Euclid – data complexity

Fabio Pasian
(INAF – OATs)

Euclid Consortium Science Ground Segment Project Office

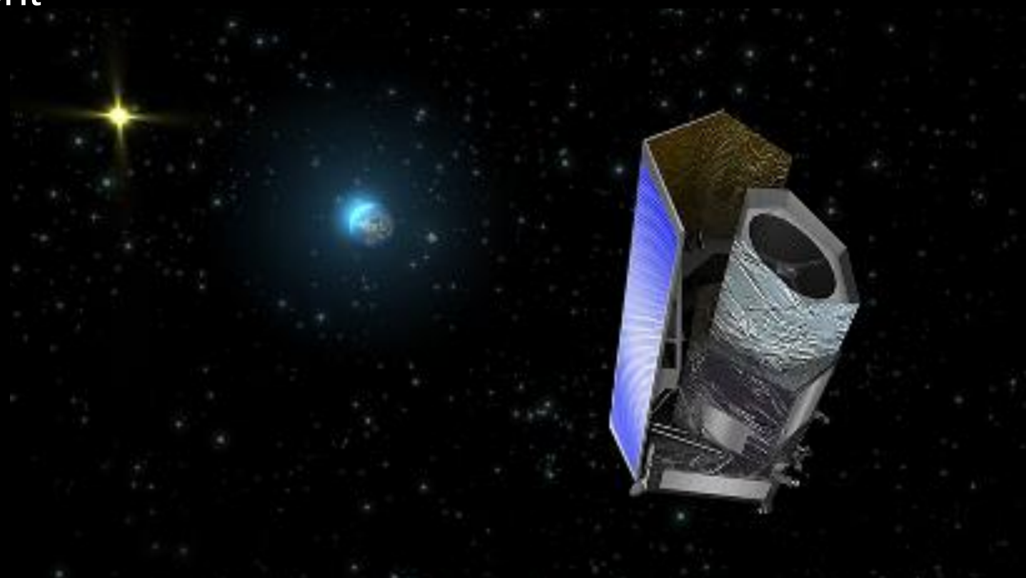
on behalf of the Euclid SGS development team

The Euclid Mission

M2 mission of the **ESA Cosmic Vision Programme** (launch Q2/2020)

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 13 European countries and USA, funded by their own national Space Agencies) → Steering Committee, Multi-Lateral Agreement



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

The Euclid Consortium

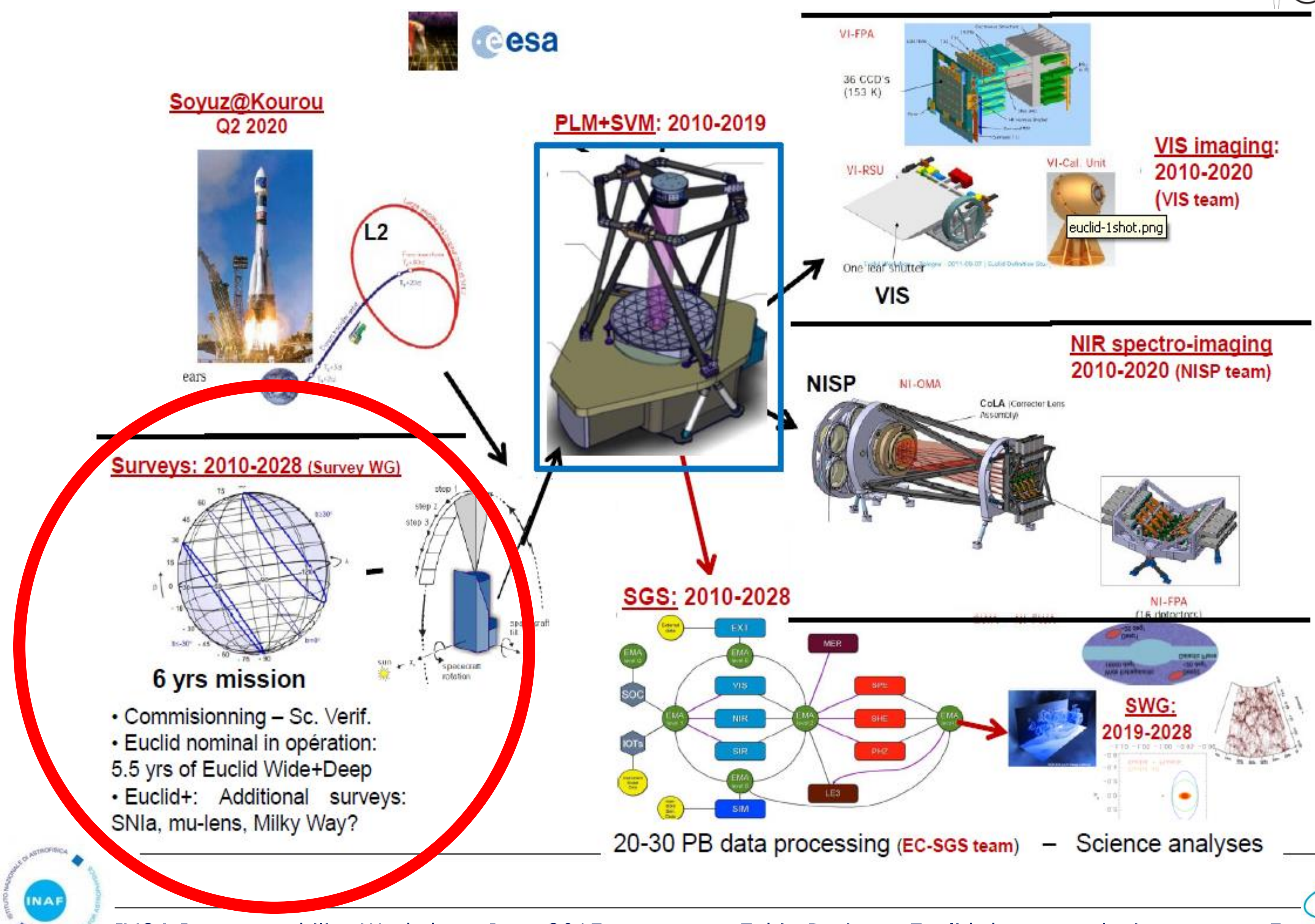
- The Euclid Consortium is in charge of:
  - building and operating the instruments (VIS and NISP)
  - developing and running the data processing within a unified Science Ground Segment (SGS)
  - performing the science analysis on the Euclid data products

- The Euclid Consortium is composed of 1300+ members
  - 350+ Consortium members signed up to participate in Science Ground Segment work (active: ~150)

Euclid science objectives

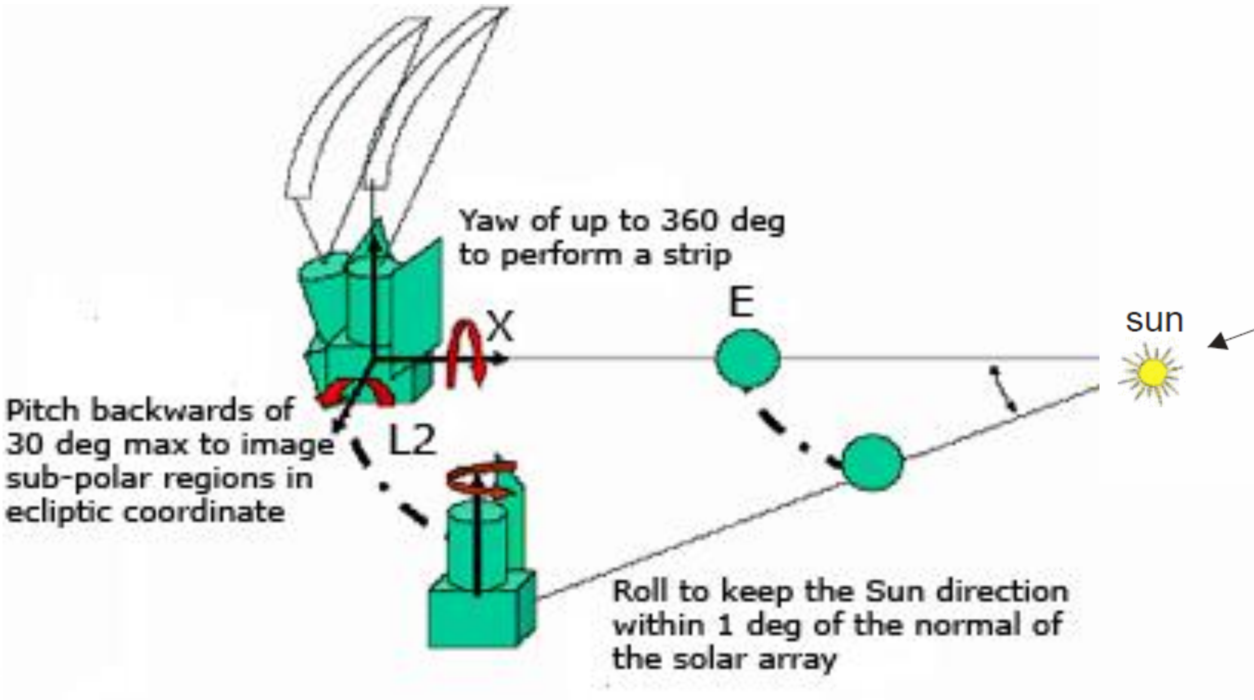
- Euclid is optimized for the measurement of its two primary cosmological probes, namely
  - weak gravitational lensing
  - galaxy clustering
- The two probes provide independent measures of both the geometry of the Universe (i.e. redshift as a function of distance) as well as the growth of cosmological structures due to gravity in an expanding Universe.
- The combination of the two probes not only enables the experiment to reach unprecedented statistical precision, but also provides a crucial cross-check of systematic effects, which become dominant at these levels of precision.
- Euclid data are of particular interest for other (legacy) science.

