



EUCLID data processing at CC-IN2P3

Quentin Le Boulic'h - CC-IN2P3
FJPPL Computing Workshop, 11.02.2016

- ▶ Euclid data processing
- ▶ CNES / IN2P3 agreement
- ▶ Euclid activity at CC-IN2P3



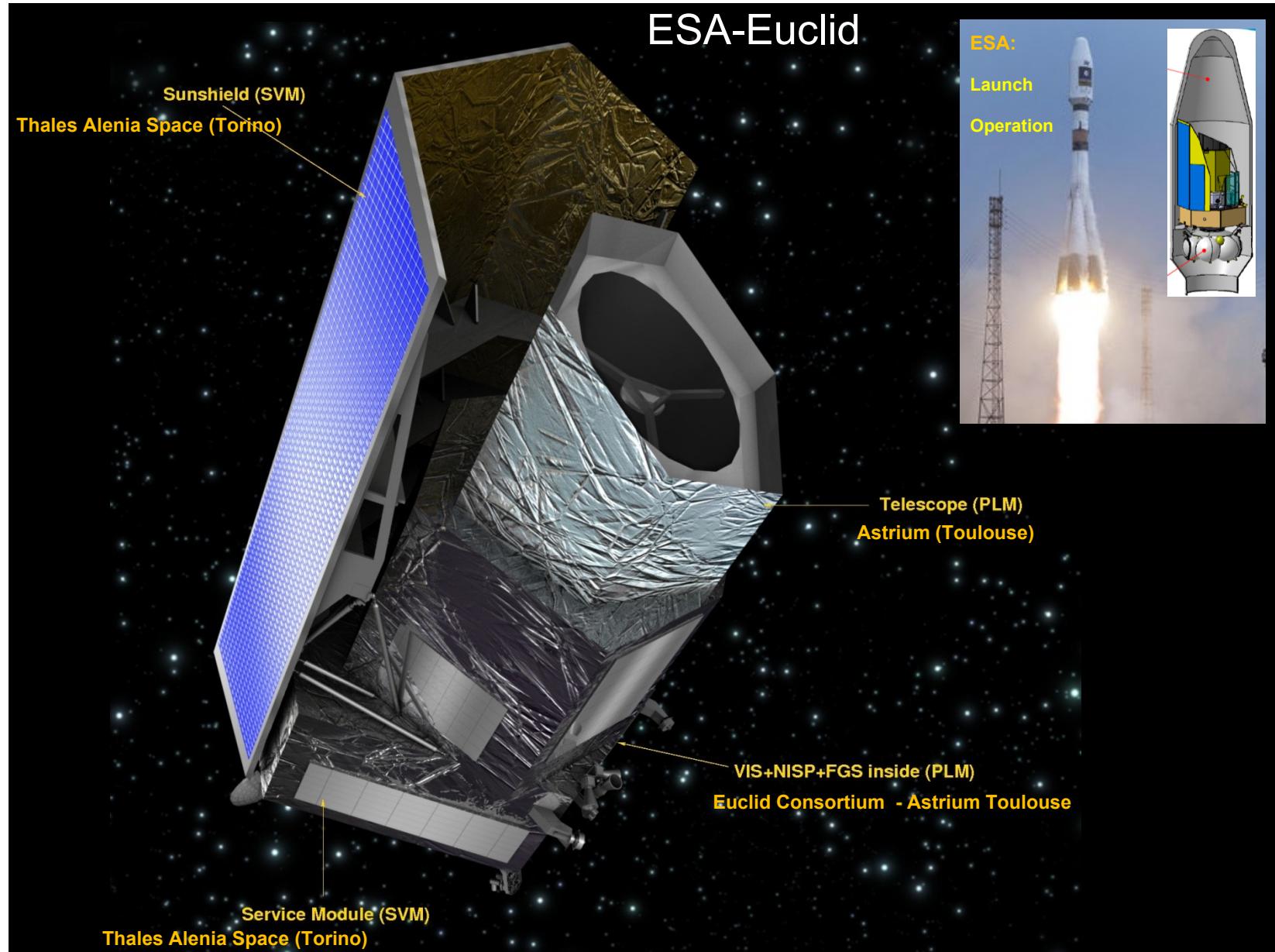
Euclid data processing

The Euclid mission

- ▶ Euclid is an ESA space mission to be launched in 2020 (6 years mission)
- ▶ Its scientific goals are:
 - ▶ Understanding the origin of Universe accelerating expansion
 - ▶ Constrain cosmological models : Dark Energy, Dark Matter and Gravity
 - ▶ Legacy for other fields of astrophysics
- ▶ Cosmological probes:
 - ▶ Weak lensing
 - ▶ Galaxy clustering (BAO,...)
 - ▶ Structure formation
 - ▶ ...
- ▶ Satellite with two instruments:
 - ▶ VIS : Visible Imager (600 MPix)
 - ▶ NISP : Near Infrared Spectrometer and Photometer (64 Mpix)



The Euclid mission: participants



The Euclid Consortium

- ▶ Provide VIS and NISP instruments
- ▶ In charge of **data processing**
- ▶ Some numbers:
 - ▶ 35% of the total Euclid mission cost
 - ▶ 14 europeans countries contributing + US (NASA + other labs)
 - ▶ More than 1200 members and 120 laboratories
- ▶ French participation:
 - ▶ Contribute to 30% with key responsibilities
 - ▶ Contribution of the CNES, IN2P3, IRFU, INSU

Data processing is done by the Science Ground Segment (SGS) :

- ▶ Development, integration and operation of the processing pipeline
- ▶ Production and archiving of the Data Releases
- ▶ Delivery of data to the scientific community

