Data Intensive Science Client Tour EMEA / IBM Team with INAF

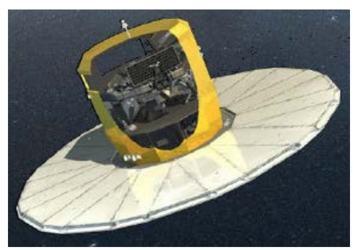
Roma - 2 December 2015

L'INAF, as R&D alone and with all R&D institutes of MIUR, is involved in the development of the e-infrastructure need for scientific research:

- ✓ **Network**
- ✓ **HPC/HTC**
- ✓ **Big Data**

All these point are strategic to allowing the INAF researchers to be involved on the future challenges like:

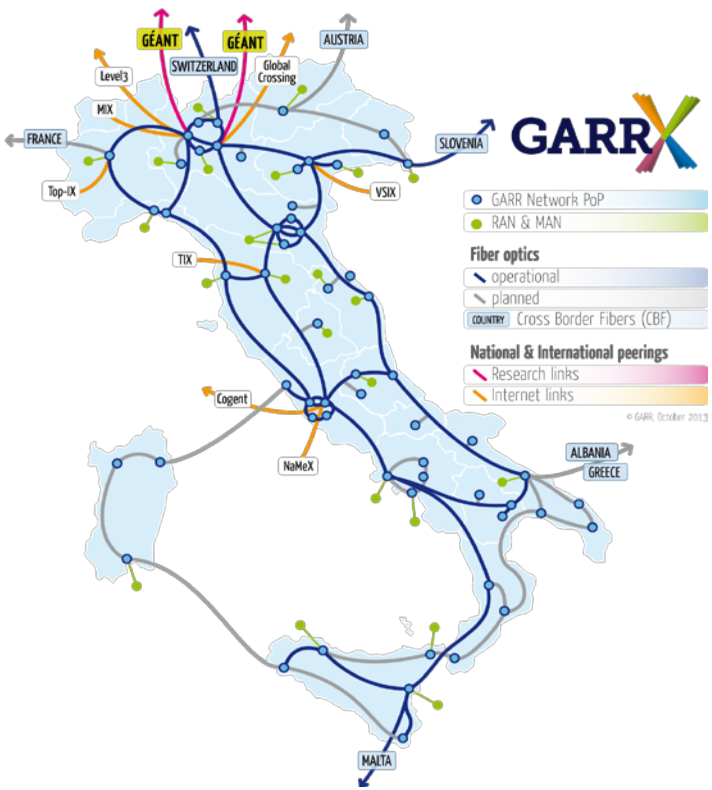
- ✓ **GAIA**
- ✓ **CTA**
- ✓ **Euclid**
- ✓ **SKA**
- ✓ **....**



H2020 Projects: INDIGO, ASTERICS, EGI

Network:

GARR is our backbone



All infrastructures are connected mainly with

➔ 1 G (less 150 Mb)
➔ Backbone 10G

Computing:

- ✓ **HPC**
  - ✓ 9/18 center are working with HPC
  - ✓ 159 +70 Billion hours @ BGQ Cineca (2013)
  - ✓ 16 research program (70 people)
  - ✓ 2.7 PB of data.

- ✓ **HTC**
  - ✓ DHTCS project (Cloud under development)
  - ✓ Cluster @ PON (Catania, Palermo, Cagliari (2010))

- ✓ **Local Cluster :**
  - ✓ ~20 “group” cluster

✓ **Open issues:**
- ✓Developing a INAF facility (Tier-2)
  - ✓Test
  - ✓Development
  - ✓Fast “answer”(qsub ORA)
- ✓ cluster INAF…allocate local resources

Data Archive:

- ✓ All INAF structures have archives
  - ✓ About 54 archives (some under development)
    - ✓ 59% public,
    - ✓ Policy INAF: dati raw are public after 1 year
- ✓ Centro Italiano Archivi Astronomici (IA2)

- ✓ GAIA (on-fly) ➔ DPAC Center (1 of 6) @ OATorino
  - ✓ 1 PB (mainly part of the DBMS, Oracle partnership)
- ✓ Euclid ➔ > 10 x GAIA (2020)
- ✓ CTA (ASTRI) ➔ > 10 TB/giorno
- ✓ SKA ➔ > 100 TB/giorno

✓Data Curation & Preservation
- ✓Standard FITS (from 1970)

✓Data Interoprability ➔ Virtual Observatory (from 2001)
- ✓IVOA – Inernational Virtual Observatory Alliance

Data Intensive Science Client Tour EMEA / IBM Team with INAF

chaired by Riccardo Smareglia, Andrea Bulgarelli

Wednesday, 2 December 2015 from **14:00** to **17:00** (UTC)
at **INAF Centrale (Sala Copernicana)**

Manage

Description **Abstract:**

INAF and IBM will like to have a short workshop to illustrate the INAF needed and roadmap about HPC and HTC.

