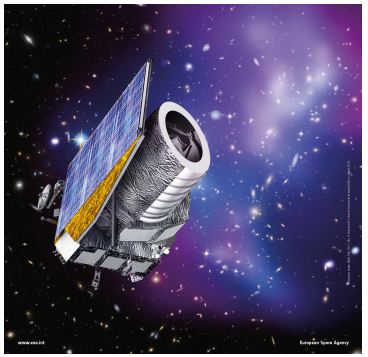
OBELICS Task 3.3
-
EUCLID resource requirements

Thomas Vuillaume

EUCLID - Introduction



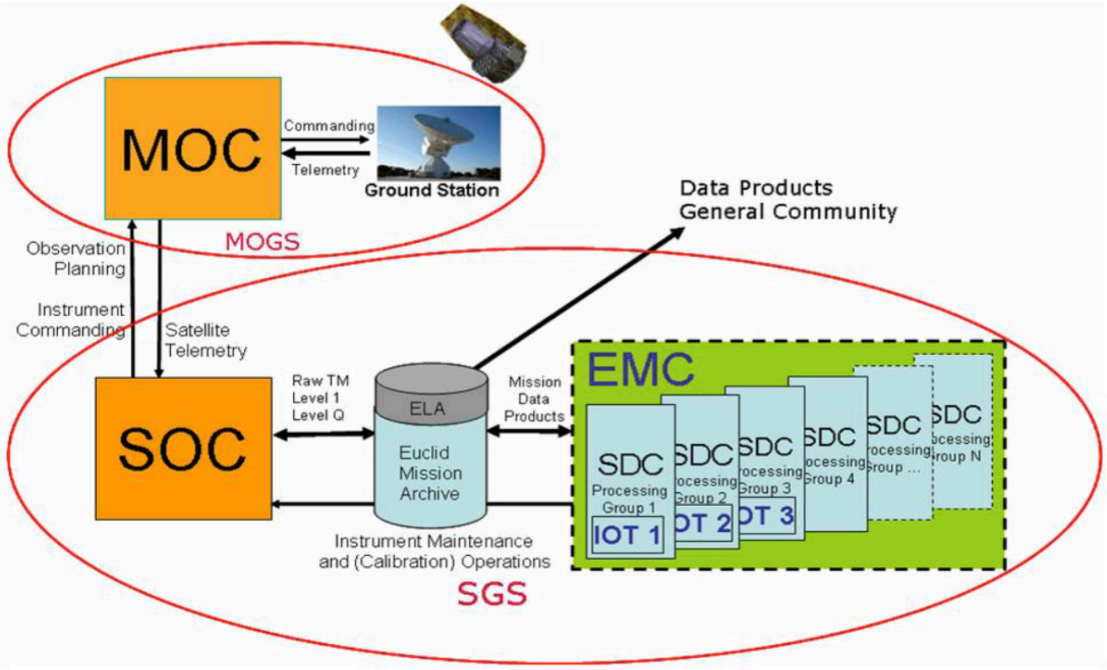
- ESA space mission, launch planned in 2019
- Located at L2
- To Study:
  - Universe expansion acceleration
  - Nature of dark energy, dark matter and gravity
  - Physics of early universe
- Deep sky survey
- Equipped with 1.2m mirror telescope and two instruments:
  - VIS (visible imager)
  - NISP (Infrared imager and spectrograph)
- **850 Gbit/day** of raw compressed science data
- Completed with ground observations

EUCLID – Data model

Sequential processing steps => data levels

- **Level 1 :** raw VIS and NISP images
- **Level 2 :** calibrated and co-added images; PSF model and distortion maps; co-added spectra
- **Level 3 :** catalogs, dark matter mass distribution, ground based information used in the derivation of the data products, …
- **Transients :** transient event data products
- **Level Q data :** products suitable for most purposes in Astronomy and not for the main goal of the mission
- **Level E data :** quality-controlled external data from existing missions and ground-based surveys (used for callibration and photometric redshift derivation)
- **Level S data :** pre-launch simulations and modeling

