ICT@Inaf 2014
- Pula (Ca) -
16-19 September 2014

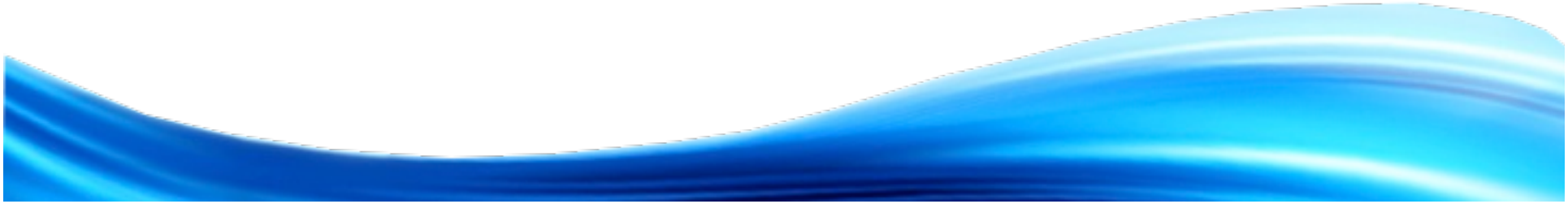
Solving a Very Large-Scale Sparse Linear System with a Parallel Algorithm in the Gaia Mission

Marilena Bandieramonte, U. Becciani, E. Sciacca, A. Vecchiato,
B. Bucciarelli, M. Lattanzi

Email: marilena.bandieramonte@oact.inaf.it

Outline

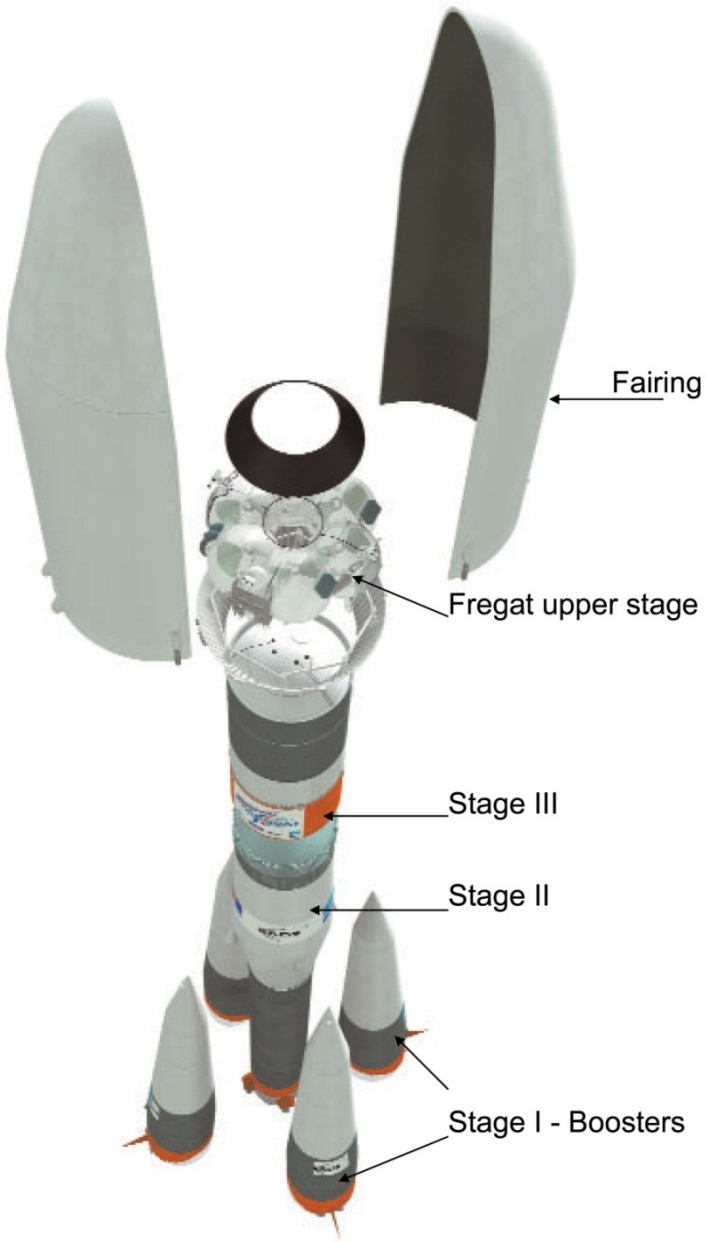
- ✓ **Gaia**: A stereoscopic census of our Galaxy
- ✓ **GSR**: The problem of the Global Sphere Reconstruction in GAIA
- ✓ **AVU-GSR**: Parallelization strategy
- ✓ **Case studies:** Results and predictions



Gaia
A Stereoscopic Census of our Galaxy

http://www.cosmos.esa.int/web/gaia

Satellite and System



- **ESA-only mission**
- **Launch**: 20 December 2013
- **Launcher**: Soyuz–Fregat from French Guiana
- **Orbit**: L2 Lissajous orbit
- **Ground stations**: Cebreros, New Norcia + Malargüe
- **Lifetime**: 5 years (1 year potential extension)
- **Downlink rate:** 4 - 8 Mbps