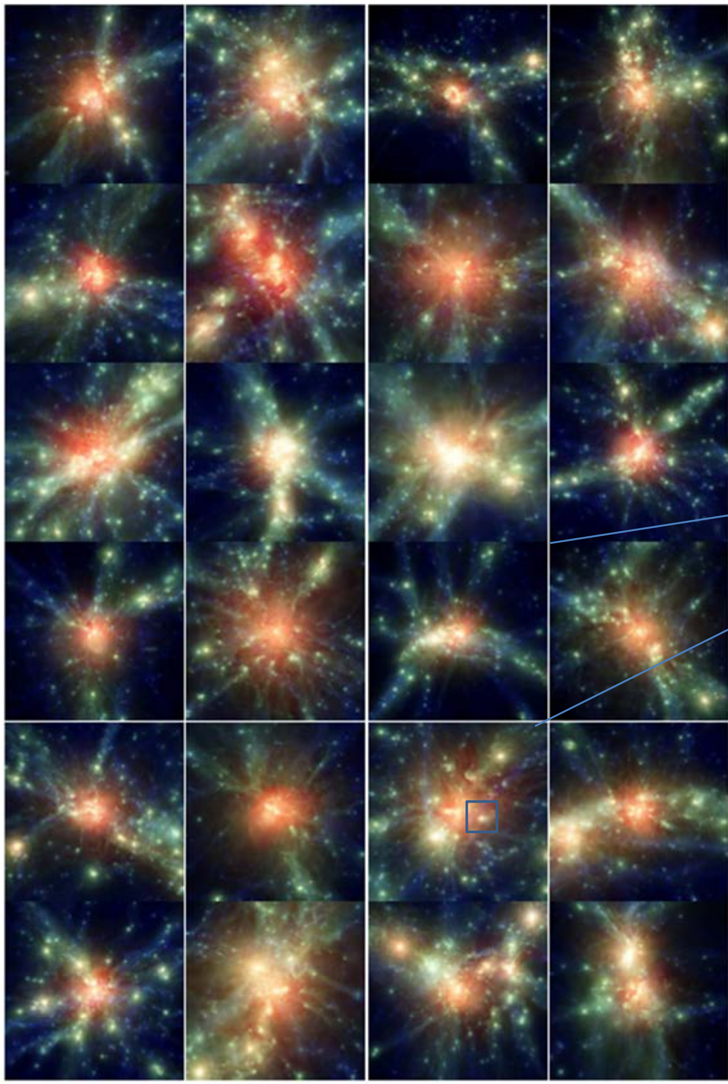
Program : (very draft)

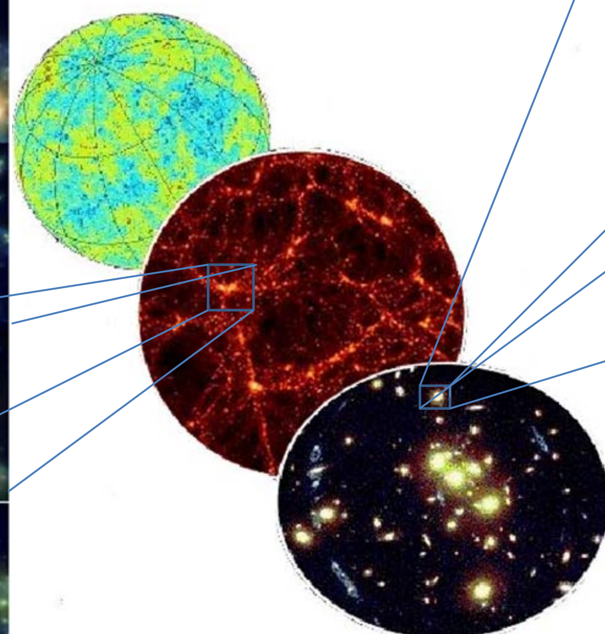
- 10' - Introduction - R. Smareglia, ICT
- 10' - OpenPower @ INAF- A. Bulgarelli
- 10' + 10' - CTA - P. Caraveo / A. Antonelli
- 10' Gravitational Wave - (Enzo Brocato) A. Antonelli
- 10' - HPC @ INAF - U. Becciani
- 10' - GAIA - A. Vecchiato
- 10' - Euclid - (F.Pasian) R. Smareglia
- 10' - Cosmolgical Simulation - (S. Borgani) R. Smareglia
- 10' - SKA - R. Smareglia
- 15' Discussion