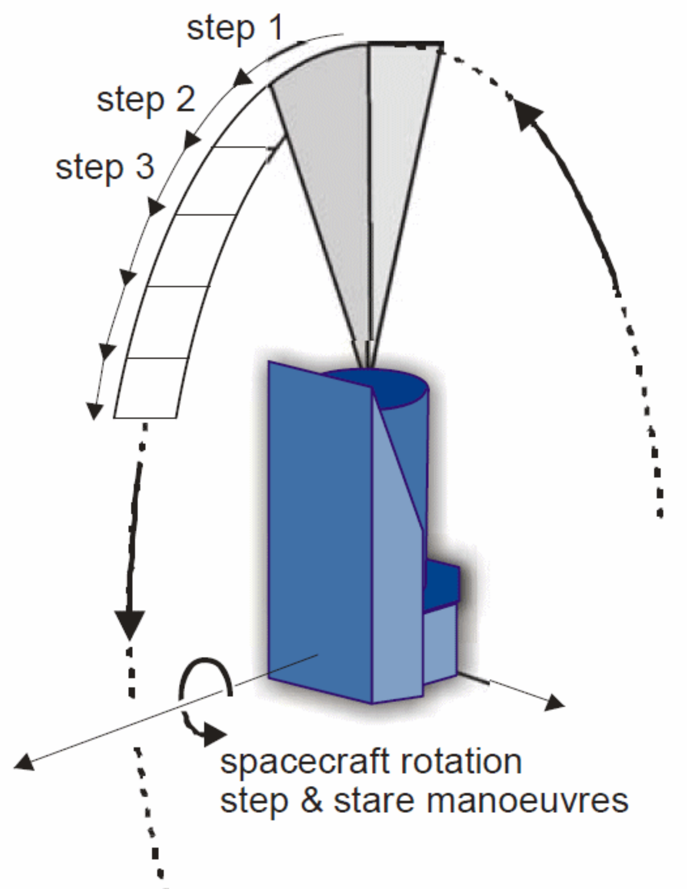
Euclid at a Glance



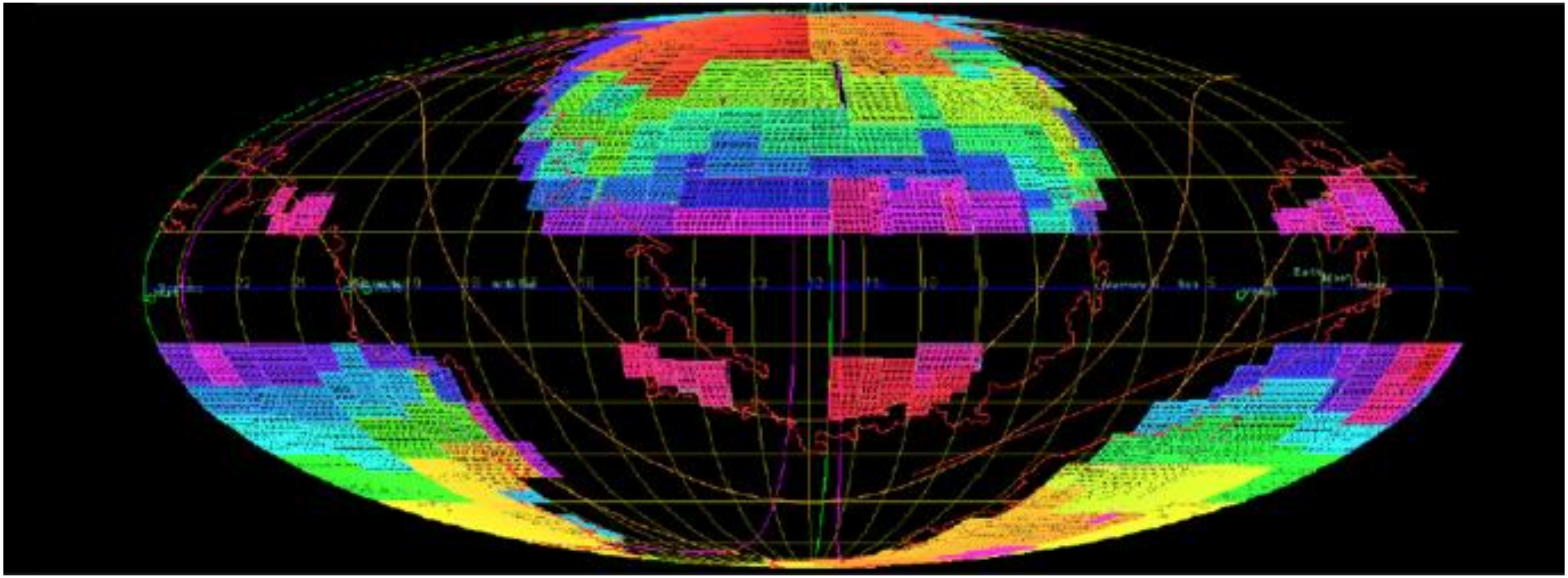
The Euclid survey (I)

- Large Lissajous orbit around the second Sun-Earth Lagrange point (L2).
- Telescope is Korsch three mirror anastigmat providing high image quality over a large field of view of more than 0.5 deg^2.





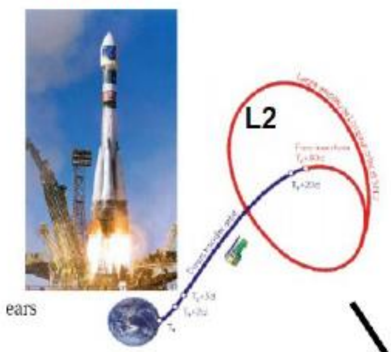
The Euclid survey (II)



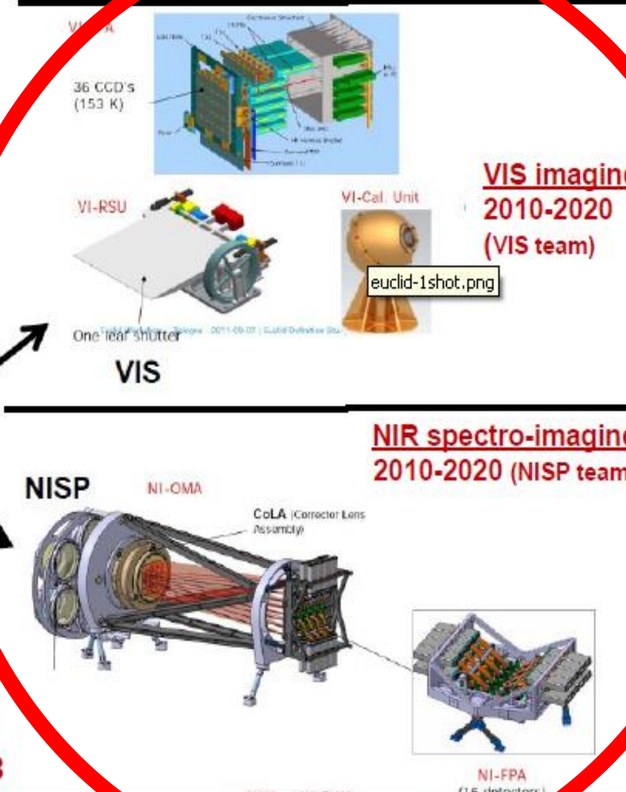
Sky coverage – Mollweide projection of the Euclid reference survey for a 6-years nominal mission – different colours indicate different years during the mission.