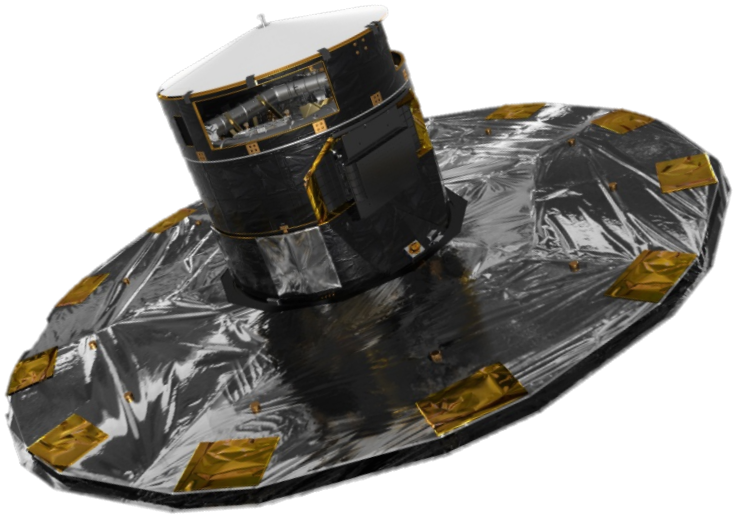
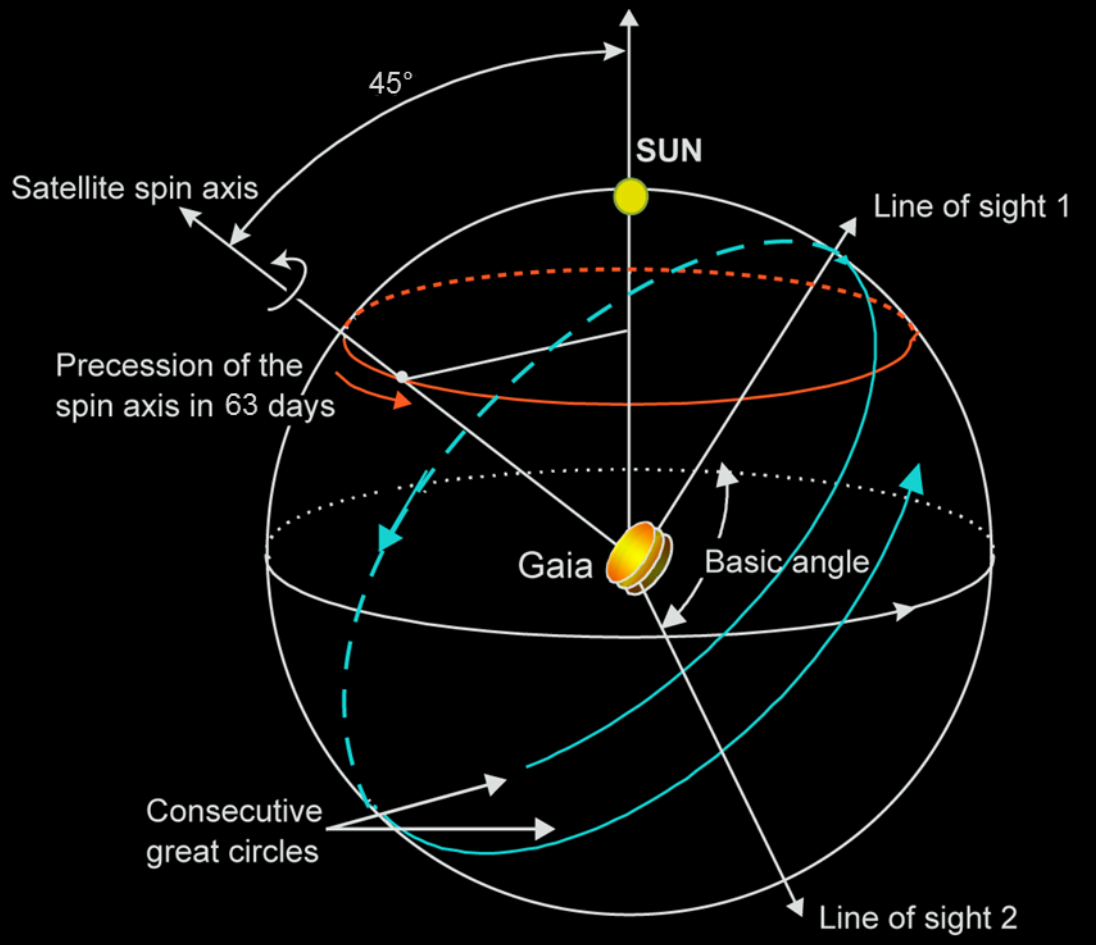


Figure Soyuz: Arianespace; figure Gaia: ESA - C. Carreau

Sky-Scanning Principle



Spin axis	45^{o} to Sun
Scan rate:	60 arcsec s^{-1}
Spin period:	6 hours

Figure ESA

Gaia: Design Considerations

- **Astrometry** (G < 20 mag):
  - **completeness to 20 mag** (on-board detection) ⇒ 10^9 stars
  - **accuracy**: **26 μarcsec at G=15 mag** (Hipparcos (1989-1993): 1 milliarcsec at 9 mag)
  - scanning satellite, two viewing directions
    ⇒ global accuracy, with optimal use of observing time
  - principle: global astrometric reduction (as for Hipparcos)

- **Photometry** (G < 20 mag):
  - astrophysical diagnostics (low-dispersion photometry) + chromaticity
    ΔTeff ~ 100 K, log g, [Fe/H] to 0.2 dex, extinction (at G=15 mag)

- **Radial velocity** (GRVS < 16 mag):
  - accuracy: 15 km s-1 at GRVS=16 mag
  - application:
    third component of space motion, perspective acceleration
    dynamics, population studies, binaries
    spectra for GRVS < 12 mag: chemistry, rotation
  - principle: slitless spectroscopy in Ca triplet (845-872 nm) at R = 11,500

One Billion Stars will provide..

- **in our Galaxy …**
  - the distance and velocity distributions of all stellar populations
  - the spatial and dynamic structure of the disk and halo
  - its formation history
  - a detailed mapping of the Galactic dark-matter distribution
  - a rigorous framework for stellar-structure and evolution theories
  - a **large-scale survey of extra-solar planets** (~7,000)
  - a **large-scale survey of Solar-system bodies** (~250,000)

- **… and beyond**
  - definitive distance standards out to the LMC/SMC
  - rapid reaction alerts for supernovae and burst sources (~6,000)
  - quasar detection, redshifts, microlensing structure (~500,000)
  - fundamental quantities to unprecedented accuracy: γ to 2×10^{-6} (2×10^{-5} present)

GSR
The Global Astrometric Sphere Reconstruction

http://www.cosmos.esa.int/web/gaia

What GSR stands for

- The Gaia DPAC (Data Processing and Analysis Consortium) decided to pursue the task of the astrometric sphere reconstruction with two independent processes within the Gaia pipeline
- One is AGIS (Astrometric Global Iterative Solution), which is run at ESAC (Spain)
- The other one is GSR, which stands for Global Sphere Reconstruction (and Comparison) and is run at the Italian Data Processing Centre
- GSR's twofold goal is to provide Gaia's with an independent way of reconstructing the Global Astrometric Sphere, and to compare its solution with the AGIS one
  - independent relativistic astrometric model
  - independent solution method for the sphere reconstruction
  - GSR depends on AGIS for the selection of the primary stars
  - AGIS depends on GSR for the comparison

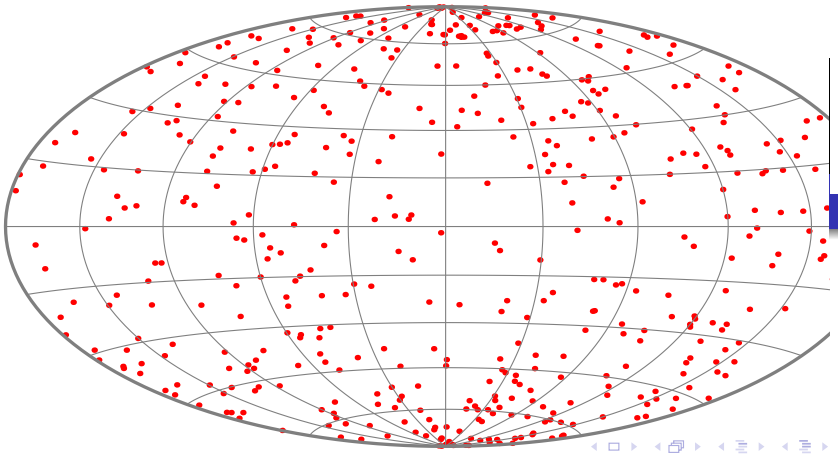
	AGIS	**GSR**
Astrometric model	GREM	RAMOD
Solution algorithm	Block-Iterative	Iterative (LSQR)
Programming language	Java	Java + C (MPI+OMP)
Primaries' selection	YES	NO
Comparison	NO	YES
# of objects	∼100 million	50 to 100 million

GSR

The problem of the Global Astrometric Sphere Reconstruction in Gaia

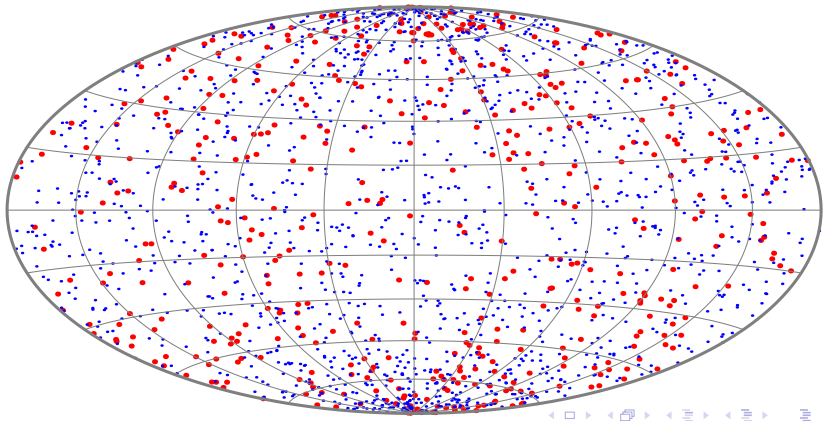
Primaries/Non-primaries

The Global Astrometric Sphere is first reconstructed with respect to a subset ($\sim 10^8$ out of $\sim 10^9$) of well-behaved stars called primaries.

