LISA (distributed) Data Processing “Center”

Antoine Petiteau

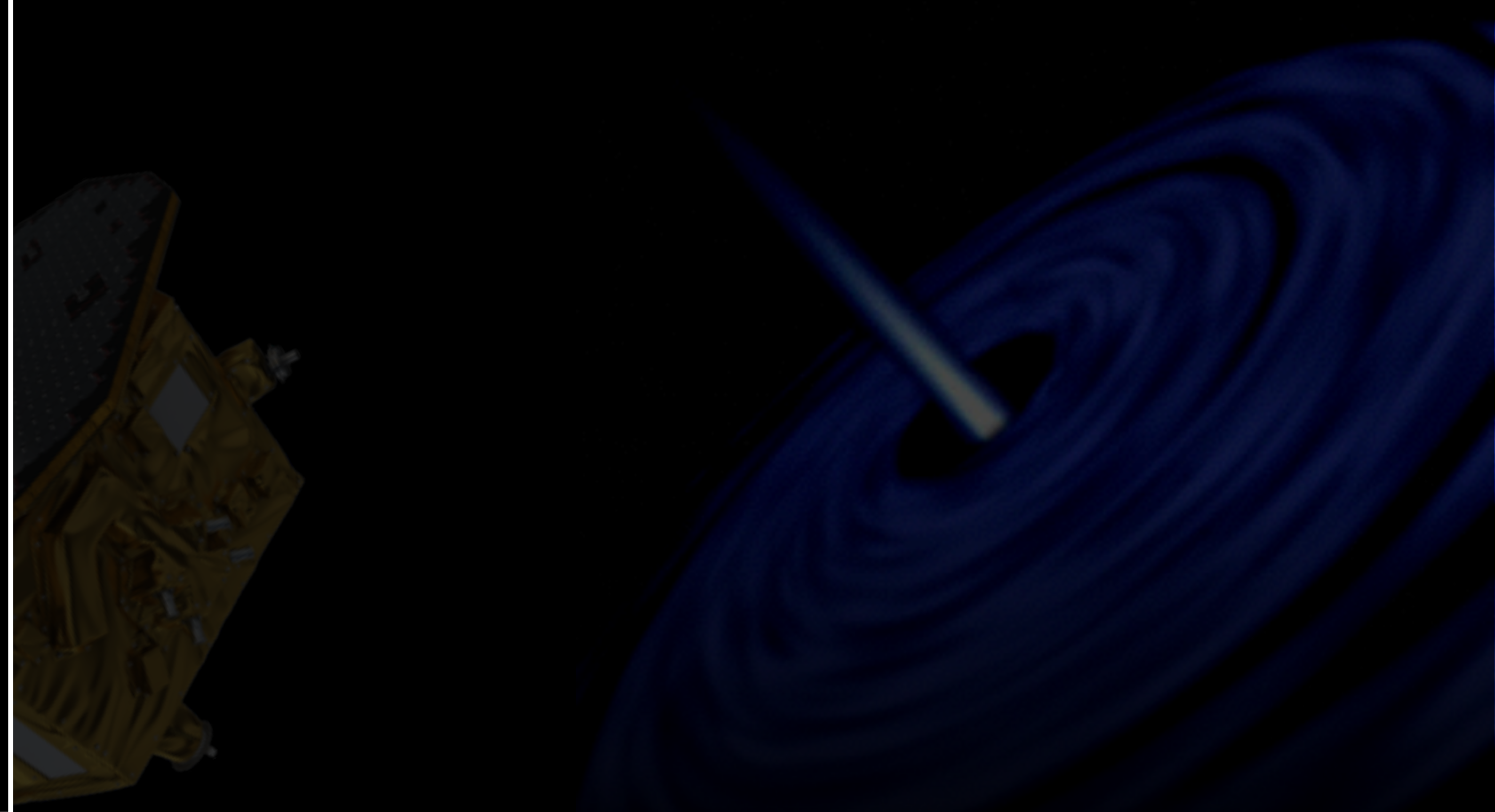
In collaboration with M. Le Jeune [1] & L. Chaoul [2]

[1] APC – Université Paris-Diderot/CNRS
[2] CNES - Toulouse

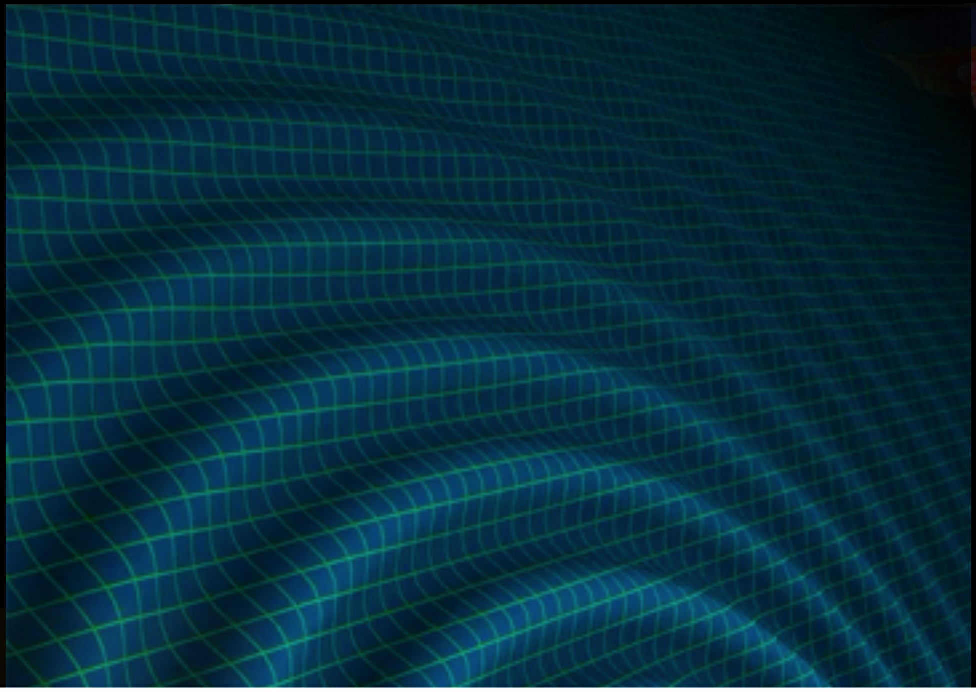
KISS workshop "The Architecture of LISA Science Analysis"
CalTech - 17th January 2018

LISA data

Gravitational wave sources emitting between 0.02mHz and 1 Hz

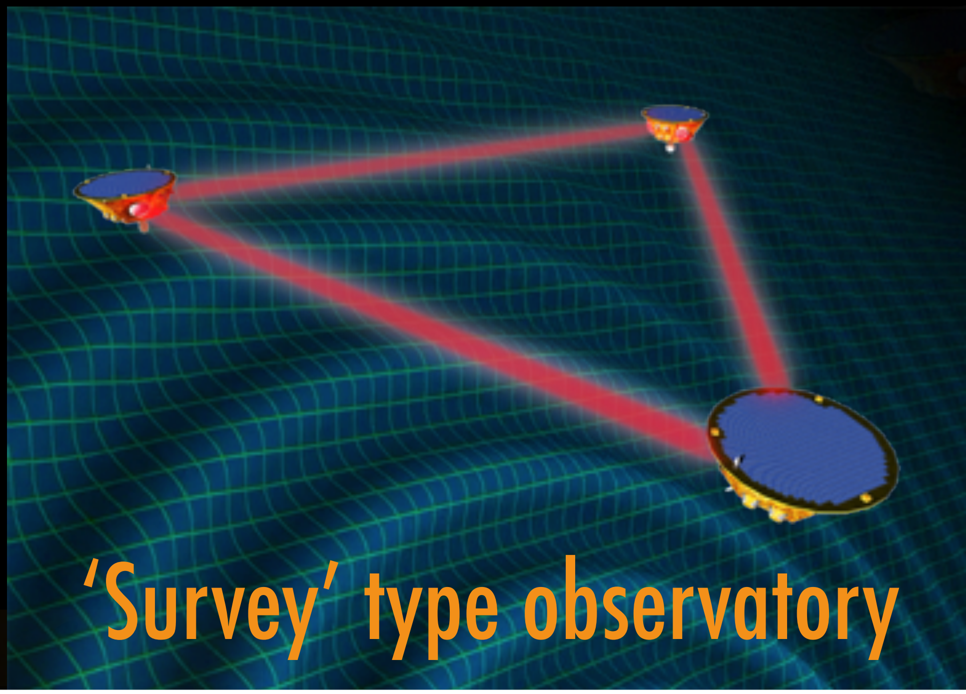


LISA data



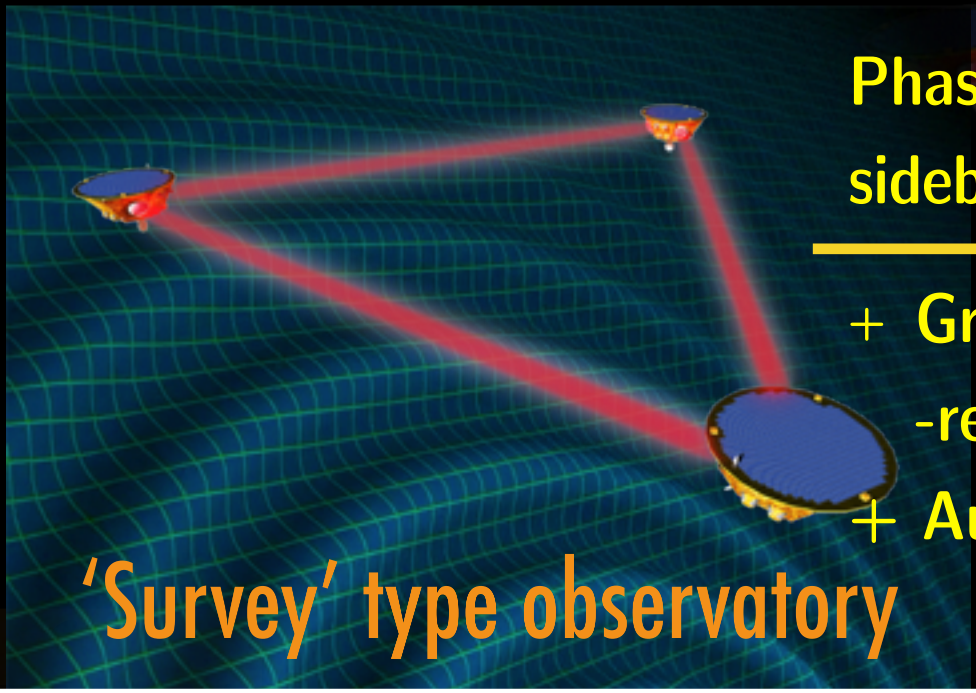
Gravitational wave sources emitting between 0.02mHz and 1 Hz

LISA data



Gravitational wave sources emitting between 0.02mHz and 1 Hz

LISA data

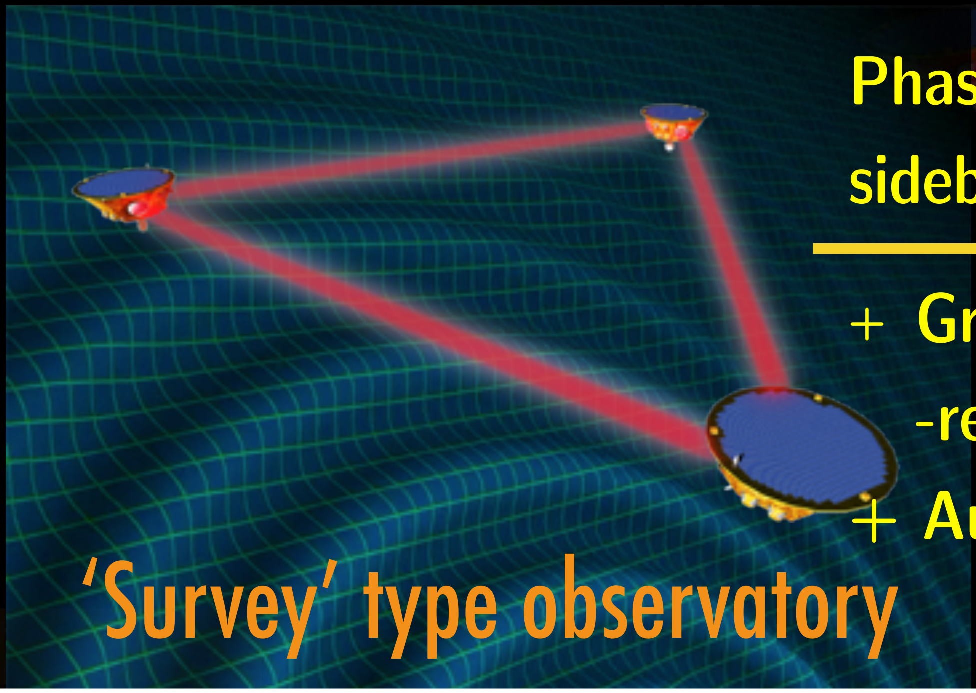


Phasemeters (carrier, sidebands, distance)

+ Gravitational Refe- -rence Sensor
+ Auxiliary channels

Gravitational wave sources emitting between 0.02mHz and 1 Hz

LISA data



Phasemeters (carrier, sidebands, distance)

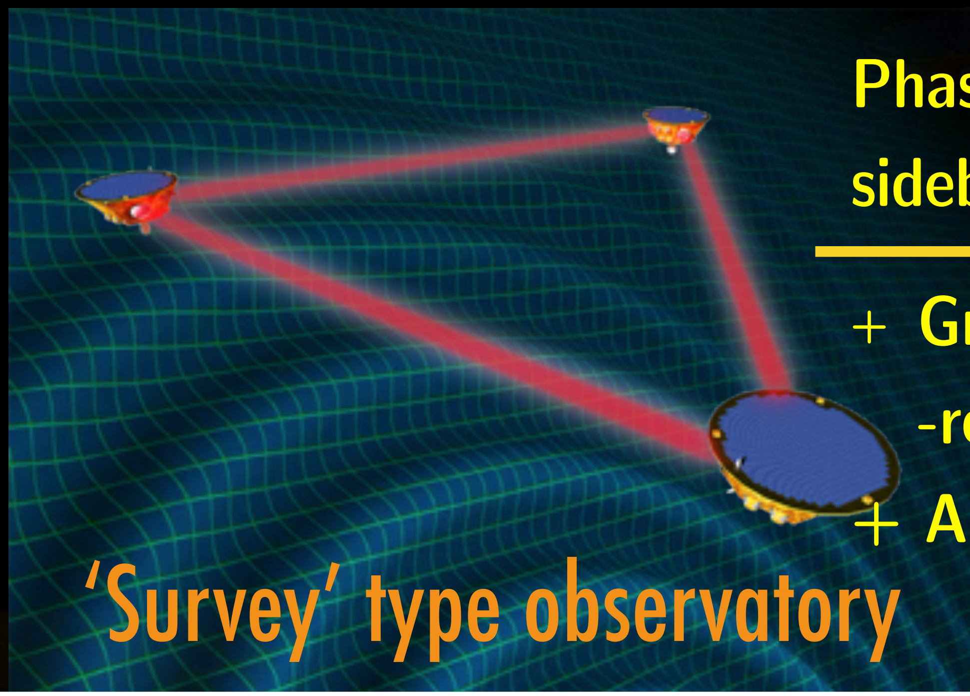
+ Gravitational Reference Sensor
+ Auxiliary channels

