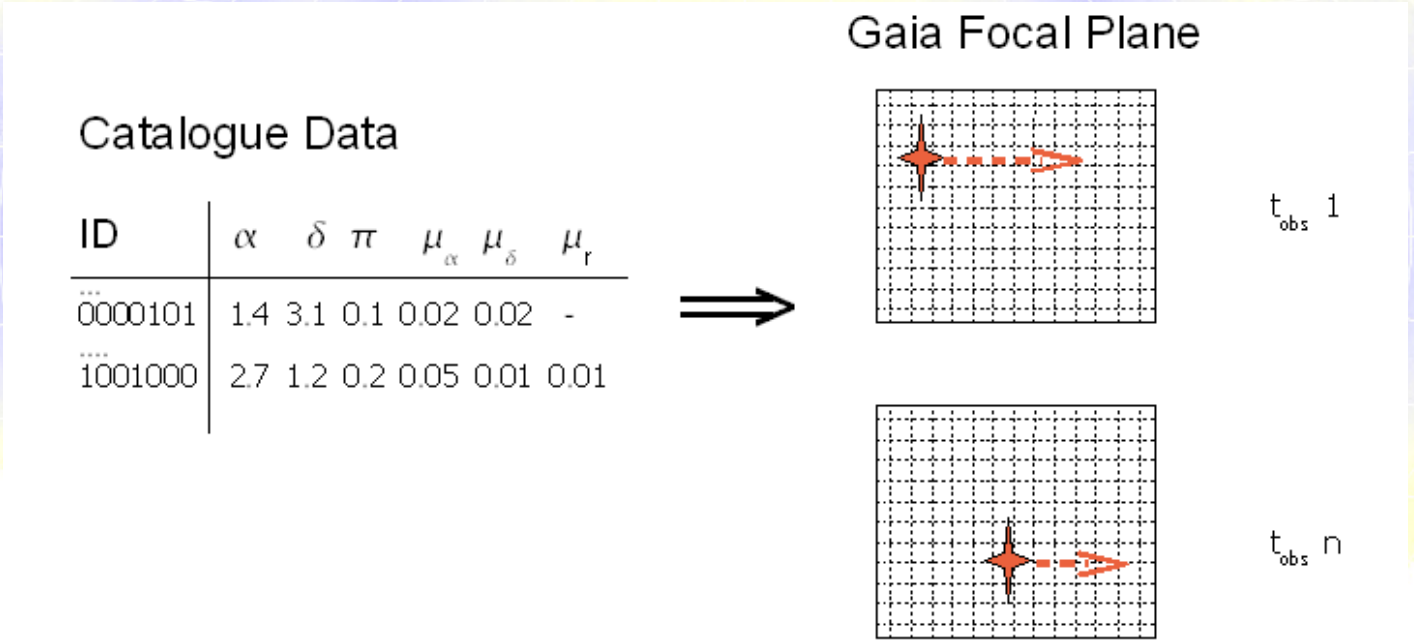
Processing Goal

Basic objective of processing is to answer:

Which catalogue parameters best predict source observations on Gaia's focal plane?

Catalogue Data

ID	α	δ	π	μ_α	μ_δ	μ_r
...						
0000101	1.4	3.1	0.1	0.02	0.02	-
....						
1001000	2.7	1.2	0.2	0.05	0.01	0.01



(idea Luri)

The best astrometric answer requires Photometry and Spectroscopy !