Euclid at a Glance

Soyuz@Kourou
Q2 2020



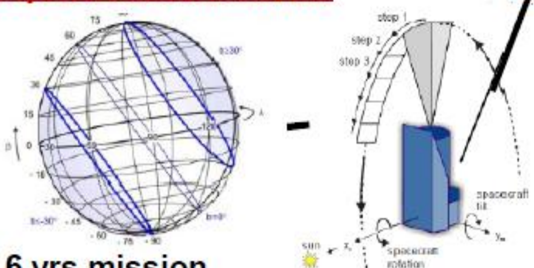
PLM+SVM: 2010-2019



VIS imaging:
2010-2020
(VIS team)

NIR spectro-imaging
2010-2020 (NISP team)

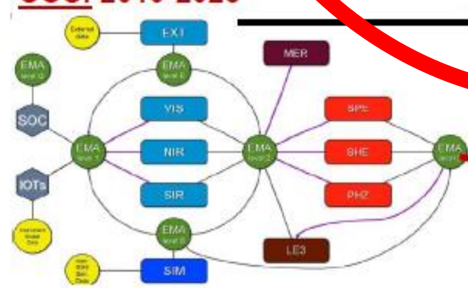
Surveys: 2010-2028 (Survey WG)



6 yrs mission

- Commisionning – Sc. Verif.
- Euclid nominal in opération: 5.5 yrs of Euclid Wide+Deep
- Euclid+: Additional surveys: SNIa, mu-lens, Milky Way?

SGS: 2010-2028



SWG:
2019-2028

20-30 PB data processing (EC-SGS team) – Science analyses

Euclid instruments

- **VIS**
  - visible band, spanning the range 550-900 nm
  - 6×6 CCDs (4k×4k pixels each) with 0.10 arcsec pixel platescale, geometric field of 0.55 deg^2 including gaps between detectors
  - designed to measure the shapes of galaxies with better than 0.16 arcsec (FWHM)
- **NISP**
  - NISP employs 4×4 HgCdTe NIR detectors (2k×2k pixels each) with 0.3 arcsec per pixel covering an area of 0.5 deg^2.
    - in photometer mode NISP collects images in three filterbands (Y, J, H) covering the wavelength range 0.92-2.0 micron
    - in the slitless spectrometry mode, dispersion through grisms in the wavelength range 1.1-2.0 micron, constant spectral resolution Δ≈250
    - two kinds of grisms with different passbands: two blue grisms (1.1-1.45 micron) two red grisms (1.45-2.0 micron); both sets of grisms are mounted in 0° and 90° → orthogonal spectra to reduce the confusion due to overlapping spectra.

Datasets size

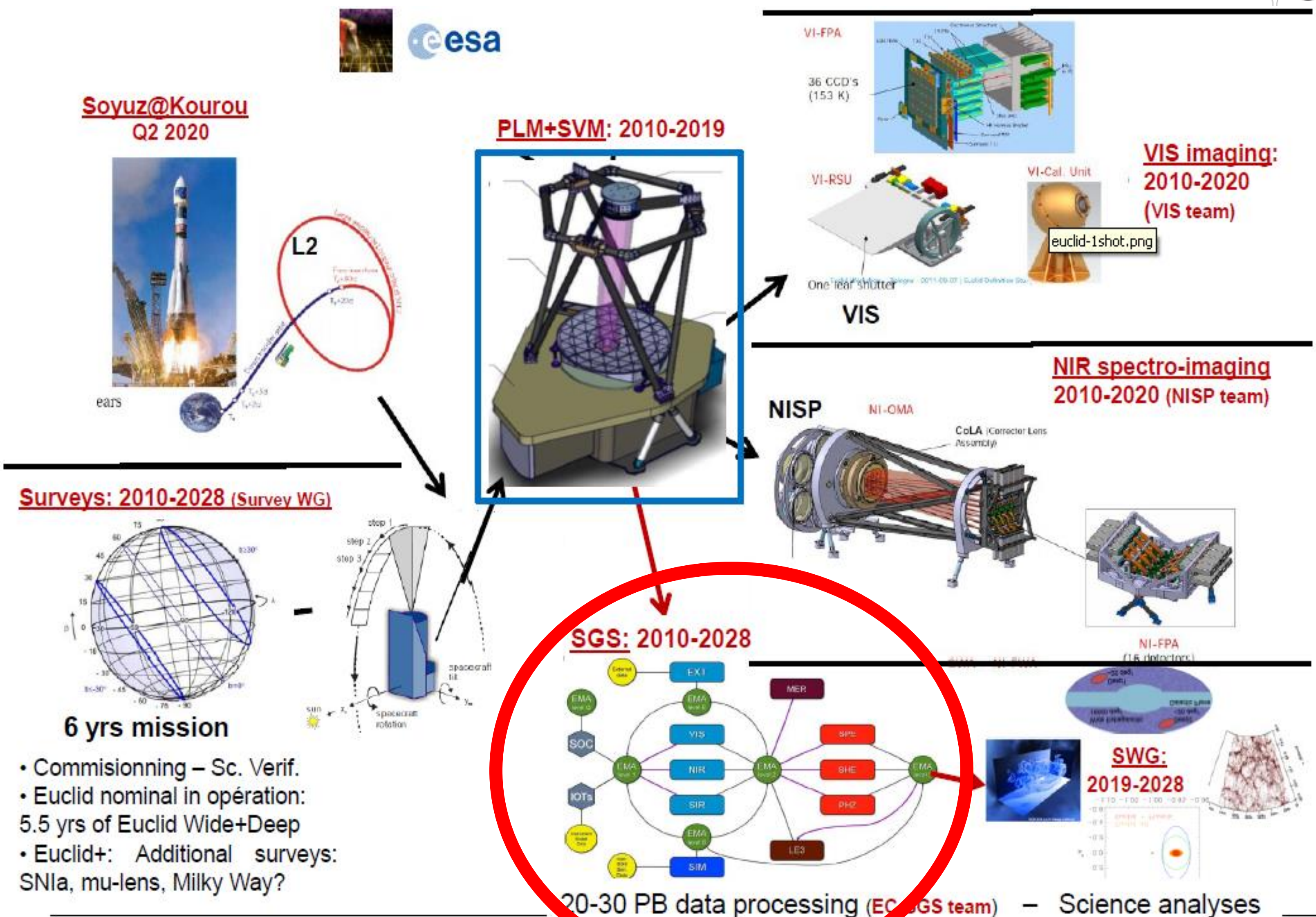
Euclid

- In nominal operations, science data are downloaded at **74 Mbps** using K-band (26 GHz) transponder – TM/TC use X-band (8 GHz)
- Daily Communication and data Transfer Period (DCTP) = **4 hrs/day**
- **≈ 100 GB/day** of compressed data downlinked
- Depending on the compression rate, **≈ 100 TB/year** of Euclid of raw science telemetry → **≥ 20 PB** of science data during nominal lifetime

