Euclid Big Data From Space

A. Zacchei

INAF-OATs

on behalf of the Euclid Collaborations

Euclid in a nutshell

- Euclid is a space-borne survey mission dedicated to investigate the origin of the Universe's accelerating expansion and the nature of dark energy, dark matter and gravity. Euclid will characterise the signatures of dark energy on the 3D distribution of cosmic structures. In 2012, Euclid was approved as the second Medium Class mission (M2) in the Cosmic Vision Programme.
- Euclid will be launched on May 2021 (TBC) and operate (baseline) for about 6 years.
- It will carry on board two instrument VIS (Visible Imager) and NISP (Near Infrared Spectrum Photometer).
- Euclid is a ESA Mission where the Data analysis and Instrument realization are in charge to a Consortium. The Consortium has the responsibility to reach the scientific goals and to provide to ESA the agreed *products.* ESA, through its archive system at ESAC, will be then in charge of making those products available.

Euclid Top Level Science Requirements

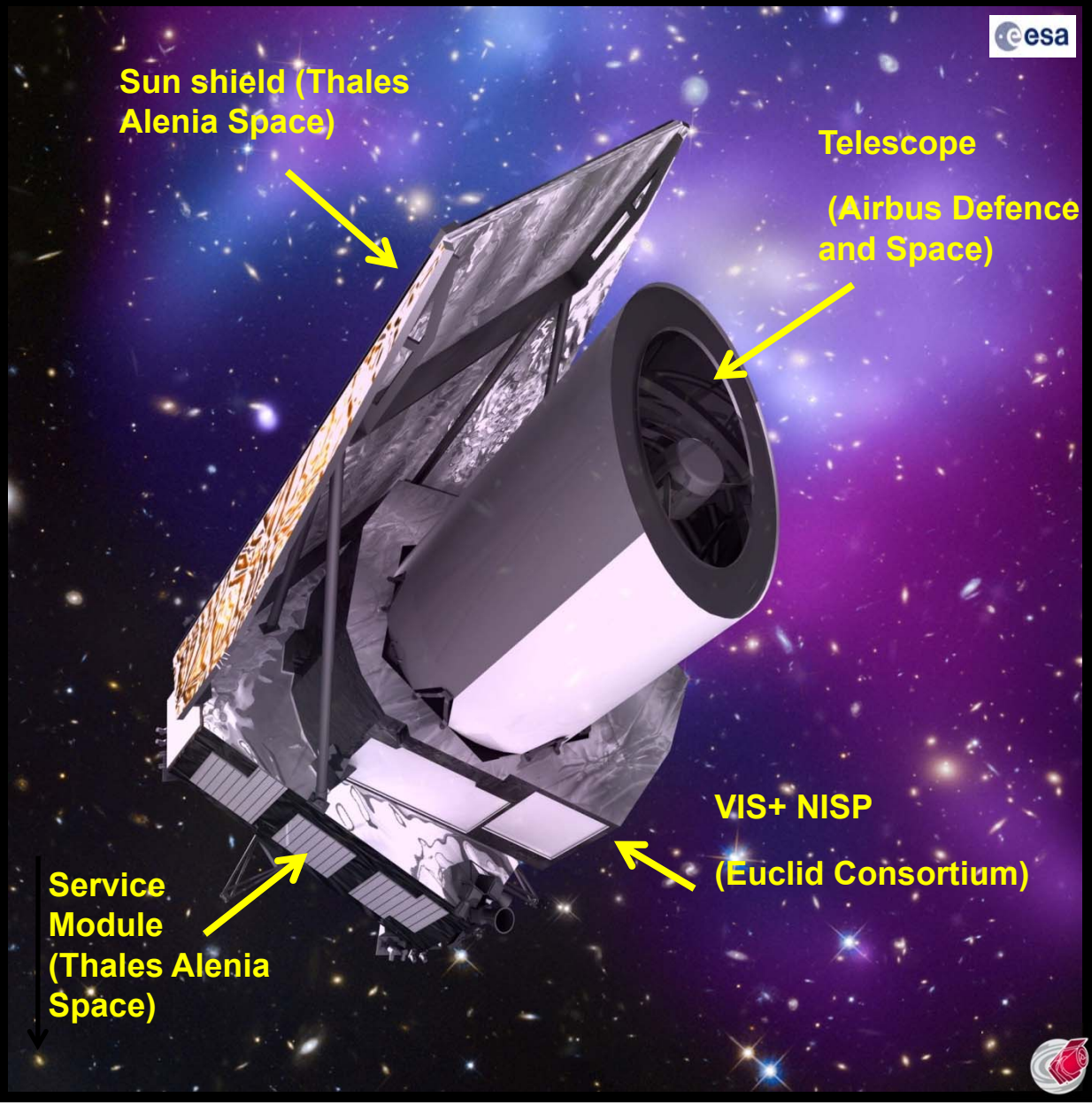
Sector	Euclid Targets
Dark Energy	• Measure the cosmic expansion history to better than 10% in redshift bins $0.7 < z < 2$. • Look for deviations from $w = -1$, indicating a dynamical dark energy.
Test Gravity	• Measure the growth index, γ, with a precision better than 0.02 • Measure the growth rate to better than 0.05 in redshift bins between $0.5 < z < 2$. • Separately constrain the two relativistic potentials Ψ, Φ. • Test the cosmological principle
Dark Matter	• Detect dark matter halos on a mass scale between 10^8 and $>10^{15}$ M_{Sun} • Measure the dark matter mass profiles on cluster and galactic scales • Measure the sum of neutrino masses, the number of neutrino species and the neutrino hierarchy with an accuracy of a few hundredths of an eV
Initial Conditions	• Measure the matter power spectrum on a large range of scales in order to extract values for the parameters σ_8 and n to a 1-sigma accuracy of 0.01. • For extended models, improve constraints on n and α wrt to Planck alone by a factor 2. • Measure a non-Gaussianity parameter : f_{NL} for local-type models with an error < +/-2.

ESA Euclid mission

- - Total mass satellite :
- 2 200 kg
- - Dimensions:
- 4,5 m x 3 m

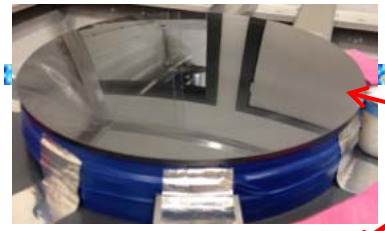
- - Launch: middle of 2021 by a Soyuz rocket from the Kourou space port
- Euclid placed in L2