Gravitational wave sources emitting between 0.02mHz and 1 Hz

LISA data

Phasemeter
sidebands, c

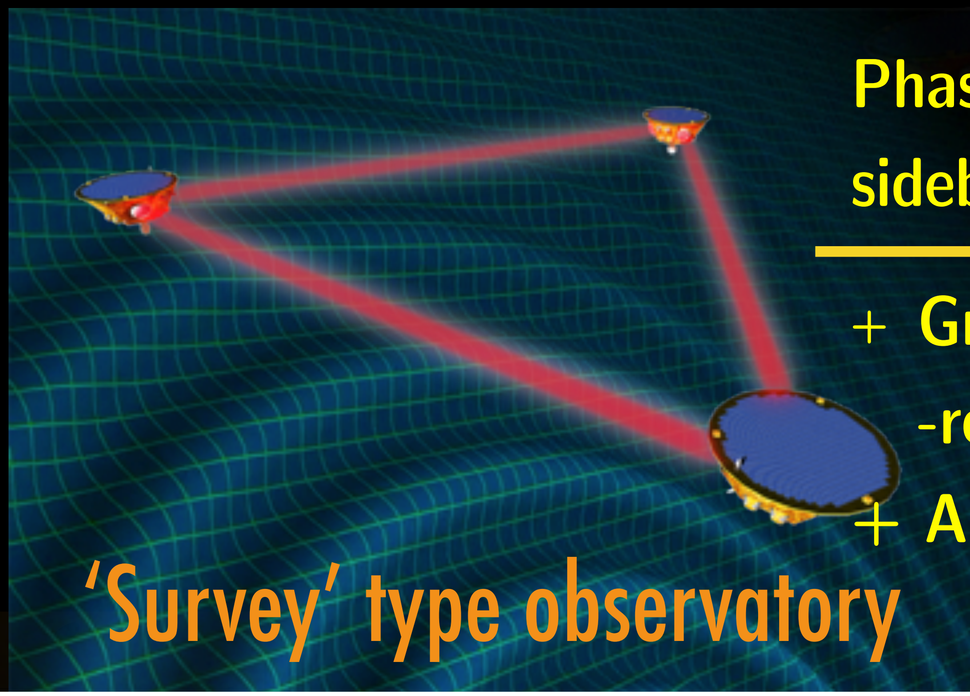
+ Gravitatio
-rence Se
+ Auxiliary



Gravitational wave sources emitting between 0.02mHz and 1 Hz

Source		Measurement	Channel Count	Sample Rate [Hz]	Bits per Channel	Rate [bits/s]
		Payload				
IFO Longitudinal		Inter-S/C IFO	2	3,0	64	384,0
		Test Mass IFO	2	3,0	64	384,0
		Test mass y IFO	0	3,0	64	0,0
		Reference IFO	2	3,0	64	384,0
		Clock Sidebands	4	3,0	64	768,0
Freq reference		error point	1	3,0	32	96,0
		feedback	2	3,0	32	192,0
		clock sidebands monitoring (local pilot tone beat)	1	3,0	32	96,0
IFO Angular		SC η,φ	4	3,0	32	384,0
		TM η,φ	4	3,0	32	384,0
		TM θ (from y IFO)	0	3,0	32	0,0
Ancillary		Time Semaphores	4	3,0	64	768,0
Optical Monitoring		PRDS metrology	4	3,0	32	384,0
			0	3,0	32	0,0
		Optical Truss	0	3,0	32	0,0
DFACS / GRS Cap. Sens.		TM x,y,z	6	1,0	32	192,0
		TM θ,η,φ	6	1,0	32	192,0
DFACS		breathing errorpoint	0	1,0	32	0,0
		breathing actuator	2	1,0	32	64,0
		TM applied torques	12	1,0	24	288,0
		TM applied forces	12	1,0	24	288,0
		SC applied torques	3	1,0	24	72,0
		SC applied forces	3	1,0	24	72,0
Science Diagnostics	Themometers	EH	16	0,1	32	51
		OB	20	0,1	32	64
		Telescope	10	0,1	32	32
		interface	10	0,1	32	32
	Magnetometers	TM	12	0,1	32	38
	radiation monitor		1			30
	FIOS output powers (Inloop and Out of Loop)		6	3,0	32	576
	pressure sensor		0	0,1	32	0
	body mic	CGAS tanks	0	3,0	32	0
		breathing mechanism	0	3,0	32	0
	RIN monitoring	2 lasers, 2 frequencies, 2 quadratures	8	3,0	32	768
				0,0		0
				0,0		0
Payload HK						1000
Total Payload						***7984***
		Platform				
Housekeeping [Based on LPF]						4000
Total Platform						***4000***
		Totals				
Raw Rate per SC						11984
Packetisation Overhead [10%]						1198
Packaged Rate per SC						**13182**
Packaged Rate for Constellation						**39546**

LISA data



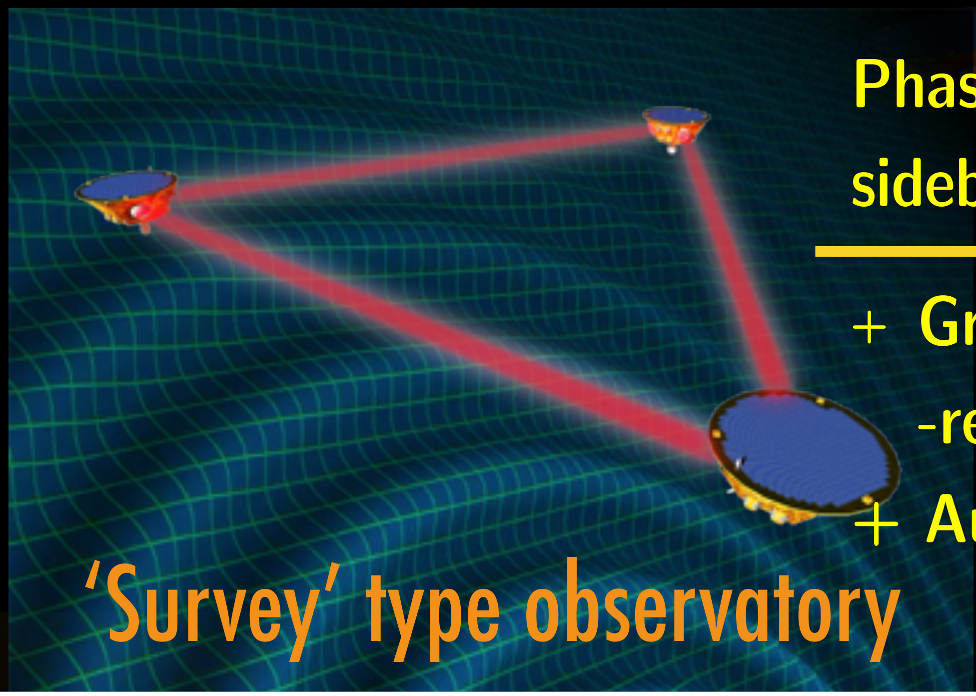
'Survey' type observatory

Phasemeters (carrier, sidebands, distance)

+ Gravitational Refe- -rence Sensor
+ Auxiliary channels

Gravitational wave sources emitting between 0.02mHz and 1 Hz

LISA data



‘Survey’ type observatory

Phasemeters (carrier, sidebands, distance)

+ Gravitational Refe- -rence Sensor
+ Auxiliary channels

Calibrations corrections

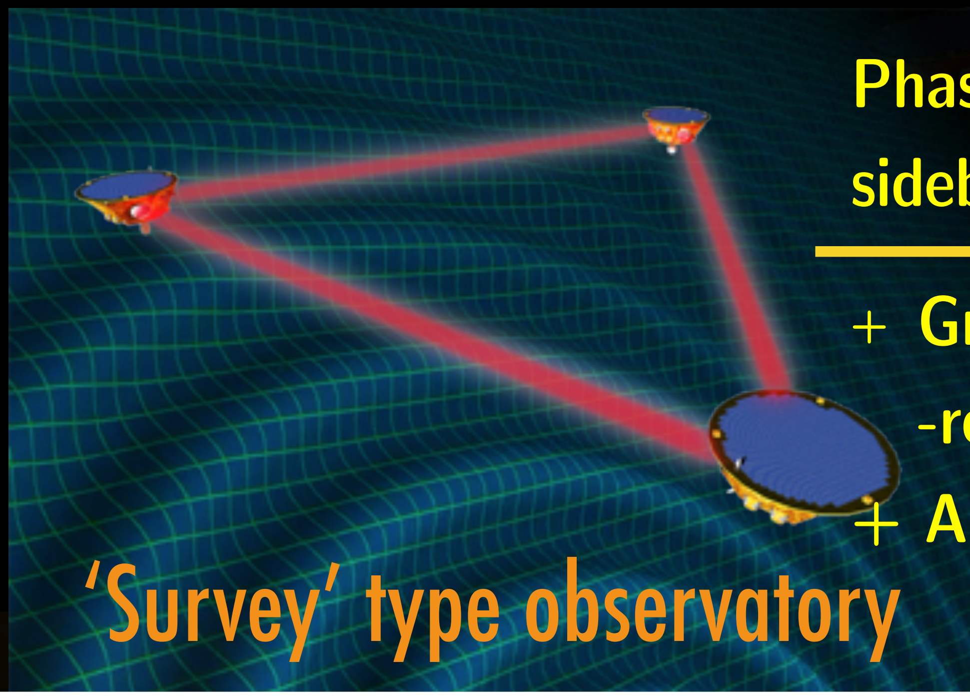
Resynchronisation (clock)

Time-Delay Interferometry reduction of laser noise

2 data channels TDI non-correlated

Gravitational wave sources emitting between 0.02mHz and 1 Hz

LISA data



'Survey' type observatory

Phasemeters (carrier, sidebands, distance)

+ Gravitational Refe- -rence Sensor
+ Auxiliary channels

Calibrations corrections

Resynchronisation (clock)

Time-Delay Interferometry reduction of laser noise

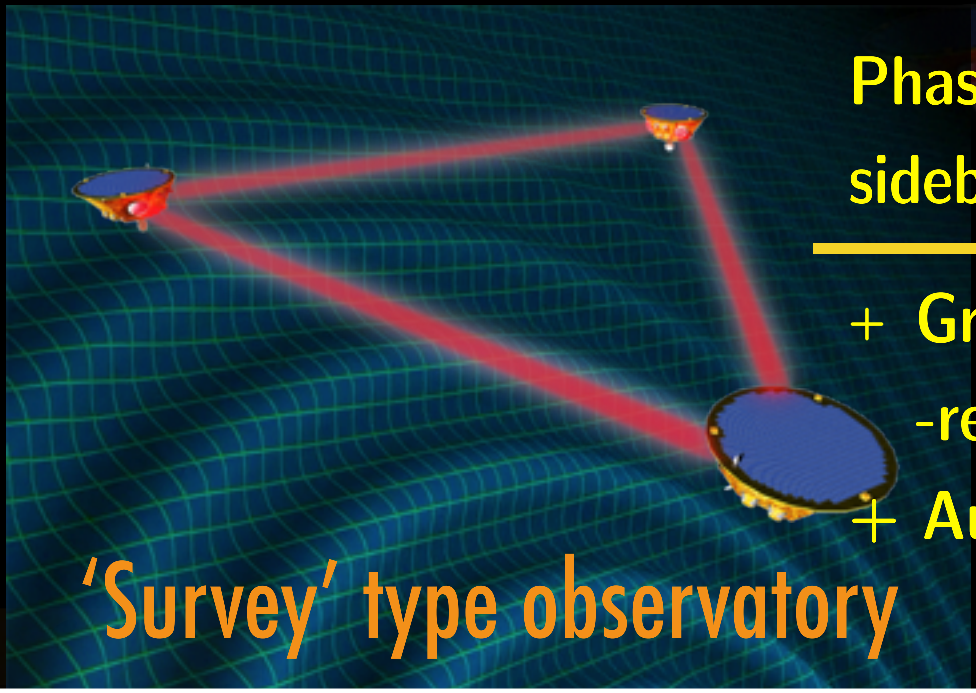
2 data channels TDI non-correlated

Data Analysis of GWs

Catalogs of GWs sources with their waveform

Gravitational wave sources emitting between 0.02mHz and 1 Hz

LISA data



'Survey' type observatory

Phasemeters (carrier, sidebands, distance)

+ Gravitational Refe- -rence Sensor

+ Auxiliary channels

L0

Calibrations corrections

Resynchronisation (clock)

Time-Delay Interferometry reduction of laser noise

L1 — 2 data channels TDI non-correlated

Data Analysis of GWs

L2

L3 — Catalogs of GWs sources with their waveform

Gravitational wave sources emitting between 0.02mHz and 1 Hz

LISA data level

- **Level L0 data**: raw science telemetry and housekeeping data.
- **Level L1 data**: TDI variables, all calibrated science data streams and auxiliary data.
- **Level L2**: intermediate waveform products such as partially regressed observable series (i.e., dataset obtained by progressively deeper subtraction of identified signals).
- **Level L3**: catalogs of identified sources, with faithful representations of posterior parameter distributions, data products related to key science questions.

LISA data volume

- Data volume to be stored:
  - Level L0: about 300 Mo per day
  - Level L1: about 600 Mo per day
  - Sub-product of the analysis: fews Go per day
  - Level L2 and L3: about 6 Go per day

LISA PROPOSAL

LISA data volume

- Data volume to be stored:
  - Level L0: 40kbps with 100% margin + 20kbps for high resolution => 1 Gb (Gigabytes) per day
  - Level L1: about 2 Gb per day
  - Sub-product of the analysis: few tens Gb per day
  - Level L2 and L3: few tens Gb per day

Phase 0 Payload

=> Storages and archives are not problematic ... but we need manage results of simulations (ex: more few Tb for LDC)

- Complexity for the DPC is mainly in data analysis because the goal is to extract the parameters for a maximum number of sources.

