Computational requirememnts for Euclid

“Level 3” data analysis

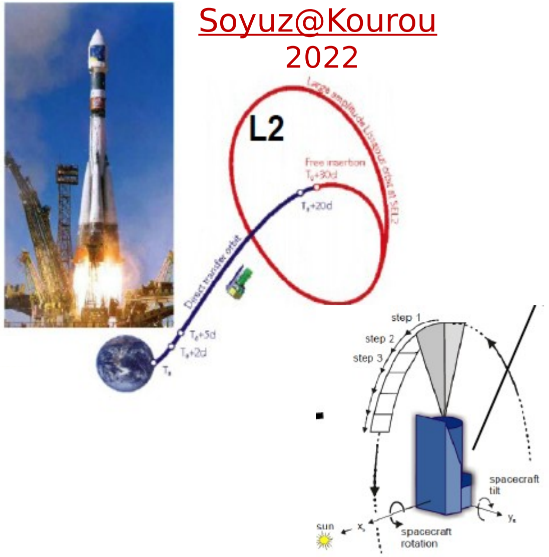
- Quick Euclid Overview
- Data analysis and its levels - Introducing Level 3
- Level 3 computational requirements - Specific examples
- Beyond level 3: science exploitation

F. Marulli, E. Branchini

on behalf of Euclid OU-LE3-WP

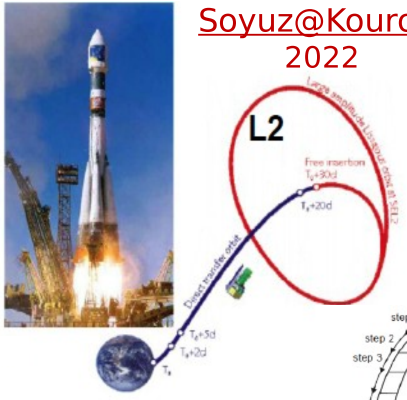
Euclid in a Nutshell

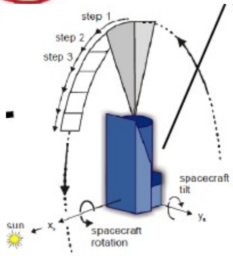
Soyuz@Kourou
2022



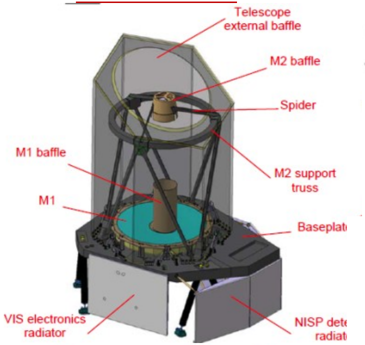
Euclid in a Nutshell

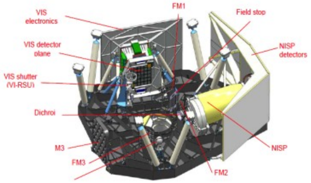
Soyuz@Kourou
2022

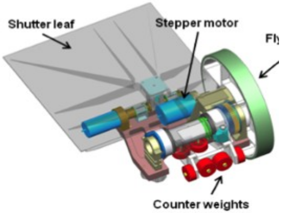


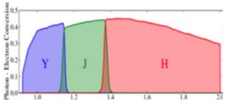


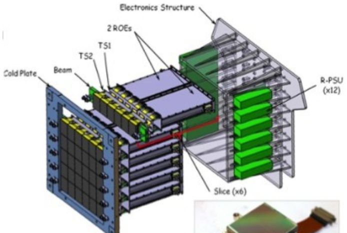
PLM+SVM:

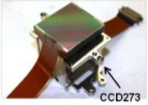


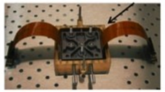






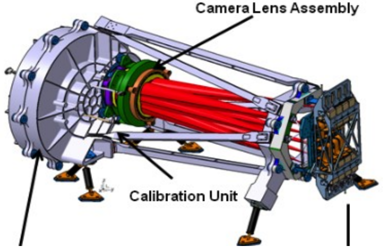




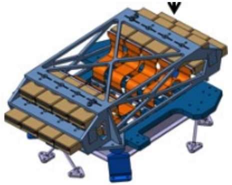


VIS imaging:

(VIS team)