- - Survey: 6 years,

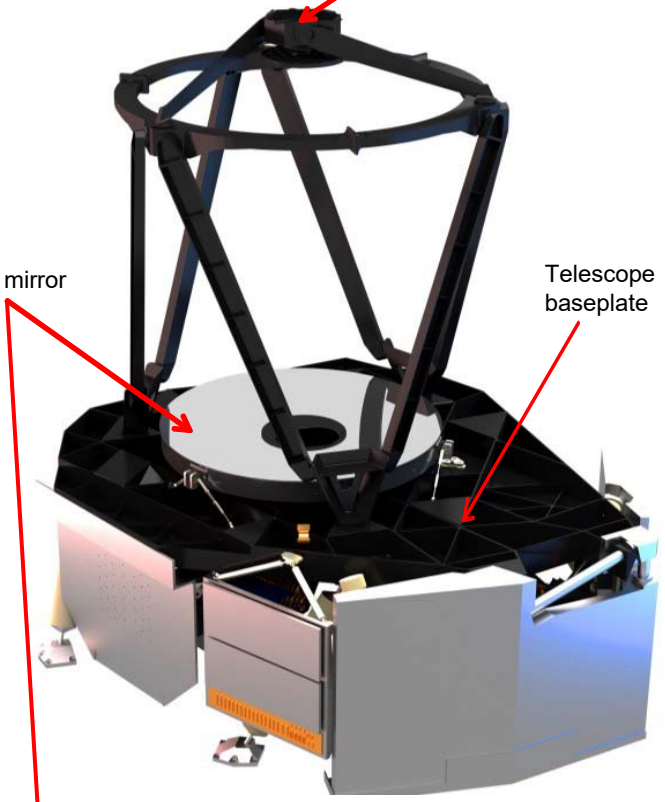


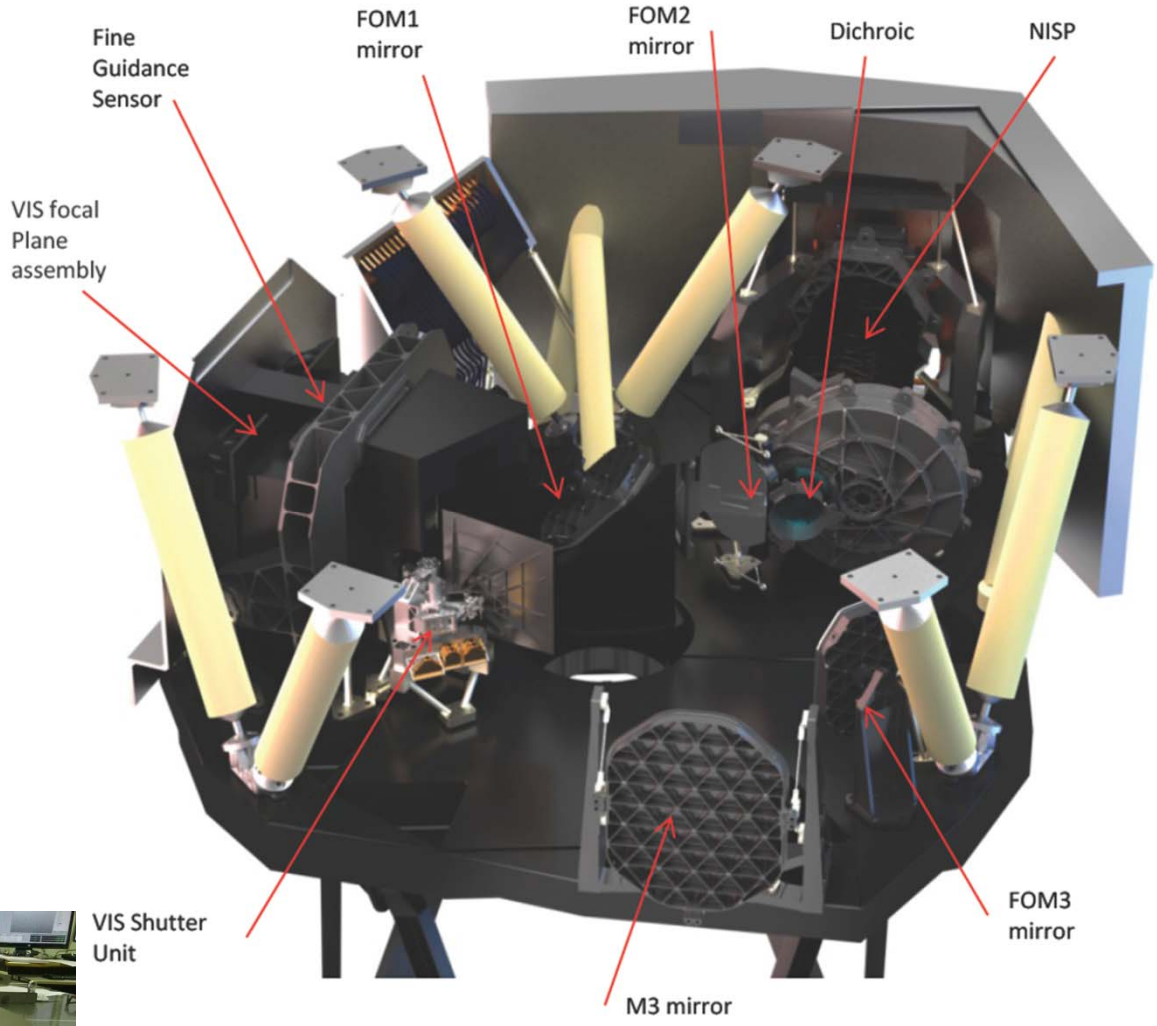
PLM, flight hardware, scientific instruments

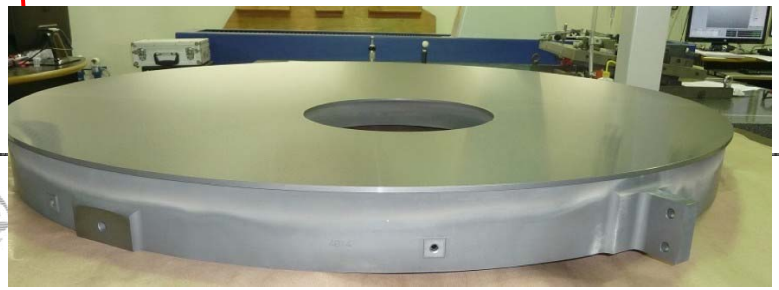
From Thales Alenia Italy, Airbus DS, ESA Project office and Euclid Consortium