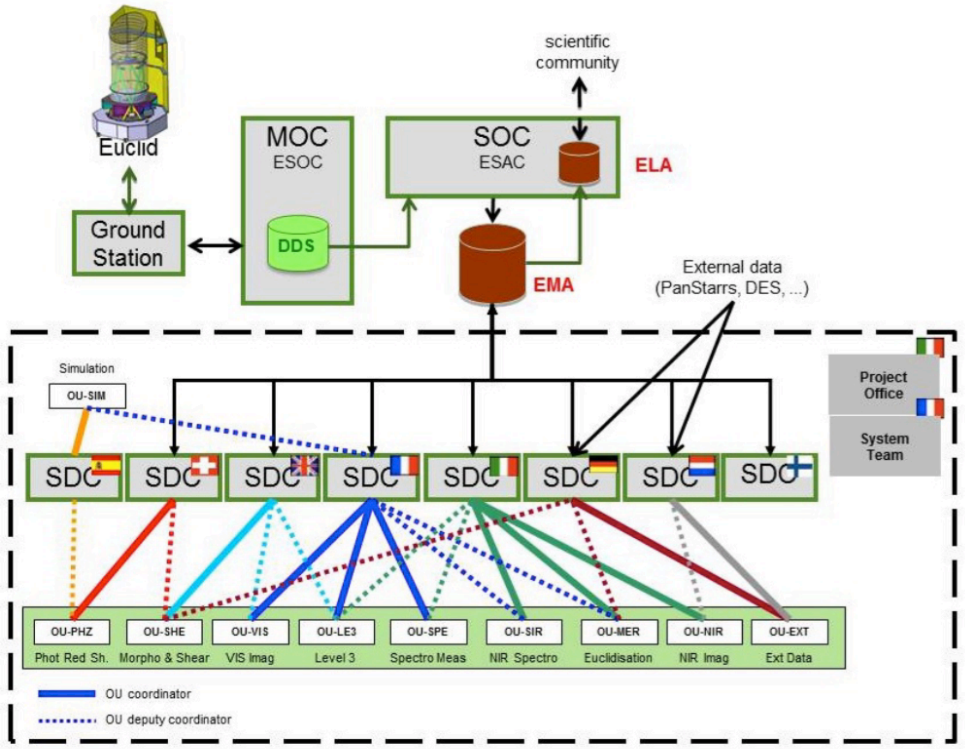
EUCLID – Data model



- Distributed Data Centers (processing + storage nodes)
- Generic SDC (virtualization)
- *"Move the code, not the data"*

* From Science Ground Segment for the ESA Euclid mission, Pasian *et al 2012*, proceedings of SPIE, vol 8451