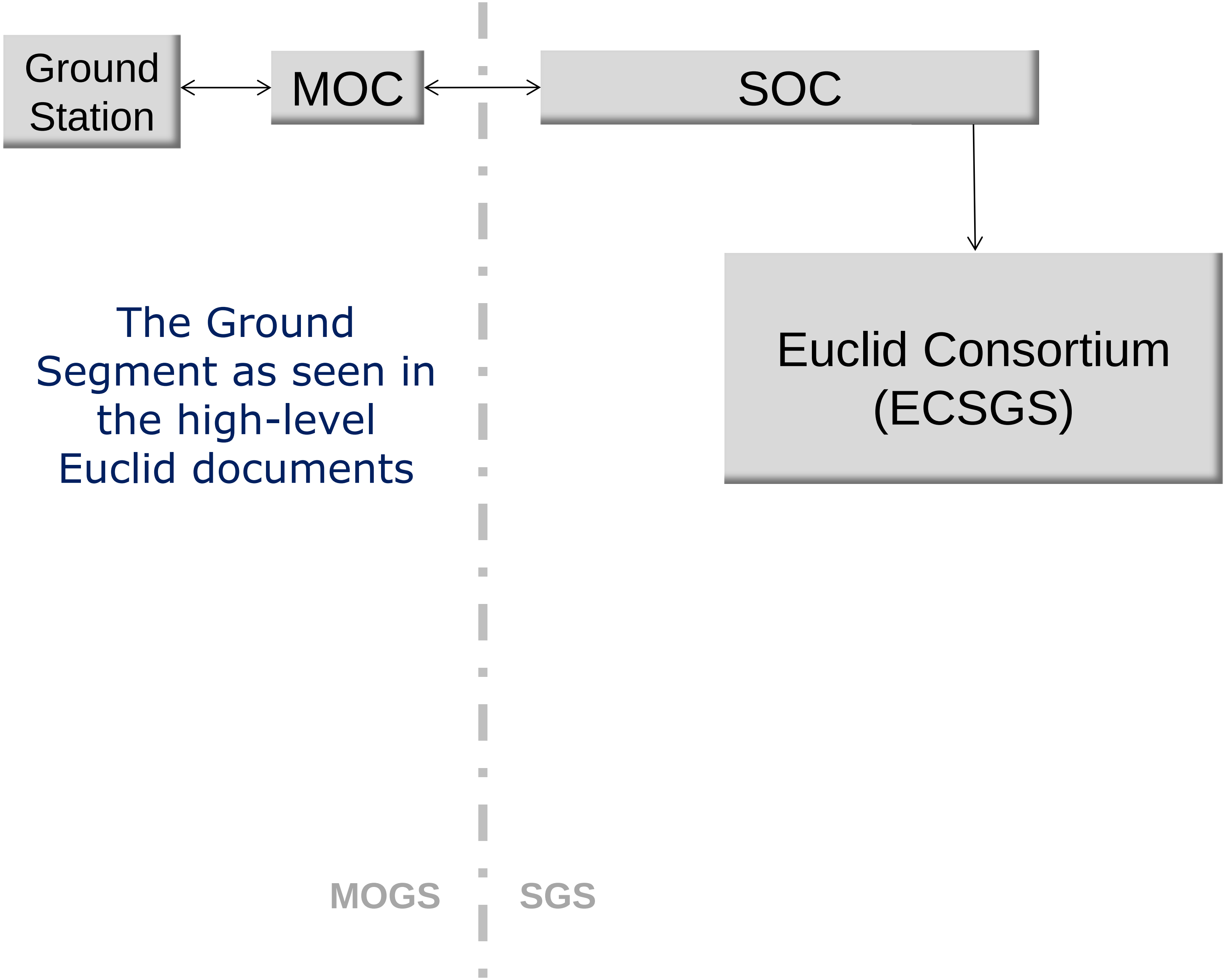
External data

- The data collected by the VIS and NISP instruments need to be complemented with **external data from ground-based surveys** to derive the photometric redshifts to the required 5% precision
- External and Euclid data are merged to obtain the survey catalogues for weak lensing, galaxy clustering and other cosmological probes
- Estimated size of currently-foreseen external data **≈ 50 PB** (bound to increase as new external surveys are added to DES and KiDS)