M2 mirror



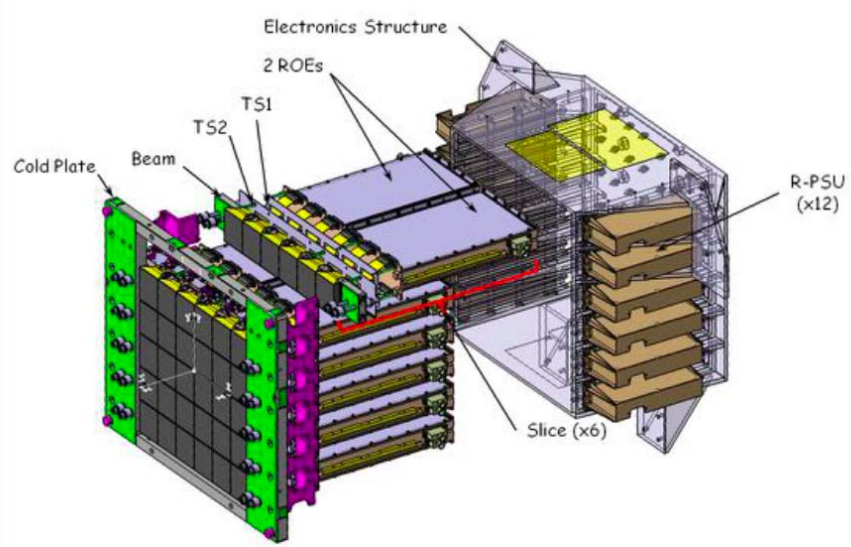




PLM CDR was in July 2017

VIS

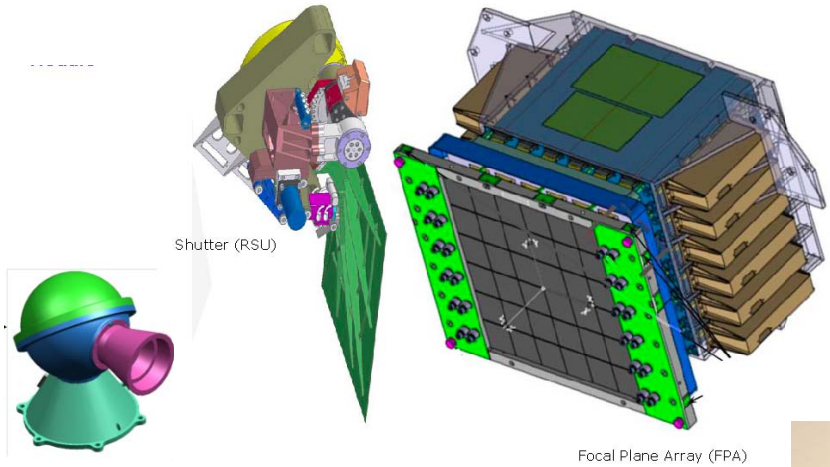
VIS CDR concluded in 2017

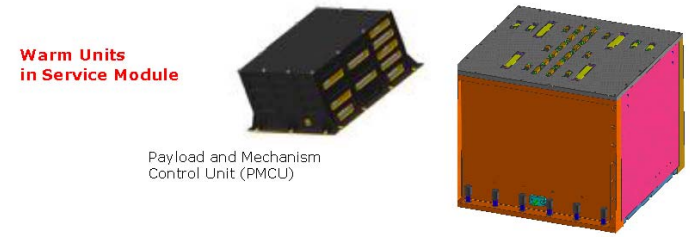


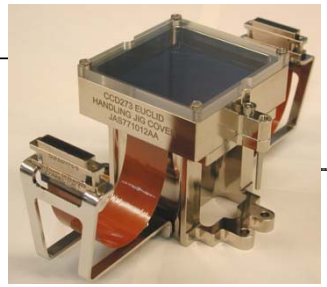
Courtesy: S. Pottinger, M. Cropper and the VIS team

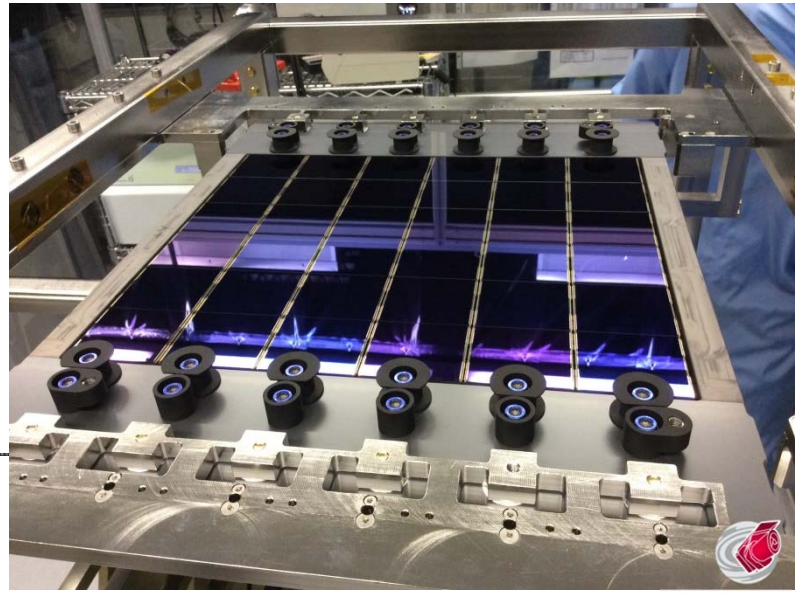
Table 1: VIS and weak lensing channel characteristics

Spectral Band	550 – 900 nm
System Point Spread Function size	≤0.18 arcsec full width half maximum at 800 nm
System PSF ellipticity	≤15% using a quadrupole definition
Field of View	>0.5 deg^2
CCD pixel sampling	0.1 arcsec
Detector cosmetics including cosmic rays	≤3% of bad pixels per exposure
Linearity post calibration	≤0.01%
Distortion post calibration	≤0.005% on a scale of 4 arcmin
Sensitivity	$m_{AB} \geq 24.5$ at 10σ in 3 exposures for galaxy size 0.3 arcsec
Straylight	≤20% of the Zodiacal light background at Ecliptic Poles
Survey area	15000 deg^2 over a nominal mission with 85% efficiency
Mission duration	6 years including commissioning
Shear systematic bias allocation	additive $\sigma_{sys} \leq 2 \times 10^{-4}$; multiplicative $\leq 2 \times 10^{-3}$