The Euclid data processing pipeline

Ground Station

MOC

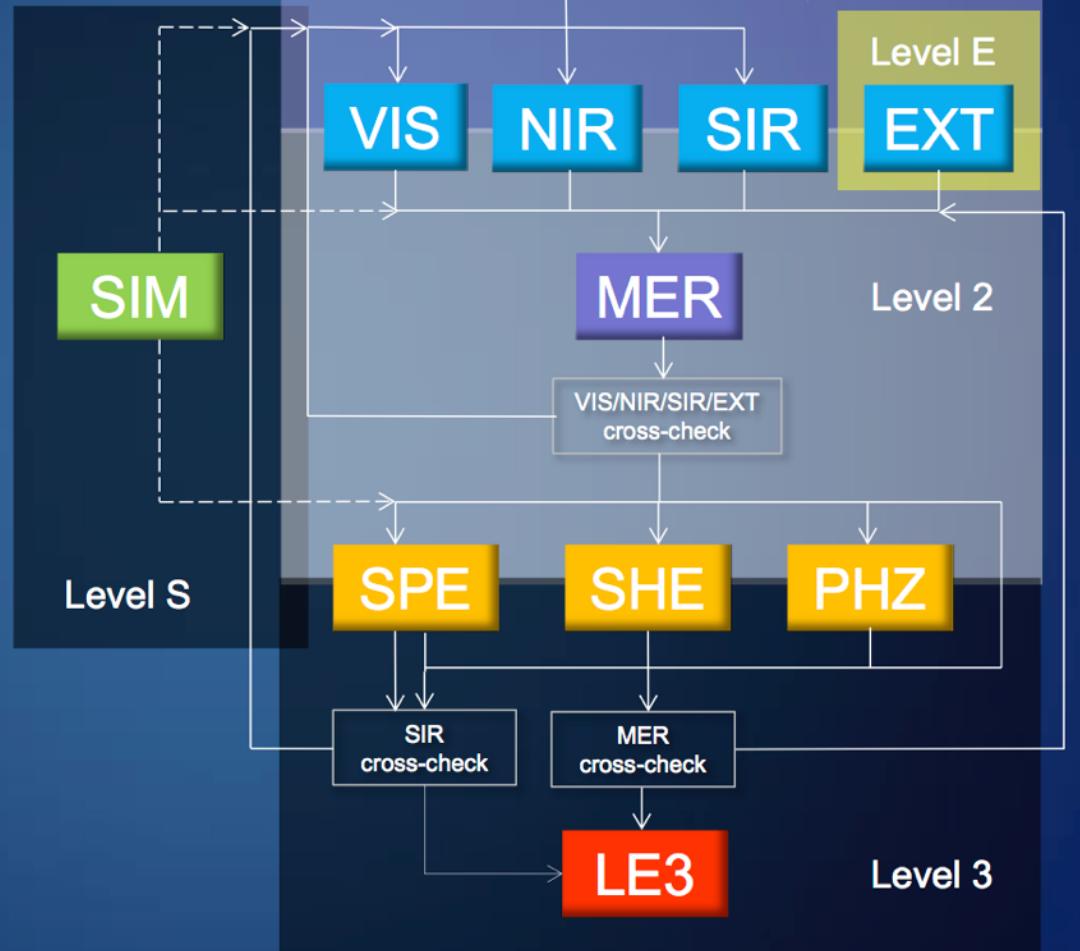
OPS

LE1

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

The coloured boxes correspond to the Processing Functions

SOC



CERN, 25 June 2014

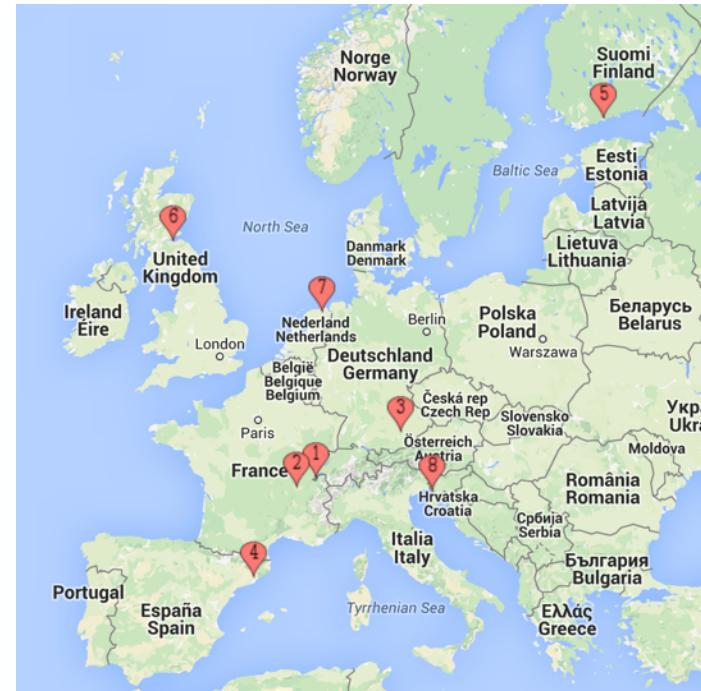
CERN-Euclid meeting

3

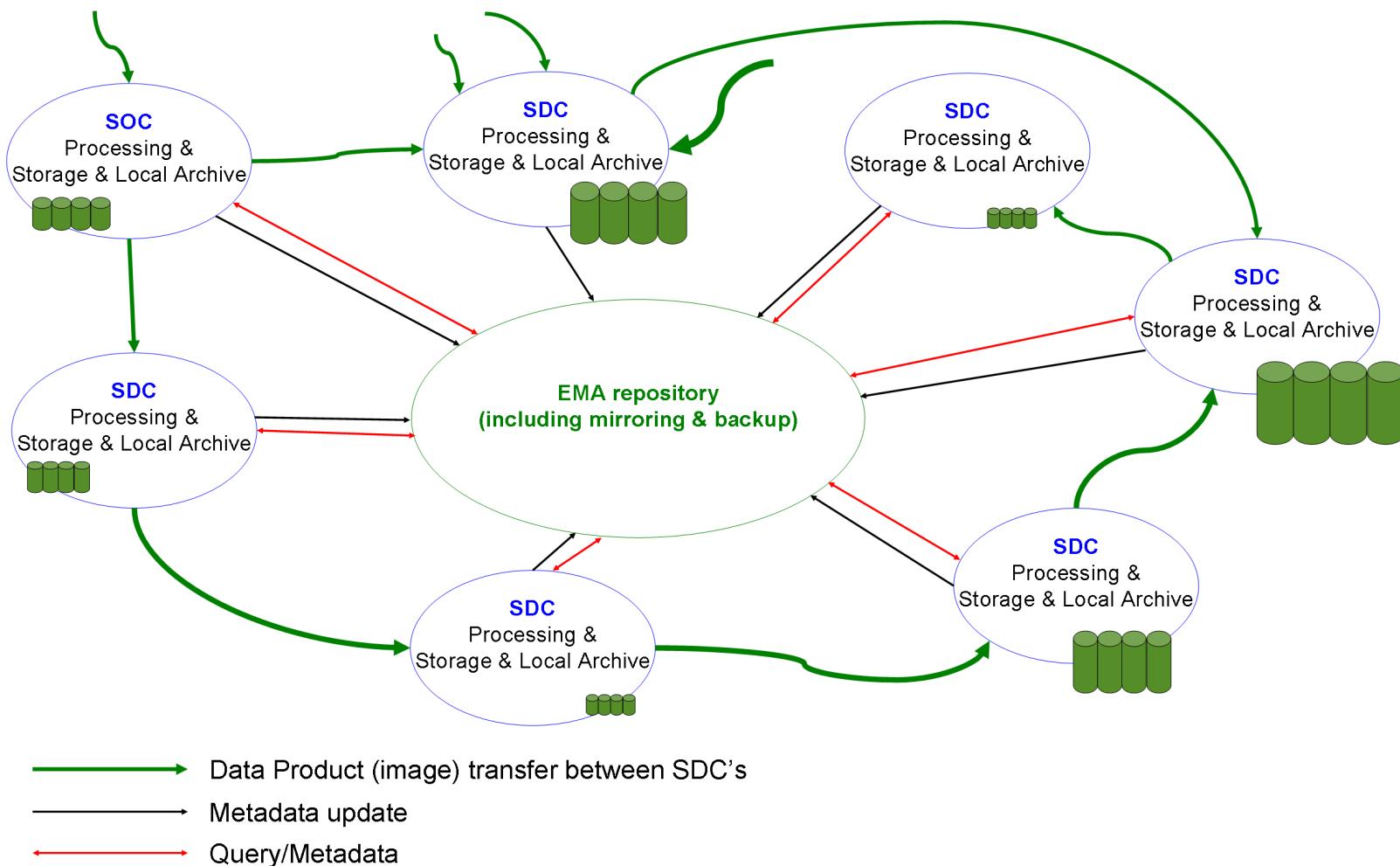


Science Data Centers

- ▶ 9 Science Data Centers
 - ▶ France : **CC-IN2P3**, Villeurbanne
 - ▶ Spain : PIC, Barcelona
 - ▶ Switzerland : University of Geneva (ISDC)
 - ▶ Finland : University of Helsinki
 - ▶ UK : Royal Observatory, Edinburgh (ROE)
 - ▶ Netherlands : University of Groningen (RUG)