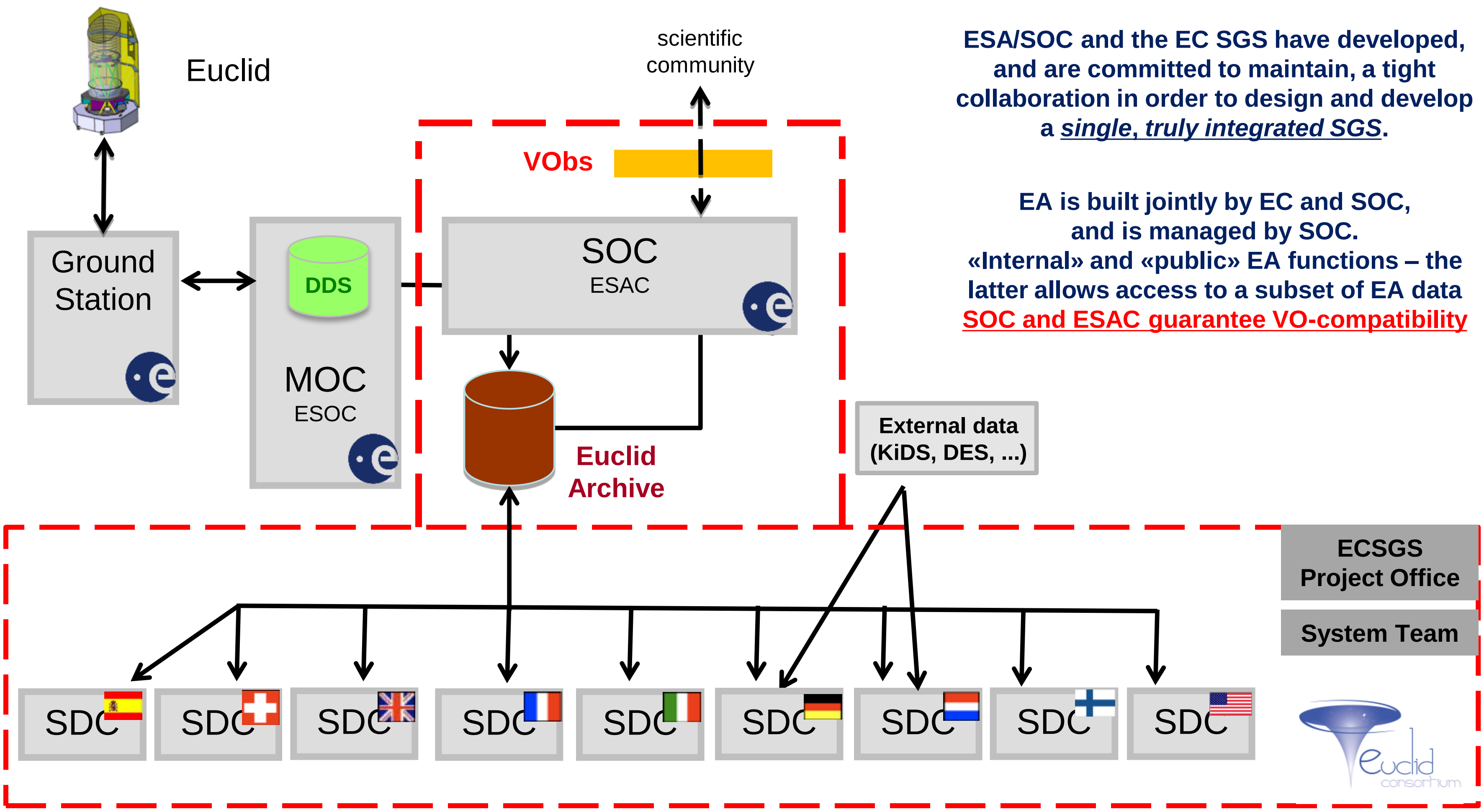
Euclid at a Glance



The Ground Segment as seen in the high-level Euclid documents



The Ground Segment at a glance

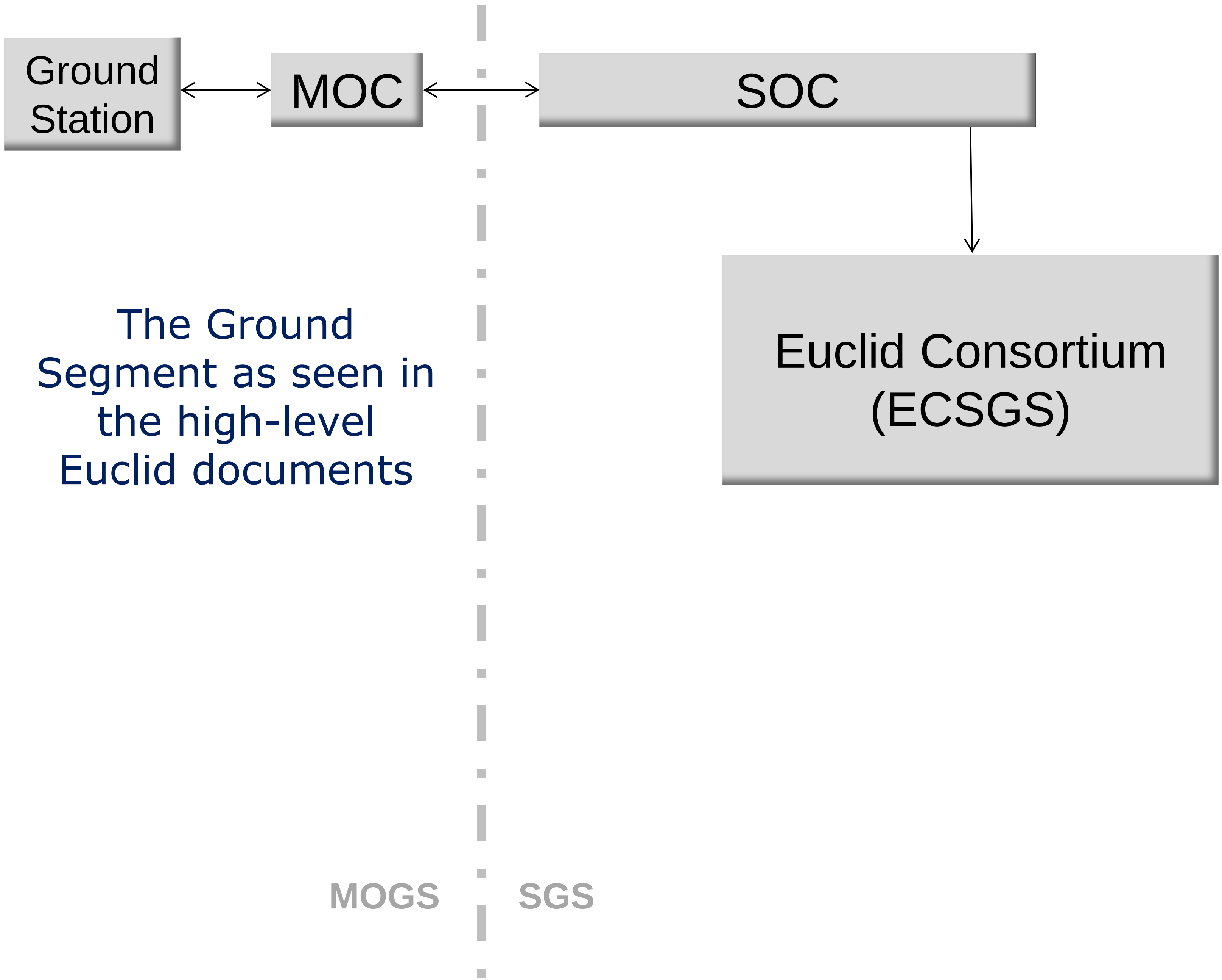


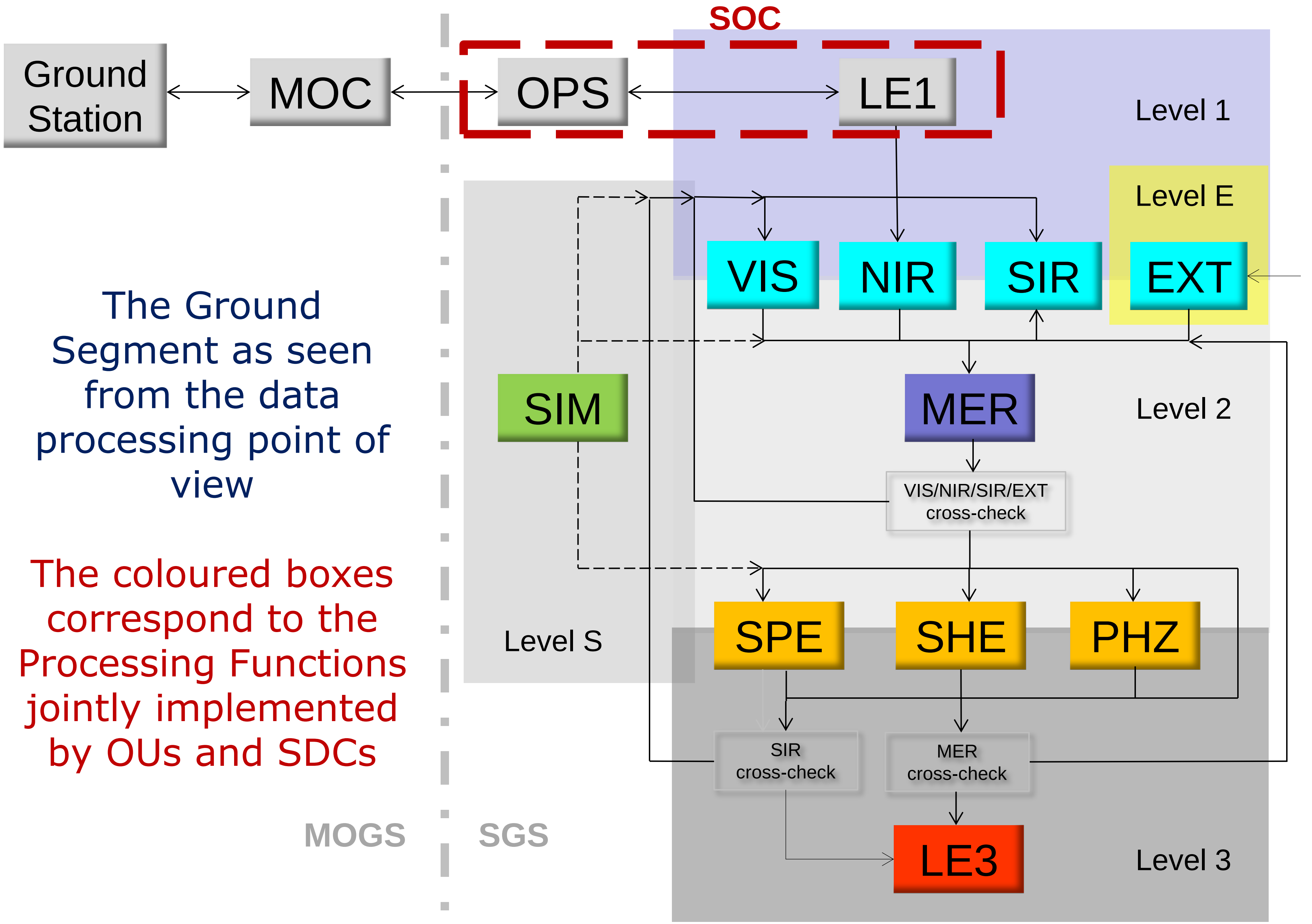
ESA/SOC and the EC SGS have developed, and are committed to maintain, a tight collaboration in order to design and develop a *single, truly integrated SGS*.

EA is built jointly by EC and SOC, and is managed by SOC.
«Internal» and «public» EA functions – the latter allows access to a subset of EA data
SOC and ESAC guarantee VO-compatibility