AGIS

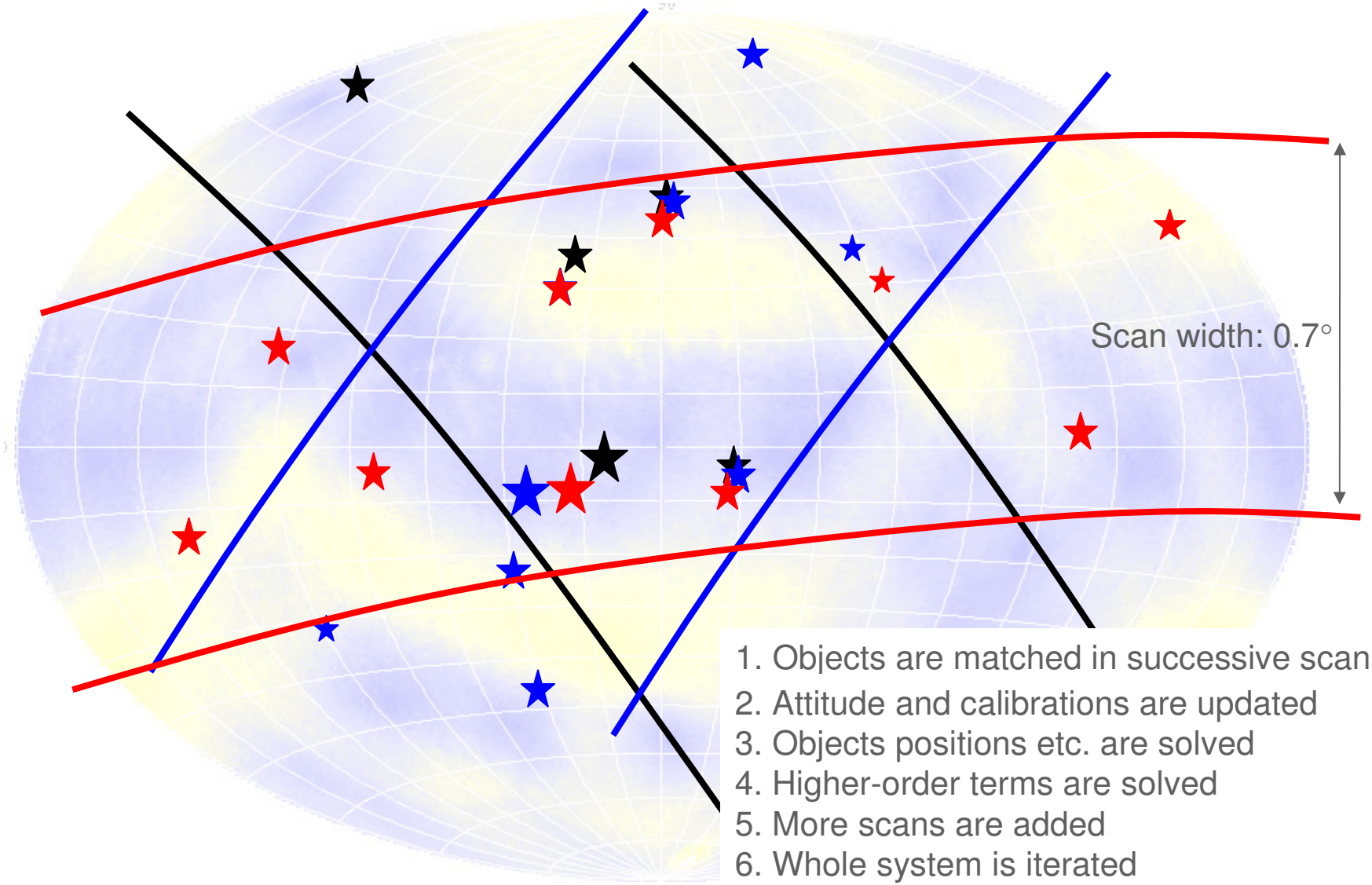
- Astrometric Global Iterative Solution (Lindegren,Lammers)
- Provide rigid independent reference frame for Gaia Observations
  - rotate to ICRS using quasars
- perhaps about 10% of all the processing
  - only deals with 10-50% of sources (well behaved stars to make the grid)
- Block iterative solution
  - using Gauss-Seidel "preconditioner" with simple iterations
  - Moved to Conjugate Gradient

The centroid of a star image is modelled as

$$\begin{pmatrix}\text{Observed}\\ \text{location}\end{pmatrix}=\begin{pmatrix}\text{Global}\\ \text{ref. frame}\end{pmatrix}+\begin{pmatrix}\text{Source}\\ \text{position}\end{pmatrix}+\begin{pmatrix}\text{instrument}\\ \text{Attitude}\end{pmatrix}+\begin{pmatrix}\text{CCD/pixel}\\ \text{offset}\end{pmatrix}+\text{noise}$$