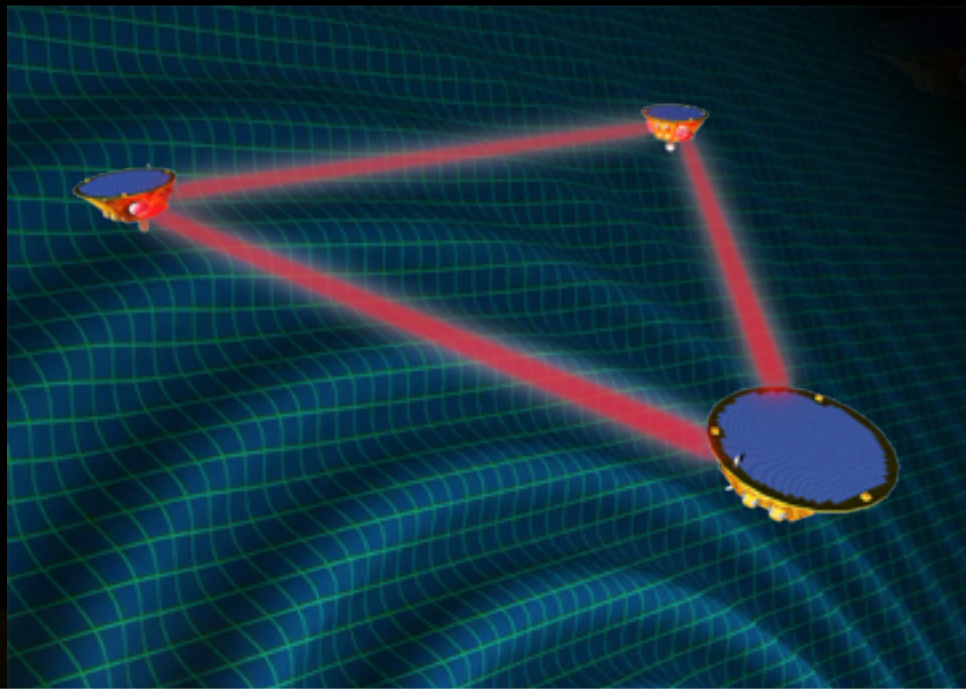
Particularities LISA data

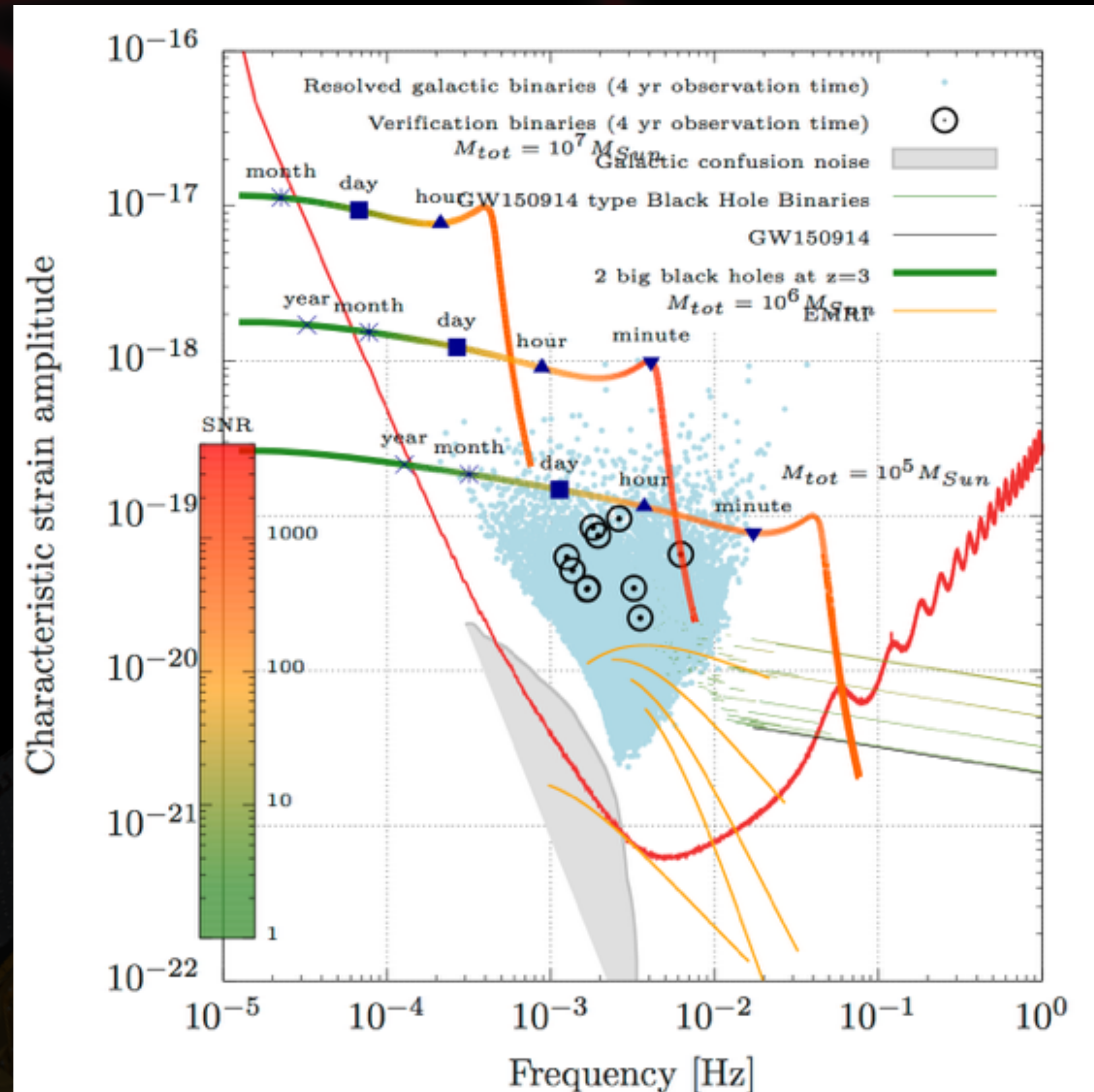
- First data of this kind
  - Discovery mission; no previous expertise on this kind of data
- Event rate is uncertain
  - Depending on the type of sources but typically from few tens to few thousands per year
- Potential unknown sources
- Transient sources + continuous sources

=> Constrains on data processing:

- Large fluctuation of computation needs
- Continuous evolution of the pipelines

LISA data



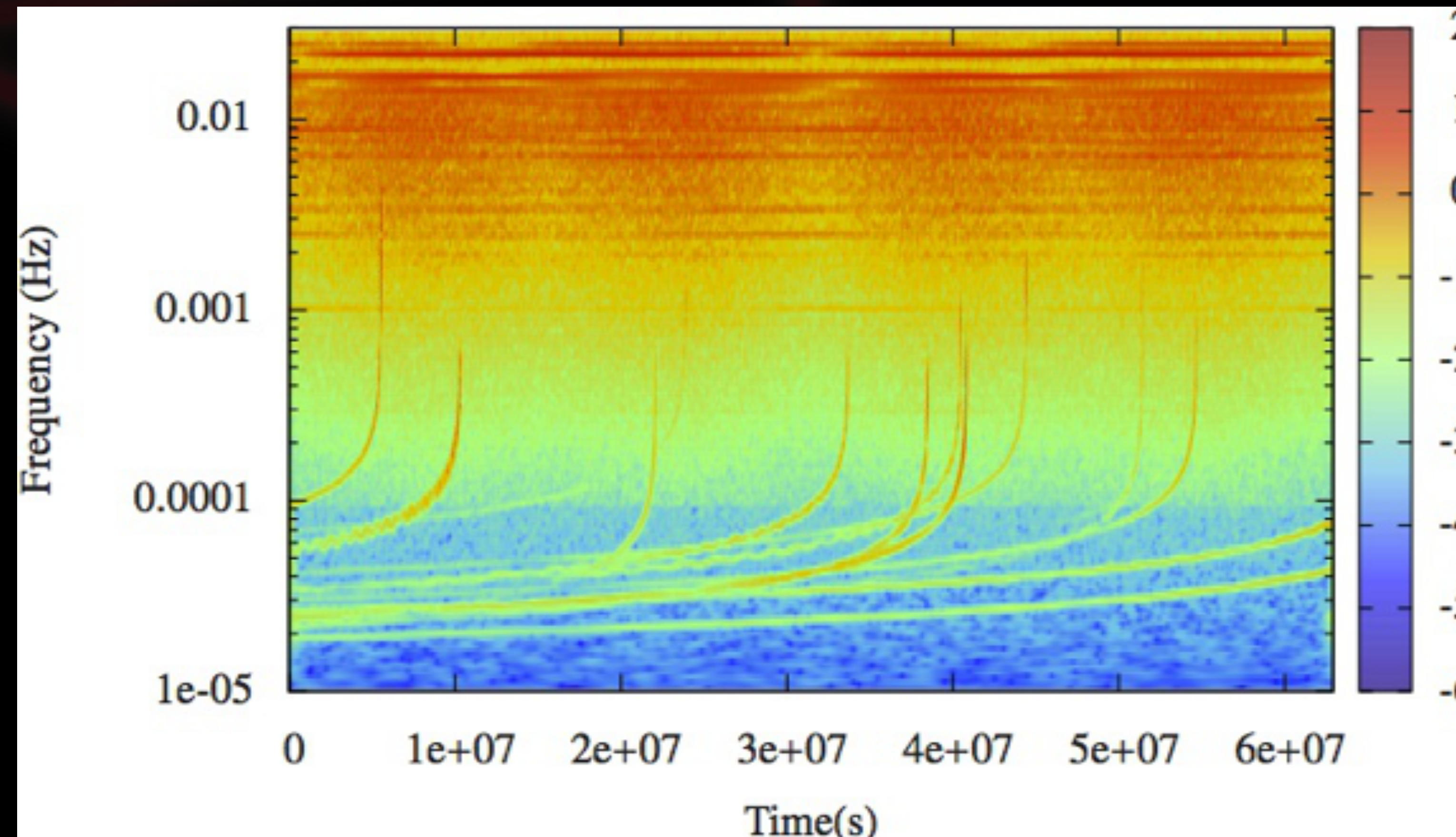


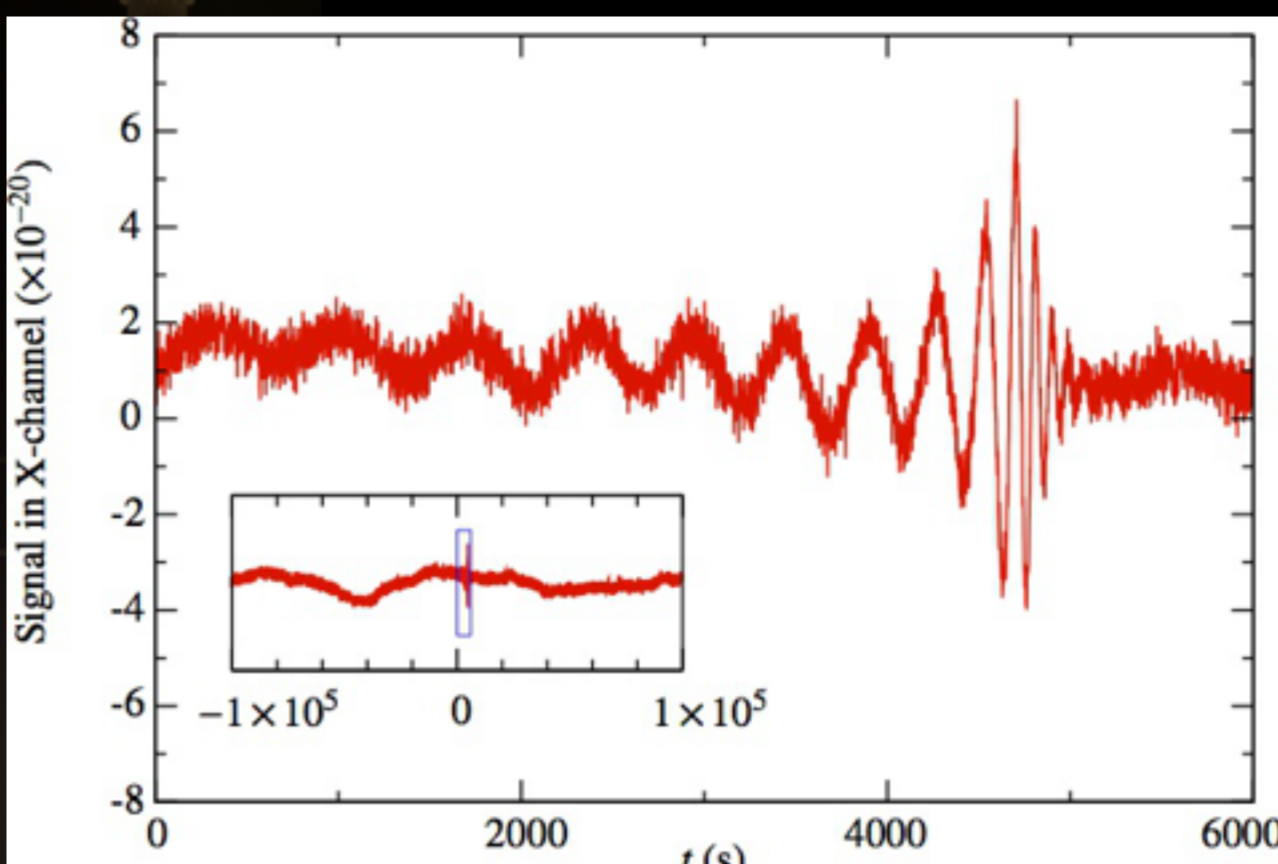
GW sources

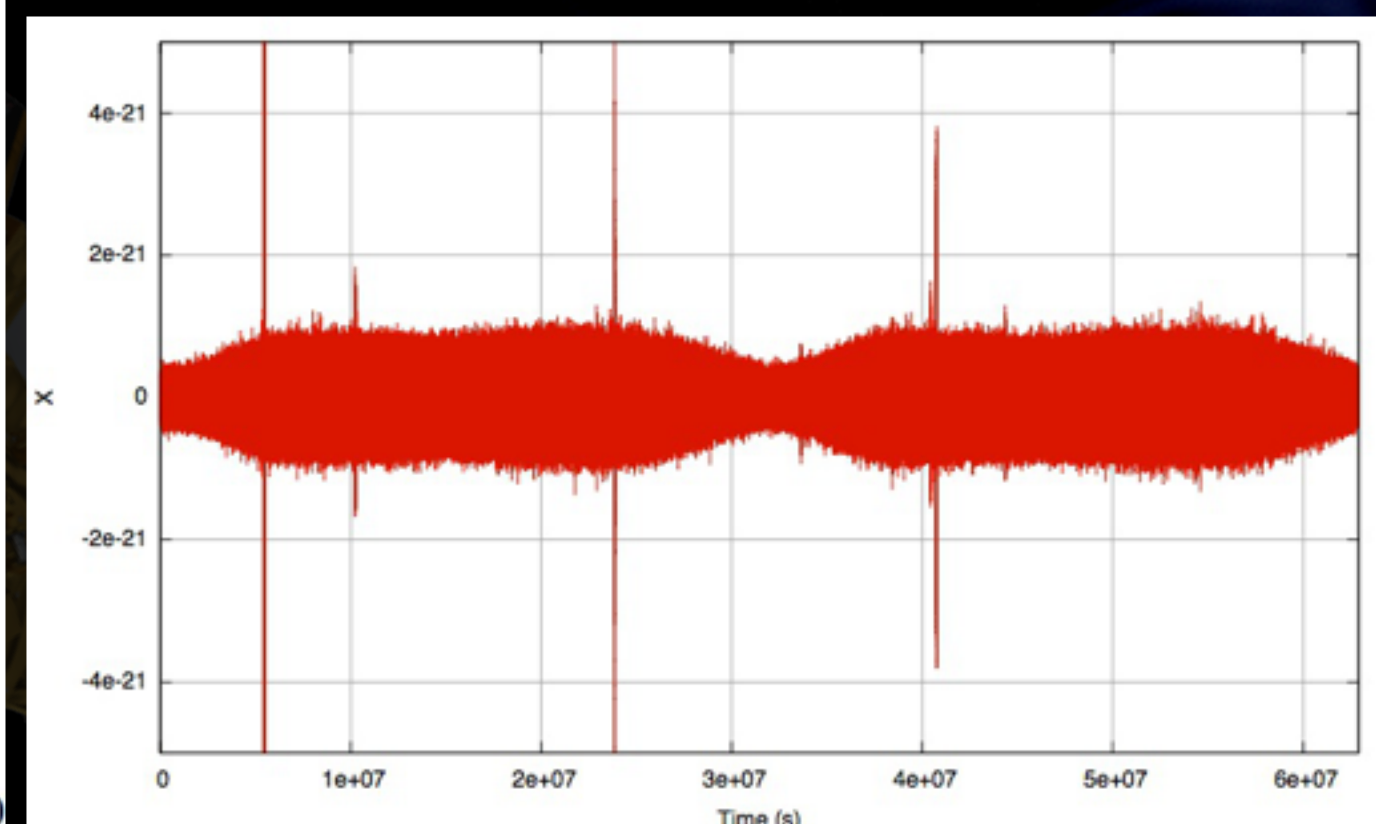
- 6×10^7 galactic binaries
- 10-100/year SMBHBs
- 10-1000/year EMRIs
- large number of Stellar Origin BH binaries (LIGO/Virgo)
- Cosmological backgrounds
- Unknown sources

GWs in LISA data

- Example of simulated data (LISACode):
  - about 100 SMBHs,
  - Galactic binaries