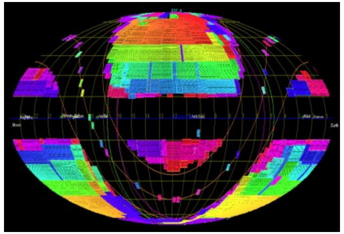


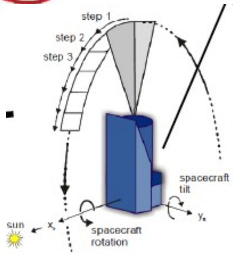
NIR spectro-imaging



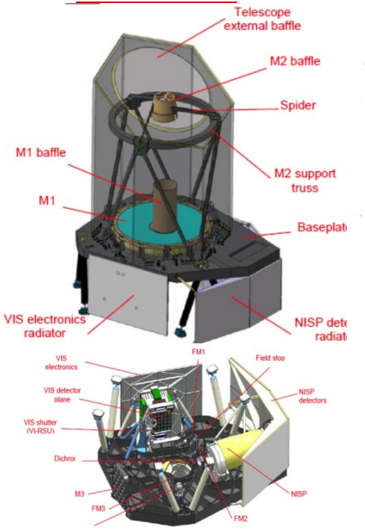
Euclid in a Nutshell

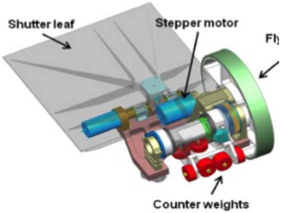
Soyuz@Kourou
2022

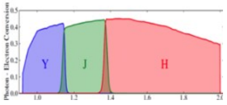


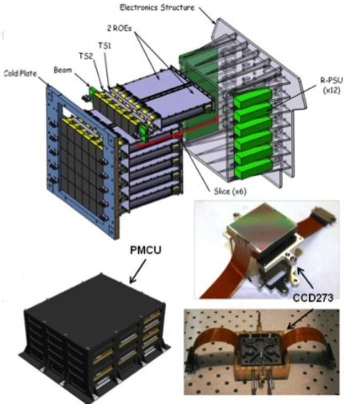


PLM+SVM:



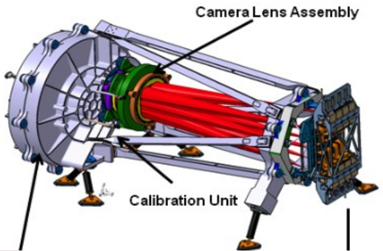




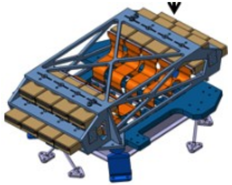


VIS imaging:

(VIS team)



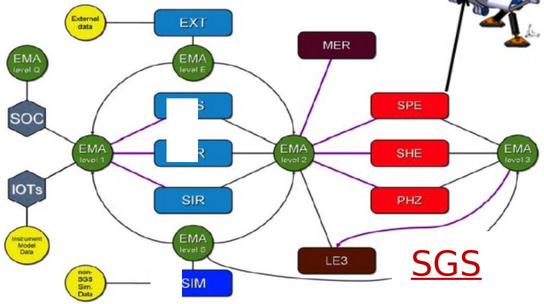
NIR spectro-imaging



Surveys: 2010-2027+ (Survey WG)

Commisioning – SV
Euclid operation:>5.5 yrs:
Euclid Wide+Deep

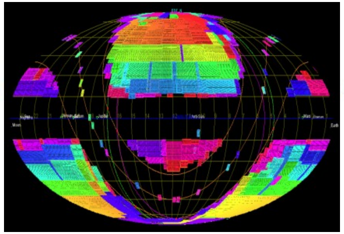
Ground data

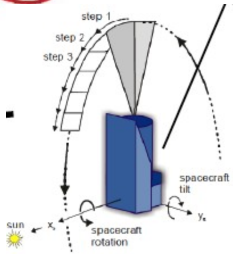


SGS

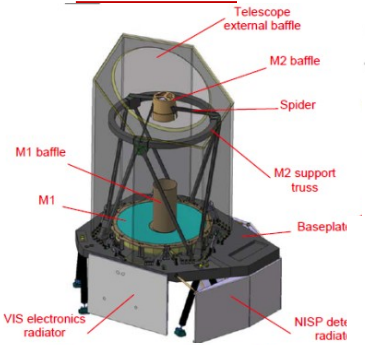
Euclid in a Nutshell

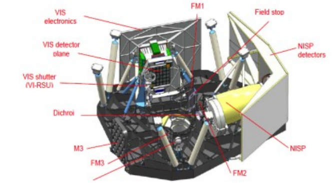
Soyuz@Kourou
2022

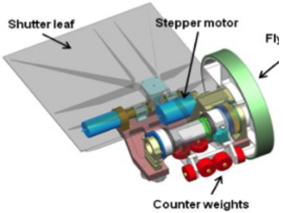


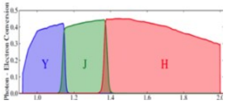


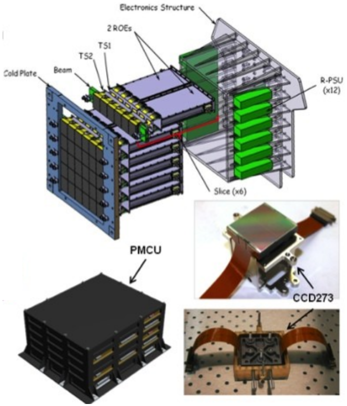
PLM+SVM:







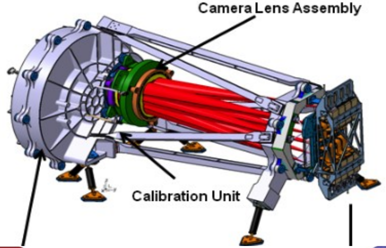


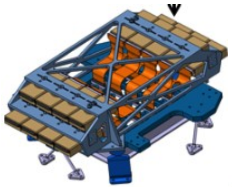


VIS imaging:

(VIS team)

NIR spectro-imaging

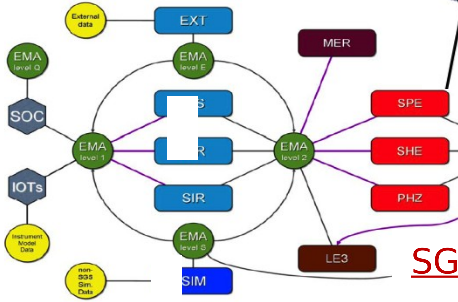




Surveys: 2010-2027+ (Survey WG)