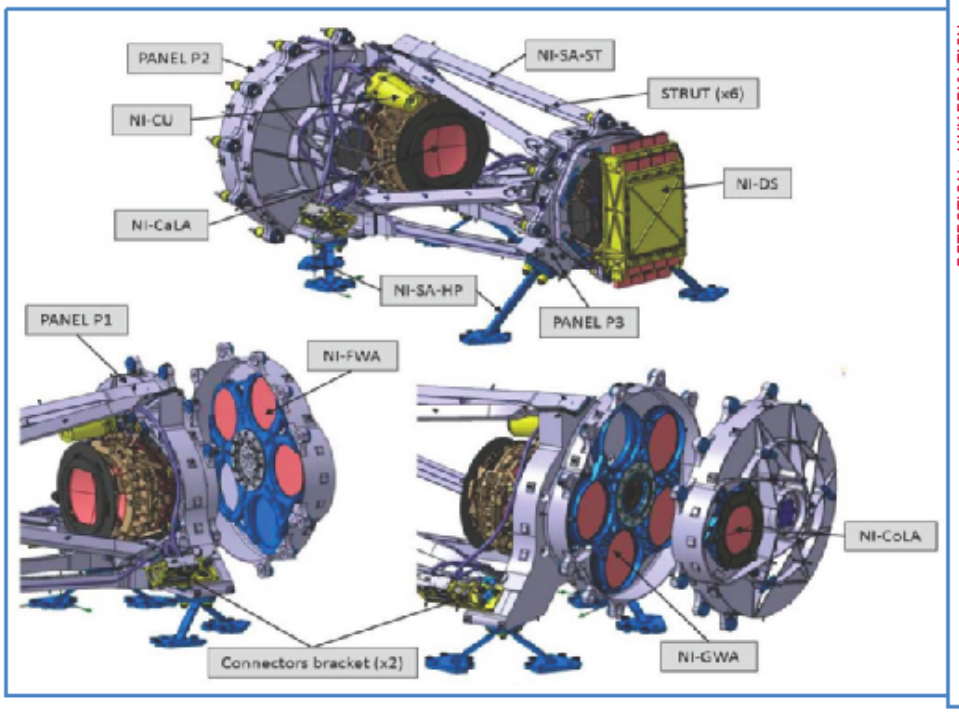


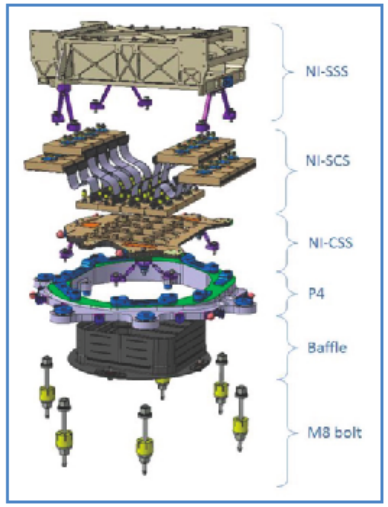


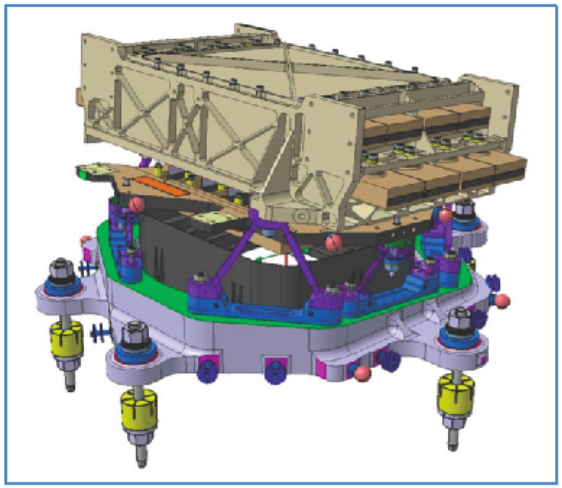
NISP

NISP CDR successful in Nov 2016

Courtesy: T. Maciaszek and the NISP team

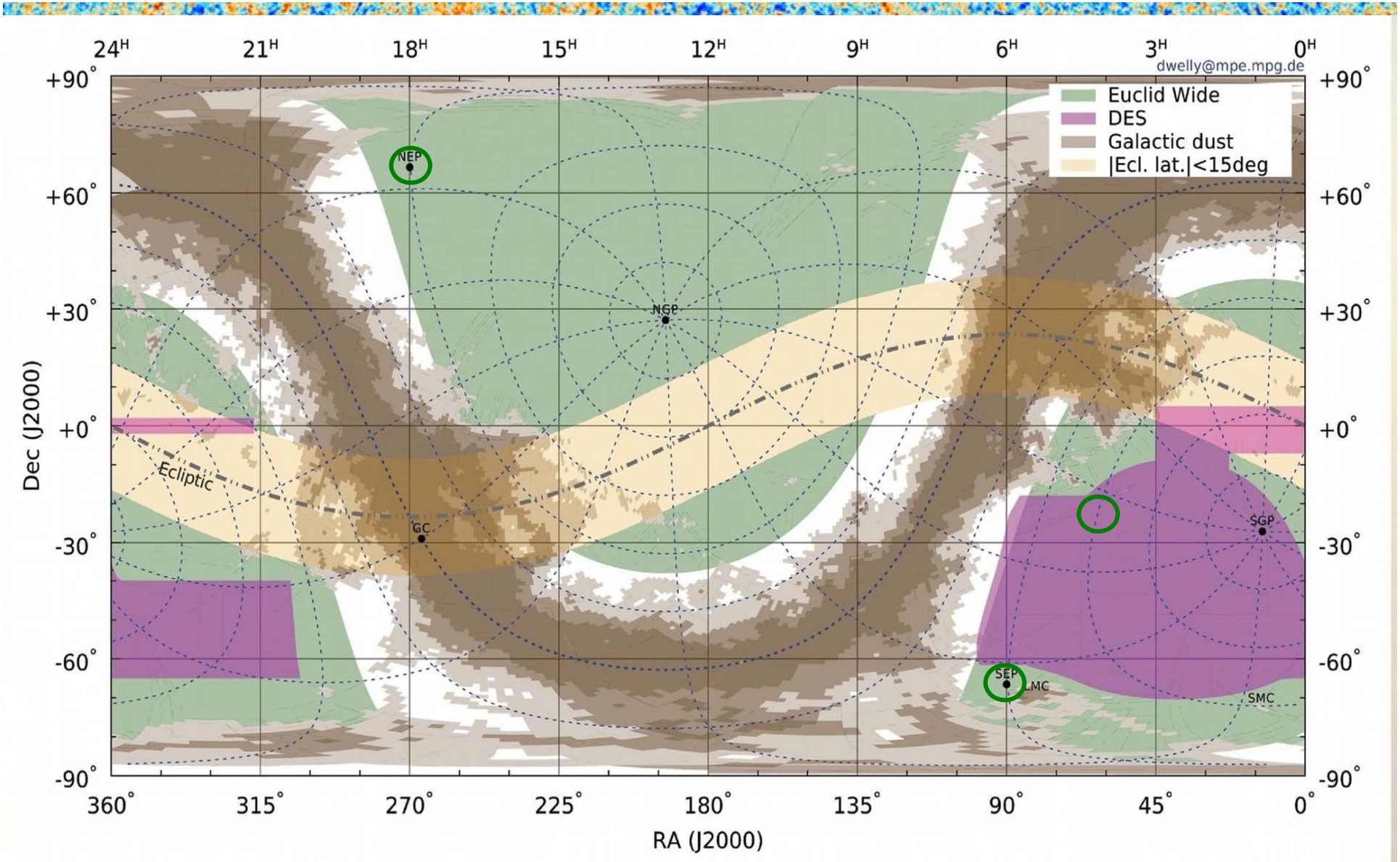






- FoV: 0.55 deg^2
- Mass : 159 kg
- Telemetry: < 290 Gbt/day
- Size: 1m x 0.5 m x 0.5 m
- 16 2kx2K H2GR detectors
- 0.3 arcsec pixel on sky
- Limiting mag, wide survey AB : 24 (5 σ)
- **3 Filters**:
- Y (950-1192nm)
- J (1192, 1544nm)
- H (1544, 2000nm)
- **4 grisms**:
- 1B (920 – 1300) , 1 orientation 0°
- 3R (1250 – 1850), 3 orientations 0° , 90° , 180°

Euclid Wide: Sky coverage



From R. Scaramella Survey WG

Euclid Wide and Deep Surveys

- **Euclid Wide:**
  - 15000 deg^2 outside the galactic and ecliptic planes
  - 12 billion sources (3-σ)
  - 1.5 billion galaxies with
    - Very accurate morphometric information (WL)
    - Visible photometry: (u), g, r, i, z , (R+I+Z) AB=24.5, 10.0 σ +
    - NIR photometry : Y, J, H AB = 24.0, 5.0σ
    - Photometric redshifts with 0.05(1+z) accuracy
  - 35 million spectroscopic redshifts of emission line galaxies with
    - 0.001 accuracy
    - Halpha galaxies within 0.7 < z < 1.85
    - Flux line: 2 . 10^{-16} erg.cm^{-2}.s^{-1} ; 3.5σ

- **Euclid Deep:**
  - 1x10 deg^2 at North Ecliptic pole + 1x20 deg^2 at South Ecliptic pole

    + 1x10 deg^2 South close to Equatorial area
  - 10 million sources (3-σ)
  - 1.5 million galaxies with
    - Very accurate morphometric information (WL)
    - Visible photometry: (u), g, r, i, z , (R+I+Z) AB=26.5, 10.0 σ +
    - NIR photometry : Y, J, H AB = 26.0, 5.0σ
    - Photometric redshifts with 0.05(1+z) accuracy
  - 150 000 spectroscopic redshifts of emission line galaxies with
    - 0.001 accuracy
    - Halpha galaxies within 0.7 < z < 1.85
    - Flux line: 5 . 10^{-17} erg.cm^{-2}.s^{-1} ; 3.5σ

The Euclid Data Processing

- The SGS (Science Ground Segment) is essentially a distributed structure.
- Euclid will produce and use a big amount of data (estimated to be at the end of the mission of the order of 100 PB). It will be then essential to avoid excessive data transfer and **develop a structure where the code will be moved instead of the data**.
- A common Data Model and common infrastructure has been built → EACH Science Data Center should be able to run the same code → the sky can be divided in patches to be analyzed in parallel in different centers to minimize data transfer between SDCs (science Data Center).
- The Data processing pipeline in Euclid will be a series of Processing Functions: designed by the OUs (Organization Units, scientist), developed in collaboration between the OUs and SDC developers , integrated by the SDCs, and running on the SDCs infrastructure.