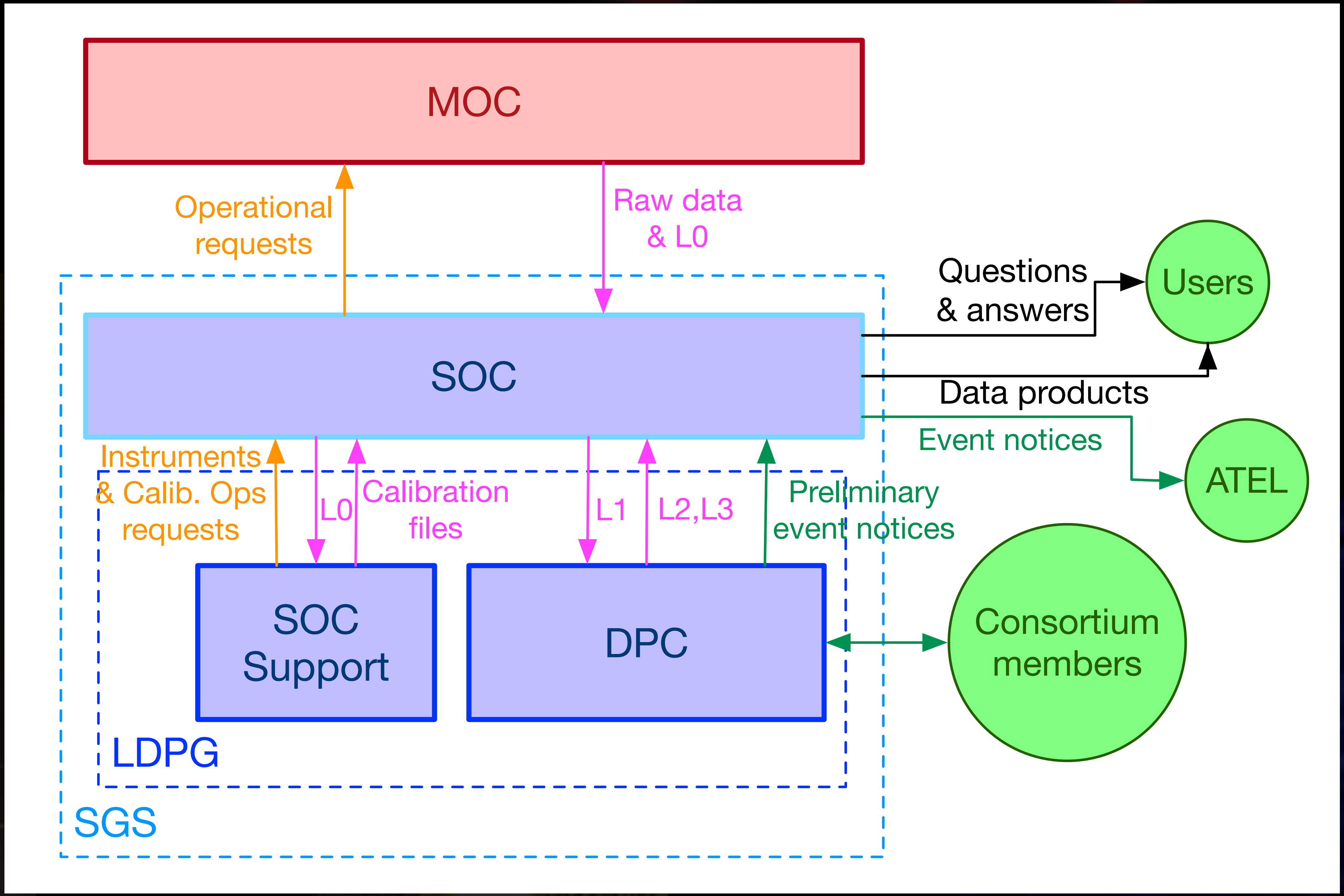




DPC: history & status

- Previous studies:
  - Before 2011, LISA yellow books
  - eLISA/NGO yellow book
  - 2014: CNES Phase 0 for eLISA/NGO
- 2015: Start of the proto-DPC
- January 2017: Proposal LISA
- April 2017: DPC kickoff meeting
- September 2017: Ground Segment meeting
- Next:
  - DPC Definition Document (in progress within the Consortiums)
  - Science Operations Assumptions Document (ESA+Consortium)

LISA Ground Segment



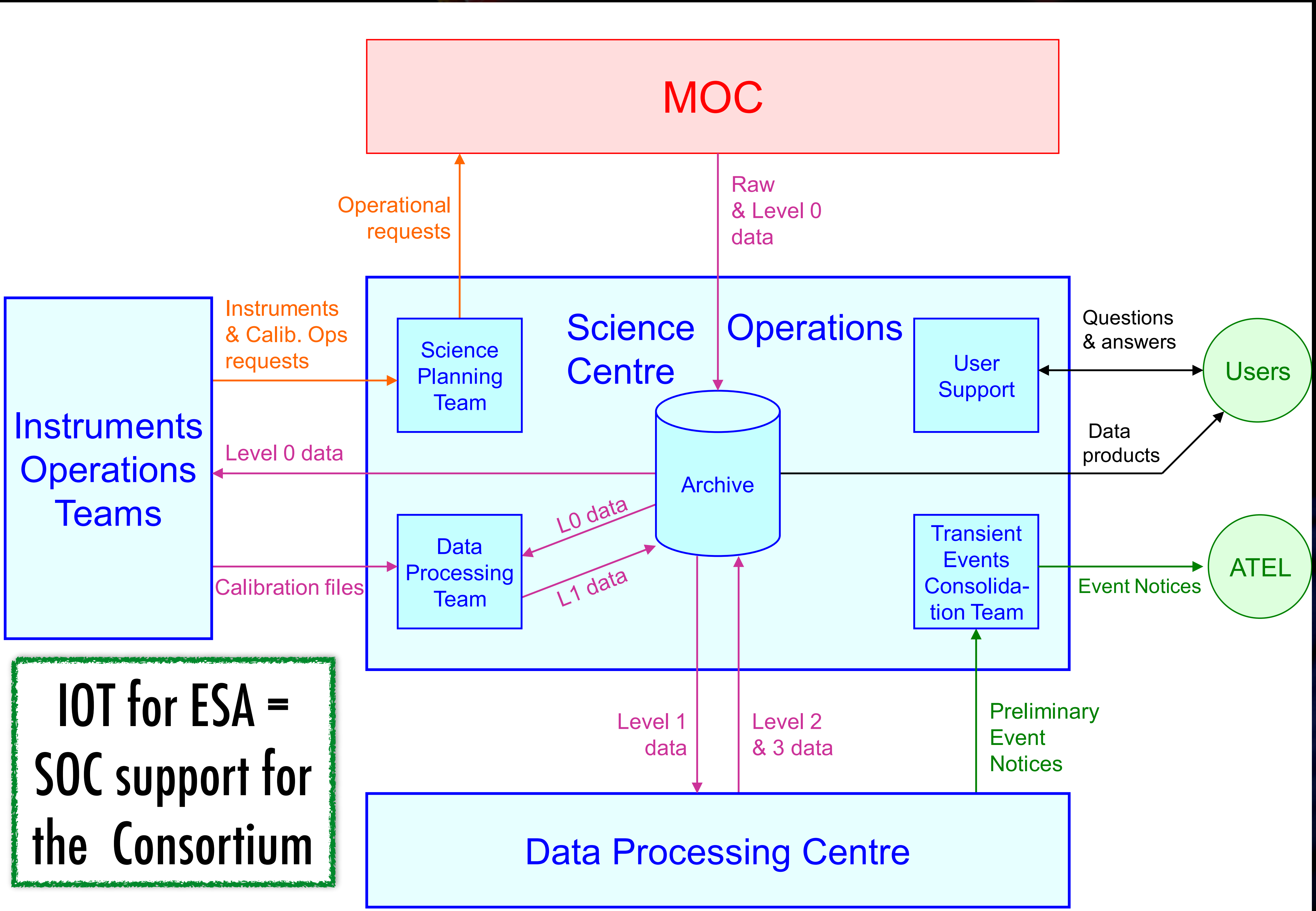
Roles of the SOC (ESA)

- SOC activities:
  - Science Planning
  - Receive L0 data from the MOC;
  - Analyse the quality of the L0 and L1 data;
  - Calibrate the data;
  - Generate alert for strong events;
  - Create L1 science products;
  - Distribute L2 and L3 science products to the DPC;
  - Receive L2 and L3 science products from DPC;
  - Receive alerts from DPC;
  - Archive;
  - Distribute alerts and periodic releases

Roles of the DPC (proposal)

- DPC activities:
  - Receive L1 data from the SOC;
  - Identify and extract waveforms;
  - Build the catalogs of sources;
  - Create L2 et L3 science products;
  - Analyse the quality of science data products;
  - Distribute data to SOC and to the scientific community of the Consortium
  - Produce periodic releases of science data products
  - Generate alerts for upcoming transients, such as mergers
  - Products selection