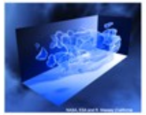
Commisioning – SV
Euclid operation:>5.5 yrs:
Euclid Wide+Deep

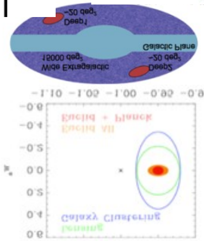
Ground data

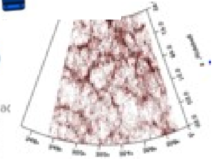


SGS

SWG:



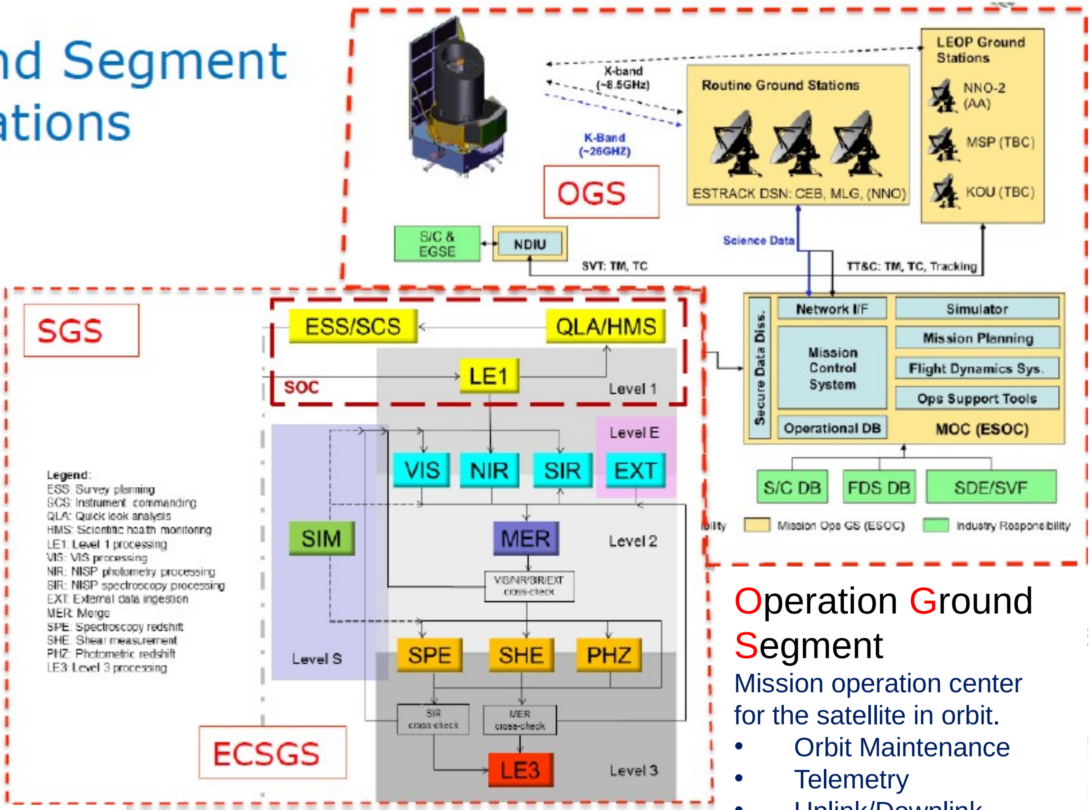




Euclid Science Ground Segment Operation

Ground Segment Operations

Science Operation Center
Data handling +
First Level data processing



Operation Ground Segment

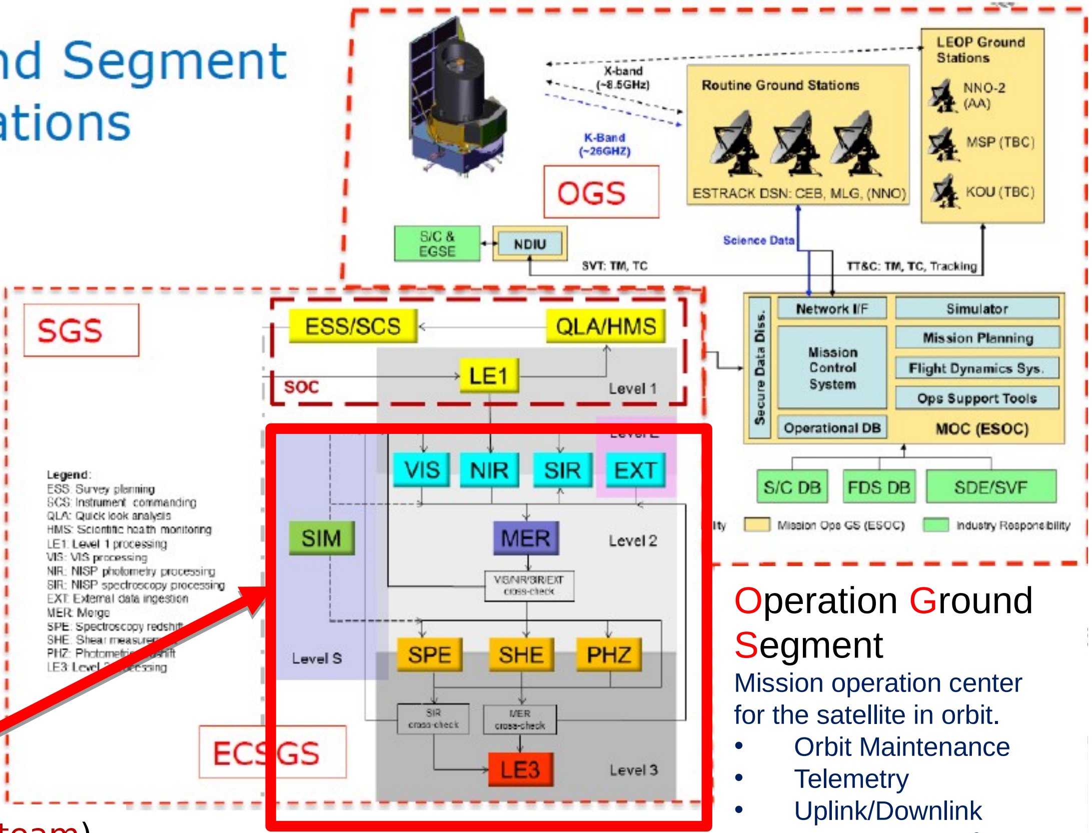
Mission operation center for the satellite in orbit.