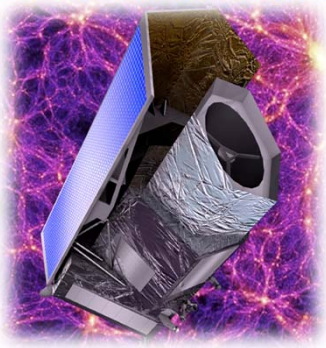
Euclid Operations: In Flight

euclid

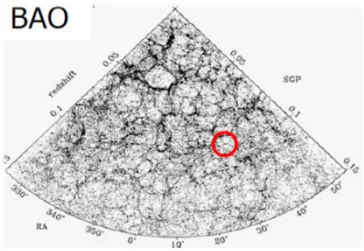
1.5 billion sources with shapes, 10 slices

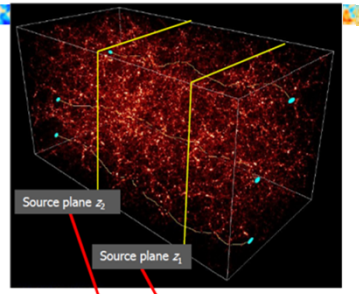
4 hrs/day
X-Band (8.0-8.4 GHz)–H/K
K-band (25.5-27 GHz)-Sci
100Mbs

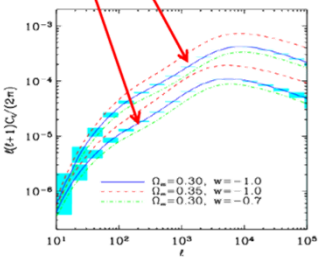
Deep Space Antenna
(Spain, Australia)



0110101
1110110







SDCs
IOTs

Mission Operation Center
(ESA/ESOC-Germany)

Science Operation Center
(ESA/ESAC–Spain)

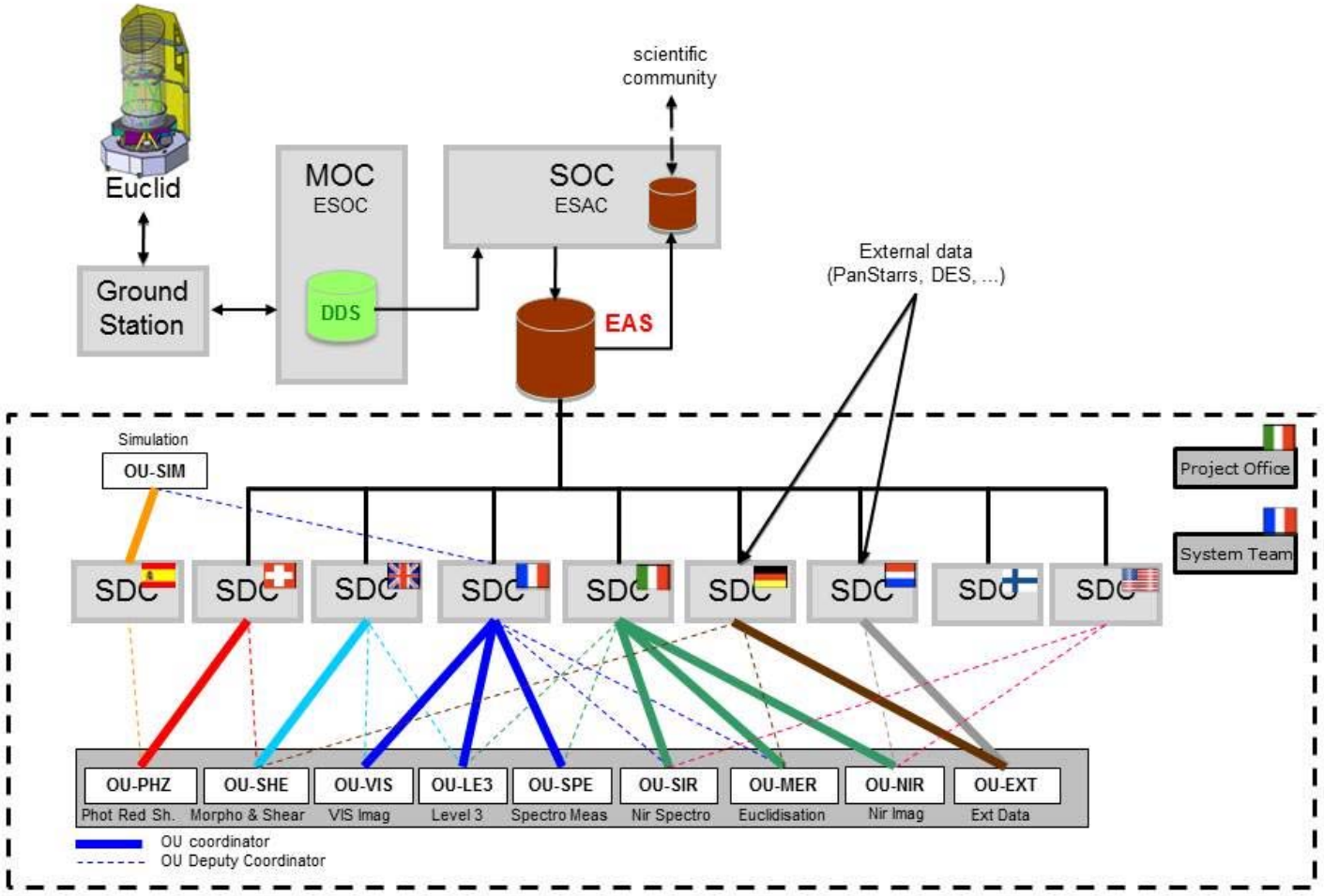
EAS
(archive)

esa euclid consortium

agenzia spaziale italiana

INAF ISTITUTO NAZIONALE DI ASTROFISICA

The Euclid Ground Segment



Working structures within the EC-SGS

For the data processing implementation phase, the EC-SGS consists mainly of two series of entities:

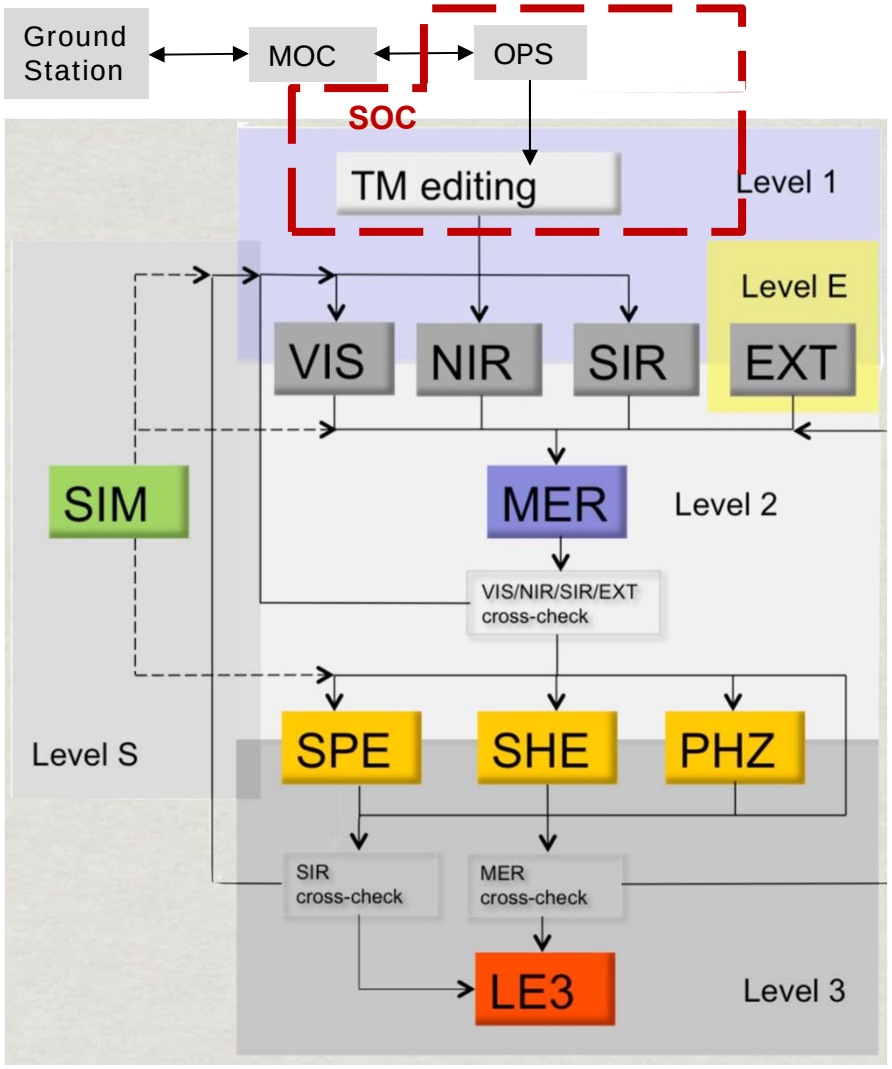
- The Organization Units (OUs)
- The Science Data Centers (SDCs)

The SDCs are built around existing national computing facilities, they gather IT support as well as developer expertise.
SDCs are in charge of the pipeline development (software & support architecture).
Production software will run in the SDCs.
The SDCs and SOC are the operational sites of the SGS.

The OUs group EC members according to their data processing expertise.
OUs are in charge of analyzing the science data processing requirements, and of producing the pre-integration version of the pipeline modules.
OUs will be in charge of future evolution of the pipeline during operations.

- Notable group coming from the SDCs: the System Team in charge of building the software infrastructure.
- OUs have been created following a plausible but a priori division of the pipeline in logical blocks.

The mapping of OUs on the pipeline



- VIS, NIR, EXT: production of fully calibrated photometric exposures from Euclid and ground-based surveys
- SIR: production of fully calibrated 1D spectra extracted from the NISP spectroscopic exposures.
- MER: production of a source catalog containing consistent photometric and spectroscopic measurements.
- PHZ: production of the photometric redshift for all catalogued sources.
- SPE: production of spectroscopic redshifts for all sources with spectra.
- SHE: measurements of galaxy shapes.
- LE3: production of all high-level science products.
- SIM: production of all the simulated data necessary to validate the data processing stages, and to calibrate observational or method biases.

Euclide Software Infrastructure

Architecture : M. Holliman (RoE)

- Infrastructure coordination with SDC's and sysAdmins
- Virtualization technology

Euclid Archive System (EAS) : A. Belikov (RuG) / S. Nieto (ESAC)

- EAS-DPS: the central metadata and data archive for Euclid (Data Processing System)
- EAS-DSS: the distributed file system deployed at the SDCs and SOC
- EAS-SAS: Science Archive System

Common tools : M. Poncet (CNES) and different Working Groups

- Scientific libraries, common libraries
- Langages : C++/PYTHON support, Coding standards, Tests, Documentation, Quality
- Continuous Development Environment development and deployment platform (CODEEN)

Infrastructure Abstraction Layer (IAL) : Martin Melchior (FHNW)

- Retrieval and execution of data processing and movement requests on SDC resources
- Implements the Interfaces between Pipelines, Euclid Archive and infrastructure

Data Model : C. Dabin (CNES)

- Rules : dictionary of types (XML), interfaces between components, data products definition and implementation
- Mission Data Base support

Monitoring and Control : A. Piemonte (MPA)

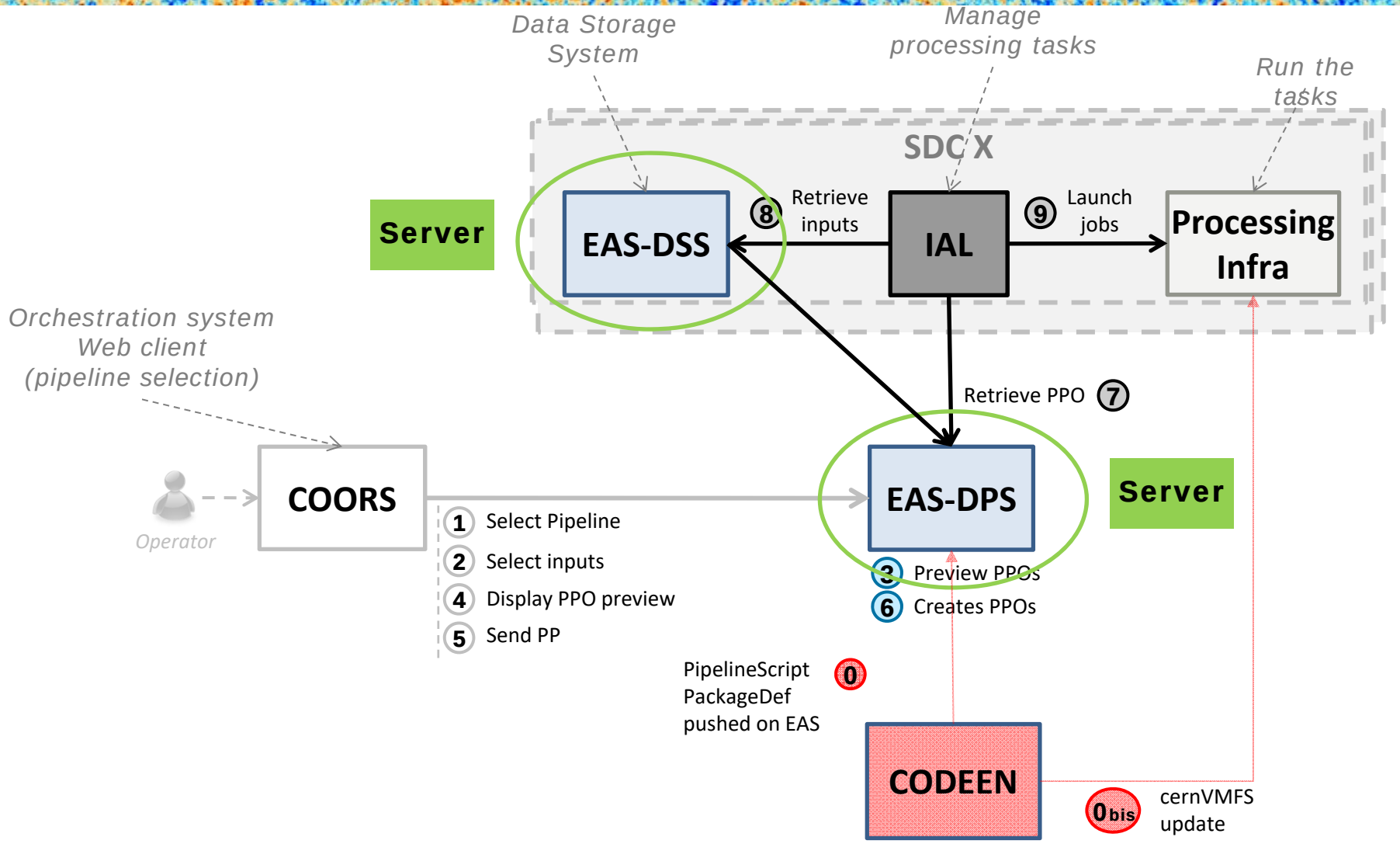
COmmon ORchestration System (COORS) : R. Blake (RoE)