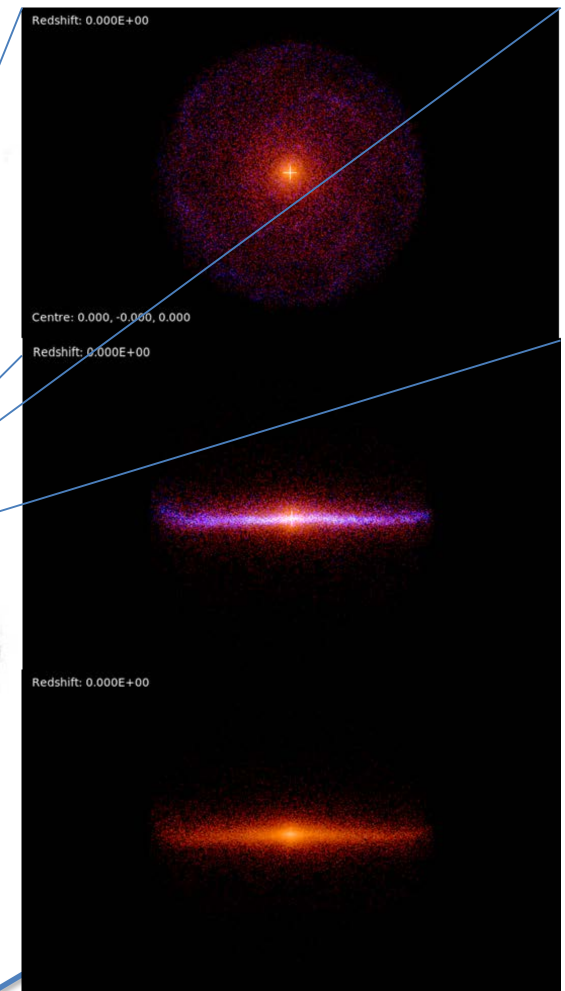
Partecipant:

- INAF
  - Nicolo' d'Amico (INAF President)
  - Paolo Vettolani (Scientific Director)
  - Riccardo Smareglia
  - Patrizia Caraveo
  - Andrea Bulgarelli
  - Ugo Becciani
  - L.Angelo Antonelli
  - A. Vecchiato
- IBM
  - ULF Troppens (Consulting IT Specialis / IBM Spectrum Scale development)
  - Martina Naughton (EMEA Business Development Manager - HPC)
  - Klaus Gottschalk (HPC Architect OpenPOWER)
  - Ulrich Oymann (Business Developer Manager EMEA HPC)
  - Burkhard Steinmacher-Burow (STSM - IBM Technical computing - OpenPOWER)
  - Kevin Gildea
  - Cecilia Carniel (IBM PowerSystem Scale Out Server)
  - Claudio Fadda (Research Senior Architect)
  - Giorgio Richelli

Cosmological Simulations: Some example **result**







10^6 – 10^{10} calculus elements (particles)

Algorithms

- **GRAVITY** – long-range, all-to-all calculus elements communication needed (in principle)
- **HYDRODYNAMICS** – short-range, but a small number of calculus elements needs many time steps
- **ASTROPHYSICAL PROCESSES** – (radiative cooling, stqar formaton, black holes evolution, energy exchanges between BH/stars and gas) partially subgrid: the exchange part needs communications

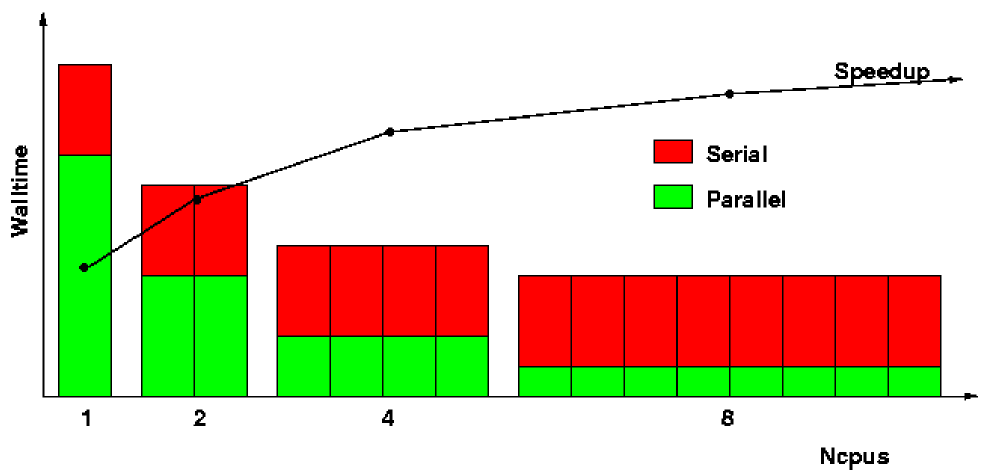
- **CODE** used by our group: **GADGET3** (V. Springel, K. Dolag et al).
- Our group has *access to the international repository*, and is among the **code developers**
- Our group often was a **beta-tester** for supercomputers installed at CINECA, since 2003