This is an institutional view of the GS

The Ground Segment as seen in the high-level Euclid documents



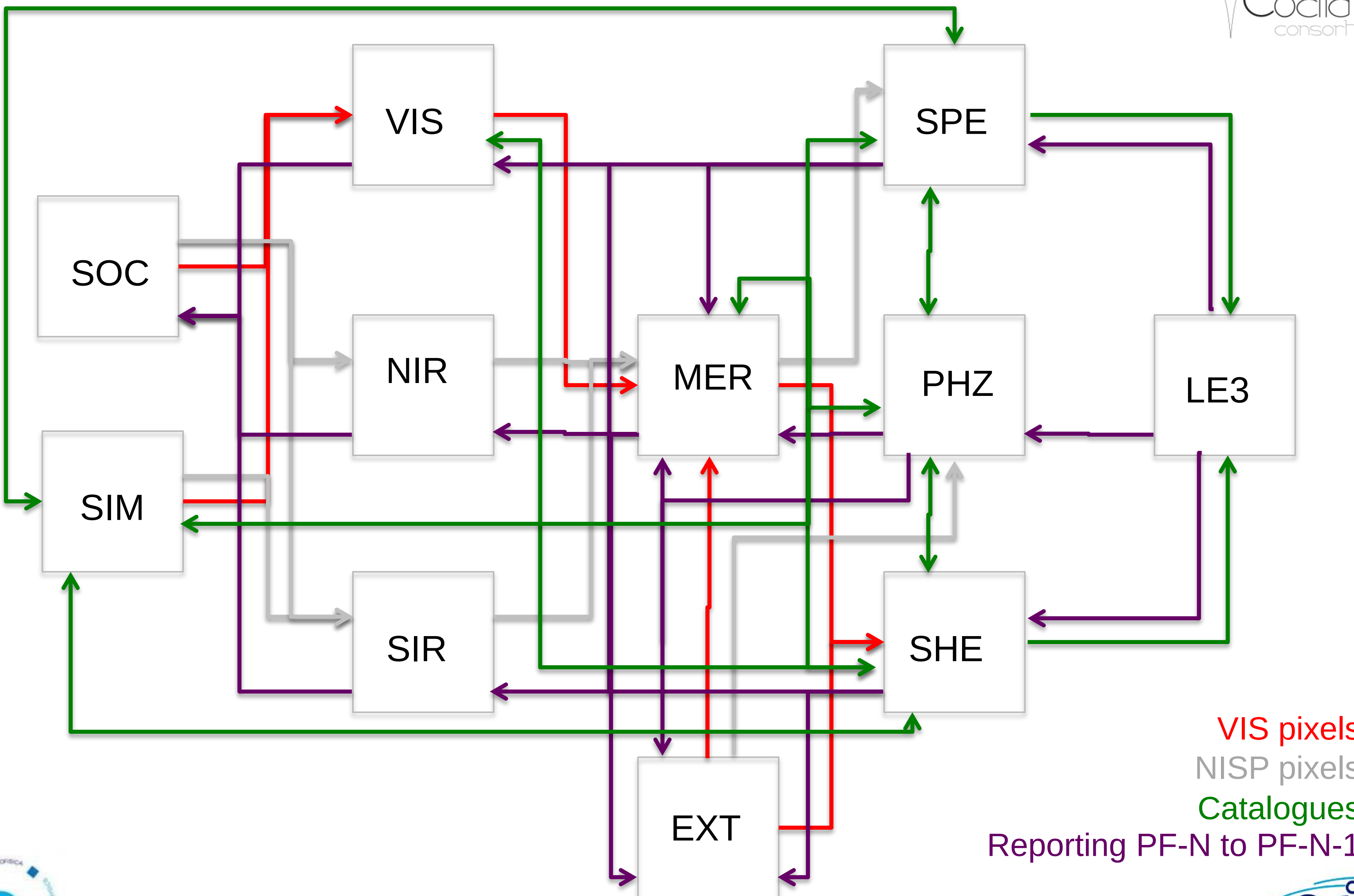


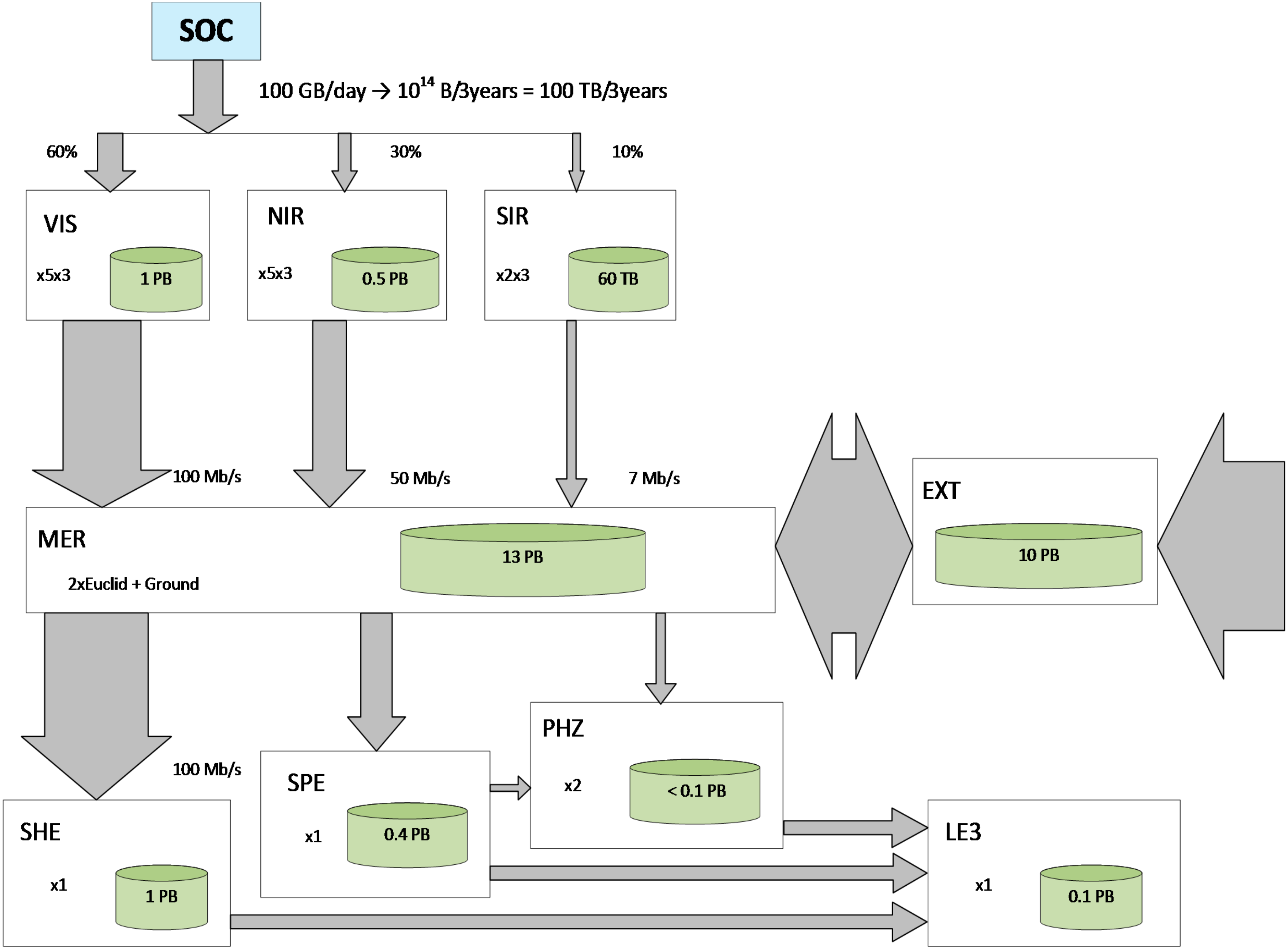
The Ground Segment as seen from the data processing point of view

The coloured boxes correspond to the Processing Functions jointly implemented by OUs and SDCs

This is a functional view of the SGS

PF-to-PF Relationships



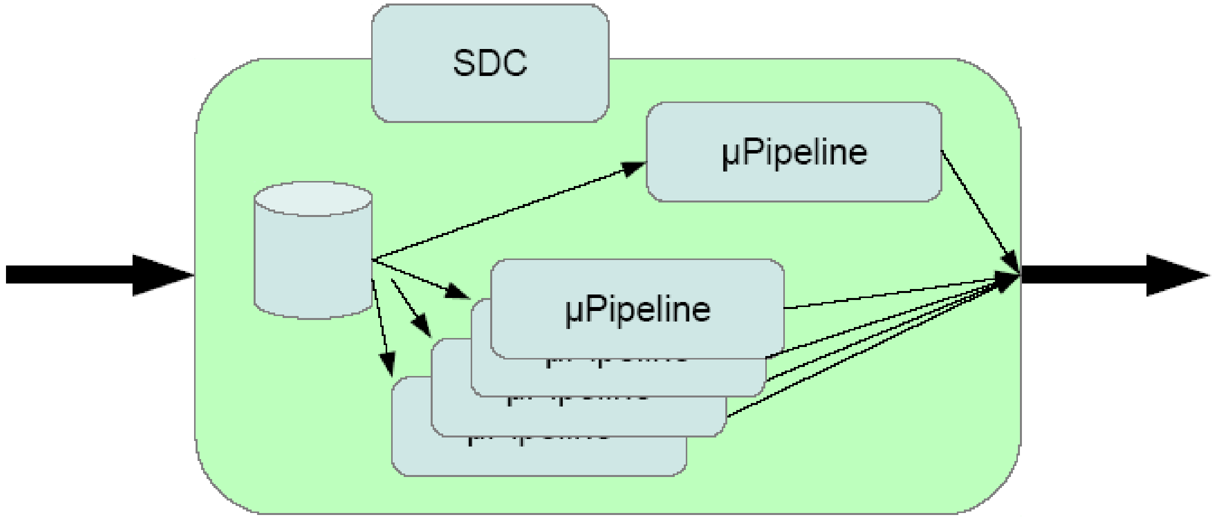


SOC
100 GB/day → 10^{14} B/3years = 100 TB/3years
60%
30%
10%
VIS
x5x3
1 PB
NIR
x5x3
0.5 PB
SIR
x2x3
60 TB
100 Mb/s
50 Mb/s
7 Mb/s
EXT
10 PB
MER
2xEuclid + Ground
13 PB
100 Mb/s
PHZ
x2
< 0.1 PB
SPE
x1
0.4 PB
SHE
x1
1 PB
LE3
x1
0.1 PB