- used for defining data processing runs and data movement activities across SDCs

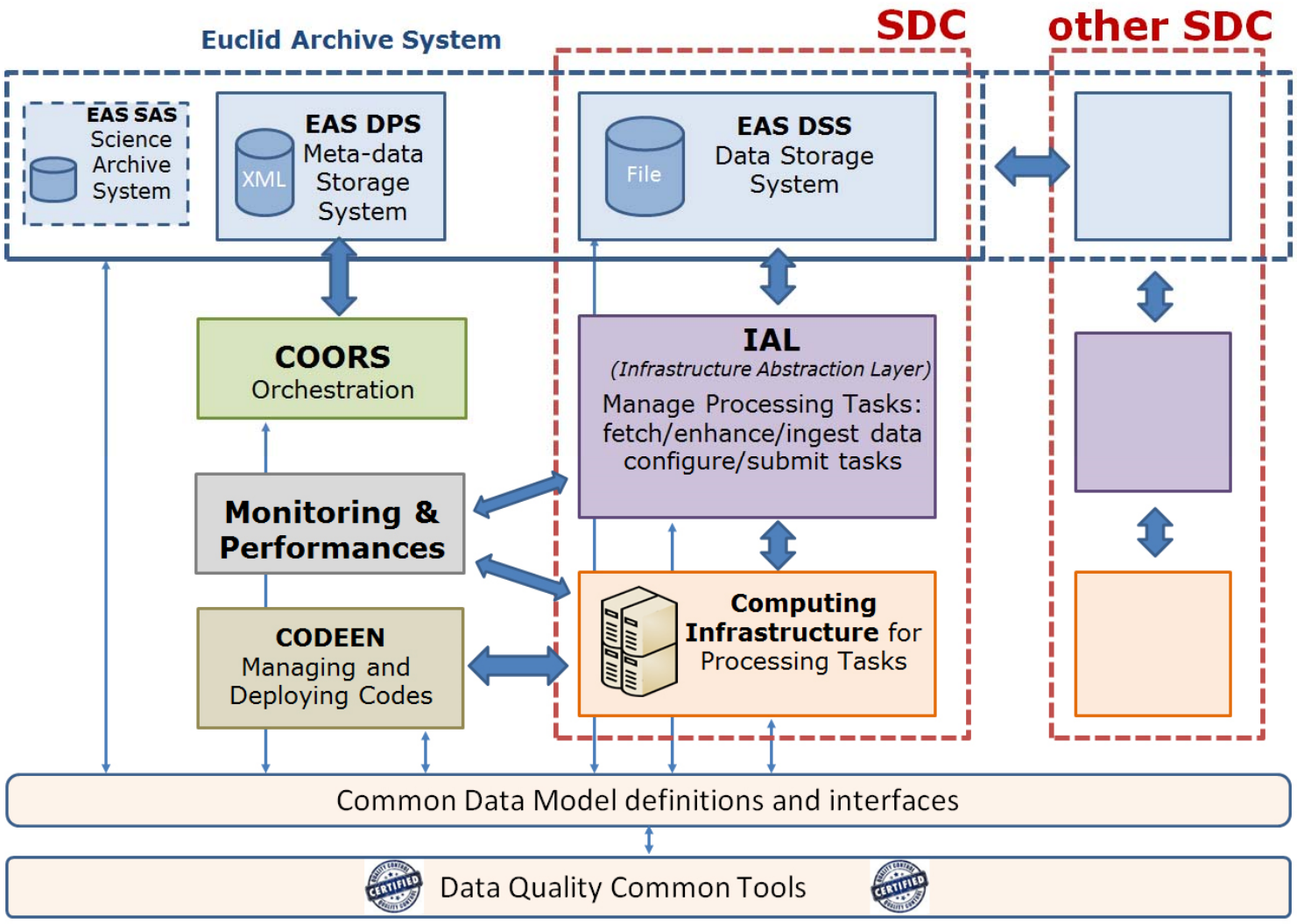
Data Quality Common Tools (DQCT) : M. Brescia (INAF) / S. Haugan (UIO Norway)

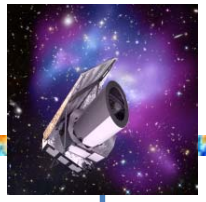
- Set of quality tools shared by the different processing functions

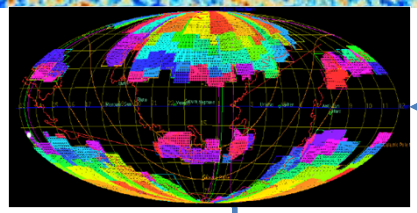
Infrastructure in the SGS

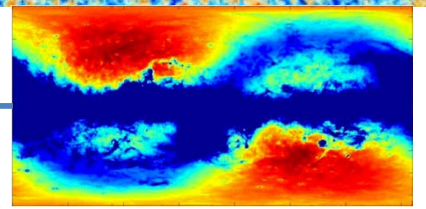


Infrastructure in the SGS









Mission Operations Centre

External Data Providers

Sky allocation through Coordinator

Science Operations	SDC-NL	SDC-DE	SDC-FR	SDC-CH	SDC-ES	SDC-US	SDC-UK	SDC-IT	SDC-FI
Public Data	Raw EXT Data	Raw EXT Data	Raw EXT Data TBC						
Level 1 Data Files	Data Files	Data Files	Data Files	Data Files	Data Files	Data Files	Data Files	Data Files	Data Files
Metadata (prime)	Metadata (backup)								

Redundant storage of the data

Development strategy

- We have a (very) complex system to build. We cannot develop all its pieces independently and then integrate them just before launch.
- We have defined a generic development roadmap materialized by **maturity gates** (corresponding to the level of code compatibility with the Euclid System).
- We have defined a **series of challenges** that put increasing sections of the SGS system to full-scale testing:
  - **Infrastructure (IT) challenges**: deal with all the generic systems of the pipeline, deployment, code compilation, orchestration of the processing.
  - **Science (SC) Challenges**: Integrate an increasing fraction of the Processing Functions, starting from the image processing upward.
  - IT Challenges lead the SC Challenges so that SC Challenges take place into the actual Euclid infrastructure, and thus also serve as integration tests for the Processing Functions.
  - Challenges are a key tool to discover hidden features of a complex system, often beyond the specified challenge objective.