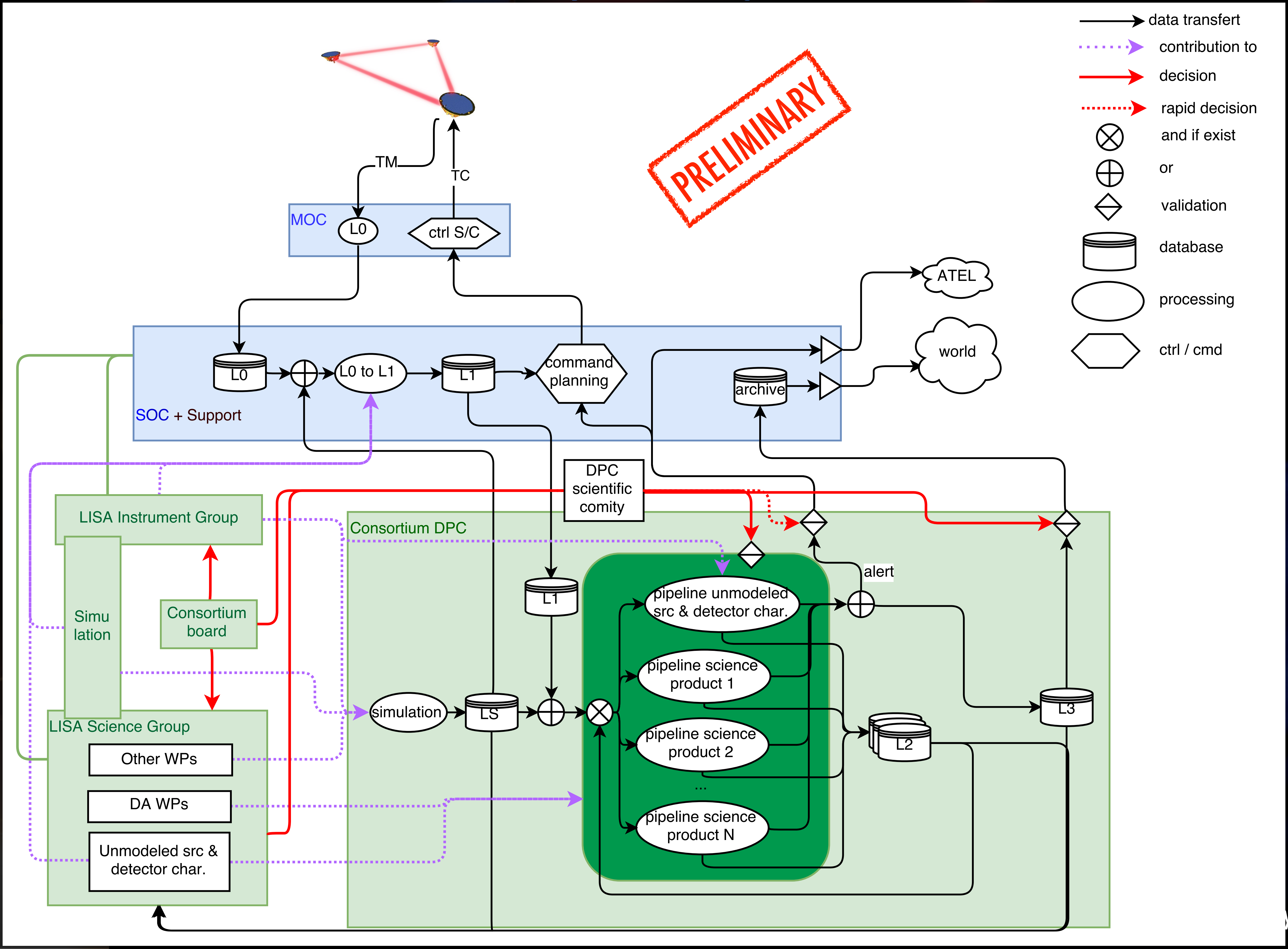
GS from SOC (ESA) point of view



IOT for ESA = SOC support for the Consortium

from Damien Texier

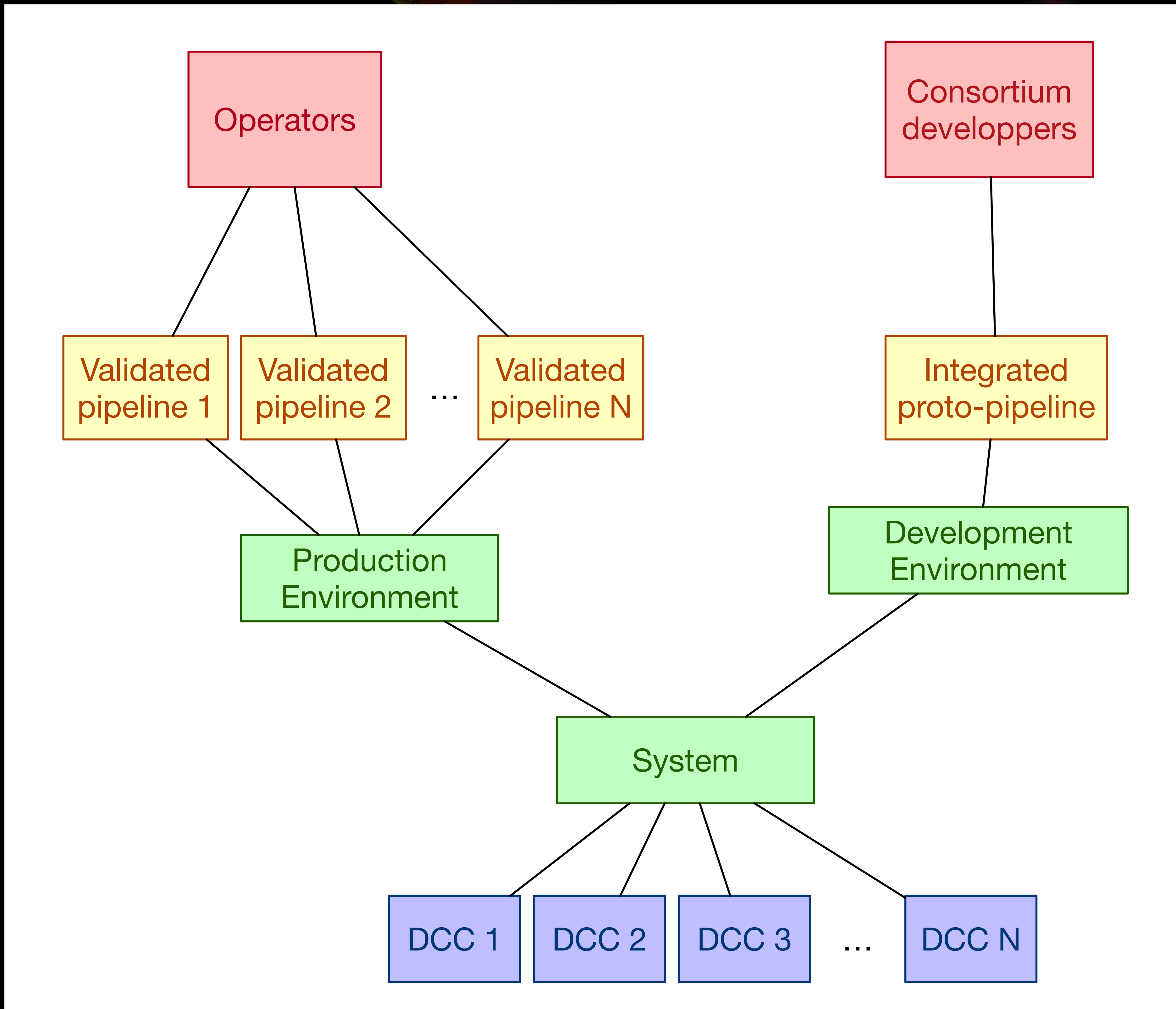
GS from DPC (ESA) point of view



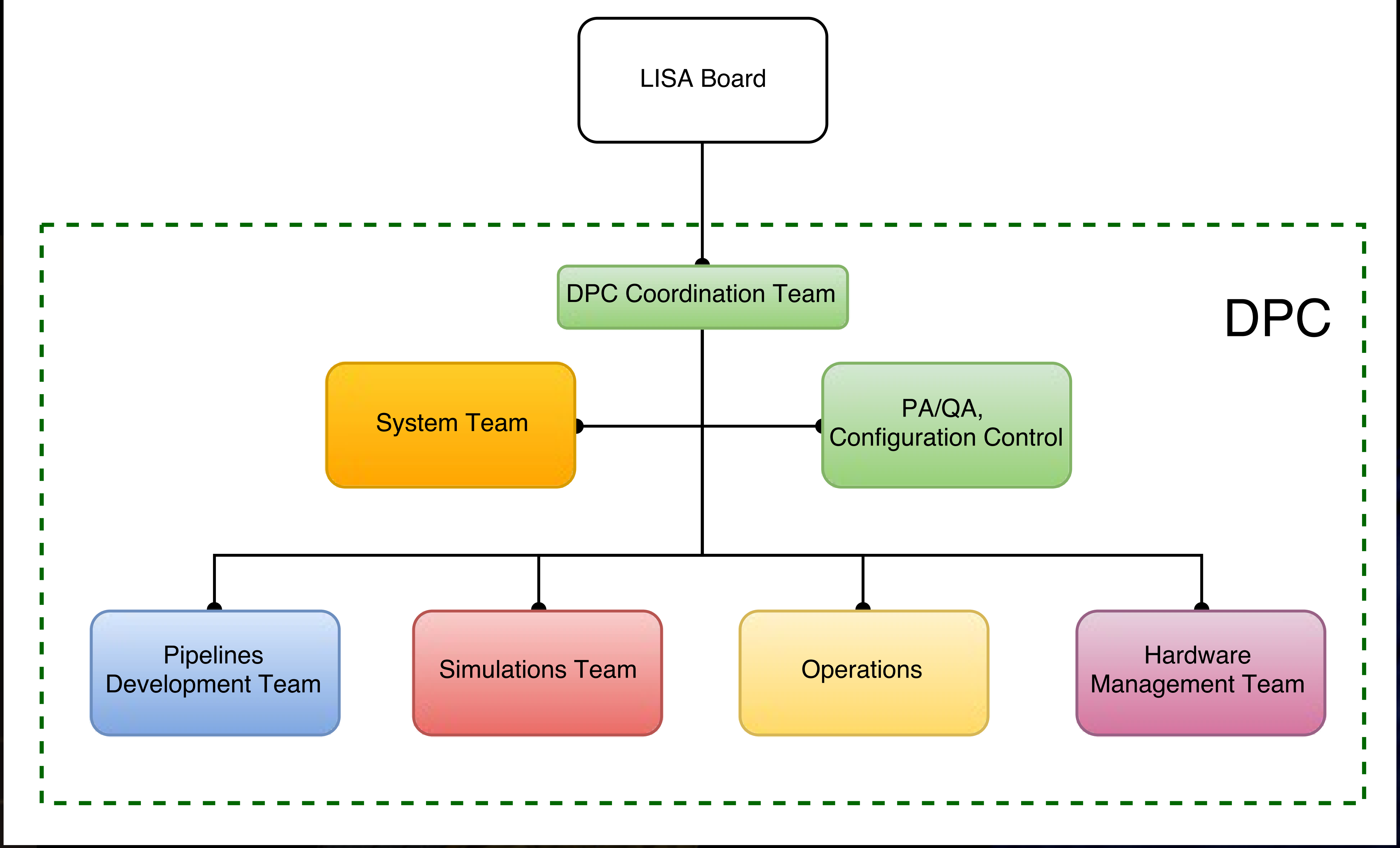
Current DPC

- DPC: unique distributed entity responsable for the data processing (driving, integration of software block, ...)
- DPC in charge of delivering L2 & L3 products + what's necessary to reproduce/refine the analysis (i.e. input data + software + its running environment + some CPU to run it).
- Data Computing Centres (DCC): hardware, computer rooms (computing and storage) taking part to the data processing activities.
- The DPC software « suite » can run on any DCC.
  - Software: codes (DA & Simu.) + services (LDAP, wiki, database) + OS.
- First solutions:
  - Separation of hardware and software: ligth virtualization, ...
  - Collaborative development: continuous integration, ...
  - Fluctuations of computing load: hybrids cluster/cloud

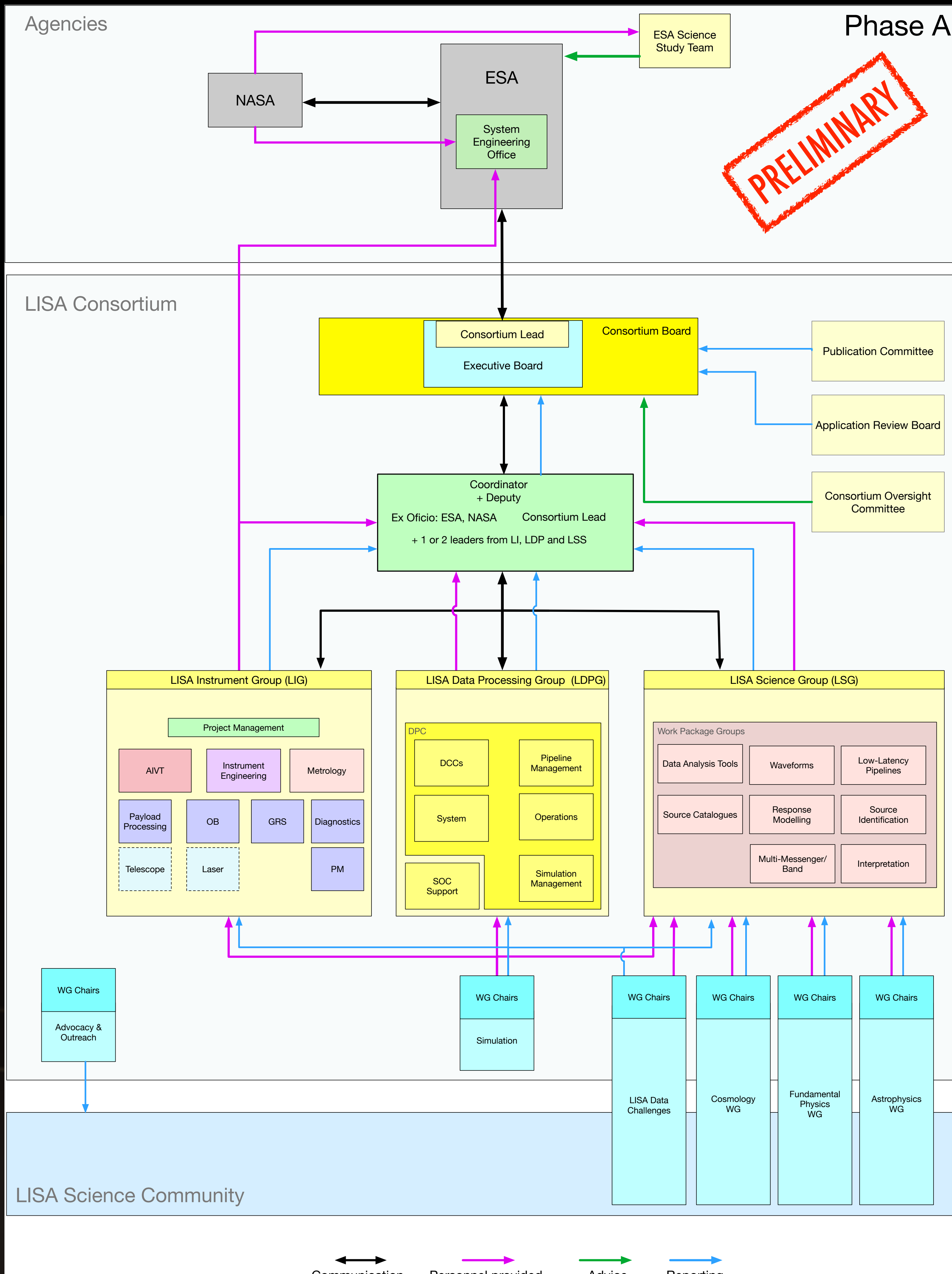
DPC system

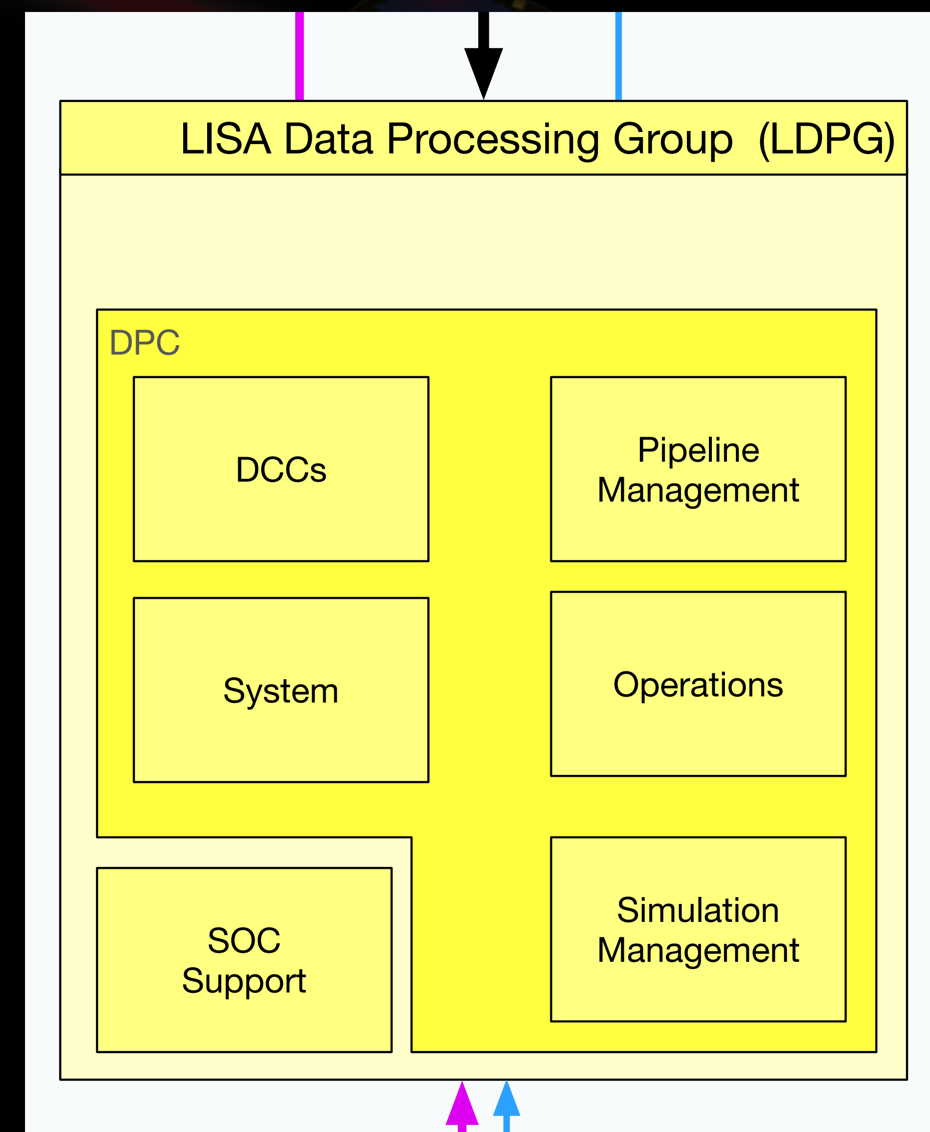

Operators
Consortium developpers
Validated pipeline 1
Validated pipeline 2
...
Validated pipeline N
Integrated proto-pipeline
Production Environment
Development Environment
System
DCC 1
DCC 2
DCC 3
...
DCC N