SGS Key Architecture Concepts

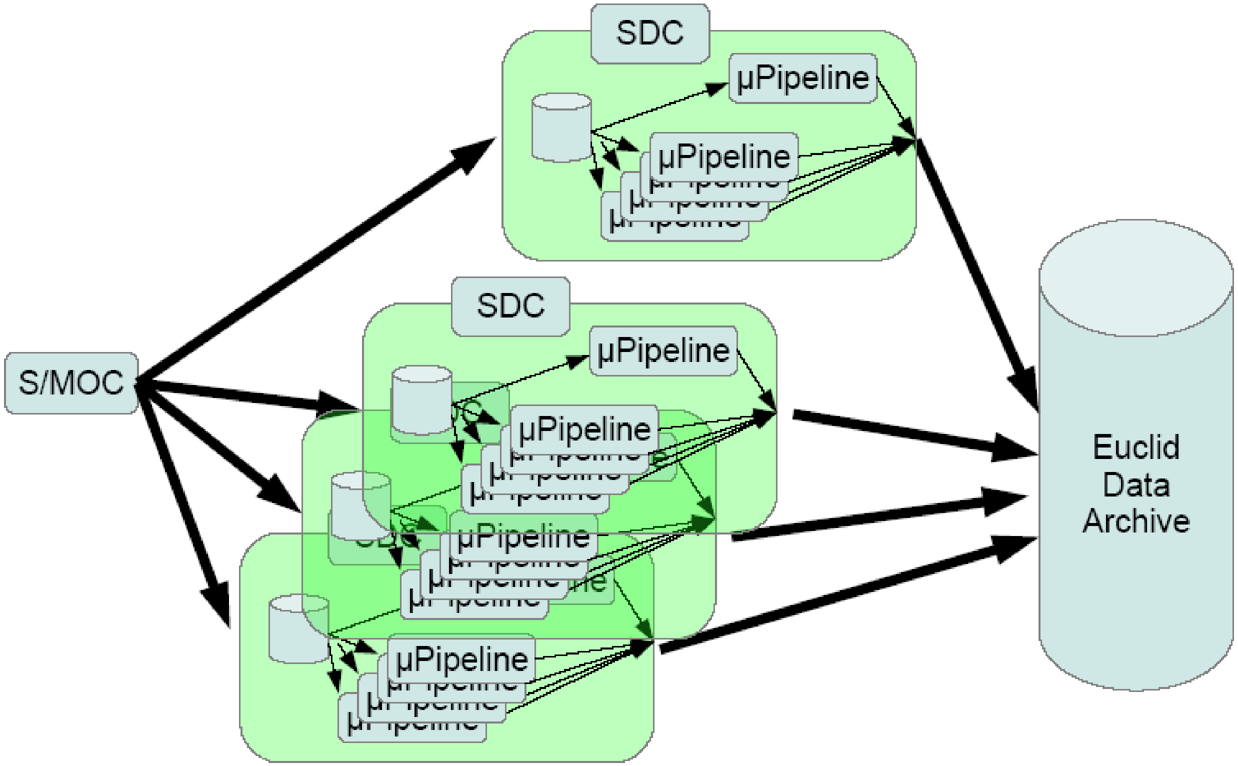
- **We need a distributed data-centric information system**
- No single dedicated SDC (political and cost reasons)
- *"Move the code, not the data"*
  - run the pipeline where the main input data is stored
- Distributed data and processing
  - each SDC is both a processing and a storage node
- No specialised SDC: any pipeline runs on any SDC
  - each SDC runs the same code through virtualisation
- Centralised information repository → separation of metadata from data

Architecture: topology (I)



- "μpipelines": full lower level processing on patches of sky built from tiles (minimal processable set of data covering a given sky area), up to the preparation of catalogues of objects (MER included)
- Higher level of processing are based on data cross-matching

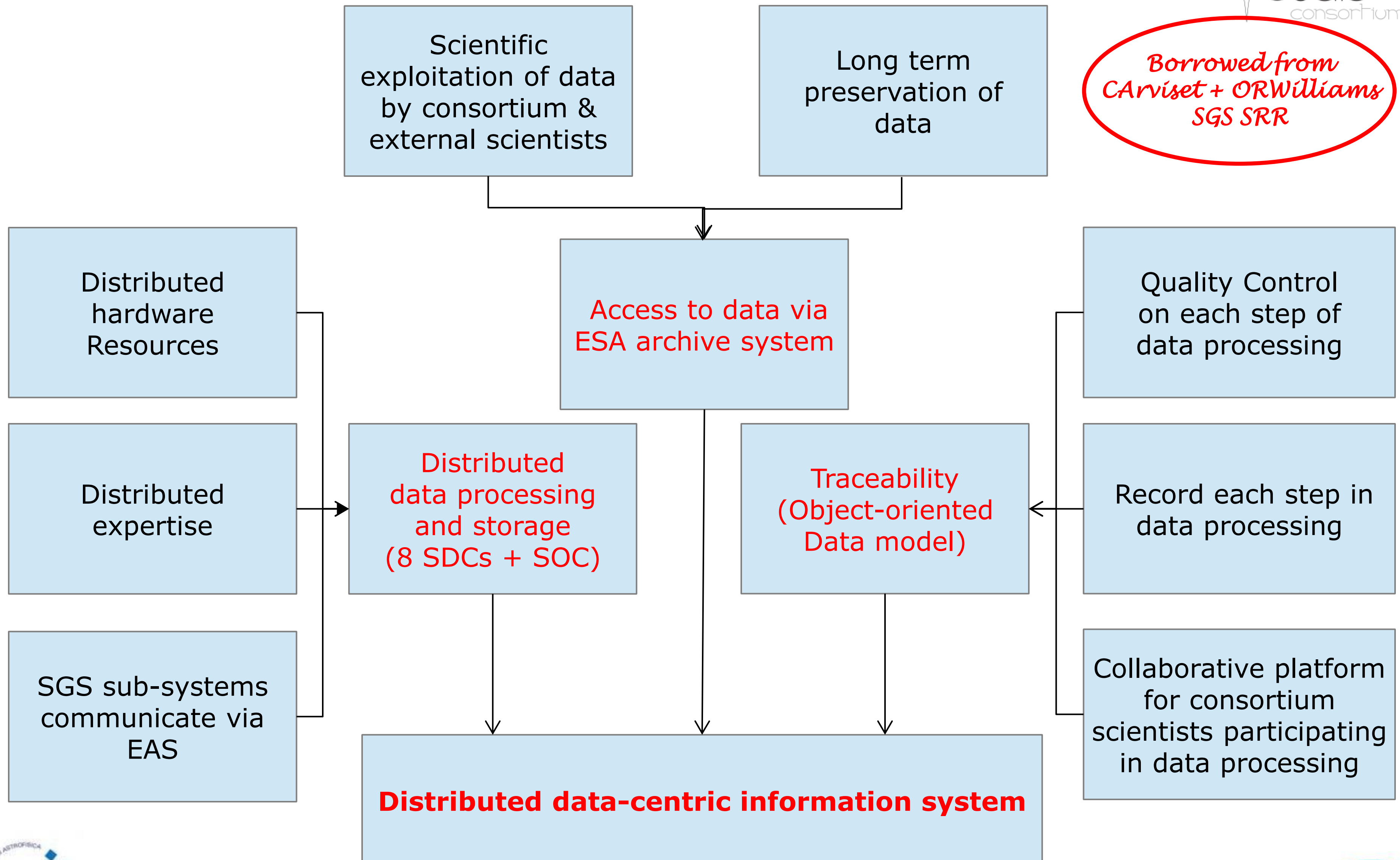
Architecture: topology (II)



The architecture deals with complexity, while being:

- Robust (predictable, able to recover from errors and unexpected behaviour)
- Reliable (produce the same results given the same input, no exceptions)
- Scalable (cope with changing demand, scale, no practical upper limit)
- Maintainable (same software everywhere)