Budget Processing : L2

per Observation	Cadence	Input (GB)	Output (GB)	Core x hour (2017)
VIS	Daily	18	80 (145)	130
	Monthly	700	5	20
NIR	Daily	5,5	12,6	12
	Monthly	55	4,5	
SIR reduc		4,5	4,5	
SIR extra		25	100	8
MER		250 (217)	200	22,5
SPE		3,4	4,56	20
PHZ		5	4,4	0,7
SHE		210 (165)	26,4 (180)	9,2 (666)
EXT (KIDS)		16,8	16,8	5
EXT (DES)		16,8	16,8	5
EXT (CFIS)		16,8	16,8	5
Total	Storage	**500 (750)**		
	Core x hour	**200 (900)**		

Legend :
in red figures from previous estimation

- **Scientific challenge 3 has confirmed PF's figures**
- **SHE has significantly revisited its output products volumetry and processing consumption**

Budget Processing : LE3 (source SDD)

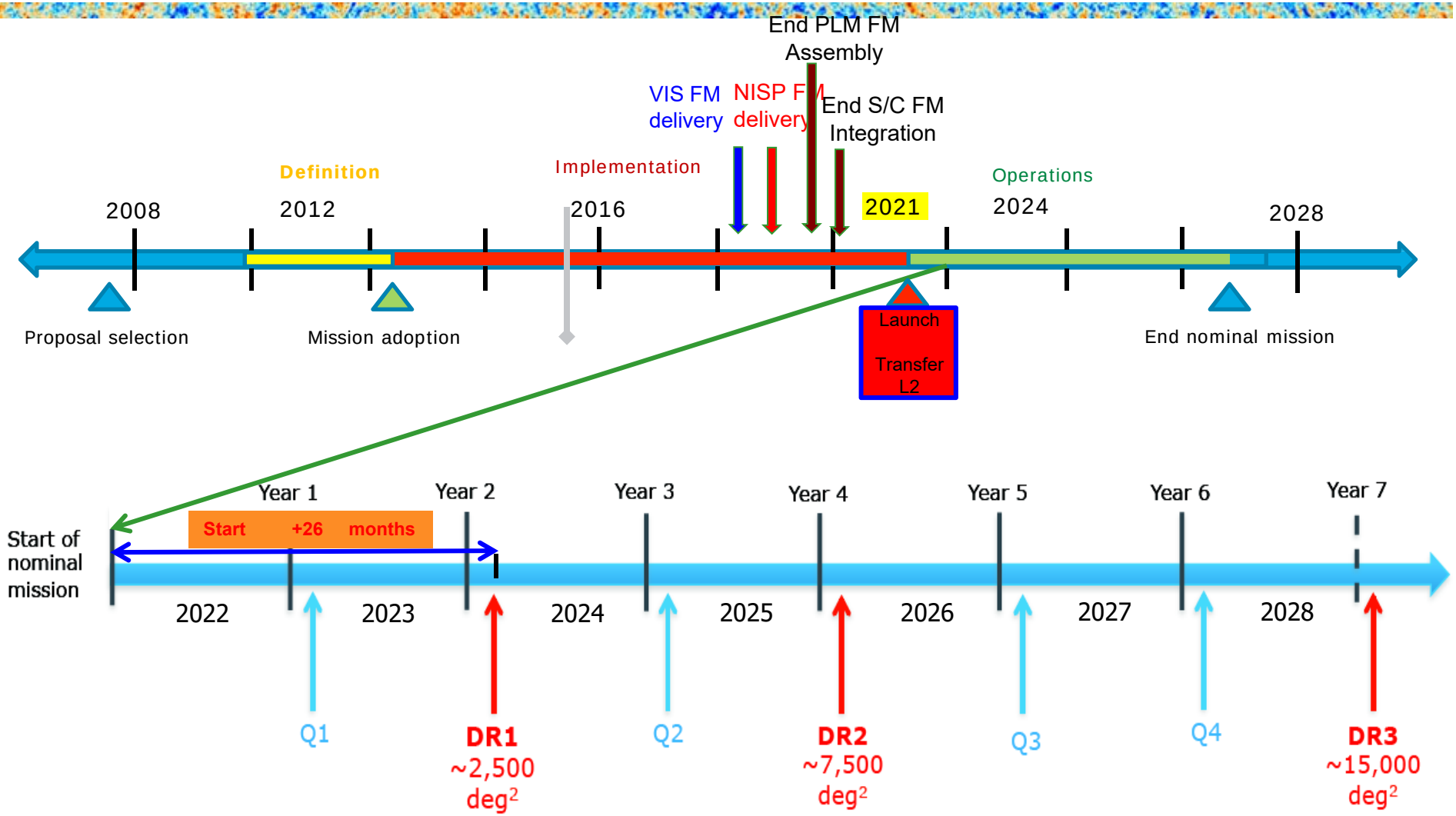
Estimation revisited :

- 2PCF-WL gain ratio = 4
- 2PCF-GC gain ratio = 18
- These significant gains (2PCF) have been achieved using profiling tools and consequently optimizing the implementation of the most critical sections of the code.
- CM GC= 10^5 * (2PCF GC)

Galaxy clustering PFs	DR1 (core x hour)	DR2 (core x hour)	DR3 (core x hour)
2PCF-GC	250		1600
CM-2PCF-GC	25M		161M
PK-GC	1,3		2
CM-PK-GC	130000		250000
BK-GC			2500
3PCF-GC	25		170
Weak Lensing PFs	DR1 (core x hour)	DR2 (core x hour)	DR3 (core x hour)
2PCF-WL	300	1700	4200
CM-2PCF-WL	300000	1M	4M
PK-WL			
CM-PK-WL			
2DMASS-WL			
3D-WL	10800	21600	56000
Total	26M		165M

Legend :
in green figures reassessed and achievable
in red figures reassessed and NOT achievable (currently)

Mission Timeline and Data Releases



Science with Euclid will start in 2023 with Q1 and in 2024 with DR1

Conclusion 1 / 2

- Euclid will release the data to the Public with an incremental approach (three/four releases are foreseen) and will be the *legacy* galaxy catalogue. It will be then very important to build it in the MOST compatible way.

- For this reason one of the requirements is to save all the data in a VO compatible format to allow easy interoperability and cross checking with others surveys. This is true NOT only for the Euclid products but is also for the intermediate products that may become part of the formal delivery.

- This was NOT true in other mission, like Planck, where we agreed to use a VO compatible structure only at the moment of creation of the public products.

Conclusions 2 / 2

- Compliance with the Virtual Observatory (VO) standards allow the data to follow the FAIR (Findable, Accessible, Interoperable, Reusable) paradigm. We can say that Planck data achieved VO-compliance due to an internal agreement, whereas in Euclid it is part of the requirements.

- To maximize the *reusability* (FAIR) of data, it is important to offer *services* (e.g. visual inspection tools, cross-matching functionalities, etc …) where the data are.

- Due to the enormous amount of data, in case of Euclid it will be almost impossible for an external user to download them locally. This means that the ESA-ESAC archive, where Euclid data will be stored and made available, should provide services.