- Orbit Maintenance
- Telemetry
- Uplink/Downlink
- Management of Ground stations

Euclid Science Ground Segment Operation

Ground Segment Operations

Science Operation Center
Data handling +
First Level data processing



This is where computational resources are most needed

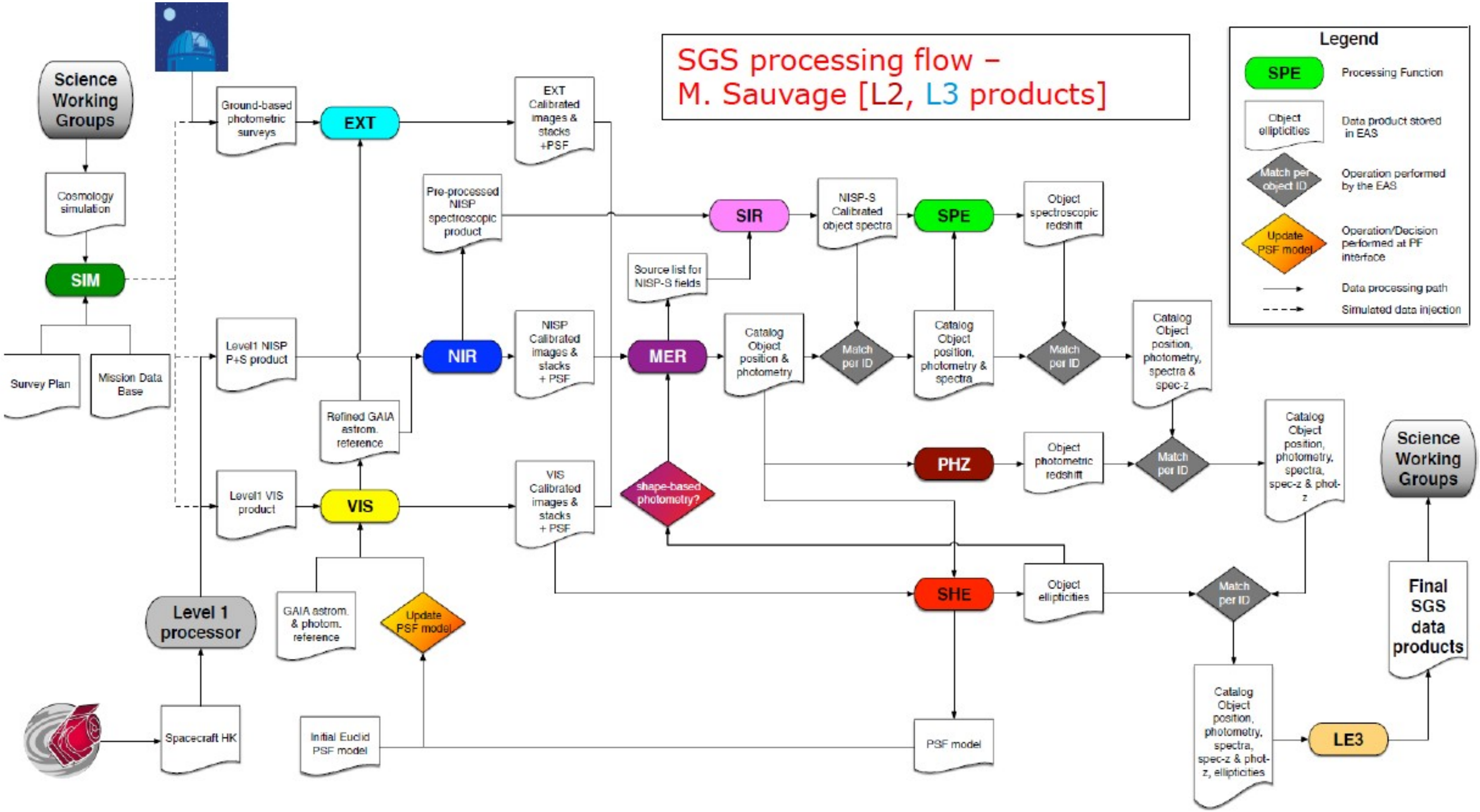
20-30 PB data processing (EC-SGS team)
Science analyses

Operation Ground Segment

Mission operation center for the satellite in orbit.