Archive: principles



Borrowed from CArviset + ORWilliams SGS SRR

Euclid Archive System

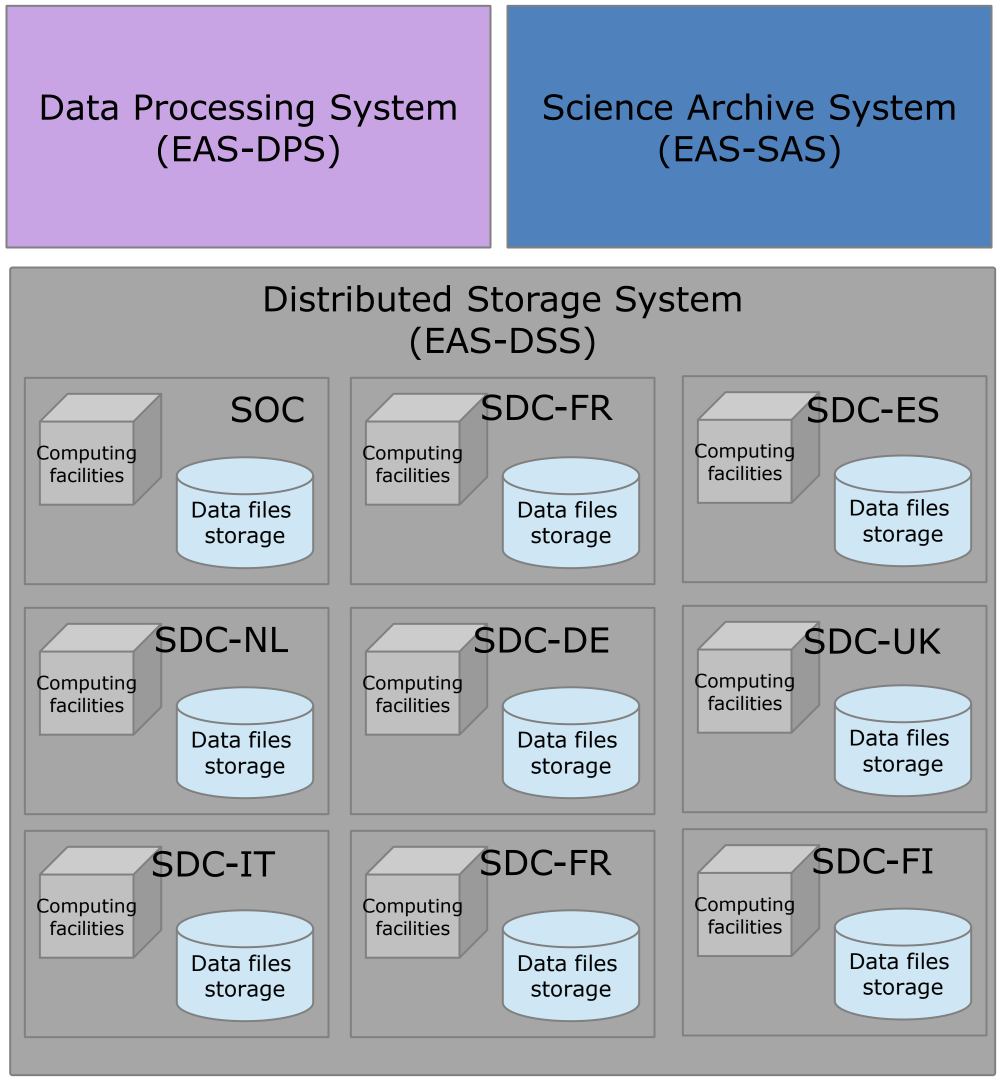
- Use cases discussed within EAUG → requirements
- Separation between EC-internal and public
  - EAS-DPS (Data Processing System) on going prototype systems are currently based on RUG expertise and is being replicated "as is" between RUG and ESAC
    - Same Oracle DB, same software systems (Python)
    - Serving data processing → Euclid Common DM
  - EAS-DSS (Distributed data Storage System)
    - Across SDCs and SOC
    - Serves data processing and science exploitation
    - Policy for science data copy from SDCs to ESAC SOC can be defined at a later stage
  - EAS-SAS (Science Archive System) will re-use ESAC Science Archives expertise e.g. Gaia Archive ➔ VO-compliance
    - Most probably different DB (e.g. PostgreSQL), Science Exploitation DM
    - Serving Science Exploitation use cases
- Proof of Concept with alternatives, lighter and performant technologies
  - EAS may well take advantage of WP3 work in ASTERICS (H2020)

EAS Design

Integral part of SGS

Euclid Common Data Model

Consortium Services



Part of ESA archives

Science Exploitation Data Model

Scientific Exploitation Services

Borrowed from CArviset + ORWilliams SGS SRR

Distributed storage system - storage nodes in each SDC and SOC
File storage with additional interfaces for Euclid services (cut-out, visualization)

SGS status – "challenges"

IV&V strategy based on challenges ensuring that the complete SGS infrastructure can deploy, install, test and run new software releases.

- Infrastructure challenges already (#3) demonstrated the capability to:
  - fetch, on the basis of the metadata provided by EAS prototype (in SDC-NL), the pipelines input data in the local SDC storage area
  - launch simulators jobs across clusters or dedicated nodes, in accordance with PPOs defined remotely (through Jenkins) or locally (by each SDC leader) – orchestration mock-up
  - produce and store output data into the local SDC storage area
  - send the appropriate metadata to EAS prototype in SDC-NL
- Next infrastructure challenge (#6): fixing and improving; performance & scalability testing; Level 1 data distribution in all SDCs; Docker and CernVM/FS as consortium level virtualization solutions
- Next "scientific" challenge (#2): SIM+VIS+NIR+SIR (and optionally SPE) pipelines prototypes on any SDC connected to the Euclid Archive on limited number (~100) of simulated frames

In summary …

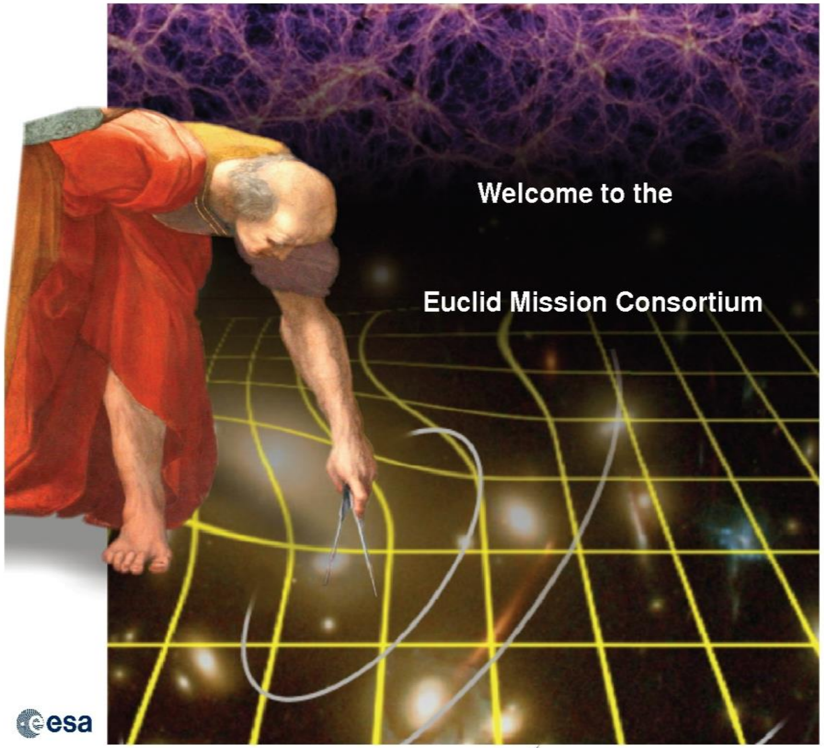
- Euclid products (for cosmology and legacy science) will be available through ESAC → VO-compliant science data access for exploitation
- The accuracy required by data processing requires complex loops and iterations across the different Processing Functions → distributed data-centric information system + distributed processing
- Two separate but interfaced/interoperable Data Models:
  - Euclid Common Data Model (e.g. containing data lineage)
  - Science Exploitation Data Model (VO-compliant)
- ASTERICS (funded H2020 project) could be beneficial for Euclid
  - improved efficiency in EAS-DPS databases
  - new efficient methods for executing distributed workflows
  - improved Auth&Auth methods/mechanisms
  - support IVOA in the implementation of extended DMs

  but also for Euro-VO and IVOA

Who is who in the Euclid SGS

- Giuseppe Racca (ESA/ESTEC) – Euclid Project Manager
- René Laureijs (ESA/ESTEC) – Euclid Project Scientist
- Yannick Mellier (IAP Paris) – Euclid Consortium Lead
- John Hoar (ESA/ESAC) – SOC Development Manager
- Euclid Consortium SGS Project Office
  - FP → Andrea Zacchei (INAF-OATs) – ECSGS Manager
  - Christophe Dabin (CNES) – System Team Lead and ECSGS Deputy
  - Marc Sauvage (CEA Saclay) – ECSGS Scientist
  - Claudio Vuerli (INAF-OATs) – PA/QA Lead
  - Oriana Mansutti (INAF-OATs) – Configuration Control Lead
  - Anna Gregorio (UniTS) – Instruments Operations Coordinator

+ national SDCs, multi-national OUs, SGS System Team ...



Thank you for your attention

fabio.pasian@inaf.it