HPC computing time

- Most of our CPU time obtained with competitive grants at CINECA (INAF-CINECA convention, ISCRA) and CASPUR
- Two PRACE projects with local PI (developement)
- Involved in several Class-A PRACE projects
- A DECI project under review

Portability, scalability…

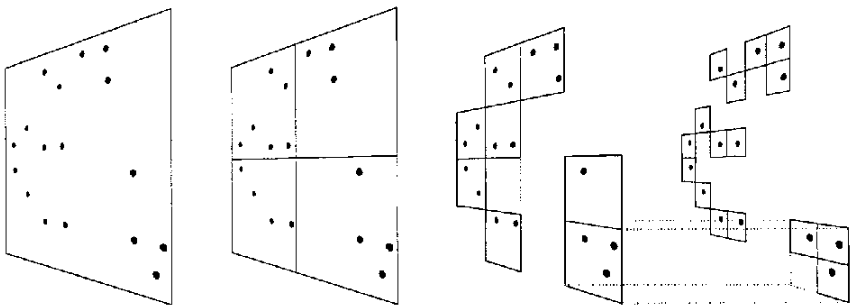
- «Trieste» group's simulations run on several machines:
  - Linux clusters (from Beowulf with a 10Mb network to bgp, raijin..)
  - Intel SP3-7
  - Server many-cores shared memory
  - SuperMUC, MareNostrum, Raijin, USC…
  - Plx, Eurora (but: no GPU)
  - …we got troubles with Fermi



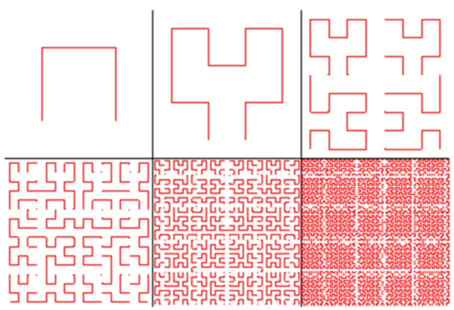
On massively parallel architectures we need extreme work-loar balance! Our kind of problem not Very well suited.

(not only us:
Eris run on 512 SP6 cores for 9 months)

Code parallelization



A tree is used for gravity computation (approximate, but less communications)



DOMAIN DECOMPOSITION using a Peano space-filling curve: work-load balance at the cost of memory unbalance

V. Springel, N. Yoshida and S. D. M. White

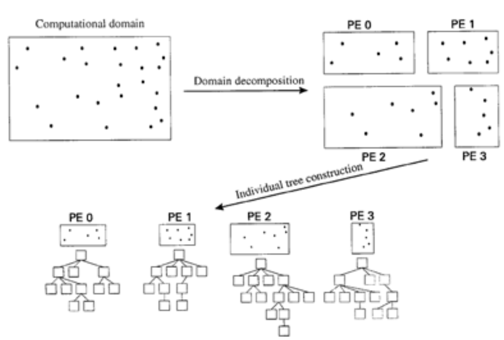


Figure 2: Schematic representation of the domain decomposition in two dimensions, and for four processors. Here, the first split occurs along the y-axis, separating the processors into two groups. They then independently carry out a second split along the x-axis. After completion of the domain decomposition, each processor element (PE) can construct its own BH tree just for the particles in its part of the computational domain.