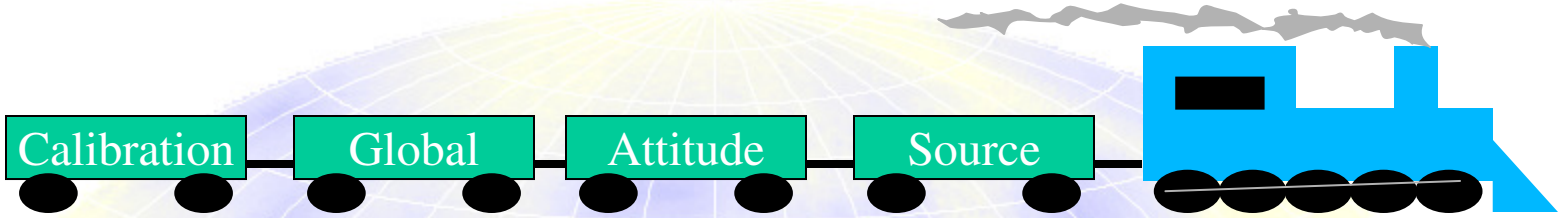
Symbolically: $O = G + S + A + C + n$

AGIS: cartoon explanation



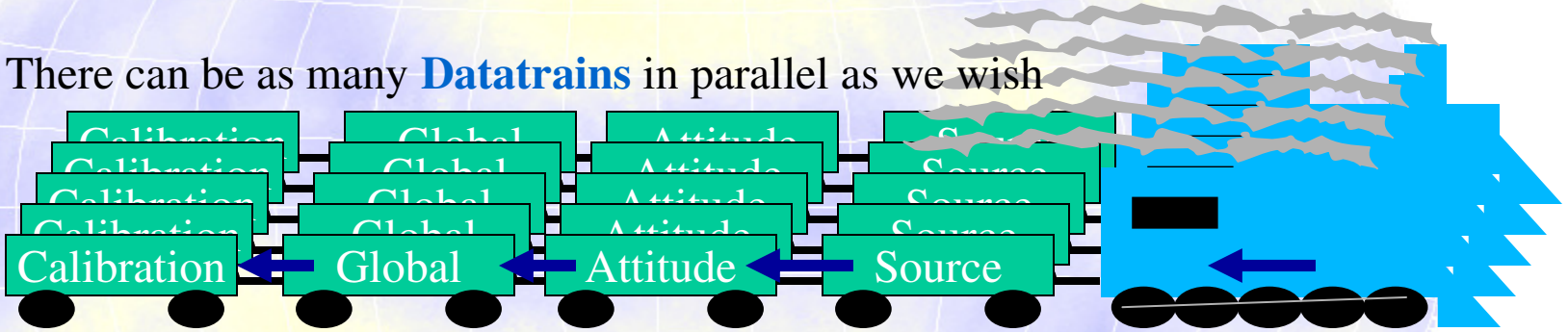
1. Objects are matched in successive scans
2. Attitude and calibrations are updated
3. Objects positions etc. are solved
4. Higher-order terms are solved
5. More scans are added
6. Whole system is iterated

AGIS Architecture



Datatrains drive through AGIS Database passing observations to algorithms.

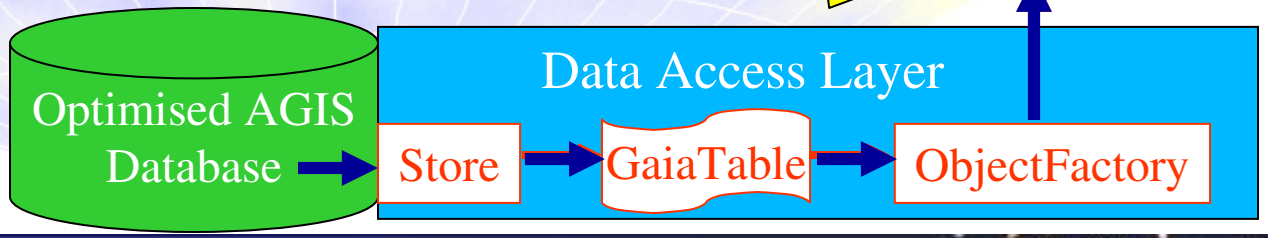
There can be as many **Datatrains** in parallel as we wish