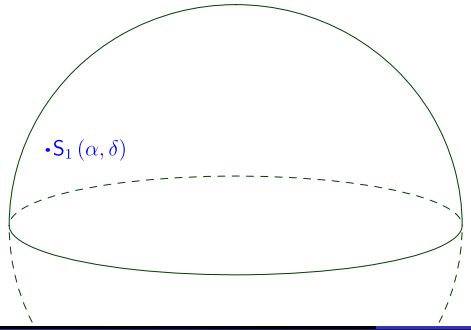


Primaries/Non-primaries

The reference frame materialized by the primaries is used by other pipeline processes to include the other stars into the Gaia sphere.



Principles of the sphere reconstruction
The ideal picture



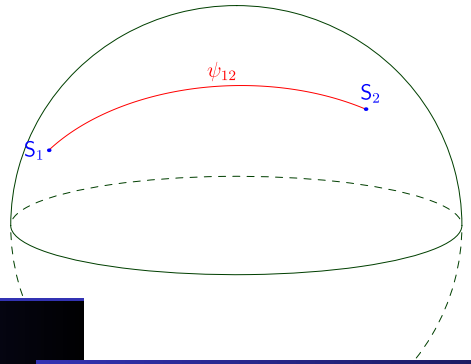
Create a "ge
network of mea

$N_* =$

$N_{\text{unk}} =$

$N_{\text{arcs}} =$

Principles of the sphere reconstruction
The ideal picture



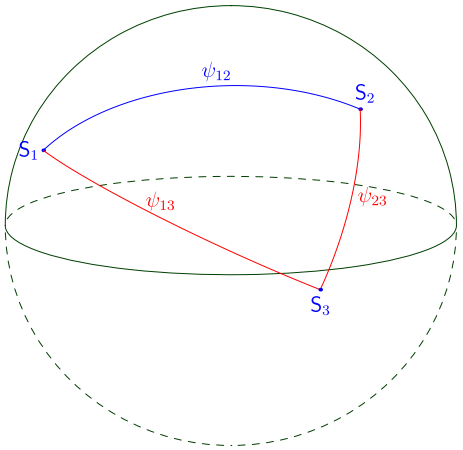
Create a "geodetic"
network of measurements

$N_* = 2$

$N_{\text{unk}} = 4$

$N_{\text{arcs}} = 1$

Principles of the sphere reconstruction
The ideal picture



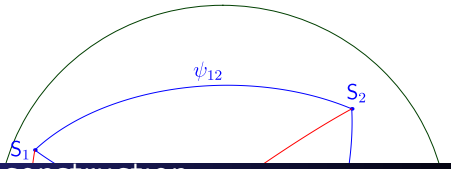
Create a "geodetic"
network of measurements

$N_* = 3$

$N_{\text{unk}} =$

$N_{\text{arcs}} =$

Principles of the sphere reconstruction
The ideal picture



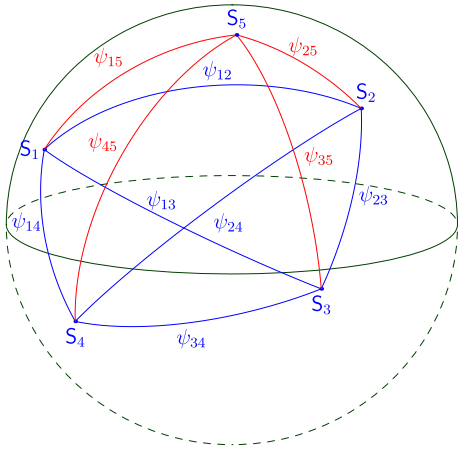
Create a "geodetic"
network of measurements

$N_* = 4$

$N_{\text{unk}} = 8$

$N_{\text{arcs}} = 6$

Principles of the sphere reconstruction
The ideal picture



Create a "geodetic"
network of measurements

$N_* = 5$

$N_{\text{unk}} = 10$

$N_{\text{arcs}} = 10$

Network closed!

Solve an Equation System

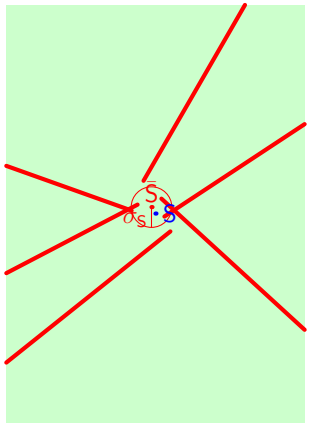
Principles of the sphere reconstruction
The (almost) real picture

Observational errors ⇒

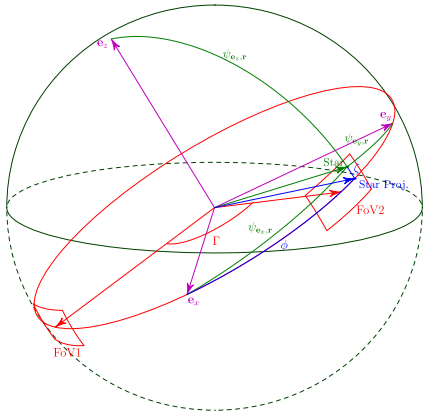
1. solution in the least-squares sense;
2. overdetermined system of equations.

$$N_{\mathrm{unk}} \sim N_* \simeq 10^8$$

$$N_{\mathrm{obs}} \sim 10^2 N_{\mathrm{unk}} \sim 10^{10}$$



Mathematical modeling: the Euclidean abscissa



- The Gaia basic observable is the abscissa ϕ between the x axis and one viewing direction

$$\cos\psi_{(\hat{a},\mathbf{r})} = \frac{\mathbf{e}_{\hat{a}} \cdot \mathbf{r}}{|\mathbf{r}|} \qquad (2)$$

$$\cos\phi = \frac{\cos\psi_{(\hat{x},\mathbf{r})}}{\sqrt{1-\cos^2\psi_{(\hat{z},\mathbf{r})}}} \qquad (3)$$

- Depends on the coordinates of one star (S) and on the satellite attitude (A) at the time of the observation
- The aberration enters in the same way as for the arcs

The Linearized system of equations (I)

- The basic equations are highly non-linear

$$\cos\phi = \frac{\cos\psi_{(\hat{x},\mathbf{r})}}{\sqrt{1-\cos^2\psi_{(\hat{z},\mathbf{r})}}} = F\left(\boldsymbol{x}^{\mathrm{S}}, \boldsymbol{x}^{\mathrm{A}}, \boldsymbol{x}^{\mathrm{C}}, \boldsymbol{x}^{\mathrm{G}}\right)$$