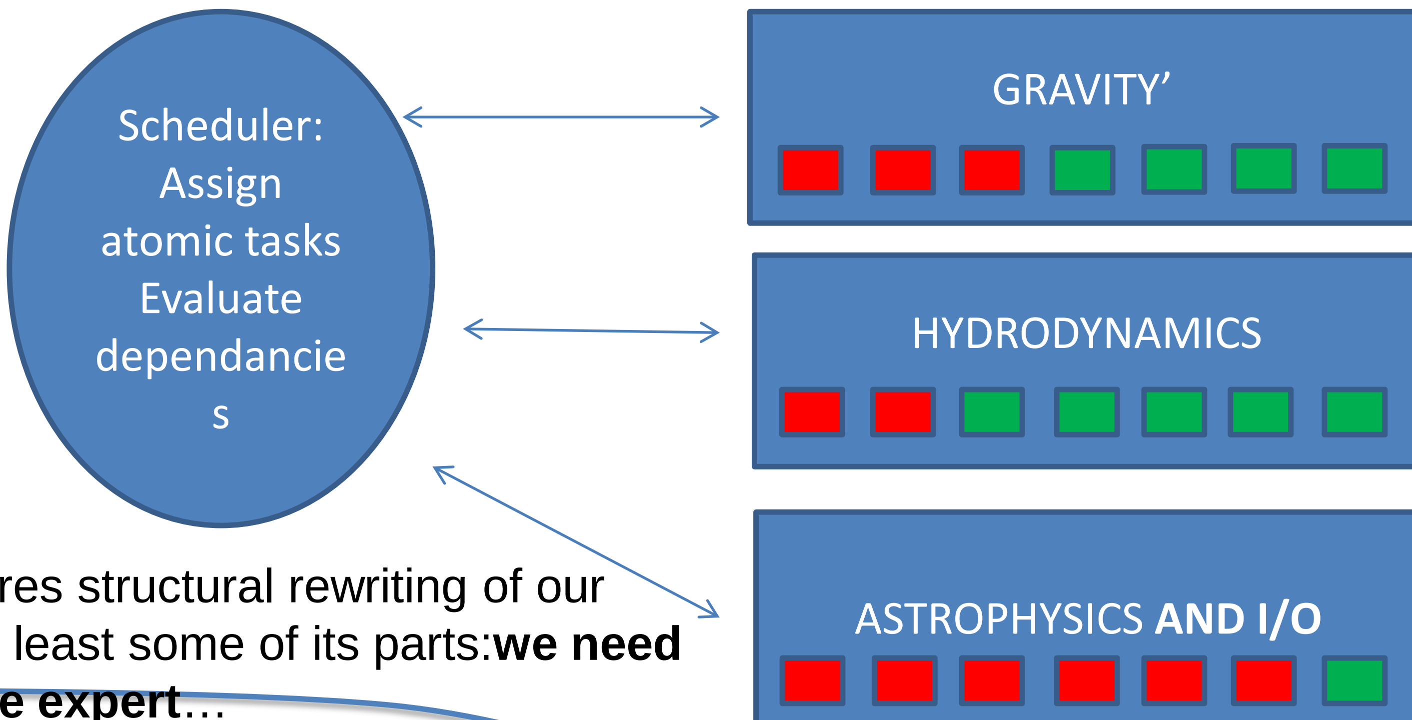
Computation assigned at single MPI tasks. Inside them, OpenMP for shared memory parallelization

Problems with the current HPC computers generation

- **Work-load** balance scheme **costly** in terms of **memory**: a **FEW** MPI tasks allowed for each computing node.
- Inside node, OpenMP parallelization not so efficientNel nodo la parallelizzazione e' fatta con OPENMP: poc
- **I/O** can be **extremely costly** on BlueGene type computers
- In single object/high resolution calculations, **our problem is intrinsecally unbalanced**: a few particles always active (maybe less particles than cores!)

Possible optimizations

- **De-syncronization** of all possible calculations, via algorithm analysis, atomic task and dependance identification, and the use of a client-server kind of scheduler



This requires structural rewriting of our code or at least some of its parts:**we need a software expert**…

Accelerators

- Historical problem with accelerators: they are effective when **flop/byte** is **high**
- ...in our case **flop/byte** is embarassingly **low**: in increasing order, gravity, hydrodynamics, astrophysics
- *Simpler solution: bring astrophysics (and/or hydro?) on accelerator and de-syncronize it*
- Problem: very good syncronization needed between accelerator and CPU calculations
- However, at least partially, a scheme as that described above has to be implemented

Hardware solutions

- In the past: **GRAPE**. Board designed to calculate gravitational interactions. Not extremely successfull.
- **Accelerators**: only solution (?), **increase bandwidth** between CPU and accelerators (or between accelerators).
- The ideal supercomputers for our kind of calculation remains ortogonal to the currend direction of HPC developement: **few CPUs, with a lot of RAM, very powerful**
- En passant, *other scientific communities have similar needs* (climatology, turbulence…)

Conclusions: possible collaborations

- «Trieste» group would benefit from a high-level training programme in which one person could deal with code optimization on specific architectures

- Our experience as hardware and software tester can be exploited

- Scientific visualization.