- Orbit Maintenance
- Telemetry
- Uplink/Downlink
- Management of Ground stations

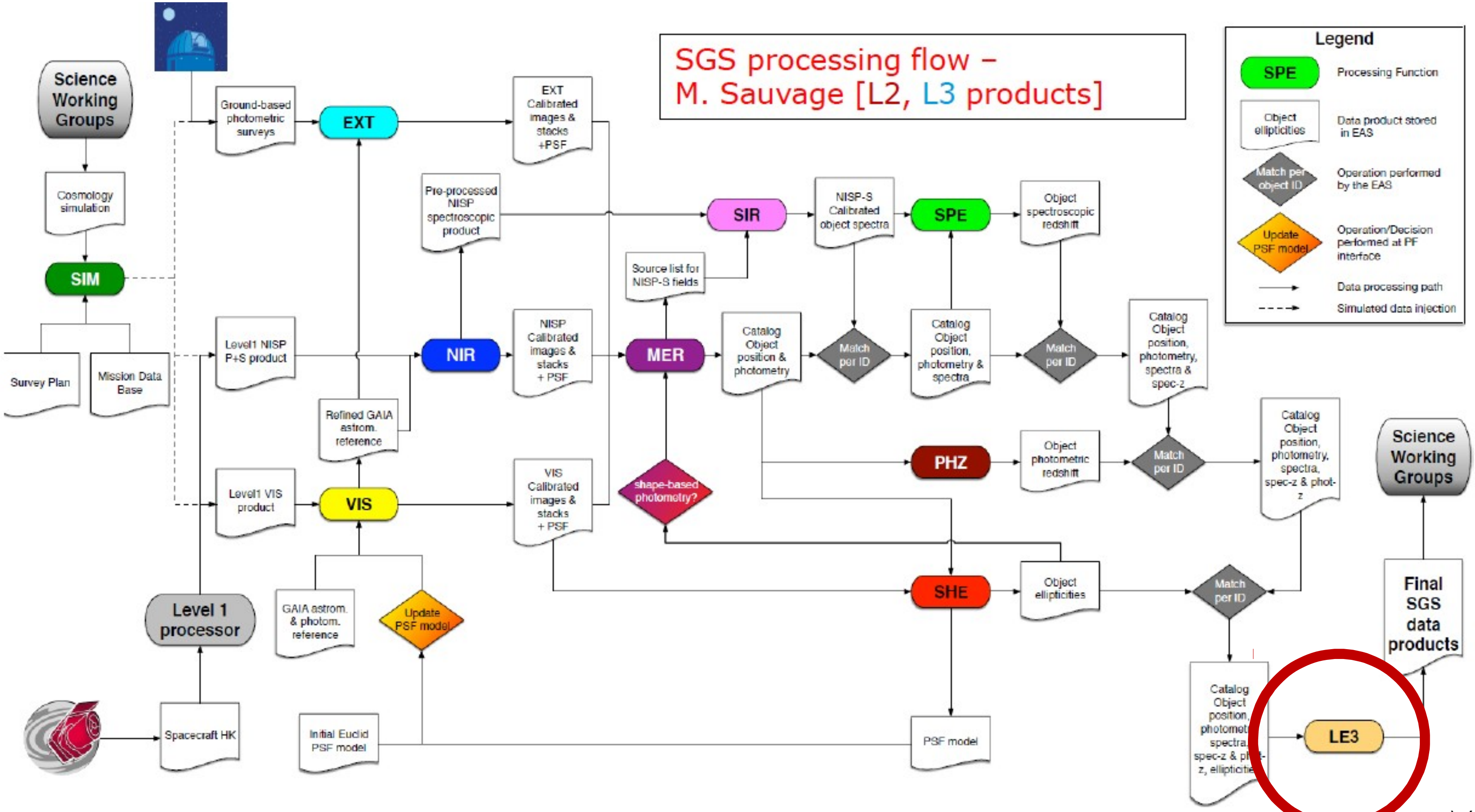
Euclid Science Ground Segment Structure



SGS processing flow –
M. Sauvage [L2, L3 products]

Science Working Groups
Cosmology simulation
SIM
Survey Plan
Mission Data Base
Ground-based photometric surveys
EXT
EXT Calibrated images & stacks +PSF
Pre-processed NISP spectroscopic product
Level1 NISP P+S product
NIR
NISP Calibrated images & stacks + PSF
Refined GAIA astrom. reference
Level1 VIS product
VIS
VIS Calibrated images & stacks + PSF
GAIA astrom. & photom. reference
Update PSF model
Level 1 processor
Spacecraft HK
Initial Euclid PSF model
Source list for NISP-S fields
SIR
NISP-S Calibrated object spectra
SPE
Object spectroscopic redshift
MER
Catalog Object position & photometry
Match per ID
Catalog Object position, photometry & spectra
Match per ID
Catalog Object position, photometry, spectra & spec-z
shape-based photometry?
PHZ
Object photometric redshift
Match per ID
Catalog Object position, photometry, spectra, spec-z & phot-z
SHE
Object ellipticities
Match per ID
PSF model
Catalog Object position, photometry, spectra, spec-z & phot-z, ellipticities
LE3
Final SGS data products
Science Working Groups
Legend
SPE
Processing Function
Object ellipticities
Data product stored in EAS
Match per object ID
Operation performed by the EAS
Update PSF model
Operation/Decision performed at PF interface
Data processing path
Simulated data injection