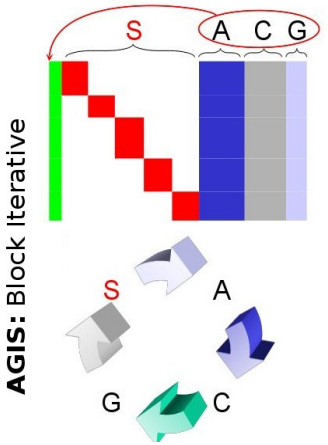
- The Equation system is quite large ($\sim 10^{10} \times 10^8$)
- Solving a large system of non-linear equations is extremely complicated because of
  - the mathematical techniques involved
  - the computational power needed

Solving the Equation System
The AGIS approach

- Linear System of Equation:
  $\mathbf{b} = A\mathbf{x}$, sparse, overdetermined

$$\mathbf{x} = \left(A^T A\right)^{-1} A^T \mathbf{b}$$

- AGIS approach: block iterative
  - each "block of unknowns" is solved separately assuming the others as known
  - after all blocks are solved, the process is repeated iteratively

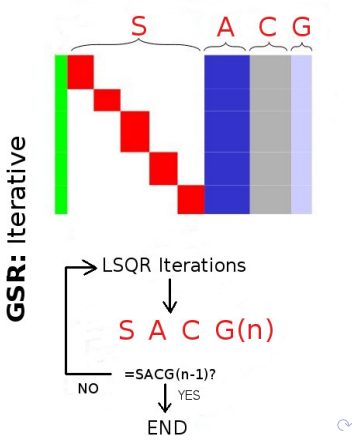


Solving the Equation System
The GSR approach

- Linear System of Equation:
  $\mathbf{b} = A\mathbf{x}$, sparse, overdetermined

$$\mathbf{x} = \left(A^T A\right)^{-1} A^T \mathbf{b}$$

- GSR approach: iterative
  - complete system solved with an iterative algorithm (LSQR)
  - if needed, the process is repeated using the previous solution as starting values



AVU-GSR
Parallelization Strategy

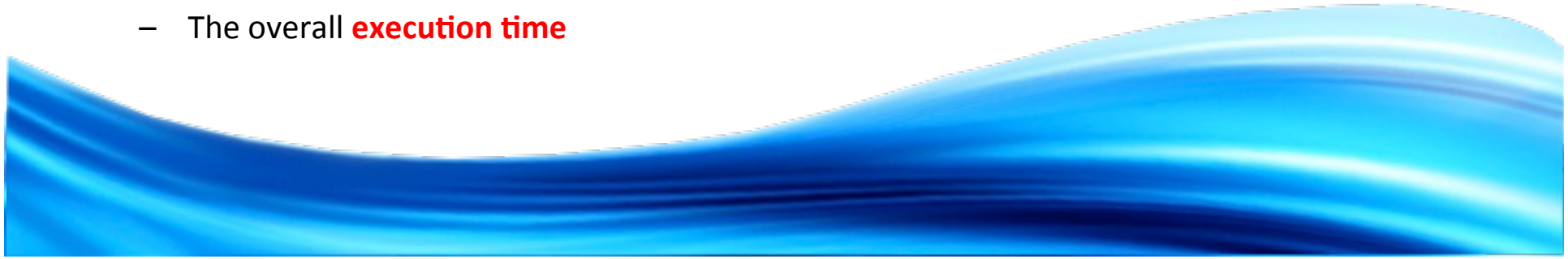
Solving a large sparse equation system

- The Solver Module uses a modified **LSQR – a conjugate gradient-based algorithm** – to solve the system of equations.

- The problem is converted into that of the solution of a large and sparse **system of linear equations:**

$$A\,x = b$$

- The target is to solve the system with the **highest number of stars** considering:
  - The computational data **dimension**
  - The overall **execution time**



The LSQR method .. In one slide

- **The LSQR method** (originally proposed by Paige and Saunders) consists of a conjugate-gradient type algorithm which is equivalent to compute, at each iteration i-th , an **approximate solution**

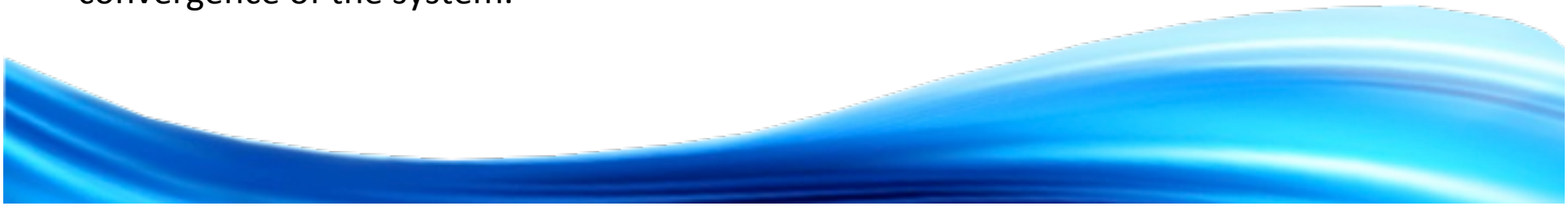
$$x^{(i)} = (A^T A)^{-1} A^T b^{(i-1)}$$

- and then evaluates the vector of residuals

$$r^{(i)} = b - Ax^{(i)}$$

- which has to be **minimized in the least-squares sense**, according to suitable convergence conditions defined by the algorithm itself.

- We adopt a PC-LSQR method that uses a **pre-conditioning technique**, which basically consists in a renormalization of the columns of A, made to improve the speed of convergence of the system.



Stopping conditions

- Three stopping conditions apply to incompatible systems like ours:

1: $||A^T r|| / (||A|| \, ||r||) <= ATOL$ (user input)

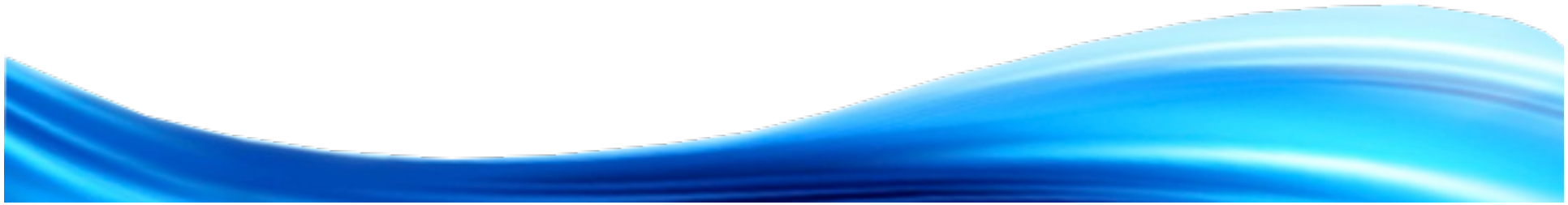
ensures that the LSQR solution is an acceptable least-squares solution for a perturbed matrix (A + E)
(ATOL = || A - E || / ||A||)

2: $|| A || \, ||x||/||b|| <= CONLIM$ (user input)

for ill-conditioned problems (not our case)