The Euclid Mission

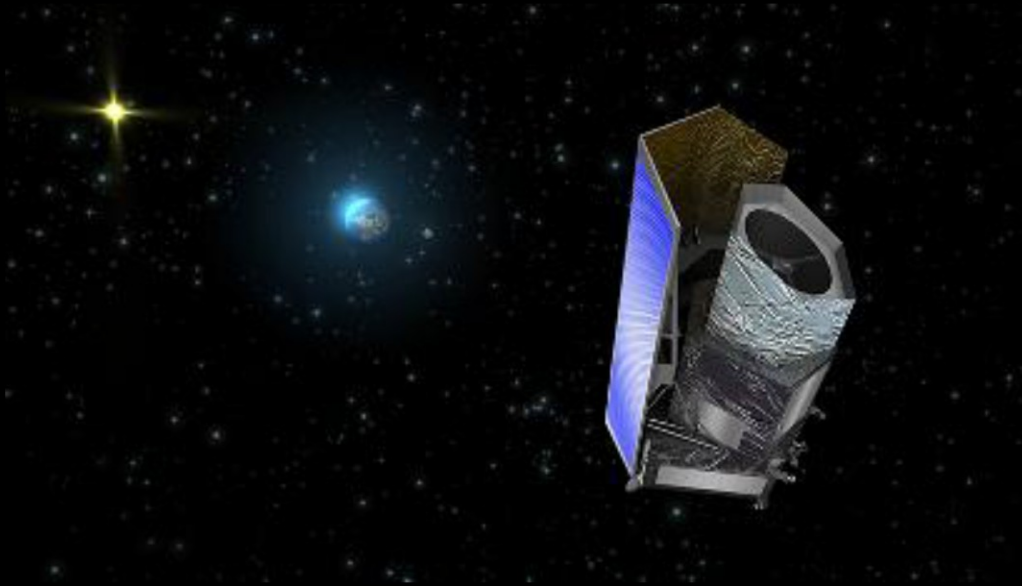
M2 mission in the framework of the **ESA Cosmic Vision Programme**

Euclid mission objective is to map the geometry and understand the nature of the dark Universe **(dark energy and dark matter)**

Actors in the mission: **ESA** and the **Euclid Consortium** (institutes from 14 European countries and USA, funded by their own national Space Agencies)

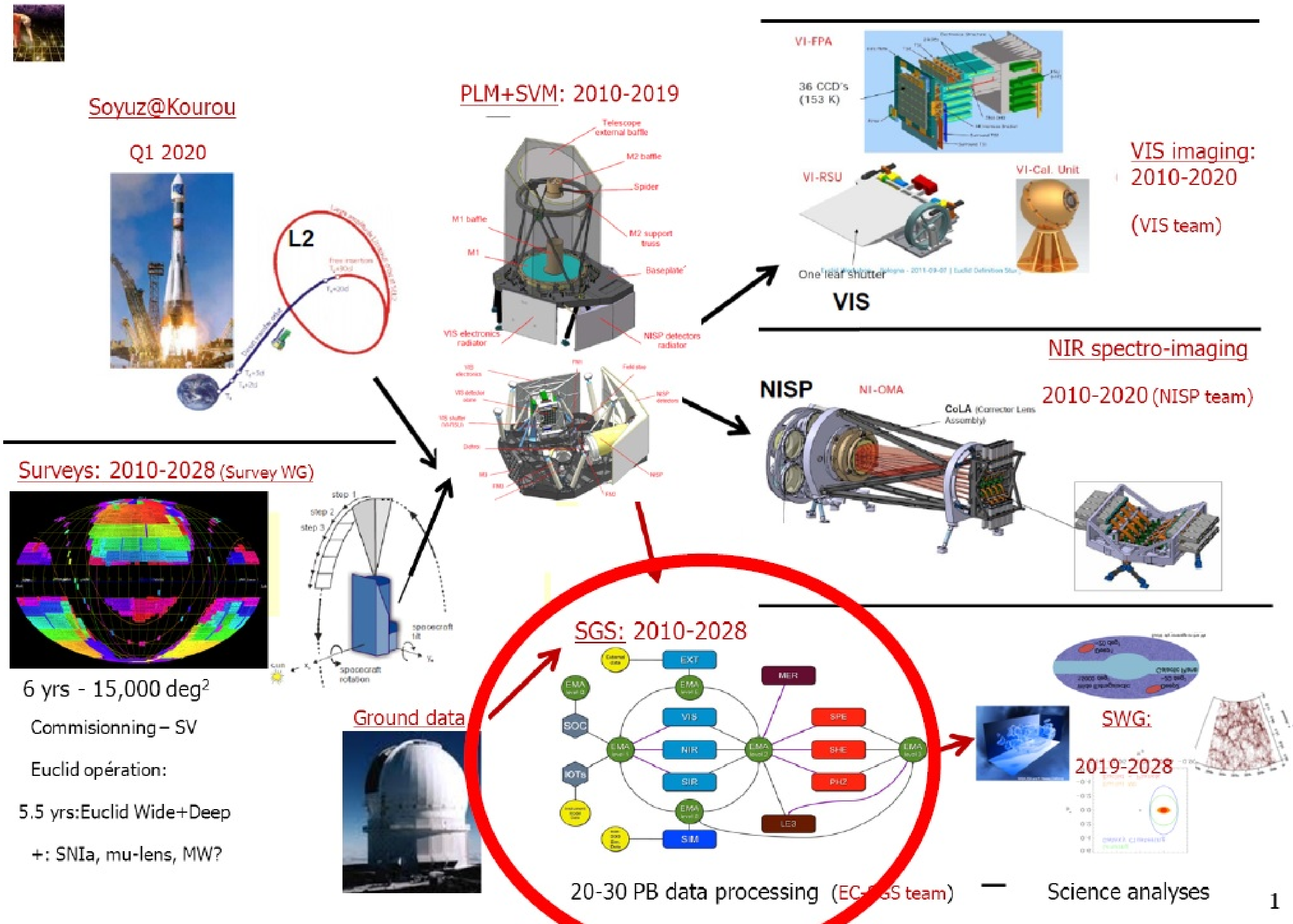
Euclid Consortium:
15 countries
100+ labs
1200+ members