Euclid Science Ground Segment Structure



SGS processing flow –
M. Sauvage [L2, L3 products]
Legend
SPE
Processing Function
Object ellipticities
Data product stored in EAS
Match per object ID
Operation performed by the EAS
Update PSF model
Operation/Decision performed at PF interface
Data processing path
Simulated data injection
Science Working Groups
Cosmology simulation
SIM
Survey Plan
Mission Data Base
Ground-based photometric surveys
EXT
EXT Calibrated images & stacks +PSF
Pre-processed NISP spectroscopic product
Level1 NISP P+S product
NIR
NISP Calibrated images & stacks + PSF
Refined GAIA astrom. reference
Level1 VIS product
VIS
VIS Calibrated images & stacks + PSF
GAIA astrom. & photom. reference
Update PSF model
Level 1 processor
Spacecraft HK
Initial Euclid PSF model
Source list for NISP-S fields
SIR
NISP-S Calibrated object spectra
SPE
Object spectroscopic redshift
MER
shape-based photometry?
Catalog Object position & photometry
Match per ID
Catalog Object position, photometry & spectra
Match per ID
Catalog Object position, photometry, spectra & spec-z
PHZ
Object photometric redshift
Match per ID
Catalog Object position, photometry, spectra, spec-z & phot-z
SHE
Object ellipticities
Match per ID
PSF model
Catalog Object position, photometry, spectra, spec-z & phot-z, ellipticities
LE3
Final SGS data products
Science Working Groups

Level 1-2 vs. Level 3

Data Level Products and computational need :

- **Level 1-2 data products** (mostly data reduction: photometry, spectroscopy, redshift estimates, shear estimates) require significant computational facilities (Euclid will analyze O(2) billions of objects). However, this type of data analysis **can be carried out independently by different computational facilities** (Science Data Centers) since disjoint patches of sky can be analyzed independently. Coordination is required at the end of the process.

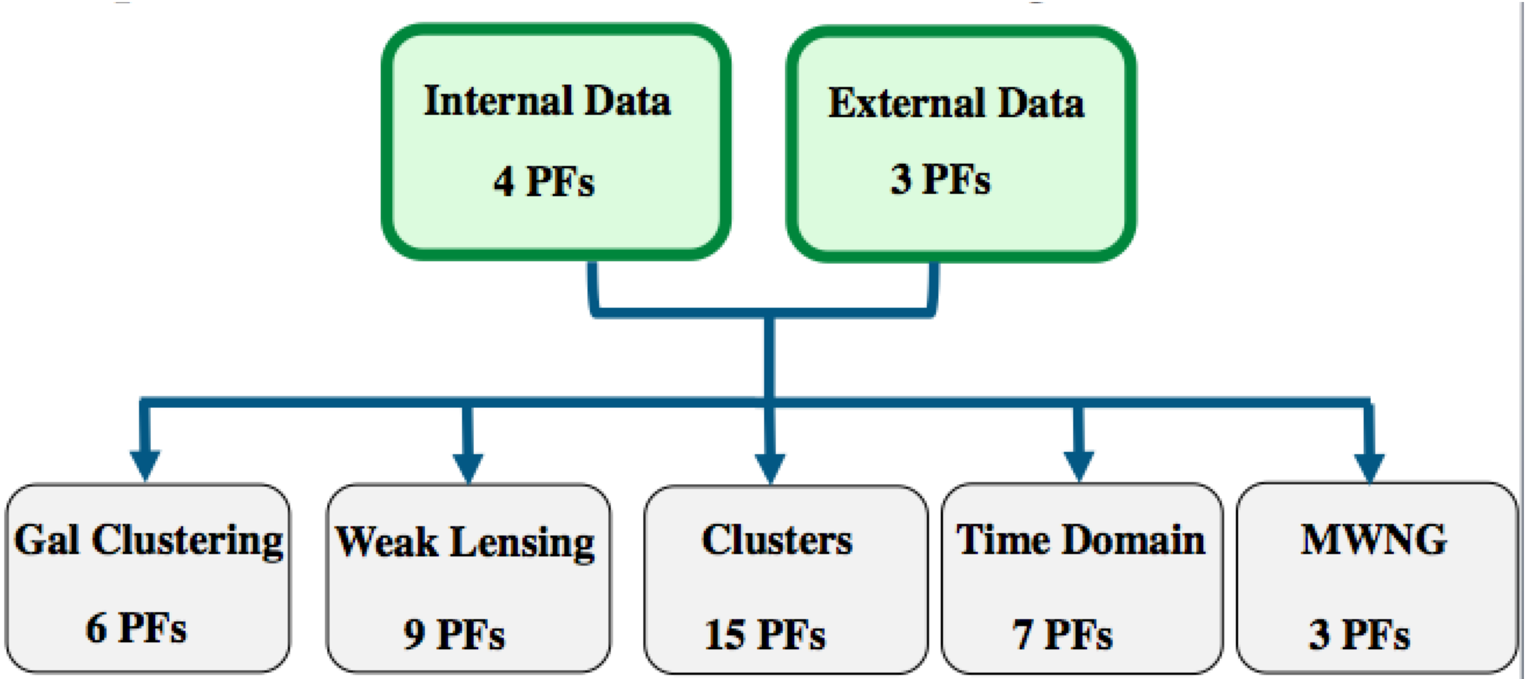
Level 2 vs. Level 3

Data Level Products and computational need :