 - ▶ Italy : Astronomical Observatory of Trieste (INAF)
 - ▶ Germany : Max Planck Institut, Garching (MPE)
 - ▶ US : IPAC, Caltech
- ▶ Some numbers:
 - ▶ RAW data : ~ 300 TB (6 years) = 150 x Planck
 - ▶ Total (including intermediate and external data) : ~ 150 PB (10^{10} objects)
 - ▶ Processing : ~ 20 000 CPU cores at maximum
 - ▶ **CC-IN2P3 : 30% of this (storage / processing)**



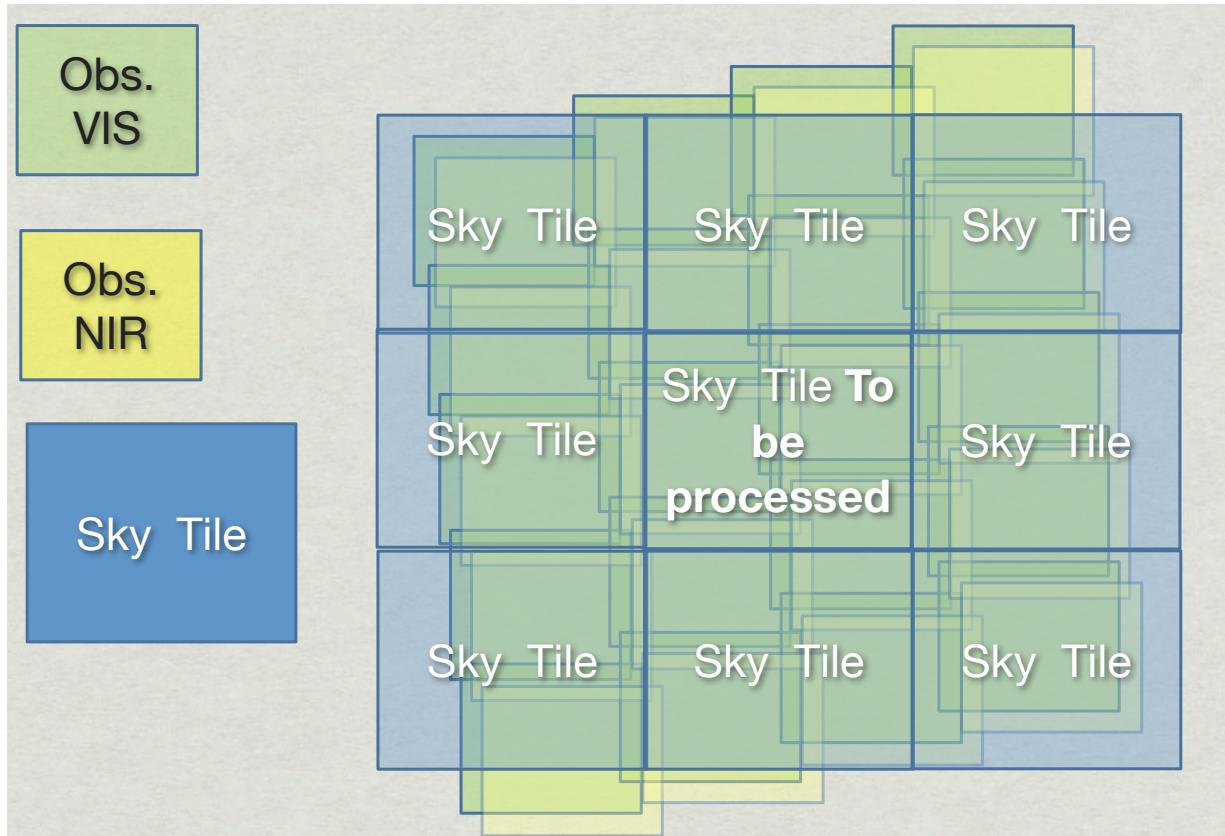
Science Data Centers



- ▶ Data distributed on the 9 computing centers (redundancy)
- ▶ Centralized metadata database

Science Data Centers

Data storage and processing distributed in computing centers following a sky division:



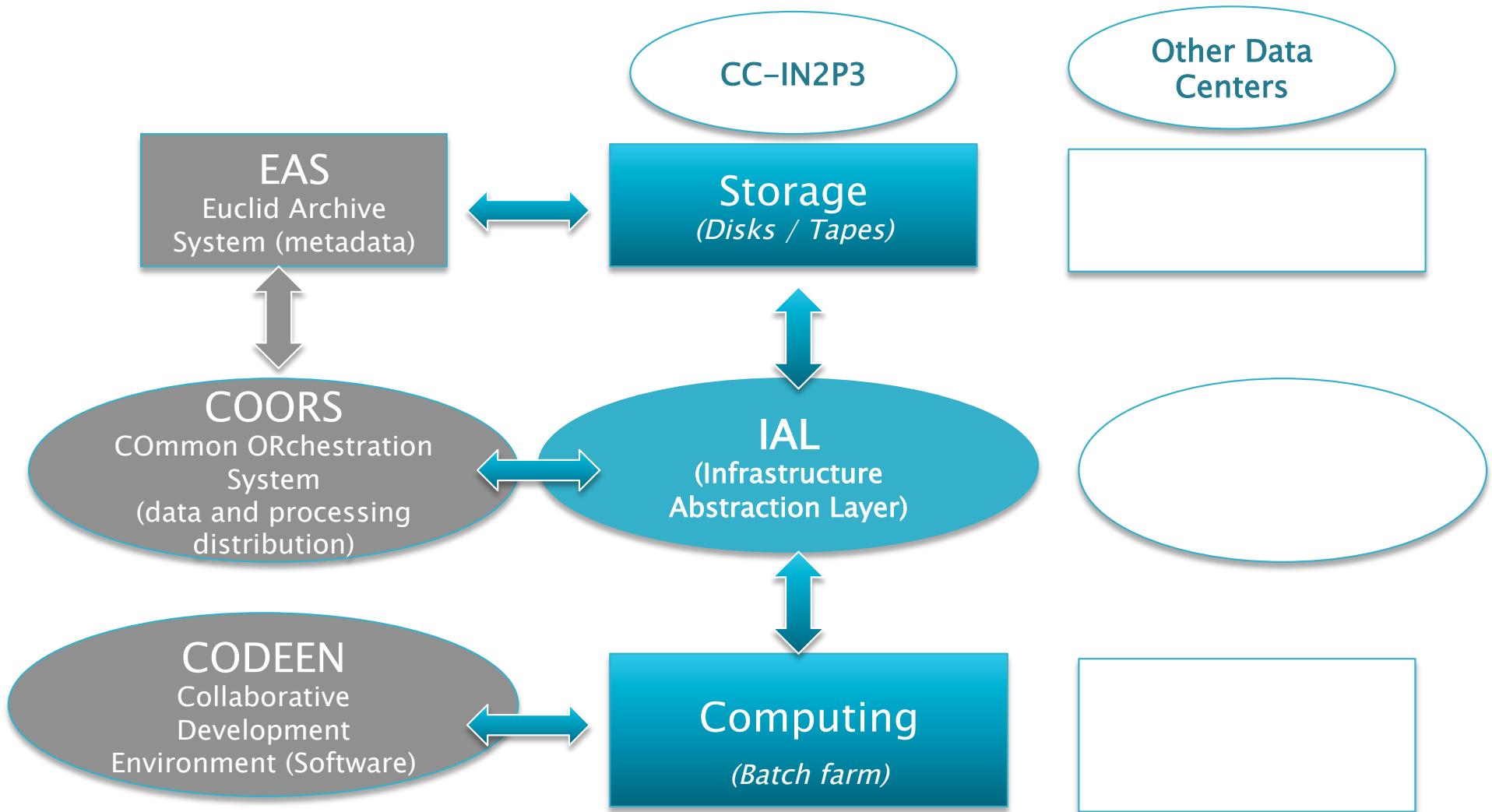
All kind of processing (Processing Functions) must run on all computing centers
(with some exceptions...)

The Euclid data processing challenges

- ▶ Challenges :
 - ▶ High volume of data to store, process, reference, make available
 - ▶ Complex processing pipeline: 2 instruments, visible / infrared, external data, massive simulations,...
 - ▶ 9 heterogeneous computing infrastructures