Biggest collaboration!

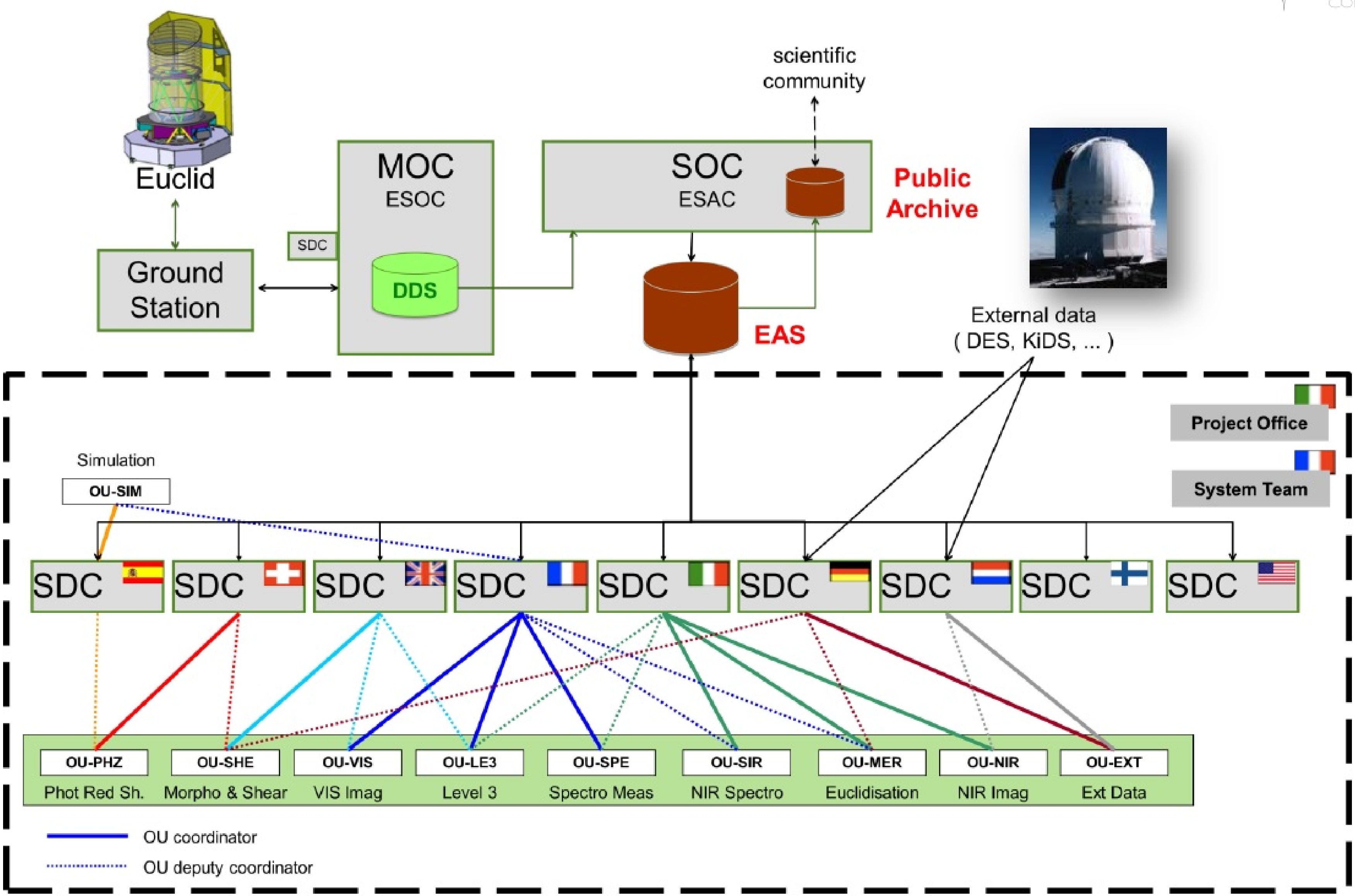


For more information see :
http://sci.esa.int/science-e/www/area/index.cfm?fareaid=102
http://www.euclid-ec.org

Euclid mission at a Glance



Euclid Ground Segment



Key Challenges

- **Federation** of 8 European + **1** US SDCs (Science Data Centers) + SOC (Science Operation Center)

- Heavy **simulations** needed before the mission

- Heavy **(re)processing** needed from raw data to science products (volume multiplied by dozens),

- Large amount of **external data** needed (ground based observations)

- Amount of **data** that the mission will generate per full release
  + 26 PBytes of data (including external data) => "'175 PB grand total
  + 1.10^{10} objects
  + => **not achievable with classical architecture**

- **accuracy and quality** contrai required at each step

Architecture key concepts

- No Dedicated Processing SDC: Any pipeline should run on any SDC (with some exceptions, e.g. Level 1, EXT ingestion, LE3)

- Distributed Data and Processing
  - Each SDC is both a processing and a storage « node »

- Move the code, not the data
  - Run the pipeline where the main input data is stored

- Separation of metadata (inventory) from data (storage)

- Kind of home made *"Map/Reduce"*
  - Lower level of processing on QoD (minimal processable set of data covering a given sky area), constituting catalogs of objects
  - Higher level of processing based on data cross-matching/correlation: need to colocate reduced set of data (whole catalog)

Conclusions

- Big challenge !
- Already active working groups on:
  + Architecture principles
  + POC Mock-up & challenges
- Working prototypes => pillars of the SGS