DPC top-level organisation

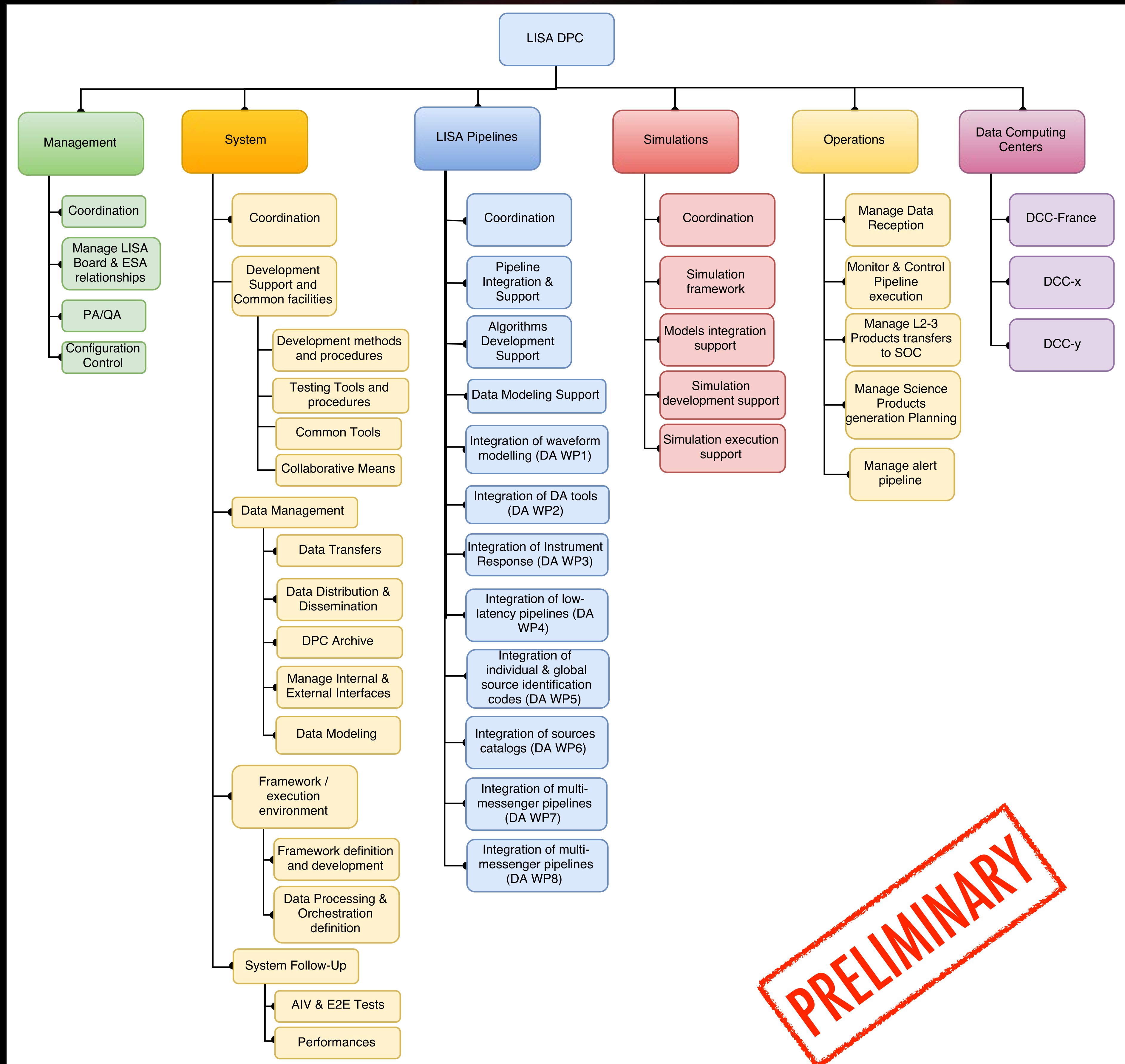


Consortium organisation

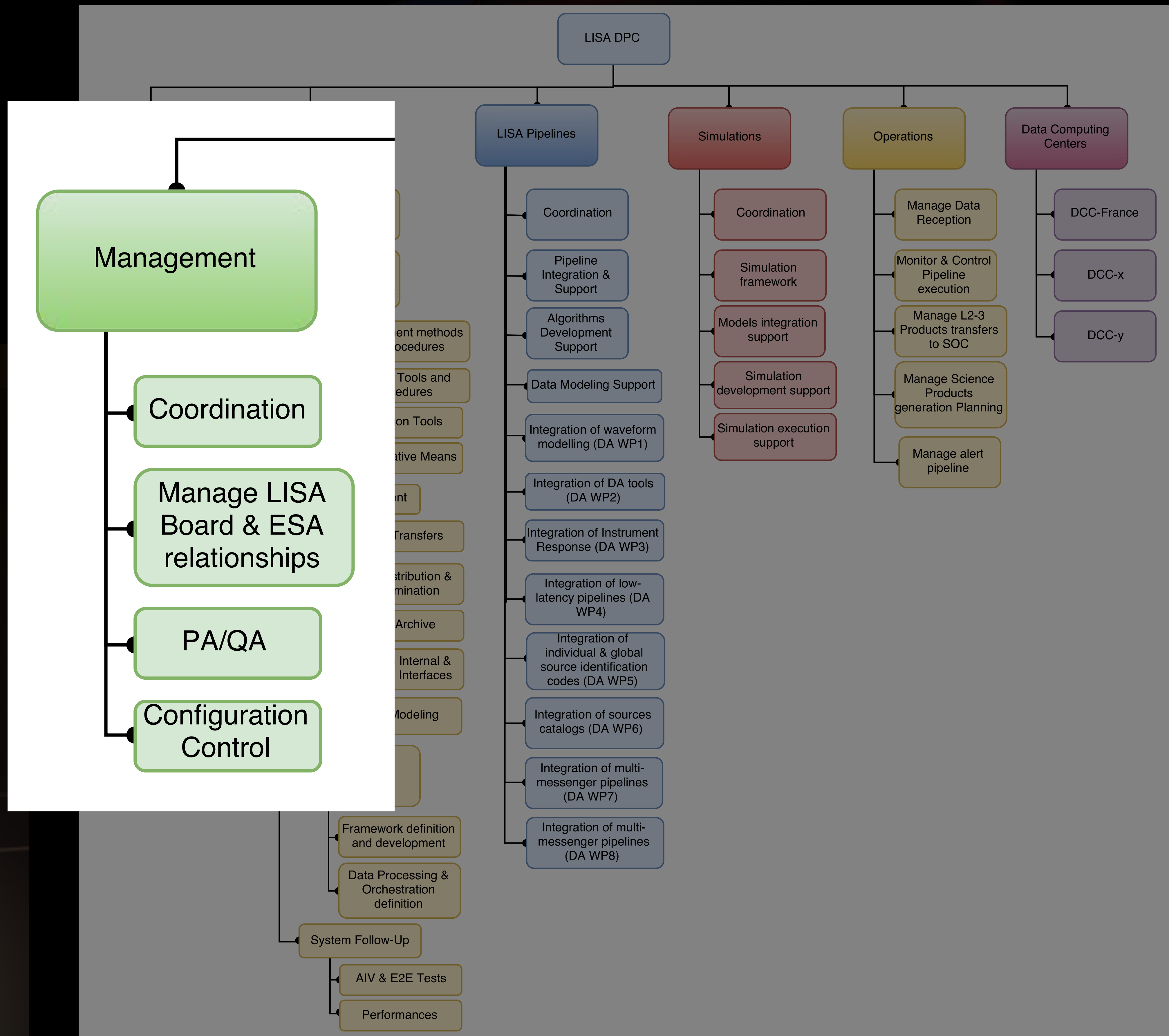




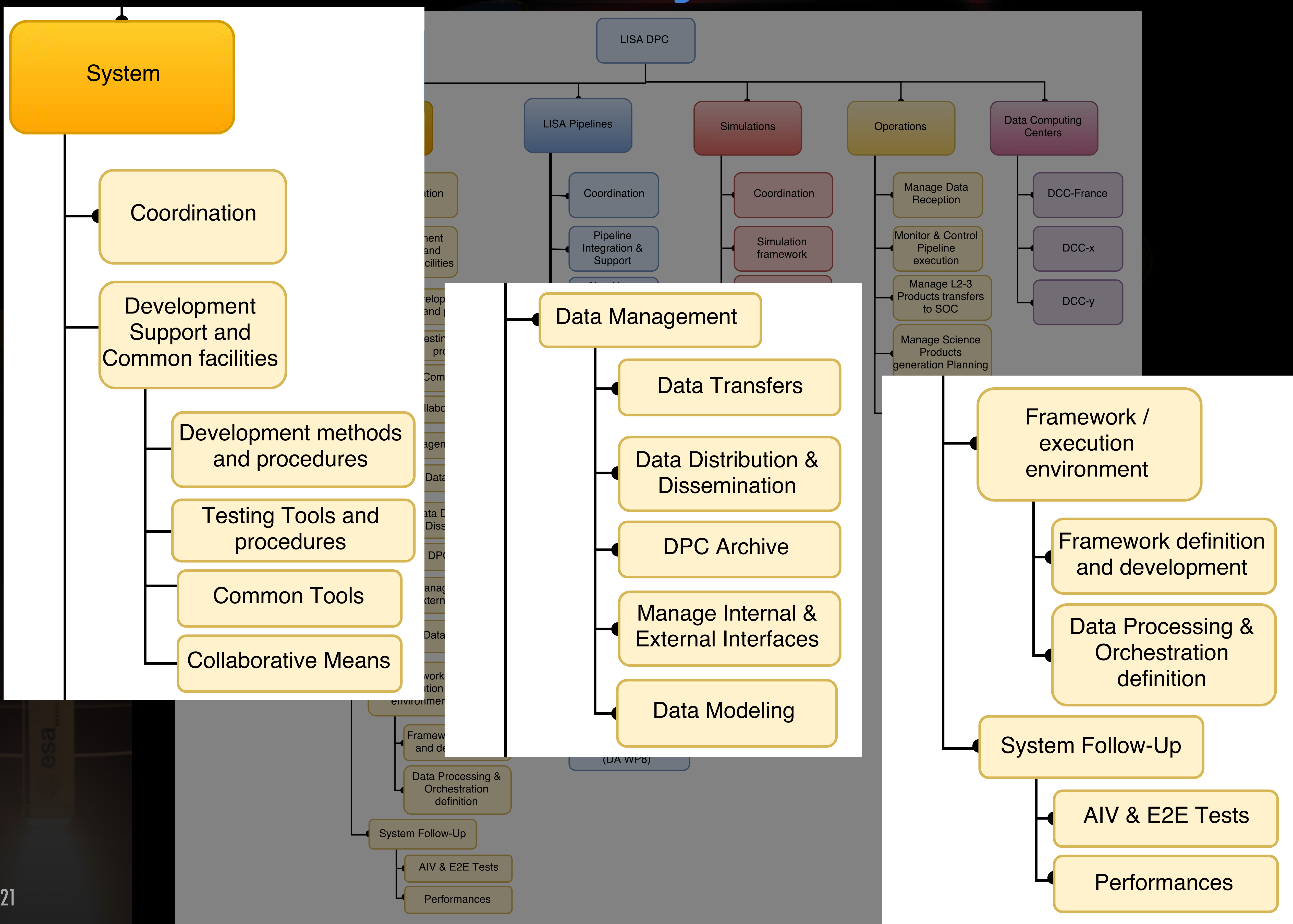
DPC Organisation



DPC Management

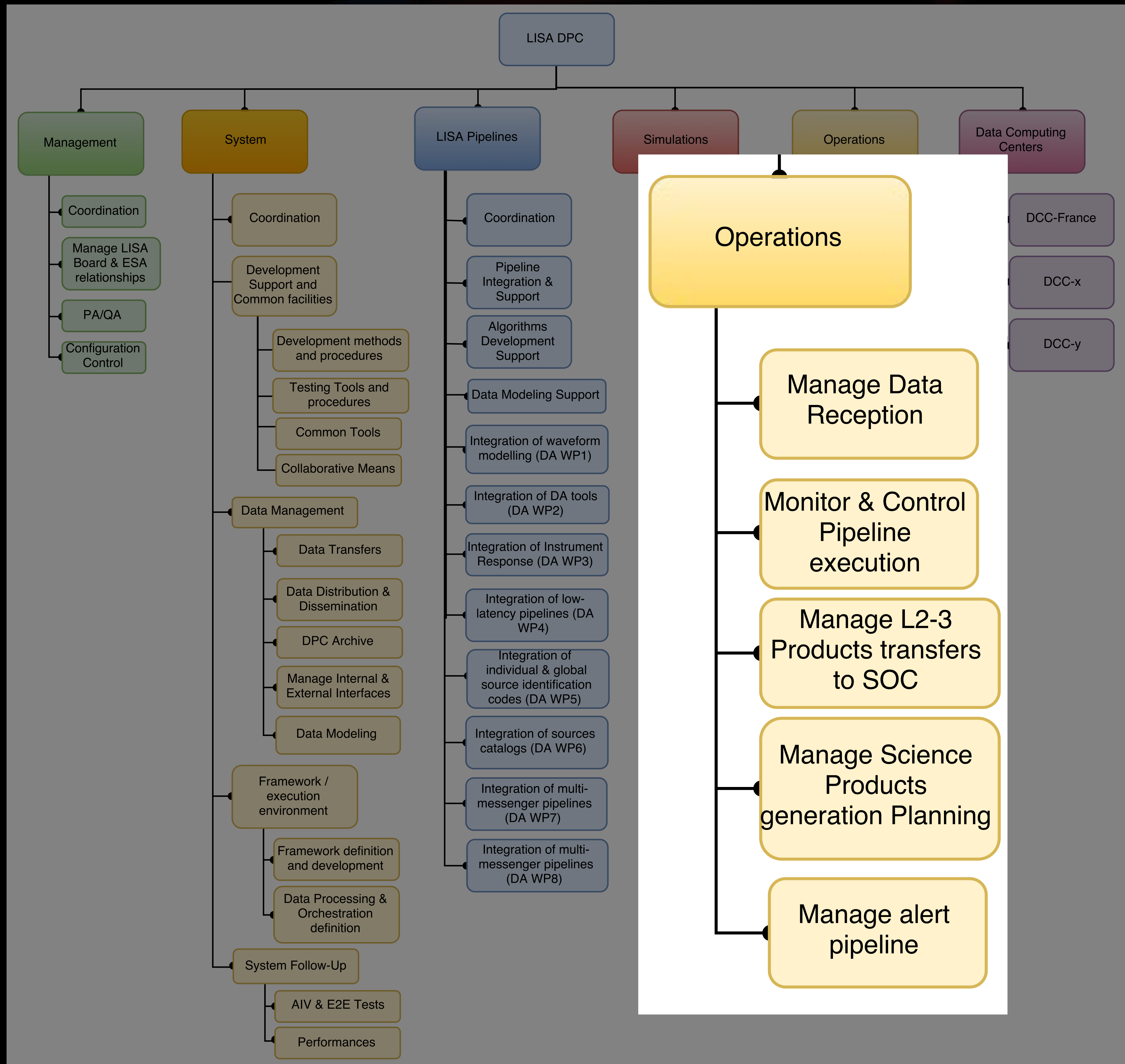


DPC System

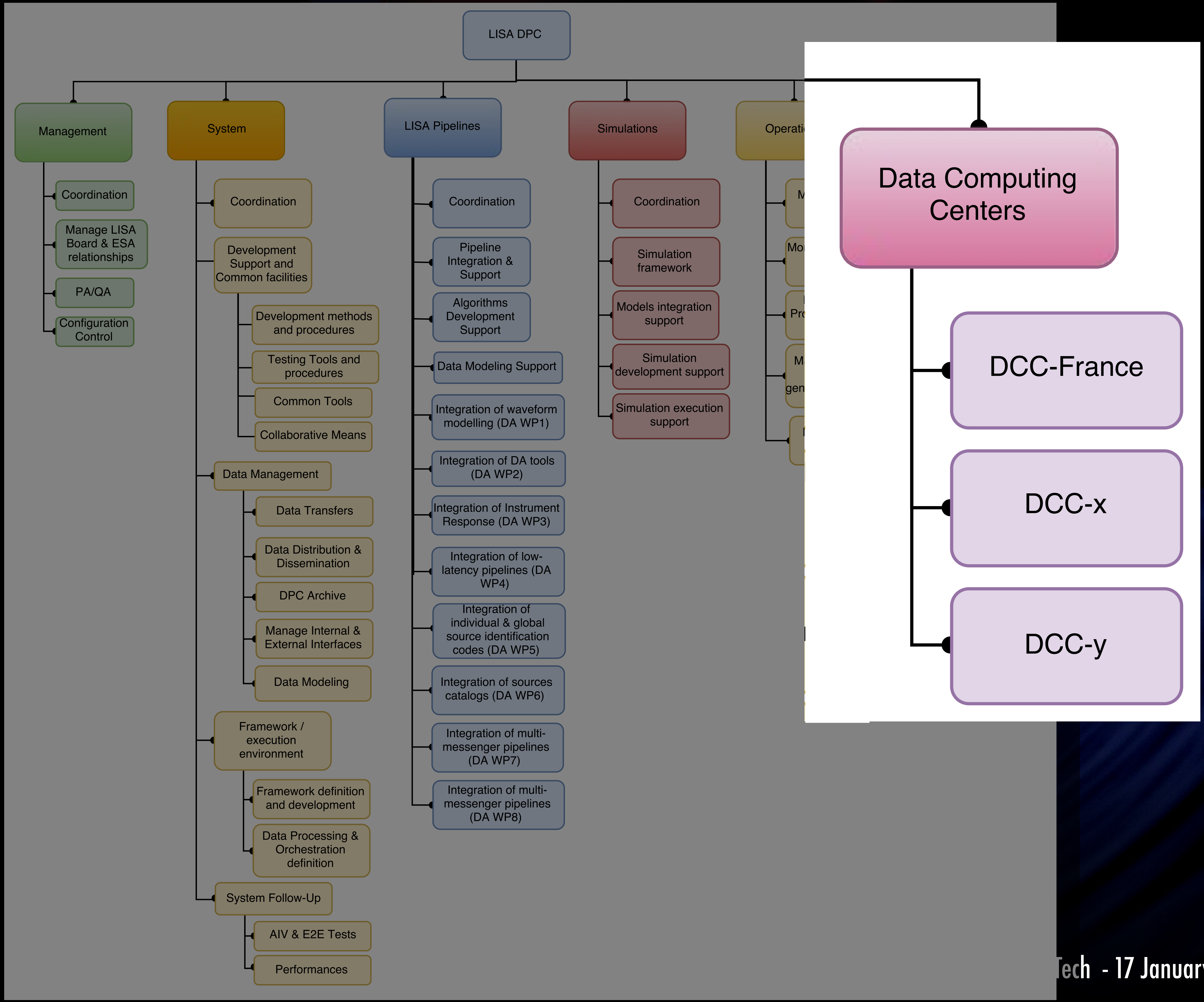


System
Coordination
Development Support and Common facilities
Development methods and procedures
Testing Tools and procedures
Common Tools
Collaborative Means
Data Management
Data Transfers
Data Distribution & Dissemination
DPC Archive
Manage Internal & External Interfaces
Data Modeling
Framework / execution environment
Framework definition and development
Data Processing & Orchestration definition
System Follow-Up
AIV & E2E Tests
Performances
LISA DPC
LISA Pipelines
Simulations
Operations
Data Computing Centers
Coordination
Pipeline Integration & Support
Coordination
Simulation framework
Manage Data Reception
Monitor & Control Pipeline execution
Manage L2-3 Products transfers to SOC
Manage Science Products generation Planning
DCC-France
DCC-x
DCC-y
(DA WP8)
Data Processing & Orchestration definition
System Follow-Up
AIV & E2E Tests
Performances