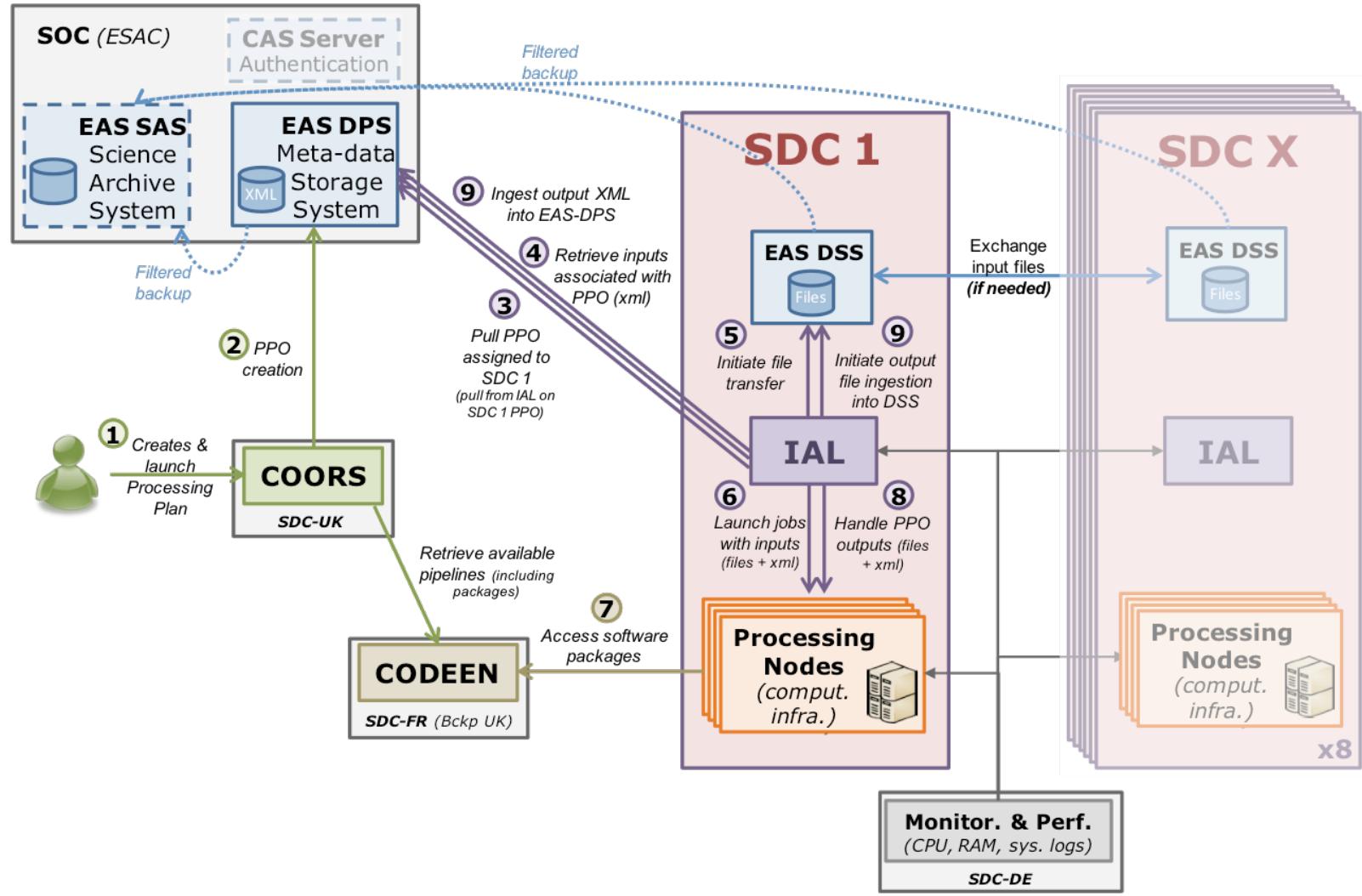
Euclid computing architecture

- ▶ The current vision of Euclid architecture is the following:



Euclid computing architecture

- The current vision of Euclid architecture is the following:





CNES / IN2P3 agreement

The CNES / IN2P3 agreement

- ▶ The agreement between CNES (french spatial agency) and IN2P3 ensure that CC-IN2P3 will provide resources to cover 30% of the computing needs for Euclid
- ▶ The agreement covers the processing until the first data release (DR1) in 2022
- ▶ It also accommodates launch delay (3 different scenarios), unexpected change in hardware cost or CC-IN2P3 participation
- ▶ The agreement does not cover:
 - ▶ HPC needs
 - ▶ Data releases 2 and 3
 - ▶ Scientific analysis
- ▶ Computing needs estimated in two ways:
 - ▶ Comparison with Planck experiment
 - ▶ « Bottom-up » calculation from processing estimations

The CNES / IN2P3 agreement: contributions

- ▶ CC-IN2P3 contribution : 2,660 M€ (mostly infrastructure)
 - ▶ Administration, maintenance and operation of the infrastructure
 - ▶ Computing, Storage and Network infrastructure
 - ▶ Electricity, cooling,...
 - ▶ Management of human resources
- ▶ CNES contribution : 1,495 M€
 - ▶ Financing of 3 engineers and travels related to Euclid :
 - ▶ 1 FTE Support : since end 2012
 - ▶ 1 FTE System : since 2015
 - ▶ 1 FTE Storage : from 2018

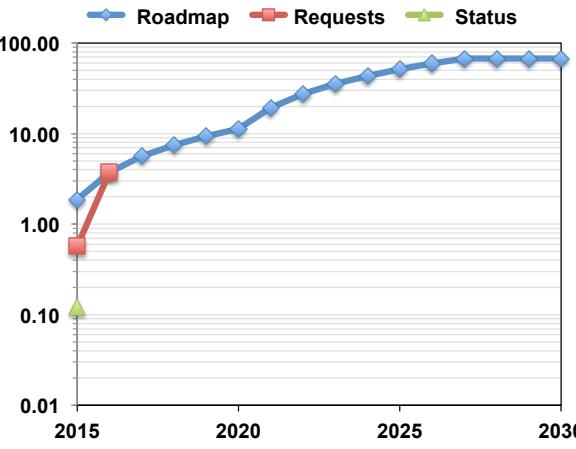
The CNES / IN2P3 agreement: roadmap

Units : CPU [kHS06] / Storage [PB]

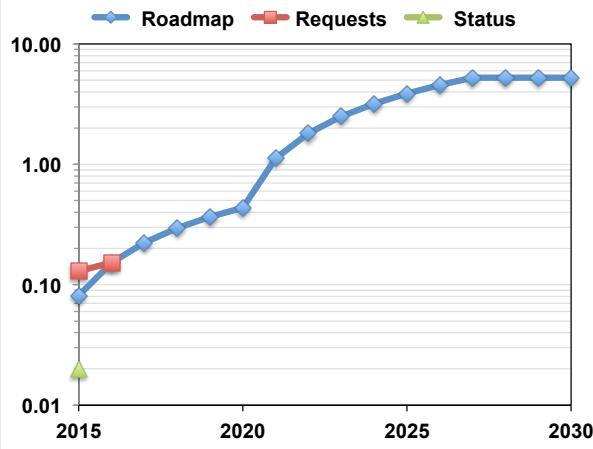
PM2

Years	Total Euclid			CC-IN2P3			CC-IN2P3 increment			Requests			Status [% requests]				
	CPU	Disk	Tape	CPU	Disk	Tape	CPU	Disk	Tape	CPU	Disk	Tape	CPU	Disk	Tape		
2015	6.24	0.271	2.706	1.87	0.08	0.81	1.87	0.08	0.81	0.570	0.13	0.10	0.122	21%	0.020	15%	0.005
2016	12.49	0.508	5.081	3.75	0.15	1.52	1.88	0.07	0.71	3.747	0.152	1.524					
2017	18.73	0.746	7.457	5.62	0.22	2.24	1.87	0.07	0.71								
2018	24.98	0.983	9.832	7.49	0.29	2.95	1.88	0.07	0.71								
2019	31.22	1.221	12.208	9.37	0.37	3.66	1.87	0.07	1.52								
2020	37.50	1.458	14.583	11.25	0.44	4.37	3.76	0.15	1.43								
2021	64.29	3.750	37.500	19.29	1.13	11.25	9.91	0.76	7.59								
2022	91.07	6.042	60.417	27.32	1.81	18.13	9.91	0.76	7.59								
2023	117.86	8.333	83.333	35.36	2.50	25.00	9.91	0.76	8.40								
2024	144.64	10.625	106.250	43.39	3.19	31.88	9.91	0.76	8.30								
2025	171.43	12.917	129.167	51.43	3.88	38.75	11.79	0.84	14.46								
2026	198.21	15.208	152.083	59.46	4.56	45.62	17.95	1.45	14.46								
2027	225.00	17.500	175.000	67.50	5.25	52.50	17.94	1.45	15.27								
2028	225.00	17.500	175.000	67.50	5.25	52.50	9.91	0.76	8.30								
2029	225.00	17.500	175.000	67.50	5.25	52.50	9.91	0.76	14.46								
2030	225.00	17.500	175.000	67.50	5.25	52.50	11.79	0.84	14.46								