- **Level 1-2 data products** (mostly data reduction: photometry, spectroscopy, redshift estimates, shear estimates) require significant computational facilities (Euclid will analyze O(2) billions of objects). However, this type of data analysis **can be carried out independently by different computational facilities** (Science Data Centers) since disjoint patches of sky can be analyzed independently. Coordination is required at the end of the process.
- **Level 3 Data Products** (generation of catalogs, their selection functions and their statistical analyses) also require large computational facilities. In most cases**, they consist of statistical analyses that need to be carried out on the full sample.** Correlation studies provide the best example. Performing a correlation study on two separate patches of sky misses a significant amount of information that can be obtained instead by analyzing a single patch with equal area. This means that most of LE3 analyses need to be performed on a single computational facility capable of high performing computing.

Level 3 in detail

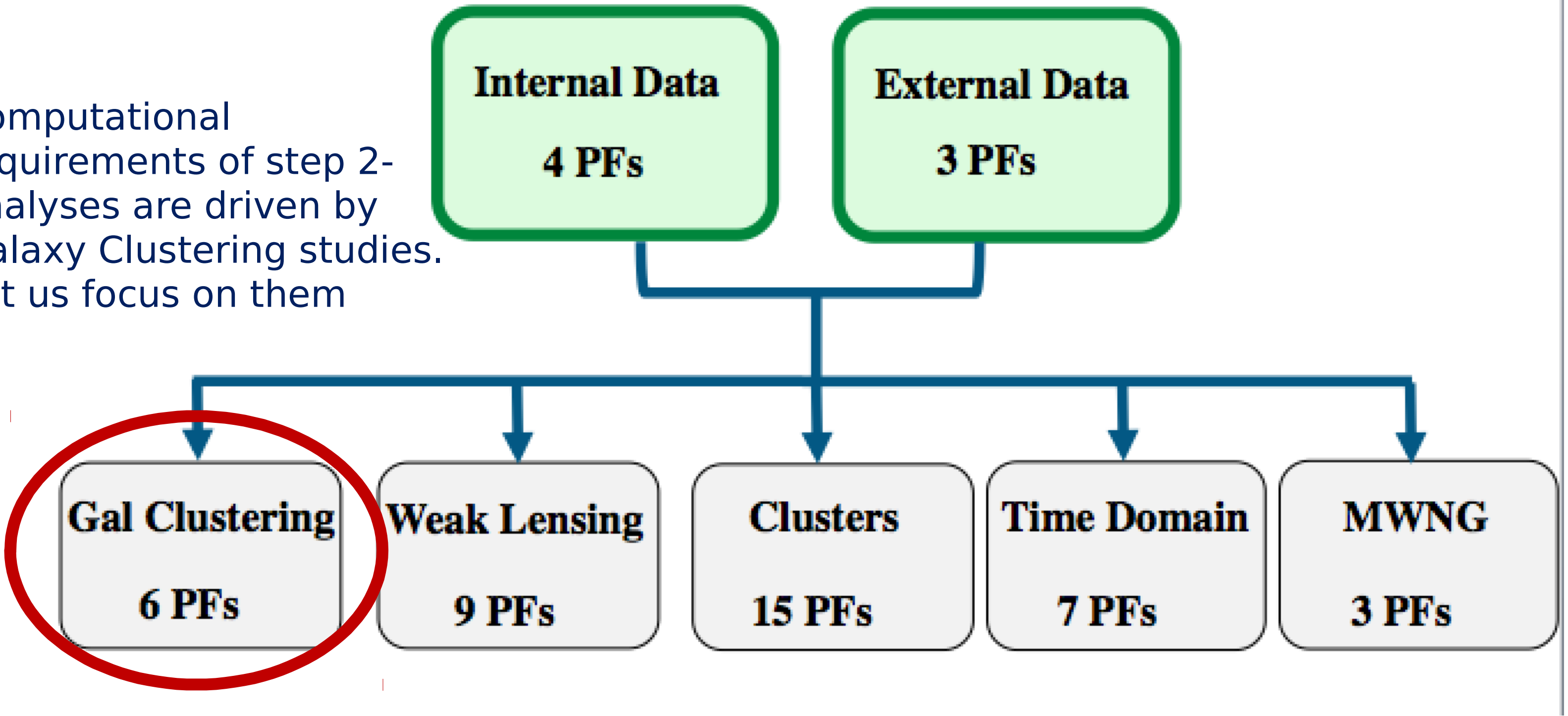
LE3 is a 2-step data analysis. Step 1: collection of LE2 info and production of catalogs and their selection functions (green boxes). Step 2: statistical analyses of the LE3 catalogs (grey boxes). These analyses are largely cosmologically-oriented.



Level 3 in detail

LE3 is a 2-step data analysis. Step 1: collection of LE2 info and production of catalogs and their selection functions (green boxes). Step 2: statistical analyses of the LE3 catalogs (grey boxes). These analyses are largely cosmologically-oriented.

Computational requirements of step 2-analyses are driven by Galaxy Clustering studies. Let us focus on them



2-point correlation function of galaxies and clusters

2PCF:

The main goal of this Processing Function is to measure the anisotropic **2-point correlation function of about 30M galaxies**. It is a very standard tool that, however, needs to be applied to an unprecedented dataset. Its estimates consist of counting pairs of galaxies and comparing these counts with the analogous quantities measured in a catalog of randomly distributed objects. These “random” objects are significantly more numerous than galaxies.
The main effort so far has been that of optimizing the estimator to reduce the CPU requirement.