EUCLID – Data model

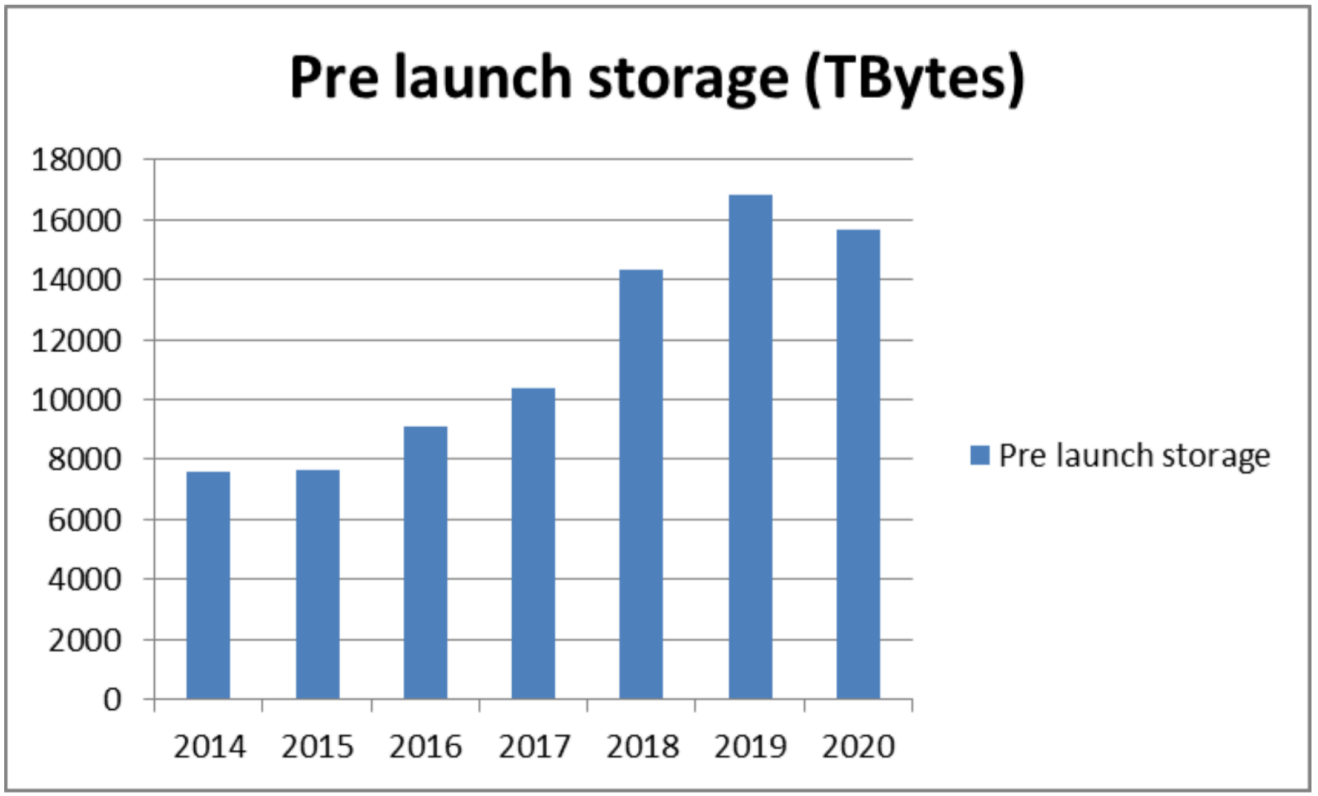


- Different Operational Units process different levels/kind of data
- Develop their own pipelines
- Processed data associated with Quality control Information to ensure traceability
- Euclid Mission Archive = ”working repo”
- Euclid Legacy Archive = final validated products

* From Science Ground Segment for the ESA Euclid mission, Pasian *et al 2012*, proceedings of SPIE, vol 8451