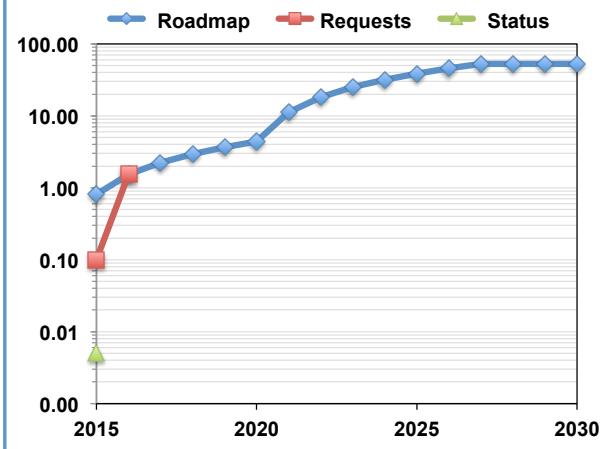
CPU [kHS06]



DISK [PB]



TAPE [PB]





Euclid activity at CC-IN2P3

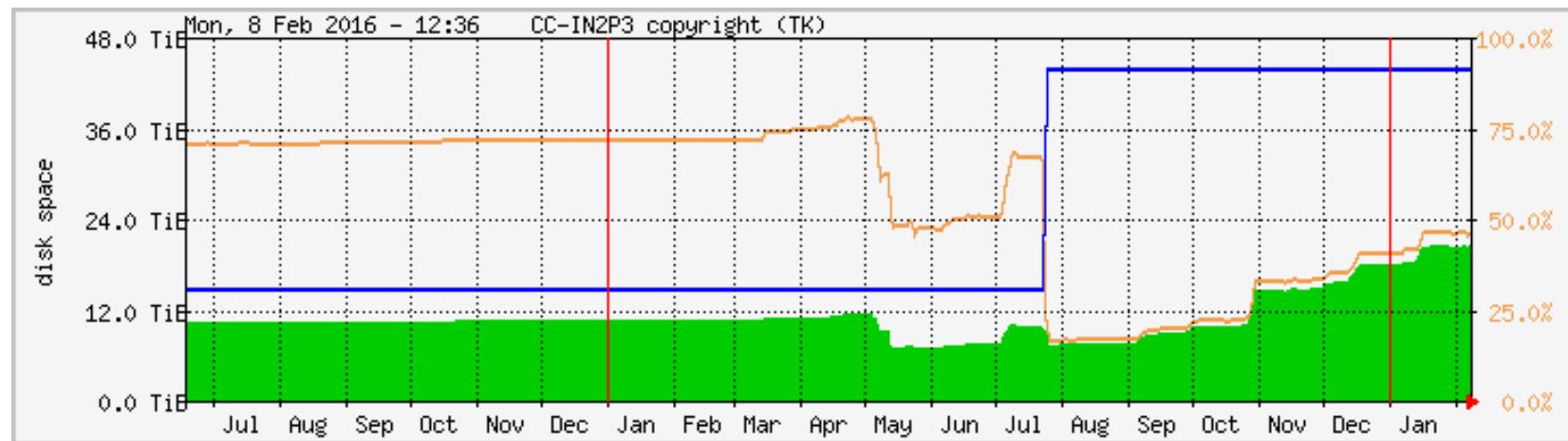
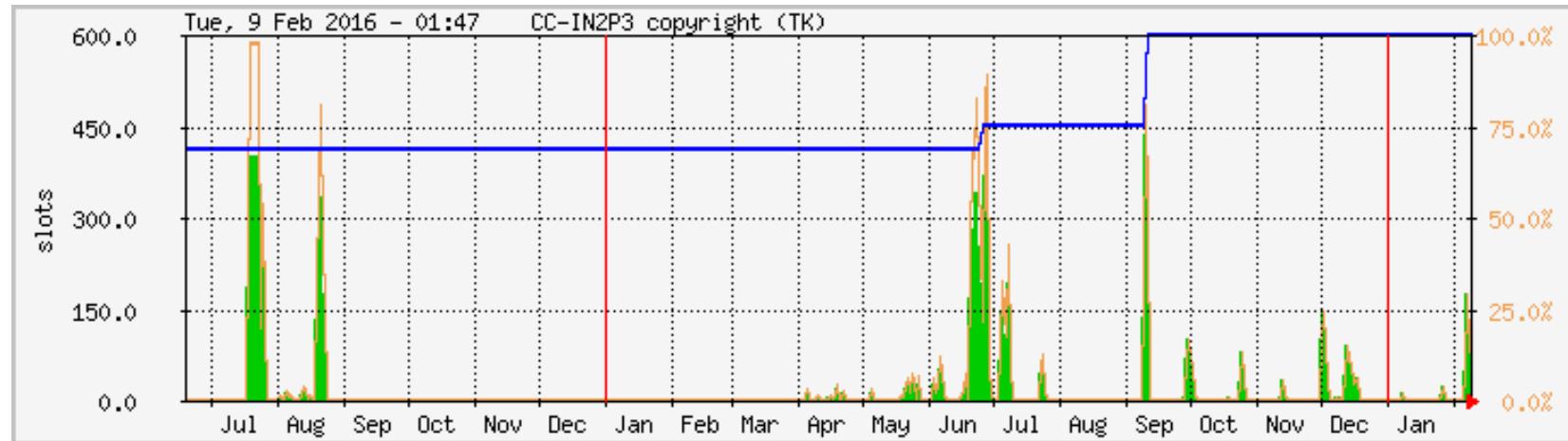
Euclid users at CC-IN2P3