- Next steps
  - Refine the architecture model according to the scientific processing requirements (granularity, triggering, volumes, ...)
  - Identify candidates implementations
  - Interleave scientific & architectural challenges

DIR
SCI

ISTITUTO NAZIONALE DI ASTROFISICA
INAF
NATIONAL INSTITUTE FOR ASTROPHYSICS

Data rates and Storage: SKA

- 2020 era radio telescope
- Very large collecting area (km^2)
- Very large field of view
- Wide frequency range (70MHz - 25 GHz)
- Large physical extent (3000+ km)
- International project
- Telescope sited in Australia and/or South Africa
- Headquarters at Jodrell Bank, UK
- Multiple pathfinders and precursors now being built around the world

SKA project

- Dishes
  - Depends on feeds, but illustrate by 2 GHz bandwidth at 8-bits
  - **64 Gb/s from each dish**

- For Phased Array feeds increased by number of beams (~20)
  - **~ 1 Tb/s**

- For Low frequency Aperture Arrays :
  - Bandwidth is 380 MHz
  - **– 240 Gb/s**

- These are from each collector into the correlator or beam former
  - **2700 dishes**
  - **– ~ 600 Tb/s**

SKA computational requirements

- SKA correlator in case of Pulsar search (PPS):
  - data rate of the pulsar search engine is expected to reach 0.6TeraSamples/sec (1sample = 4*8 bit)
  - SKA Pulsar Search input is approximately **1PetaBytes** on each cycle of observation which lasts up to 600s
  - It is expected to observe in pulsar surveys for 1 day → **144 PB** of raw data
  - No possibility to handle with this amount: from 1 PB raw → some hundreds MB of correlated data for each cycle → **14 TB/day**
  - Particular case of massive objects:pipeline performances required → **10 PetaOps/s** for acceleration process