In its current version the code **requires ~7500 core-hours*** to estimate the 2-point correlation of the full Euclid Spectroscopic catalog of 30M objects. The required **RAM is 150 GB**.

As anticipated, this correlation analysis cannot be carried out independently on different sky patches analysed at different computational facilities. They require the use of a single machine in which the full catalog is stored and analysed.

*Obtained using computer nodes with two Intel Xeon E5-2680 v3 with clock speed 2.50 GHz. Each node had 24 physical cores

Power spectrum of galaxies and clusters

PK:

The goal of this Processing Function is analogous to that of 2PCF, except for the fact that the analysis is carried out in Fourier space. This strategy significantly reduces the CPU requirement. The processing function estimates the anisotropic **galaxy power spectrum and its moments**. Like in the 2PCF case, a catalog of random objects is required. However, the most computationally demanding step is related to memory usage and to the creation and storage of a large cubic grid at the vertex of which physical quantities are specified.

In its current version the code **requires ~10 core-hours*** to estimate the power spectrum of the full Euclid Spectroscopic catalog of 30M objects. The required **RAM is ~120 GB**.

As for the 2PCF case, this analysis cannot be carried out independently on different sky patches analysed at different computational facilities. They require the use of a single machine in which the full catalog is stored and analysed.

*Obtained using a workstation with 3.20 GHz processors.

Covariance matrices

Covariance Matrices (CM-2PCF and CM-PK):

Together with the measurement of 2-point statistics, one has to provide an estimate of the errors, i.e. one needs to measure the covariance matrices. This is a very challenging task, since many independent measurements of these two-point statistics need to be performed to build a covariance matrix. So, in practice, the CPU time required to measure a covariance matrix can be obtained by multiplying the CPU time needed to measure the two-point function with the number of individual measurements required. It is therefore fundamental to minimize the number of these individual measurements. This can be done using theoretical priors.
The most recent, conservative assessments suggest that for Euclid one will need ~5000 such measurements.
Therefore the CPU requirement to estimate the covariance matrix of the 2-point correlation function is **~5000x7500=37.5M core-hours***. The good news is that the measurements of the individual 2-point functions can be performed separately and simultaneously at different facilities, so that the RAM requirement does not change **(120 GB)**.
The computational requirements for the covariance matrix of the power spectrum are comparatively negligible.

*Obtained using computer nodes with two Intel Xeon E5-2680 v3 with clock speed 2.50 GHz. Each node had 24 physical cores

Higher order statistics

Bispectrum (BK-GC) and three-point function (3PCF):

The computation of higher order statistics (we only consider 3-point statistics) is even more challenging than that of 2-point statistics. In this case, being able to measure such quantities in a sample with the same size as Euclid is mandatory to use fast and necessarily approximated estimators that, however, need to guarantee the expected accuracy.
After much development, the current codes achieve this goal. Of course at the price of considerable computational requirements that, for the 3-point correlation function, consists in CPU requirement, whereas for the bispectrum in Fourier space consists of RAM requirements.

The current Euclid code for the 3-point function **require ~9x10^5 core-hours** to analyse the full Euclid sample, whereas for the bispectrum only 200 core-hours are needed. In terms of RAM, the 3-point function requires **~300 GB**, whereas the bispectrum needs **~1.1 TB**.

Additional computational requirements

Ground Segment Computational Budget and Beyond:

The computational requirements previously illustrated (and in general, all computational needs for Level 3 and Level 2 analyses) are accounted for in the global computational budget of the Ground Segment. In other words, the Euclid Consortium should be able to provide the required computational facilities to produce all required data products, including LE3 statistics.
However, there is plenty of additional activities, largely related to science exploitation, that are not included in this computational budget and for which computational facilities should be provided by external entities.
There are basically three types of activities that will require additional (and significant) computational facilities:

- **Activities that generate ancillary datasets required for science analysis.**
- **Activites that produce LE3-like datasets, currently not included in the computational budget.**
- **Activities related to scientific exploitation of the data products from LE3.**

We shall provide a few examples in the next slides.

Additional datasets

Additional LE3-like datasets :

The current computational budget accounts for a limited number of LE3 datasets. For example: it assumes that one Euclid spectroscopic catalog will be produced and its 2-point correlation function will be computed.

But what if one wants to extract some specific subset of objects (i.e. red vs. blue galaxies), and compute their correlation function or their cross-correlation? Or what if one wants to manipulate the catalog to enhance some cosmological feature of particular interest like the BAO peak in the 2-point correlation function?

None of this is currently expected to be produced by the official Euclid pipeline.

It appears that external computational facilities would be needed to guarantee their production. Clearly, the ability to provide such facility would almost automatically imply that the scientific exploitation of that particular product will be performed by and granted to the provider.

Scientific exploitation

Scientific exploitation of LE3 data products :

Within the Euclid Consortium the scientific exploitation of the data products produced by the Ground Segment is responsibility of the Scientific Working Group, that provides the know-how to extract the scientific information.

However, it is unclear if they will also be able to provide the computational resources to do that. Indeed, some of these analyses are computationally intensive ane require significant computational facilities.

Once again, providing such facilities would allow one to lead some specific scientific project and enjoy the best share of scientific return.