- ▶ 78 accounts in Euclid group

Last login	Number of accounts
2011	1
2012	1
2013	6
2014	9
2015	31
2016	30

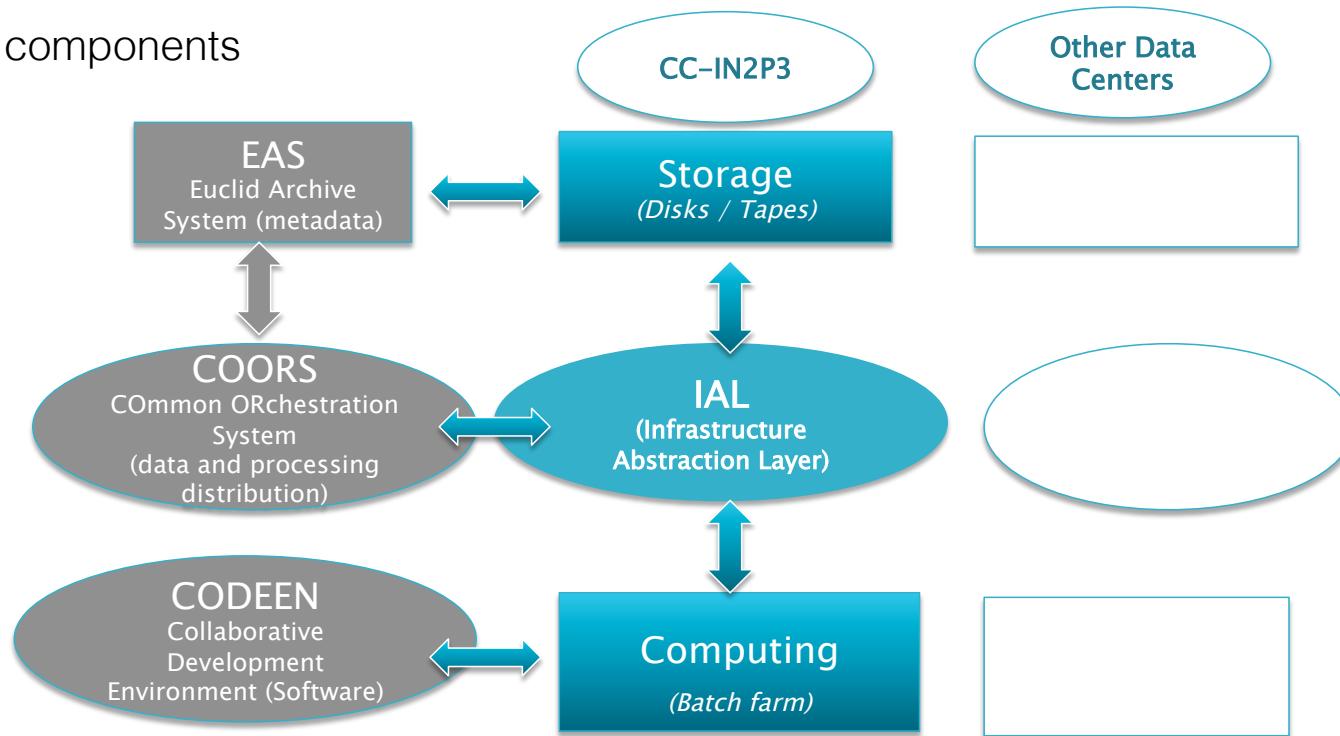
Past activity

- ▶ So far visible Euclid activity is quite low and irregular
- ▶ Mostly simulations



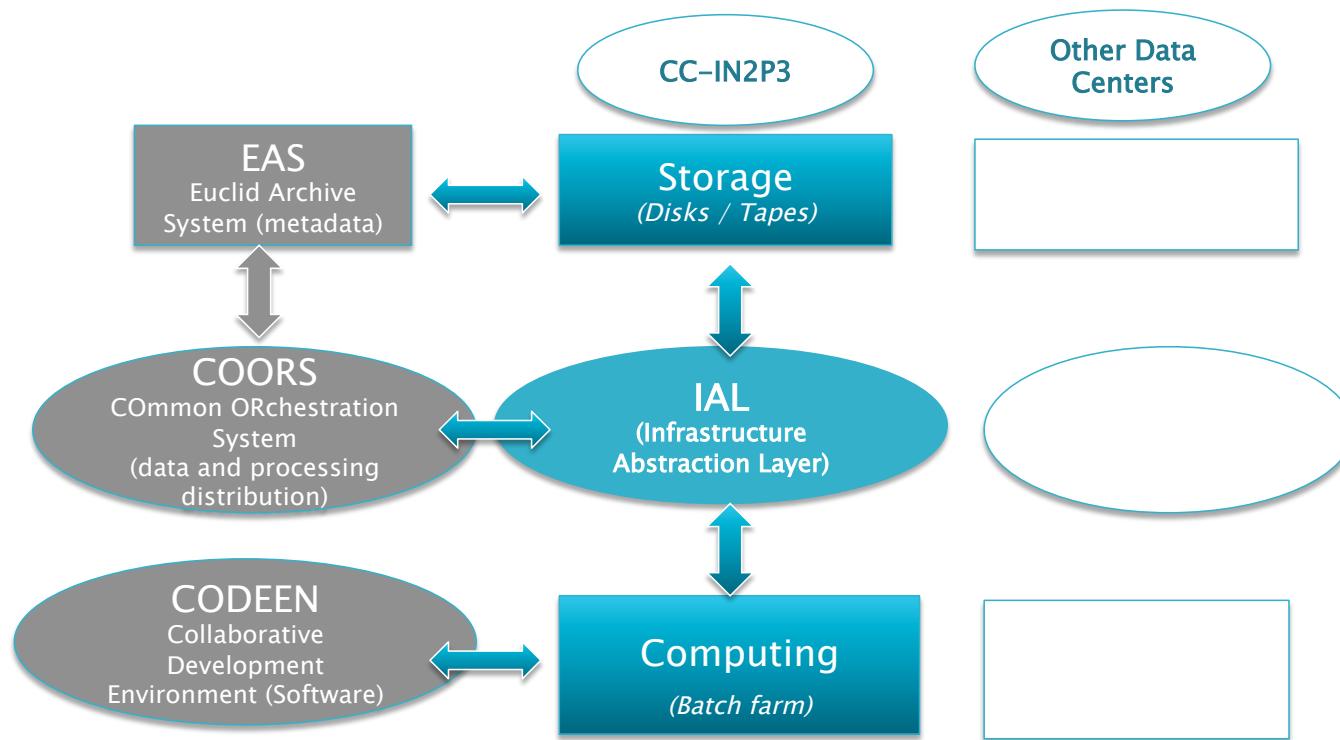
Past activity

- ▶ A lot of work done since the last 2 years on the distributed architecture for Euclid
 - ▶ Euclid internal developments (metadata database, orchestration system, software deployment, Infrastructure Abstraction Layer, Data Storage System , monitoring, ...)
 - ▶ Work at CC-IN2P3 (and other SDCs) to deploy, integrate and test these components
 - ▶ Progresses organized around « Data Challenges » with defined objectives
 - ▶ At each Data Challenge we deploy new components and new versions of components