ElemetaryTakers

AstroElementary

NB: Algorithm does not access data directly

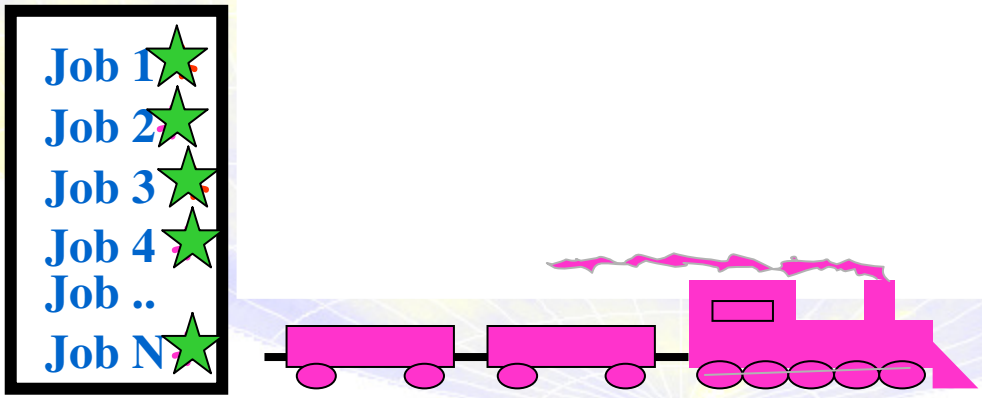


Scheduling

- Very simple ..
  - Keep all machines busy all the time!
    - Busy = CPU ~90%
  - Post jobs on whiteboard

    Trains/Workers mark Jobs – and do them

    mark finished – repeat until all done



Some important architecture points

From the outset we try in Gaia to:

- Keep it just as simple as possible
- Isolate algorithms form Data
  - Already tried to virtualize algorithms
- Let Data drive the system (DataTrain)
  - Algorithms mostly not allowed to 'query'
  - Specific data access patterns Data orgainised accordingly
  - Similar to Ferris-Wheel idea (Szalay) but no hopping on/off!
- Access any piece of data on disk exactly once
  - preload some data on each node
  - E.g. 5 years attitude quaternions fit in 150-250Mb
- Be distributed
  - try to avoid large memory processes
  - But then gain in some cases it makes sense ..

Notes on AGIS Implementation

- Highly distributed usually running on >40 nodes has run on >100 (1400 threads).
- Only uses Java no special MPI libraries needed – new languages come with almost all you need.
  - Hard part is breaking problem in distributable parts – no language really helps with that.
- Truly portable – can run on laptops desktops, clusters and even Amazon cloud.

Virtualization

- Started looking at virtualization ~2007
- Seemed ideal for the multiple test setups needed (VMWARE)
- Agreed Cloud experiment 2009 (with Parsons)
- Had to be convincing
- RUN AGIS obvious choice
  - Already 4 years in development
  - In Java
    - so its portable right!

AGIS on the cloud

- Took ~20 days to get running (Parsons,Olias).
  - Used 64Bit EC2 images Large, Extra Large and High CPU Large
  - Main problem DB config (But oracle image available)
    - Oracle ASM Image based on Oracle Database 11g Release 1 Enterprise Edition - 64 Bit (Large instance) -ami-7ecb2f17
  - Also found scalability problem in our code (never had one hundred nodes before)
    - only 4 lines of code to change
- It ran at similar performance to our in house cheap cluster.
  - E2C indeed is no super computer
- AGIS image was straightforward to construct but was time consuming – better get it correct !
  - Self configuring Image based on Ubuntu 8.04 LTS Hardy Server 64-Bit (Large, Extra Large and High CPU Large Instances) - ami-e257b08b
- Availability of large number of nodes very interesting
  - not affordable in house today.

NEXT 1000 nodes

Cost effectiveness of E2C

- AGIS runs intermittently with growing Data volume.
  - Unless we end up doing much more and running 100%
- Estimate 2015 ~1.1MEuro (machine) + 1Meuro (energy bill less ?) = ~2Meuro
  - In fact staggered spending for machines
  - buy machines as data volume increase
- Estimate on Amazon at today prices with 10 intermittent runs ~400Keuro
  - Possibility to use more nodes and finish faster !
- Reckon you still need in house machine to avoid wasting time testing on E2C
- Old nut, Vendor lock-in ? Need standards

Conclusion

- Complex algorithms
- Distributed processing
  - Six European wide DPCs
  - Local algorithms must be distributed
- Large quantity of data
  - All data accessed repeatedly
- No users – no security
- Naïve approaches have proved impossibly slow
- Requires much Thought and Work.
- Seems Cloud Compatible (at least in parts)
  - Especially for testing and early runs

Questions ??

Ariane V188 carrying Herschel and Planck (May 14 2009)