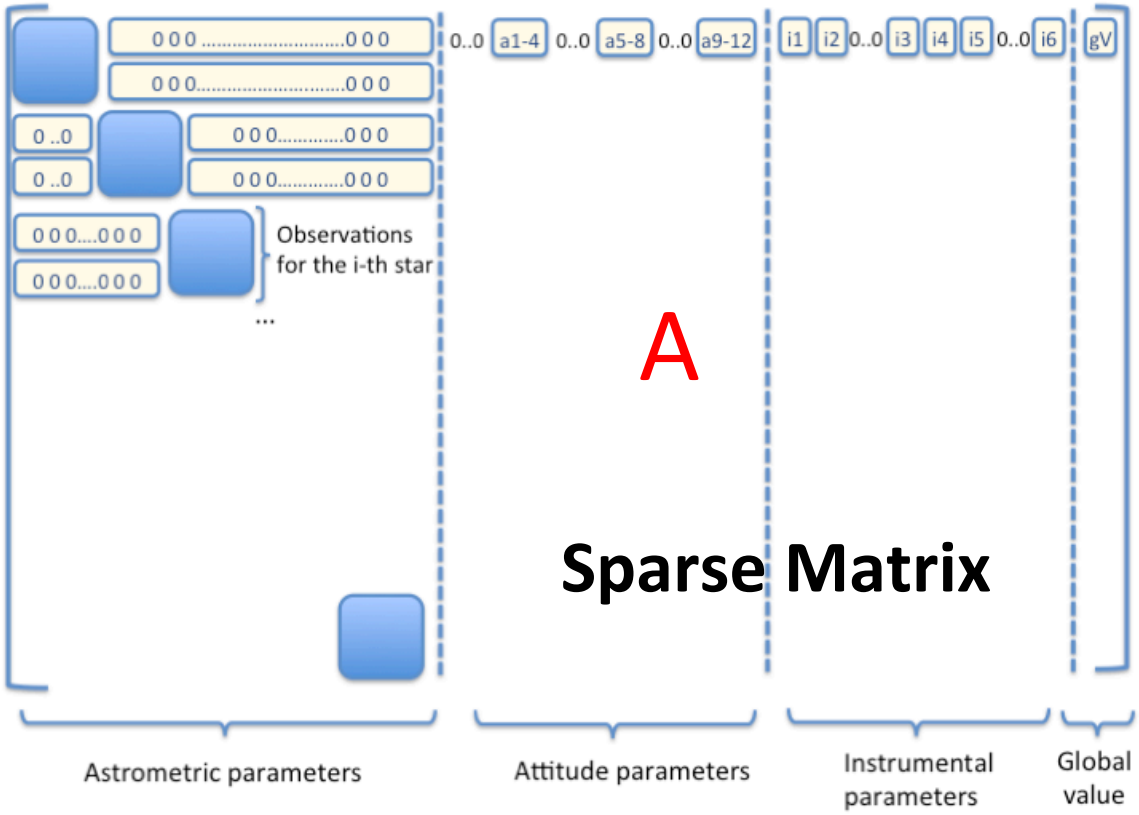
3: N_iter >= NITER (user input)

maximum number of iterations reached



The global problem

For each observation, the Total Matrix stores the astrometric, attitude, instrumental parameters and a Global Value coefficient



← *24 non-zero coefficients*

10^8 stars, each star observed **≈$7*10^2$** times

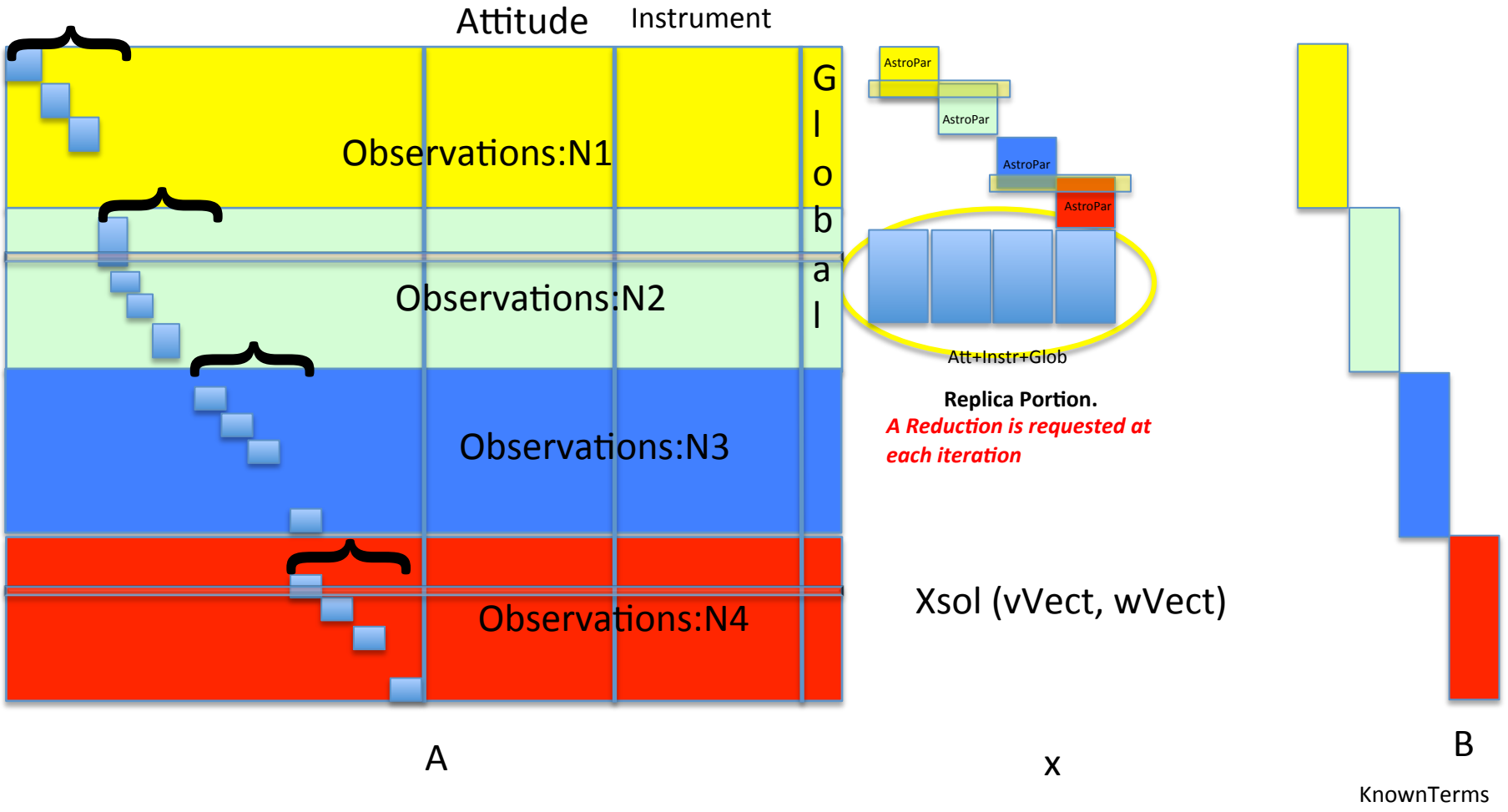
➔ ≈10^{11} obs with ≈30 parameters for each obs

➔ **≈$3*10^{12}$** coeff.

➔ **24** TBytes Ram

Each observation has: **5** astrometric coeff., **12** attitude coeff. (4x3 *equally spaced blocks*) , **6** instrumental coeff., **1** relativistic gamma coeff.