Past activity

- ▶ Example: during Data Challenge 4 (end 2014 / beginning 2015) we deployed:
 - ▶ The Data Storage System (data storage and transferts between SDCs)
 - ▶ Ressources monitoring
 - ▶ New version of the IAL



- ▶ At CC-IN2P3 we also experiment alternative technologies (i.e. not in the baseline architecture):
 - ▶ Cloud computing:
 - ▶ A prototype of virtual cluster was setup and used for Data Challenge 4.
 - ▶ Will be used again for Data Challenge 6 (ongoing).
 - ▶ CernVM-FS (software deployment technology):
 - ▶ A testbed architecture was setup for Data Challenge 4 (Stratum 0 / Stratum 1 / proxy / Clients).
 - ▶ The architecture has been validated and is being used in several SDCs for Data Challenge 6.

Ongoing activity

- ▶ Data Challenge 6:
 - ▶ Deployment and test of new versions of DSS / IAL / monitoring
 - ▶ Work to provide dedicated CentOS 7 working nodes (required for this Challenge)
 - ▶ Work on the CernVM-FS architecture and the support of SDCs
 - ▶ Data and processing will be soon distributed into the SDCs to evaluate performances of the architecture

Expected 2016 activity

- ▶ We expect more and more activity from users: simulations, challenges, tests
- ▶ Scientific Challenge 2
 - ▶ Real prototypes of scientific codes will run on simulated raw data
 - ▶ More intensive jobs (computing and storage)
 - ▶ Will require CentOS 7 resources
 - ▶ Integrated in the Euclid architecture
- ▶ Investigation:
 - ▶ Data storage and access (I/O performances ?)
 - ▶ Multicore
 - ▶ HPC

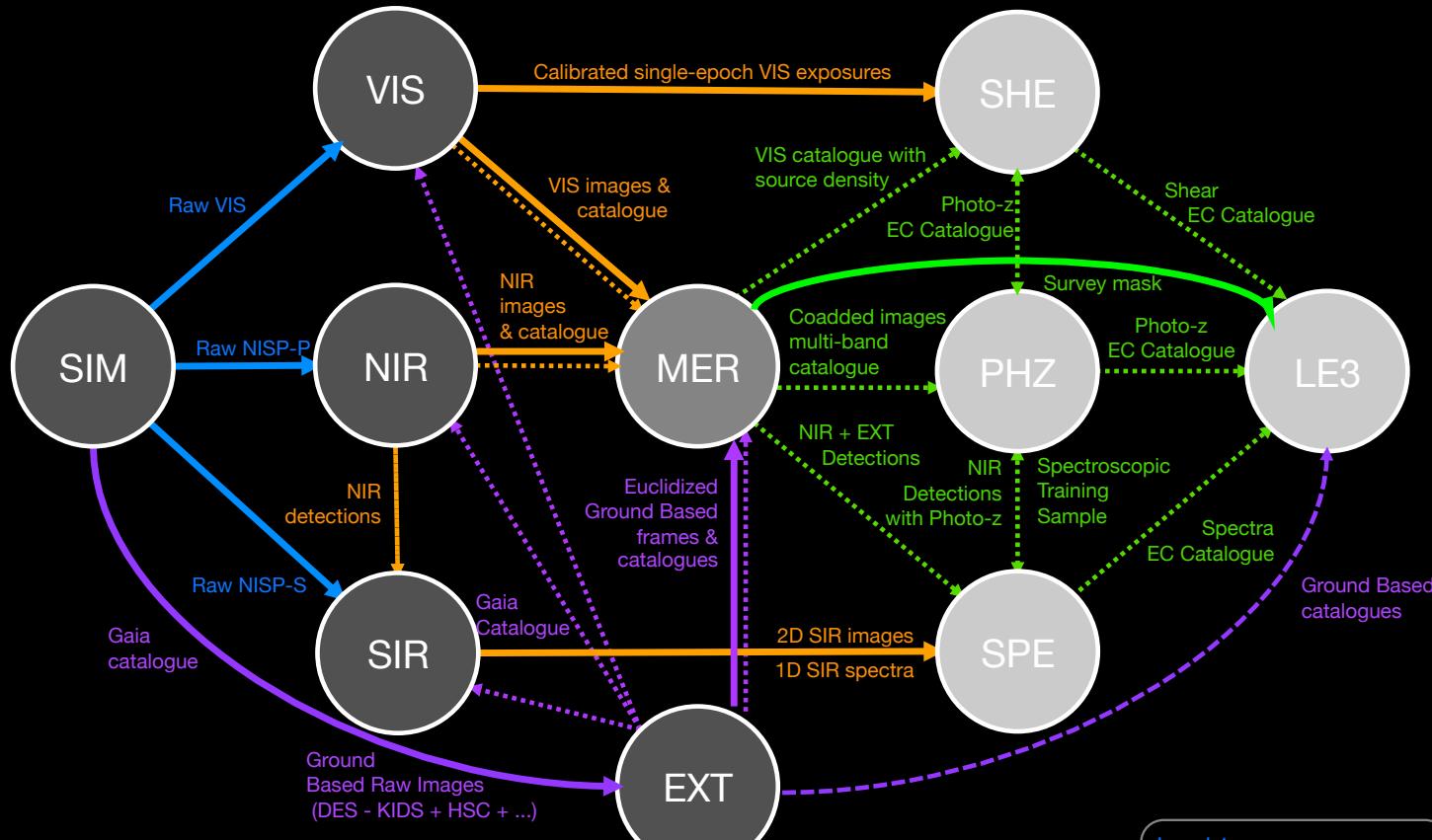


Thank you for your attention!



Backup

1. Euclid OUs overview



Level 1	Image data
Level 2	Catalogue data
Level 3	
Level EXT	

Virtual cluster architecture