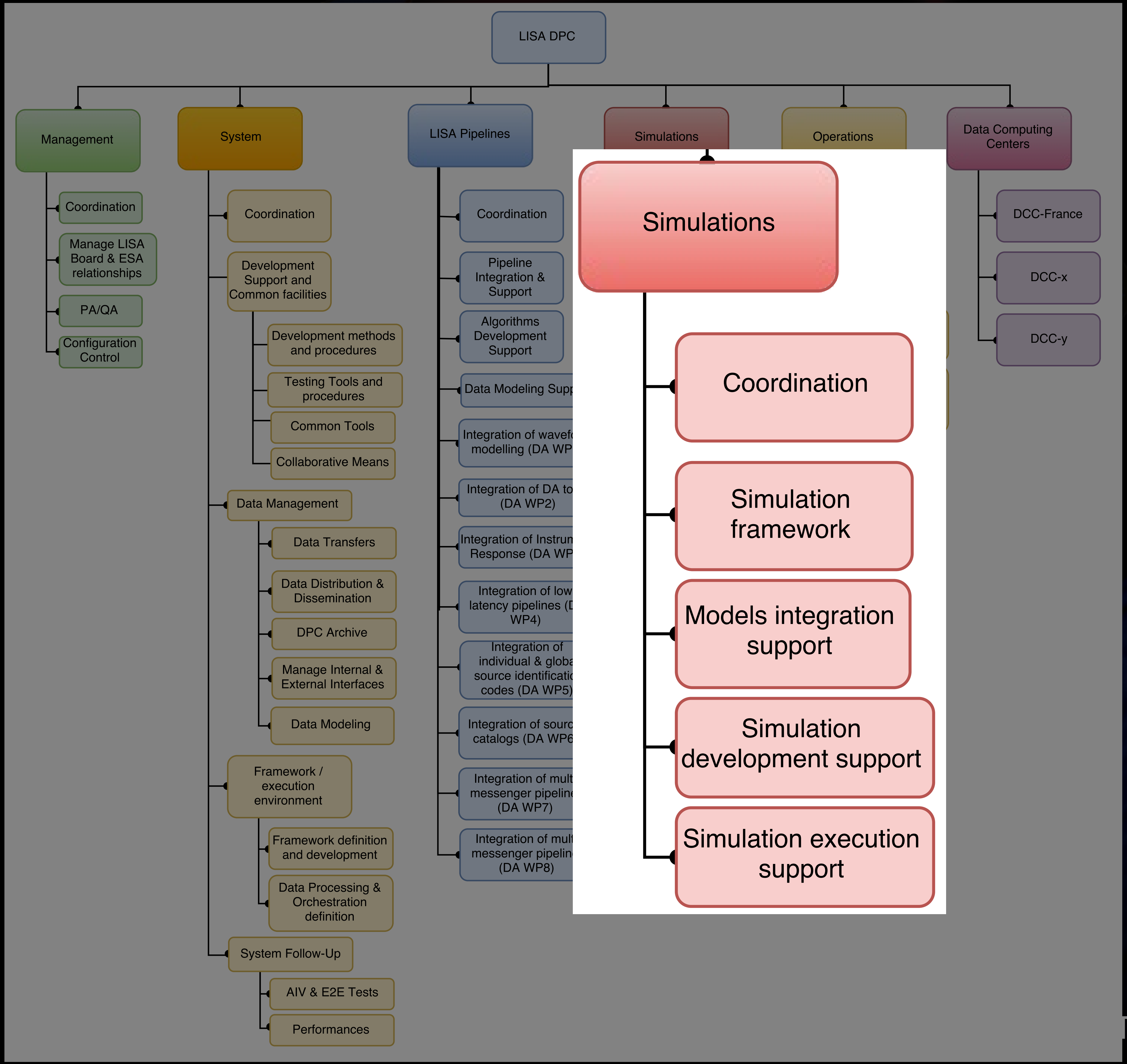
DPC Operations



Data Computing Centers



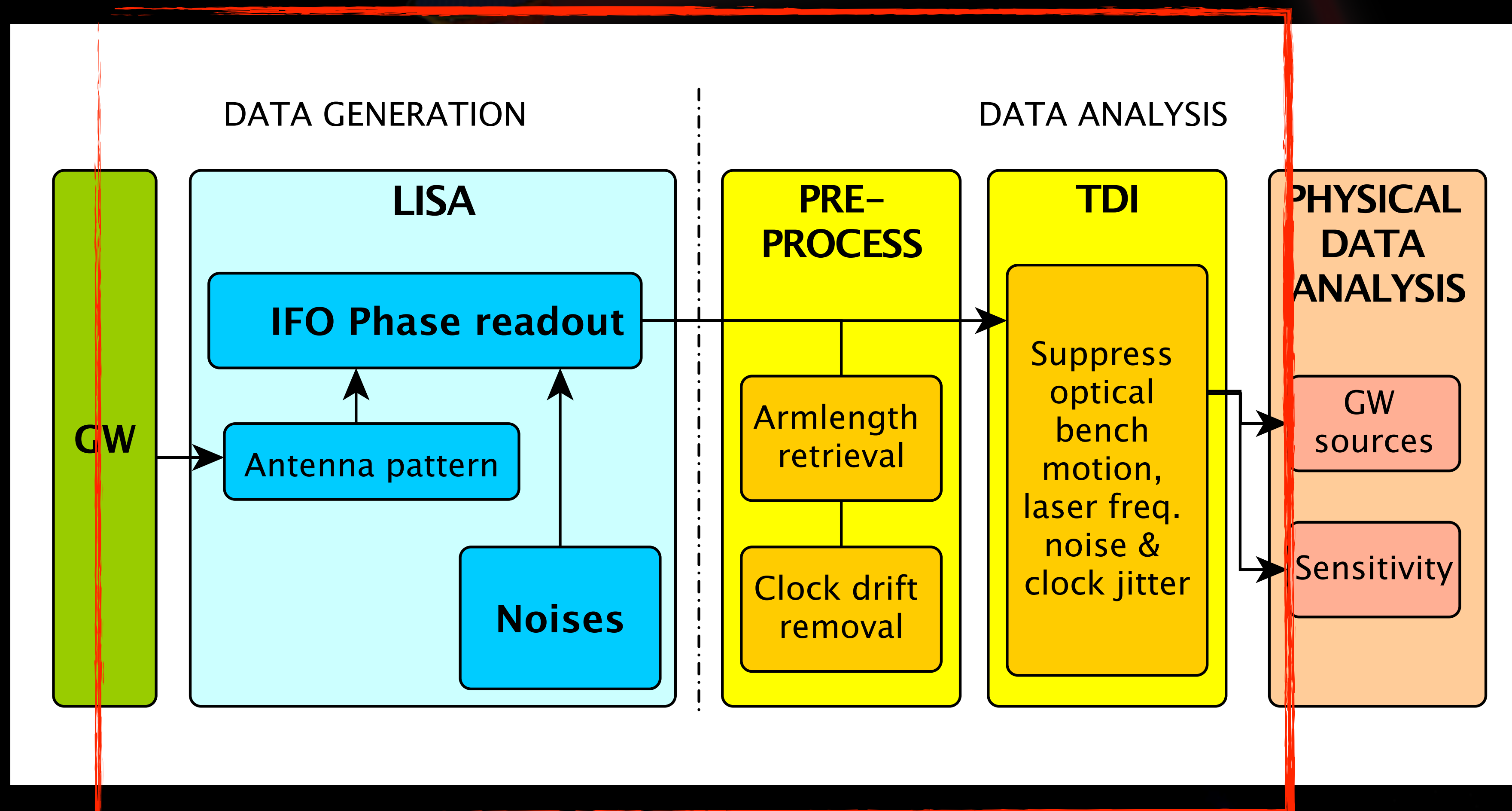
DPC Simulations



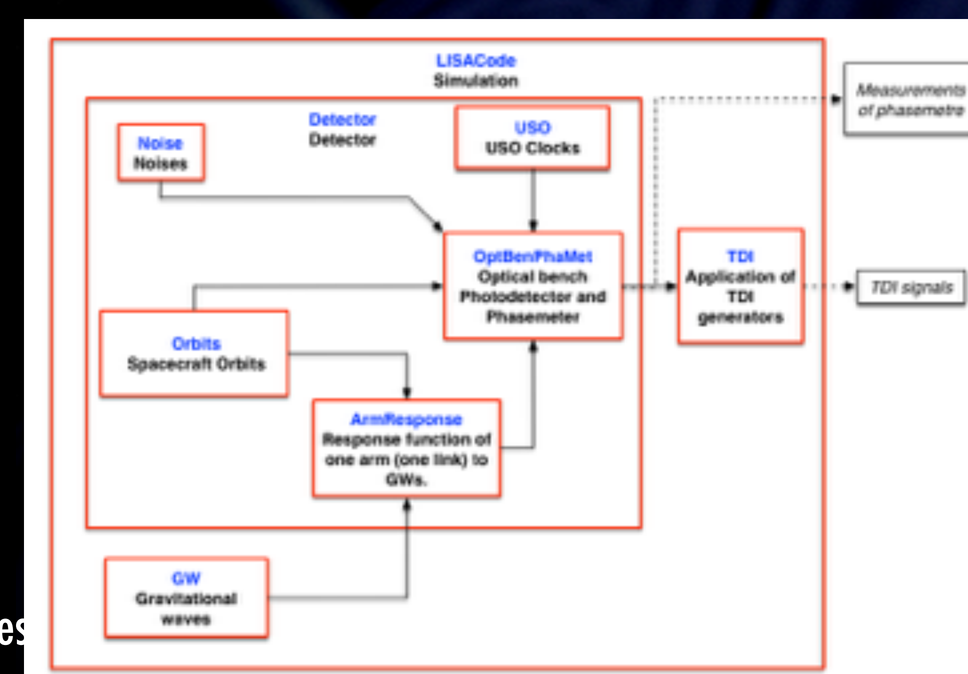
Simulation

- **Goals**:
  - End-to-end simulation → the mission simulator
  - "Quick performance" study for various configurations → final design (required for phase A)
  - Accompany the hardware developments (industries & labs.)
  - Tool(s) for performance controls
- **First requirements**:
  - Close modeling of the instrument subsystems
  - Waveform generation for various GW sources
  - Noise generation using various types of representation
  - Data pre-processing (distinct from simulation)
  - Modularity
  - Computation speed (> 10-100 times faster than reality)
  - Open-source

Simulation

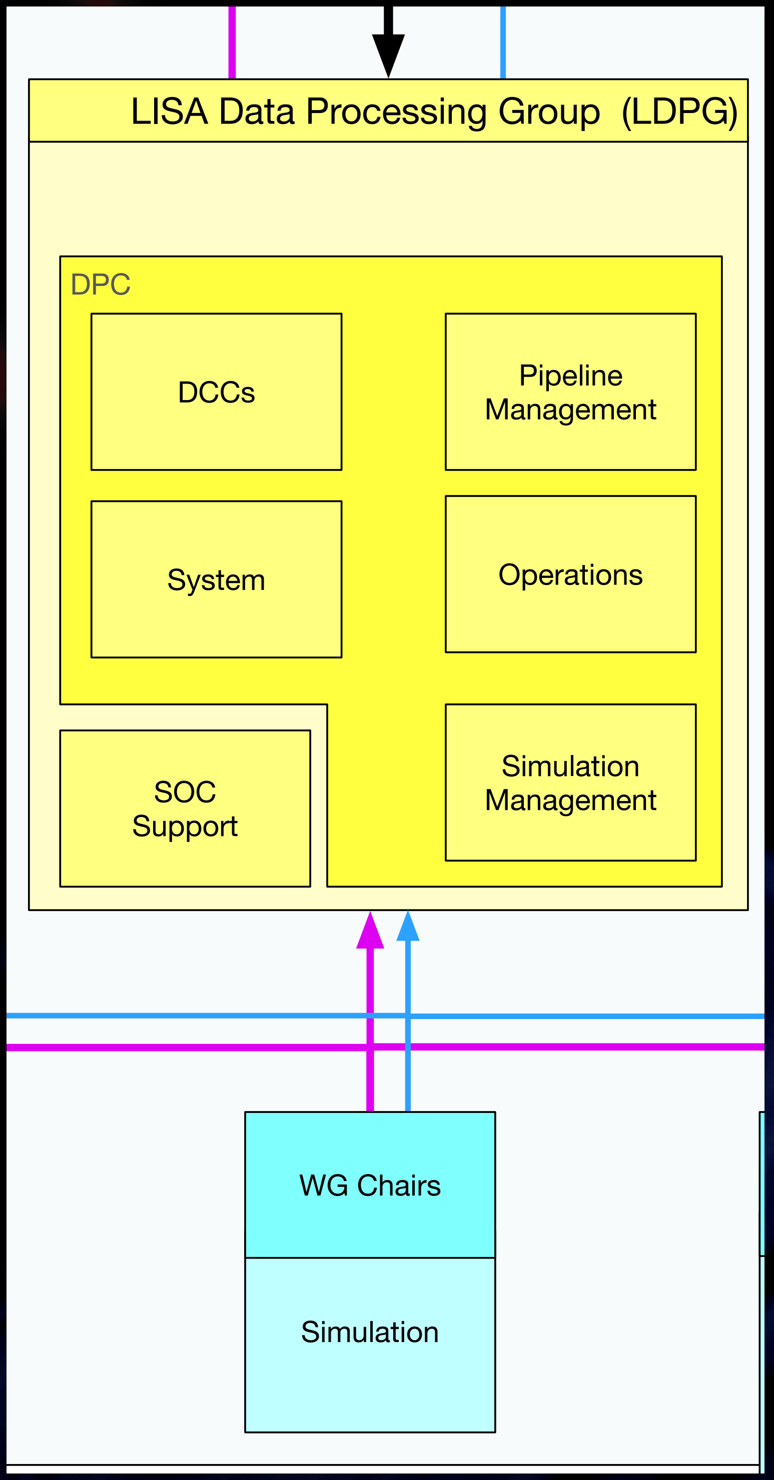


- LISACode is the starting point of the end to end simulator and evolve now in LISANode
- Complementary simulators:
  - TDISim (check TDI)
  - LISADyn (3D dynamic)
  - Synthetic LISA