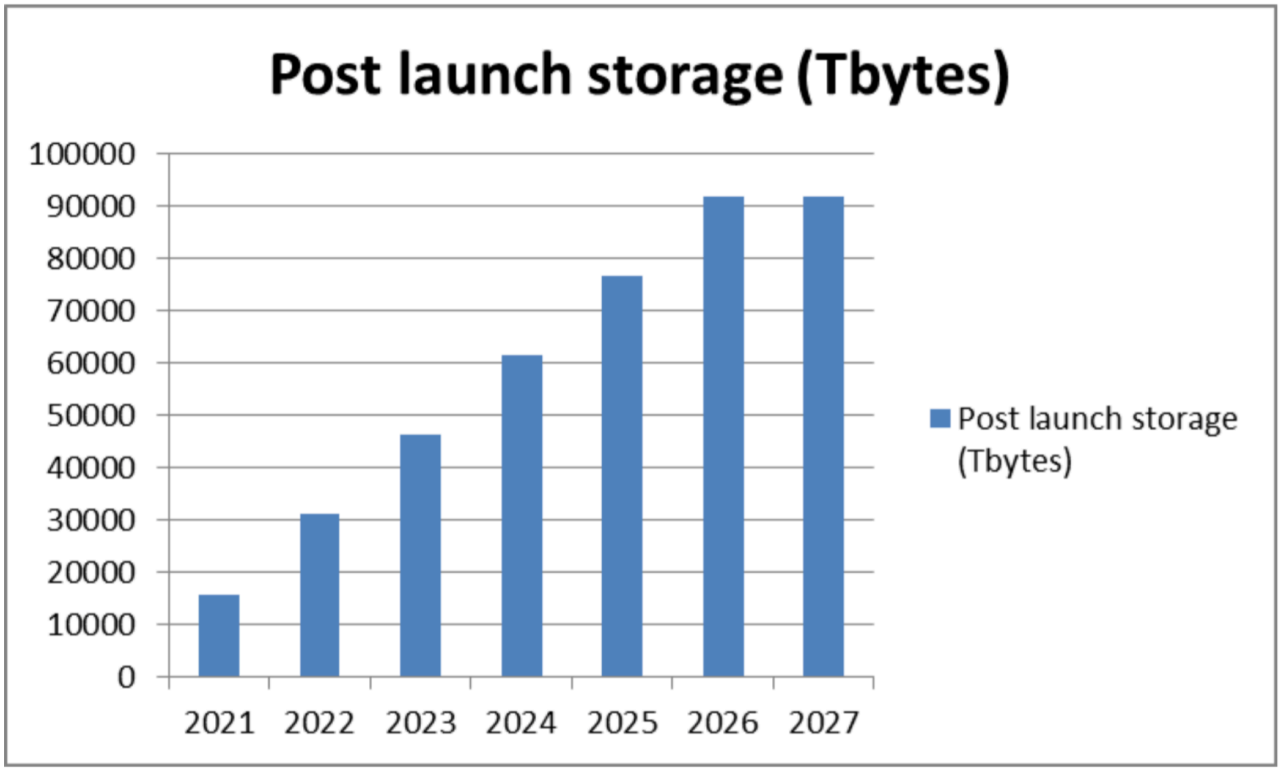
EUCLID – Storage Needs

18 PB

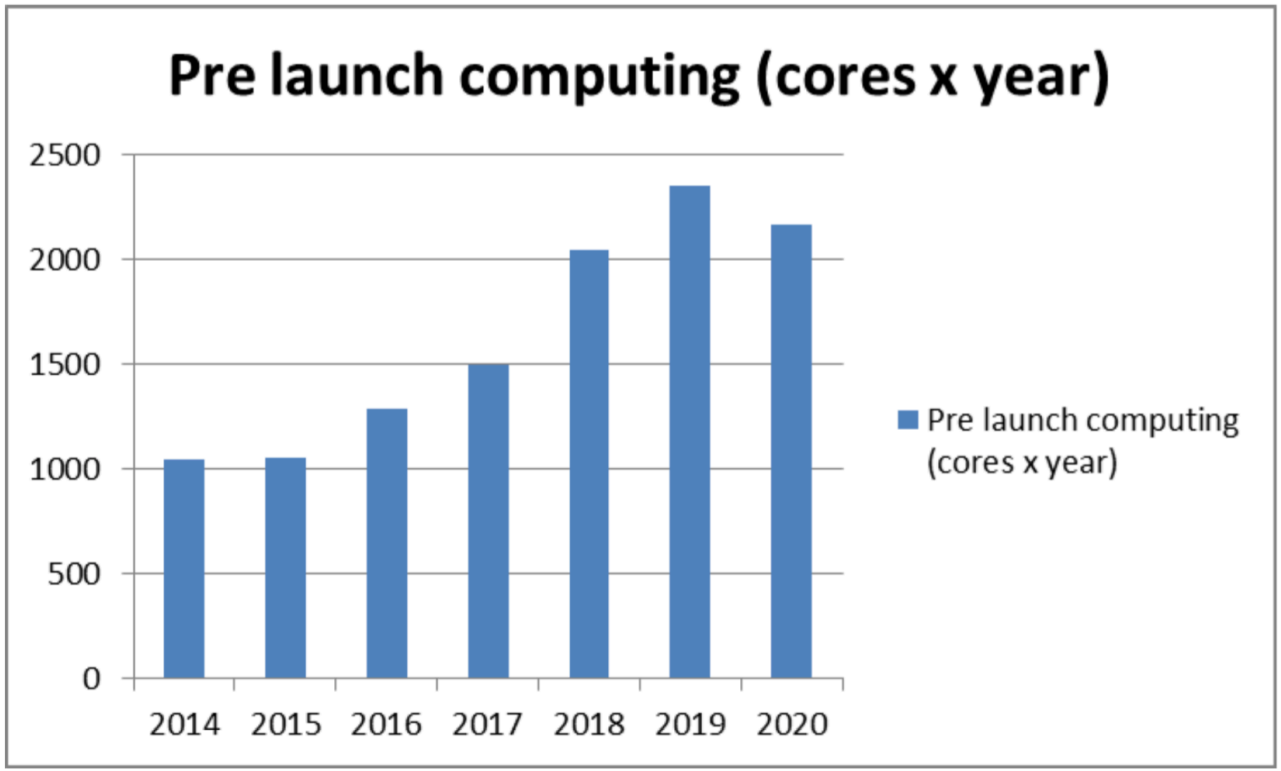


EUCLID – Storage Needs

100 PB

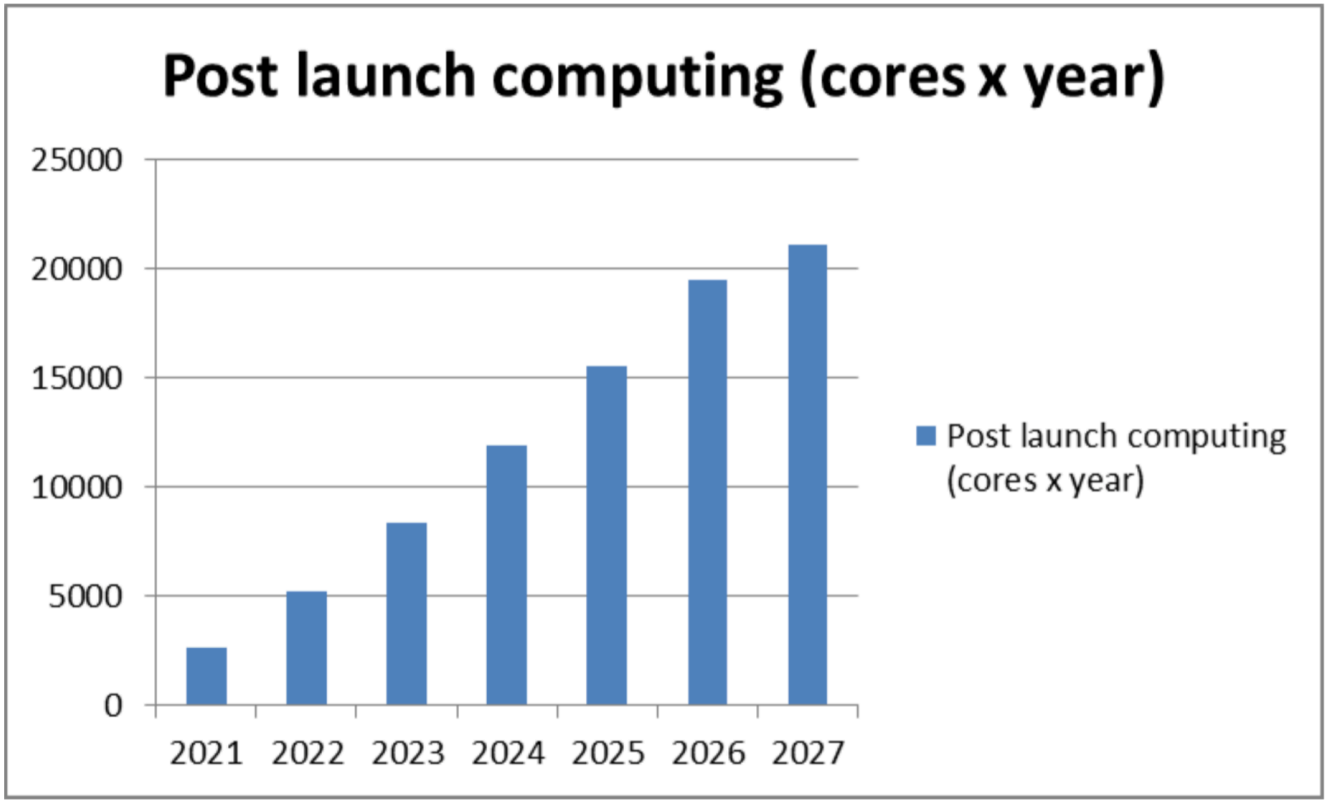


EUCLID – Storage Needs



=> 3 months with 9000 CPU cores (to compare with CTA)

EUCLID – Storage Needs



=> 27 months with 9000 CPU cores (to compare with CTA)

EUCLID – Database

- Central metadata data base that hosts all the data product metadata processed by the different SDC's
- Distributed file storage on SDC's hosting the data product files
- Common Orchestration that implements the data storage distribution and processing distribution in the various SDC's
- Central infrastructure used to ensure continuous integration of Euclid Software and embedded tools (build, tests, quality check, documentation, packaging deployment, dashboards, VM's provisioning, libraries and executable deployment etc...)

EUCLID – Data transfer

Bottleneck between SDCs and long term archive

- Data releases: 5 PB for DR1, 15 PB for DR2 and 32 PB for DR3
- transfer rate ≃ 2.5 Gb/s

⇒ DR1 requires 200 days to be transferred to the long term archive 😡

⇒ possible alternatives:

- reduction factor of at least 10, achieved by removing raw exposures
- using the SDC's storage infrastructure as long term archive

EUCLID – Database

- 10^{10} objects (with ~1500 parameters characterizing each object)

$\Rightarrow$ huge constraints on the database

- sources spectra implemented in files (probably fits format)
- some metadata (a few KB/source) describing these spectra in the Euclid metadata database

$\Rightarrow$ 808 GB/hour to ingest in the database in 2027 !

EUCLID – Processing

Processing :

- On-the-fly processing needs: continuous, constant and known
- Data Release: periodic, intensive and known
- Reprocessing: discontinuous and unknown (data analysis and validation, algorithm improvements, unknown systematic effects to be assessed and corrected...)