Data distribution



The main goal is to keep the computation as local as possible

The Parallel Code MPI and OpenMP

The code parallelization is based on a mixed paradigm:

➔ MPI:

- Each PE runs on a computing node.
- The input System is distributed in separated files
- Each PE reads (in parallel) the assigned vector portion (files)
  - ➢ *The reading phase is executed in parallel assuming shared home directory*

- Each PE computes the equations of the PC-LSQR method

$$x^{(i)} = (A^T A)^{-1} A^T b^{(i-1)} \quad \text{and} \quad r^{(i)} = b - Ax^{(i)}$$

on the portion of the global System and vectors **without MPI communications**

- At the end of each iteration step, global data reductions (collective sums) are requested to obtain the global result

With the adopted strategy for data distribution, we mainly need only some synchronism points (MPI_Barrier), some reduction operations (MPI_Allreduce), and few data communications (MPI_Bcast). The code scalability is therefore very high.

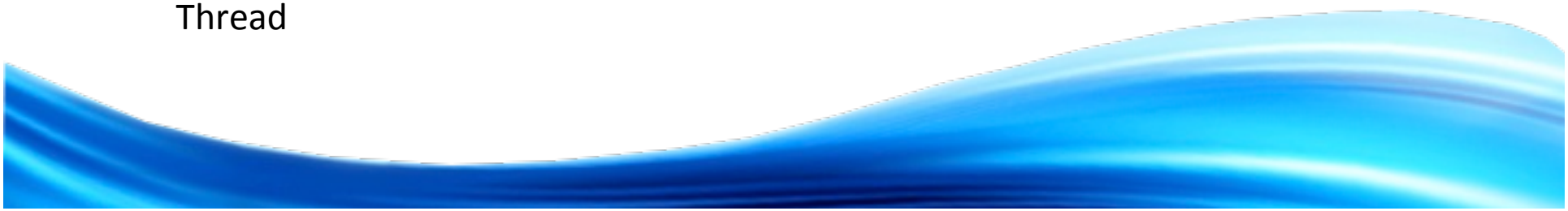
The Parallel Code MPI and OpenMP

➔ OpenMP

Each PE executes the matrix multiply op. using local data portion.

On each PE the OpenMP paradigm starts N-Threads that cooperate for the matrix multiplication operations

- The operation $r^{(i)} = b - Ax^{(i)}$ is executed in Multi-Thread on each PE with the higher gain (total indexes independency at each cycle).
- The operation $x^{(i)} = (A^TA)\text{-}1\ AT\ b^{(i-1)}$ is also executed in Multi-Thread on each PE but one critical region must be set (there are the same indexes for some cycles).
- Other setting operations (zeroing, initialization etc.) are also executed in Multi-Thread



Infrastructure MoU INAF - Cineca
2013 - 2021

Cineca will support INAF - AVU GSR Solver Module, searching a solution for 100 Million Star ….

Infrastructure MoU INAF - Cineca

Architecture: 10 BGQ Frames
Processor Type: IBM PowerA2, 1.6 GHz
Computing Cores: 163840
Computing Nodes: 10240
RAM: 1GByte / core
Disk Space: 2.6 PByte of scratch space
Peak Performance: 2PFlop/s

Basic compute element: **compute node**
PowerA2 chip with 16 cores,16 GB of RAM and the network connections

Infrastructure MoU INAF - Cineca

Phase 1

April 2013 - December 2013

➔ Preliminary tests of the parallel Solver Module and test phase.

➔ Procedure definition for the data transfer between ALTEC/DPCT and Cineca
➔ Available resource: **5 Million CPU core hours**

Phase 2

January 2014 - December 2015

➔ Test phase *(cont.)*

➔ System Solutions for **GSR-*Cycles 1-4***

➔ Available Resource: **25-30 Million CPU core hours**

Infrastructure MoU INAF - Cineca

Phase 3

January 2016 - February 2021

- A more powerfull system will be available in 2016 ***(MPI+OMP+GPUs)***
- Solutions for **GSR-Cycles 5-10**
- Available Resource: **40 – 70 Million CPU core hours** ***(… or much more)***

Others

- At least **100 TB Disk Space**
- **Technical support** by Cineca staff
- ***High priority queues***

Memory Request vs Execution Time

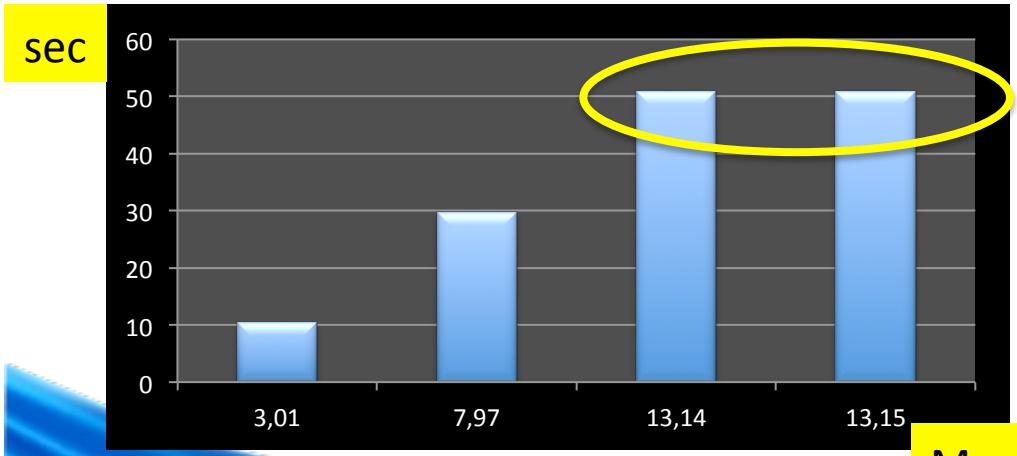
Case Study 1
GSR-Cycle 1
(72 obs. each star)

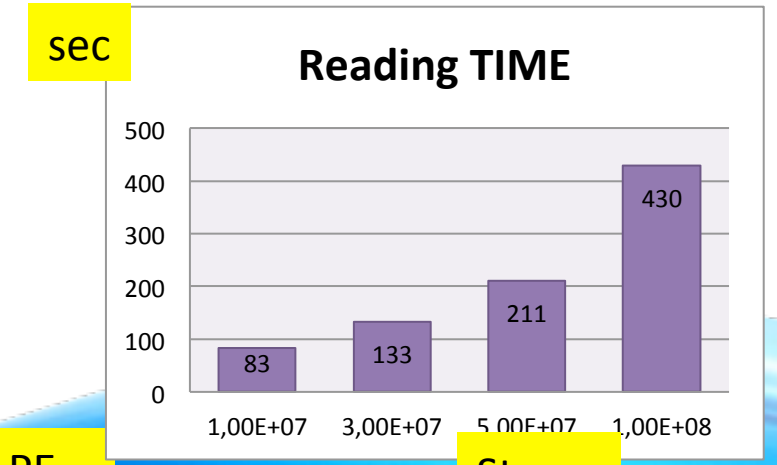
Searching the Identity solution.
Parameter normalized between -1.0 – 1.0

PE ➔ computing node

STARS	Reading TIME (sec)	Iterations Time	Single Iteration (sec)	N. Iter.	#PEs	Mem x PE (GB)
1.00E+07	83	1164	10	112	64	3.01
3.00E+07	133	3182	30	107	64	7.97
5.00E+07	211	5236	51	103	64	13.14
1.00E+08	430	5332	51	105	128	13.15

← Global PE Memory Usage





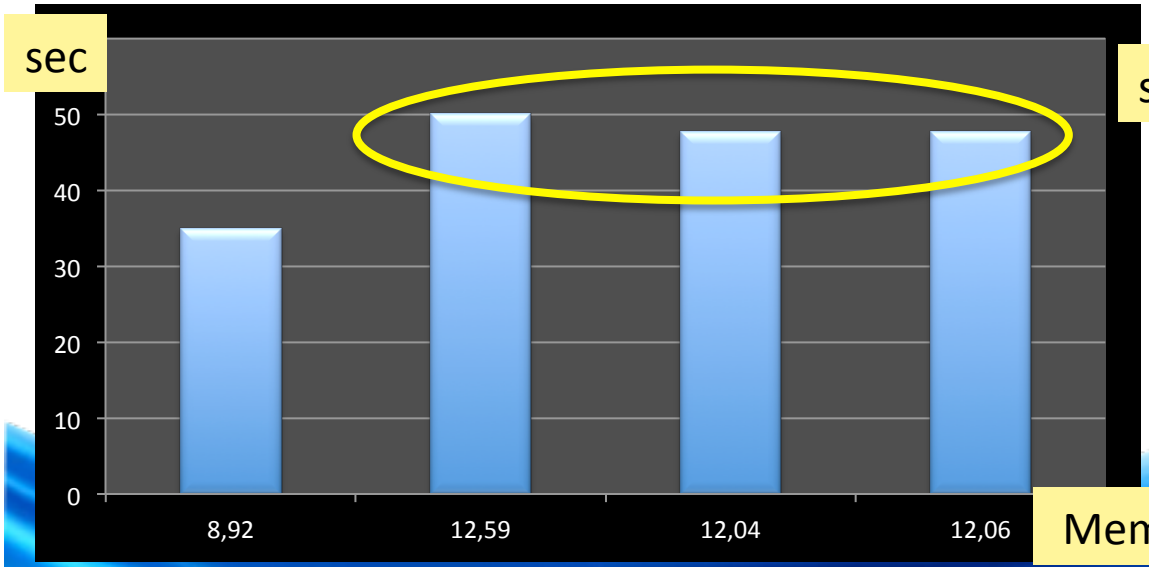
Memory Request vs Execution Time

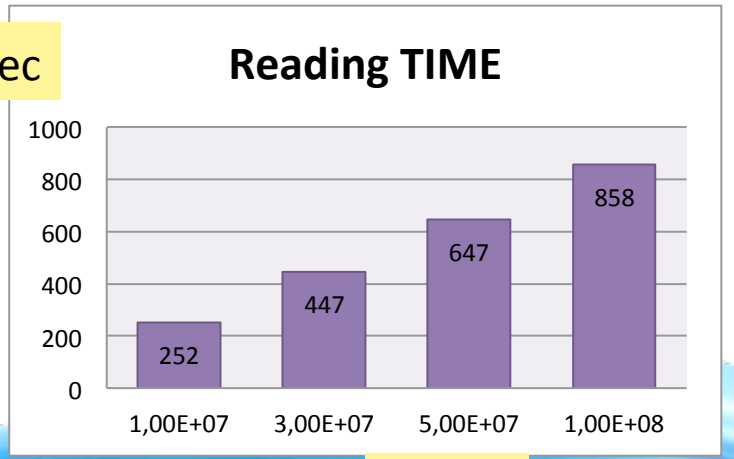
Case Study 2
GSR-Cycle 6
(432 obs. each star)

Searching the Identity solution.
Parameter normalized between -1.0 – 1.0

STARS	Reading TIME	Iterations Time	Iteration/ sec	N. Iter.	#PES	Mem x PE (GB)
1.00E+07	252	3706	35	106	128	8.92
3.00E+07	447	4947	50	99	256	12.59
5.00E+07	647	4768	48	100	512	12.04
1.00E+08	858	4721	48	99	1024	12.06

← Global PE Memory Usage





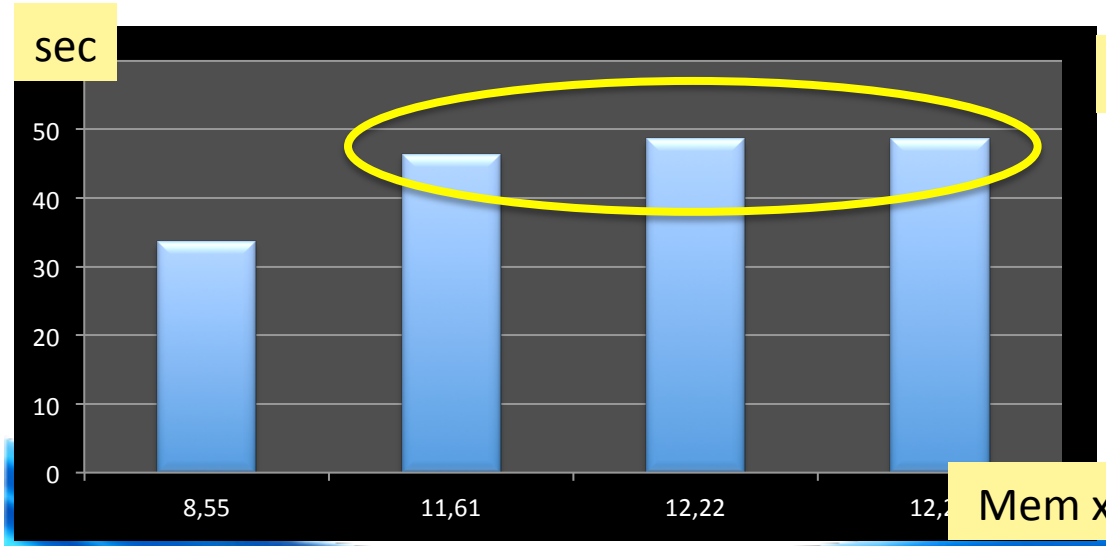
Memory Request vs Execution Time

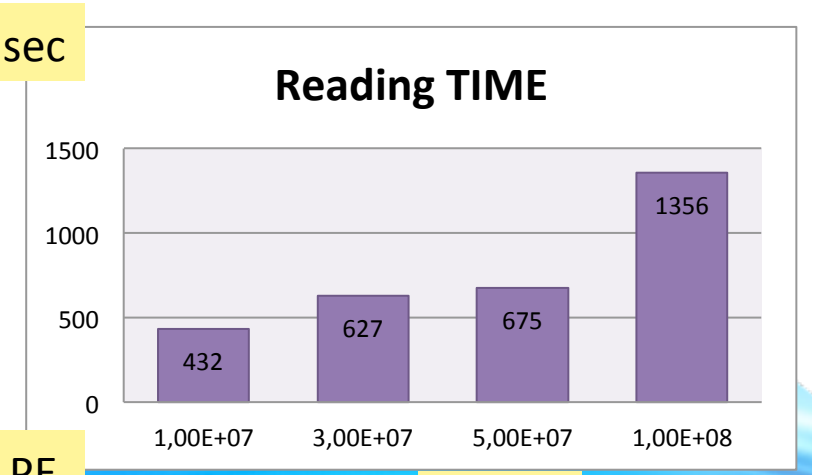
Case Study 3
GSR-Cycle 10
(720 obs. each star)

Searching the Identity solution. Parameter normalized between -1.0 – 1.0

STARS	Reading TIME	Iteration Time	Iteration/ sec	N. Iter.	#PES	Mem x PE (GB)
1.00E+07	432	3466	34	103	256	8.55
3.00E+07	627	4590	46	99	512	11.61
5.00E+07	675	4856	49	100	1024	12.22
1.00E+08	1356	4762	49	98	2048	12.25

← Global PE Memory Usage



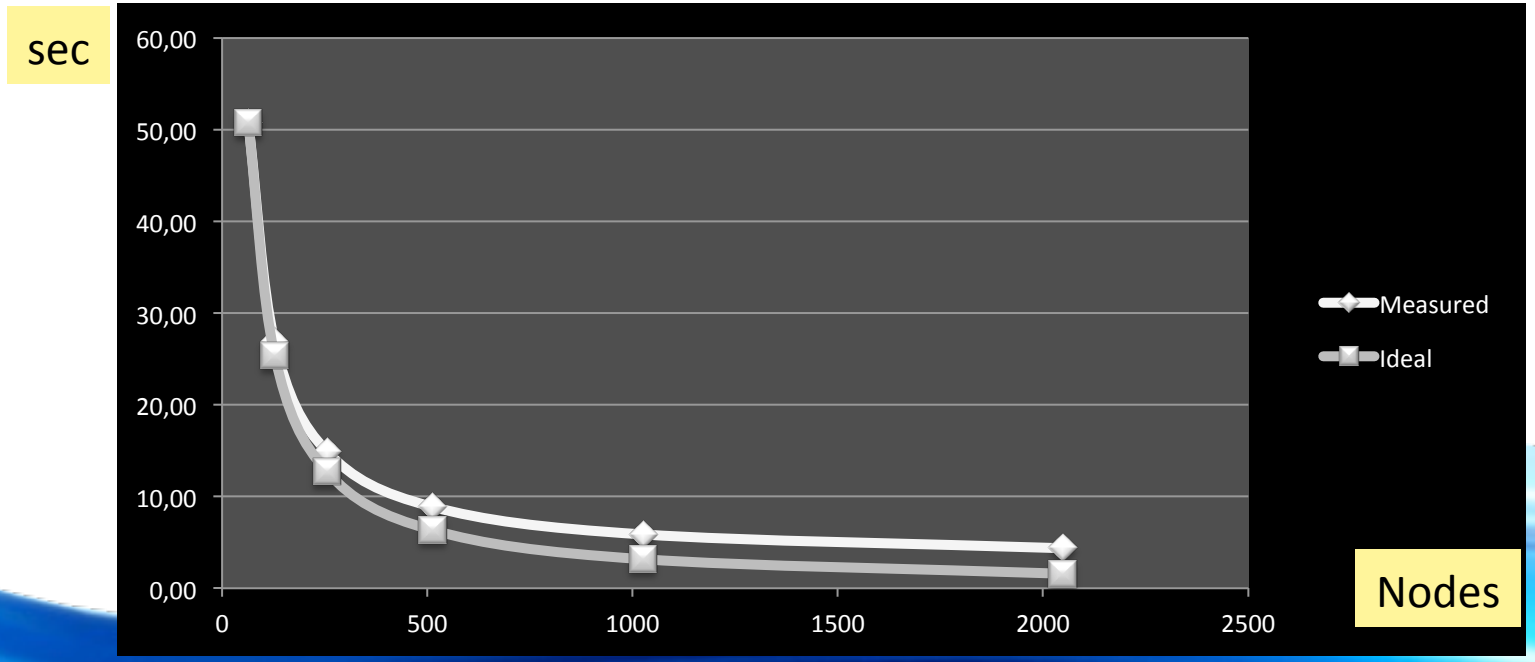


Code Scalability

Node	Measured Ex time	Ideal	Mem x PE (GB)
64	50.76	50.76	13.1
128	26.85	25.38	6.95
256	14.87	12.69	3.85
512	8.89	6.35	2.3
1024	5.87	3.17	1.52
2048	4.37	1.59	1.13

Case Study
BGQ Fermi Cineca System

Code Scalability with 50M Stars at Cycle 1



Global Execution expected on BGQ

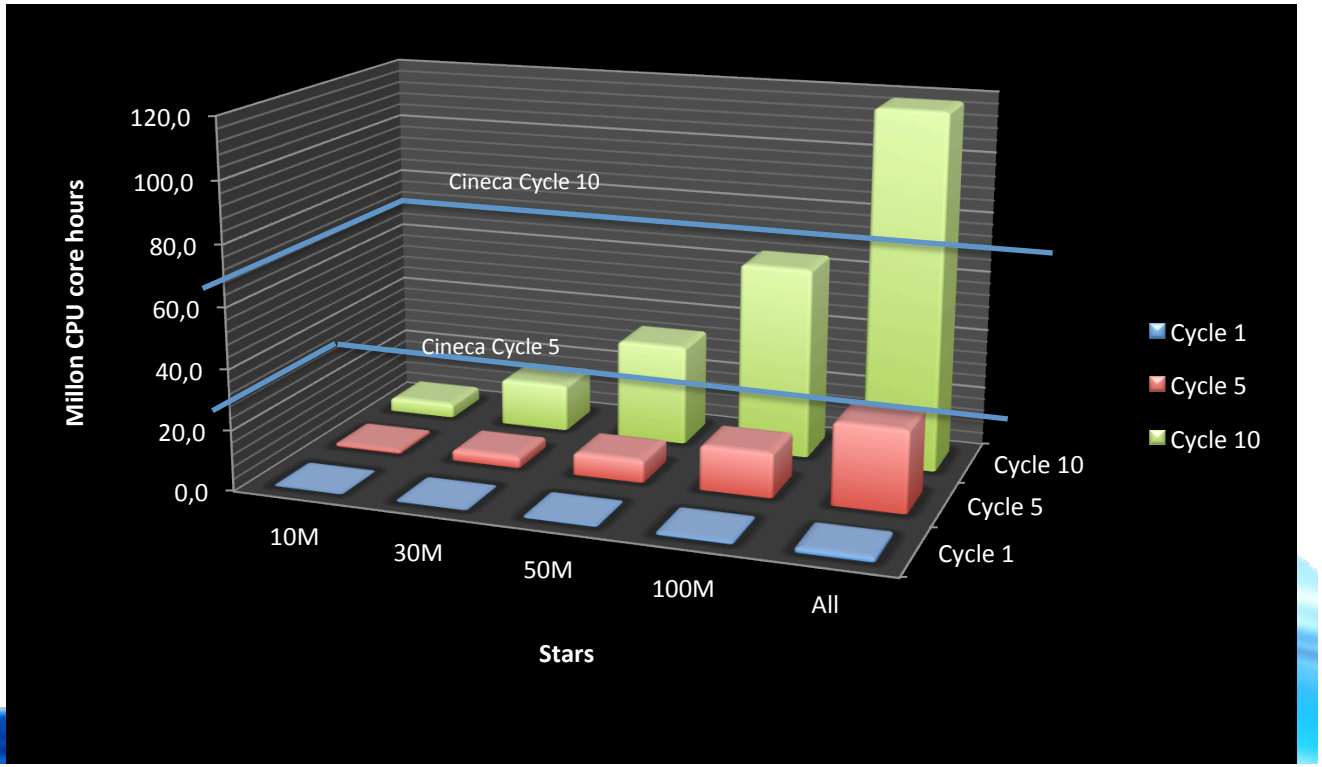
	Cycle 1	Cycle 5	Cycle 10
10M	0.1	1.2	4.8
30M	0.2	3.7	15.8
50M	0.4	7.5	33.6
100M	0.8	14.8	63.9
All	1.5	27.2	118.1

Global CPU core Million hours.

- Convergence hypotesis: **4,000** iterations
- Number of solutions executed for each cycle: **7**

Cineca MoU

- Cycle 5 ➔ 30 Million hours
- Cycle 10 ➔ 70 Million hours



Grazie per l'attenzione

Restiamo umani.