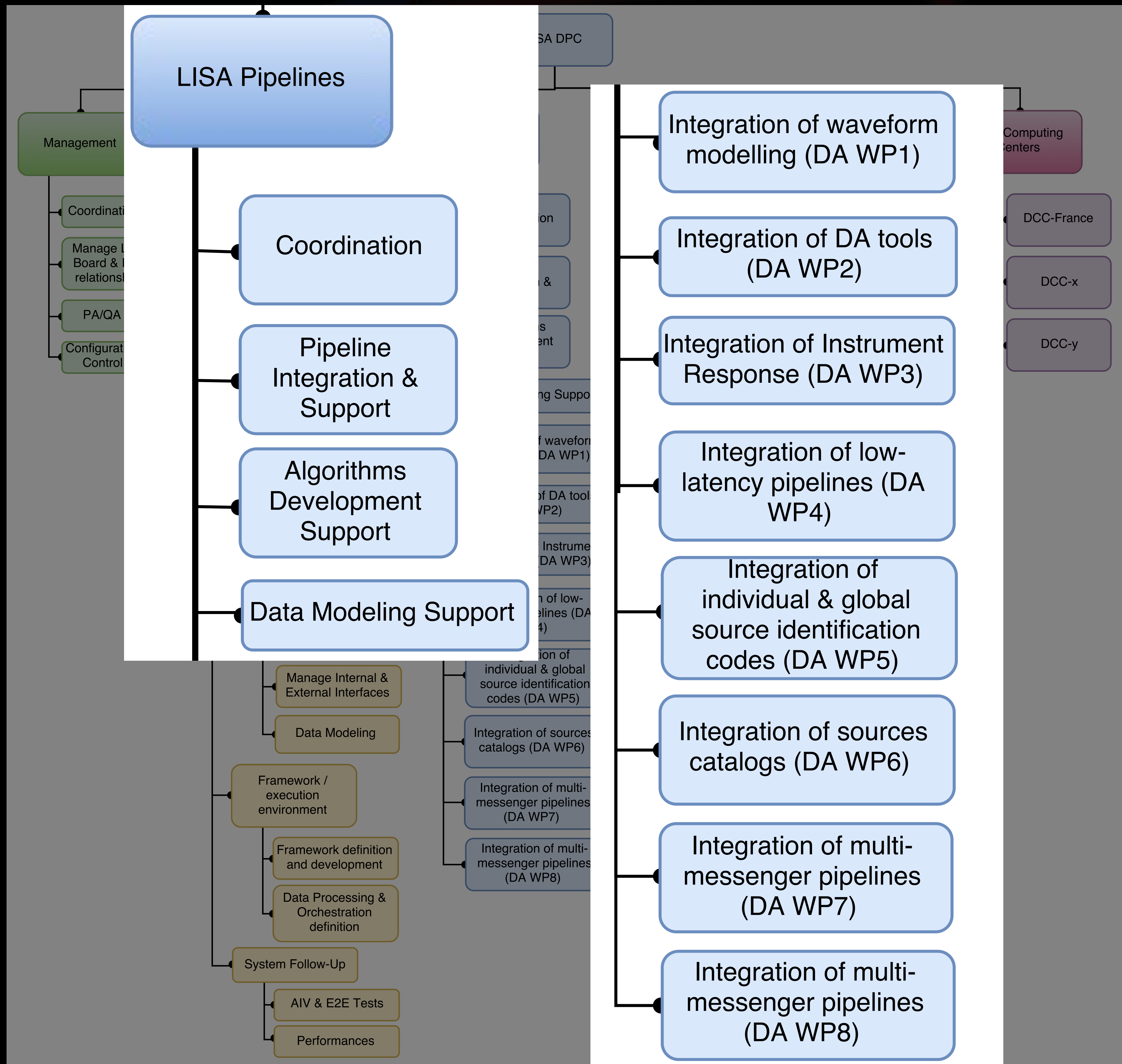


DPC / Simulations

- Strong link with the simulation working group
- Develop and maintain the framework for simulations
- Integration of the models, methods, etc developed by the LISA Instrument Group and LISA Science Group.
- Support developments and executions



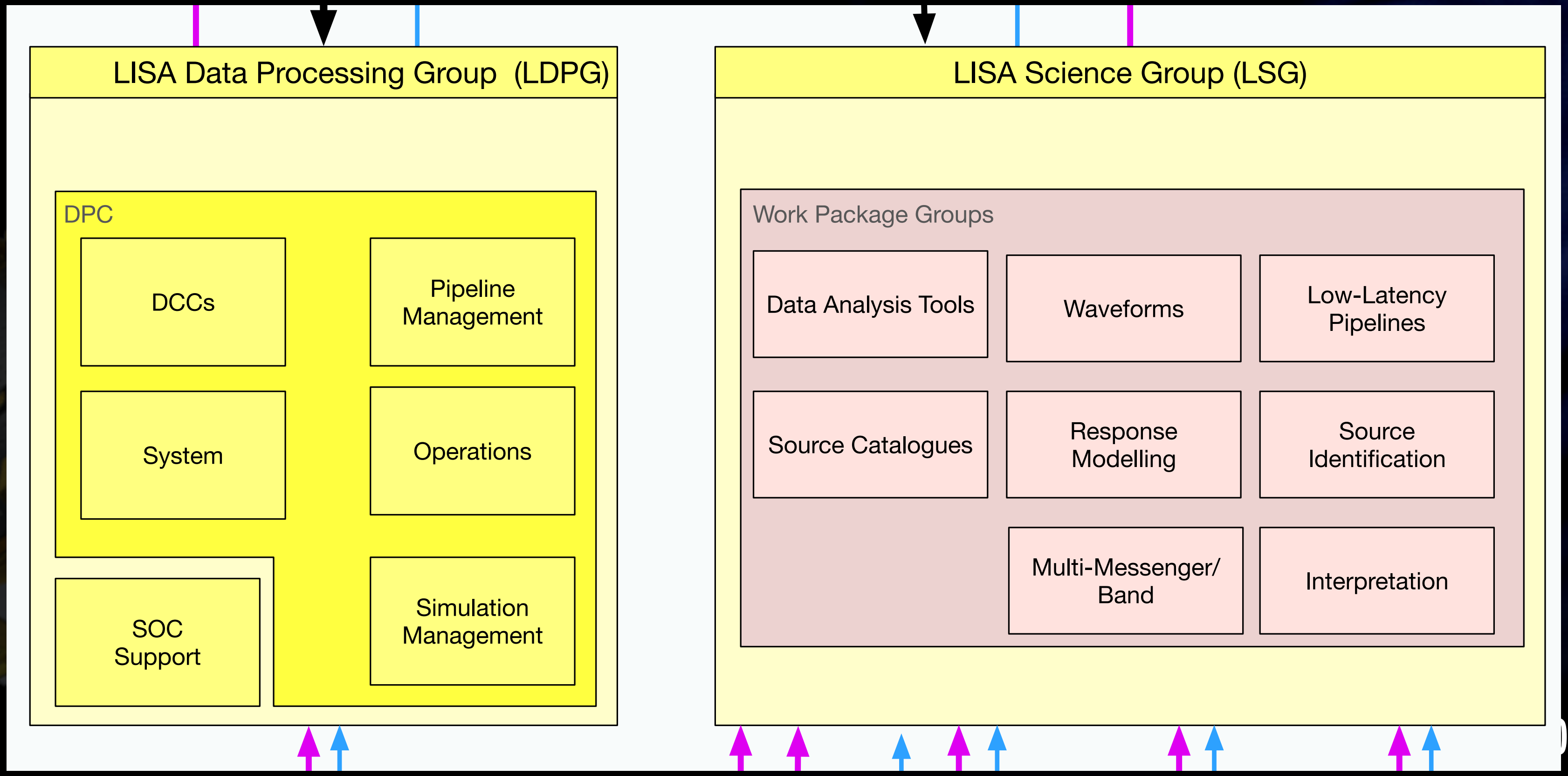
DPC Pipelines



DPC Pipelines

- Integration of algorithms & pipelines: support & optimisation
- Methods and prototype pipelines will be developed by the LISA Science Group and integrated in the DPC software with the support of the DPC teams
- Work Packages Science/Data Analysis are identified => integration of the developed codes, tools, prototype of pipelines, ...

  Some WP almost purely DPC

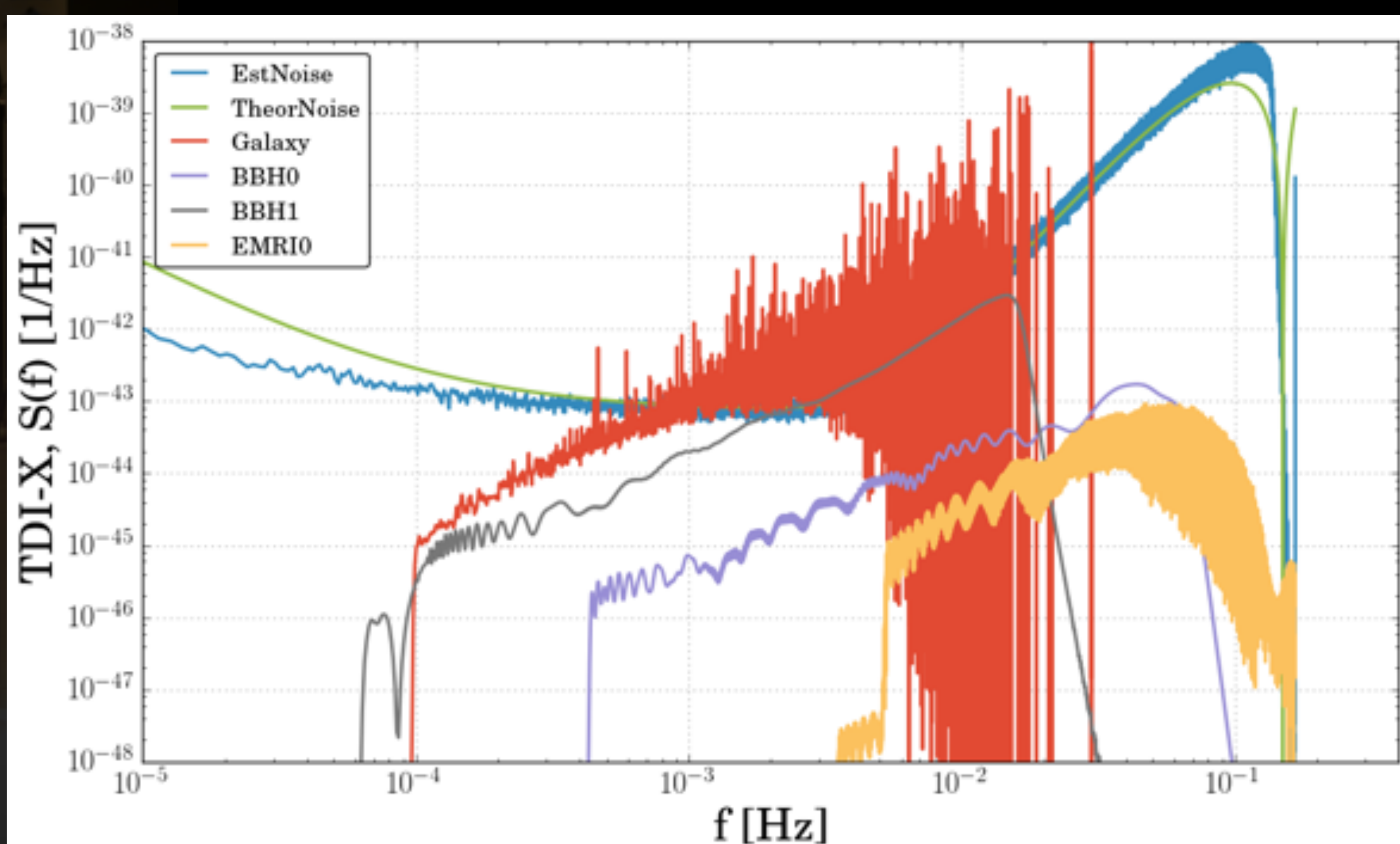


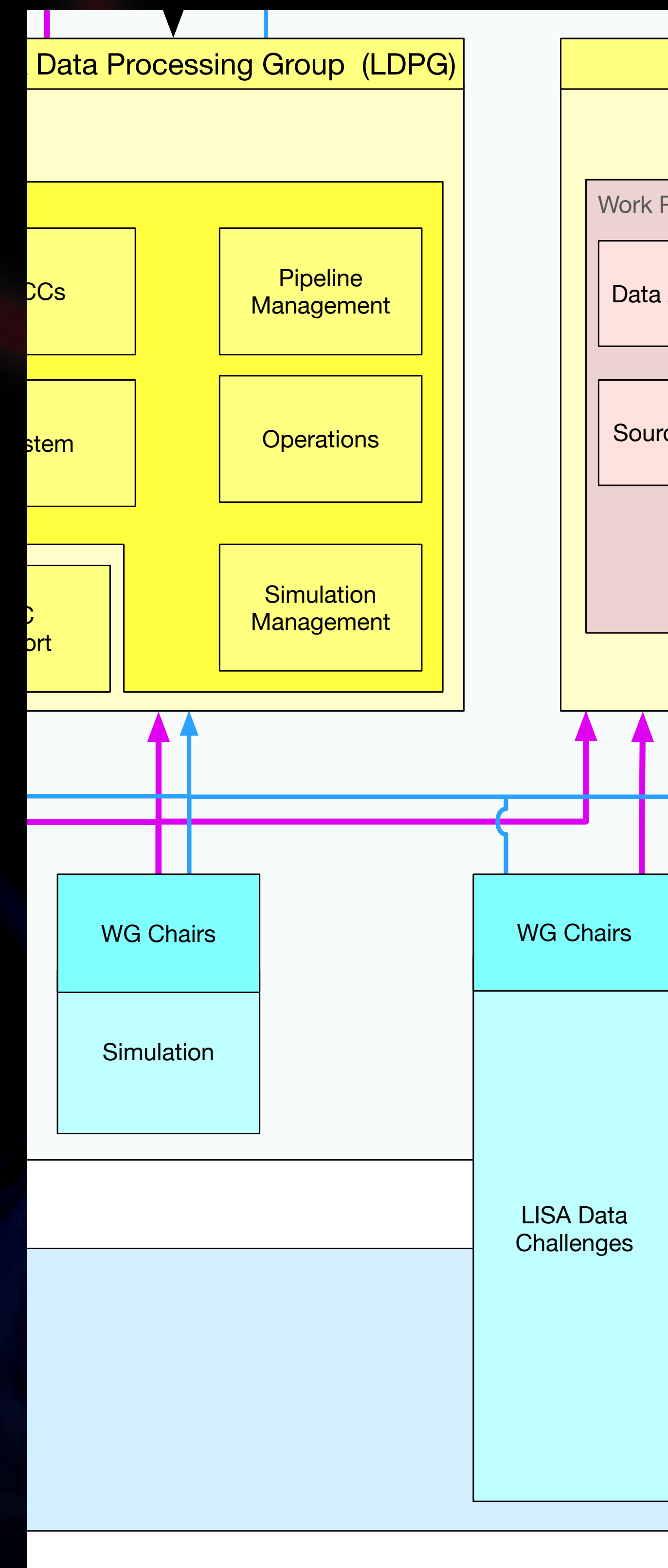
DPC Pipelines

- From the 63 Work Packages some are strongly linked to the DPC:
  1. Waveform modelling
  2. Data analysis tools
  3. Instrument response modeling
  4. Low-latency pipelines
  5. Individual and global source identification codes
  6. Sources catalogues
  7. Multi-messenger, multi-band
  8. Interpretation, key-science projects

DPC/LDC

- Mock LDC: 2005→2011
- 2017: start of the LDC
  - Develop data analysis
  - Design the pipelines of the mission
- Example of the potential data for LDC1





LISA DPC today: proto-DPC

- The proto-DPC is a set of tools provided to ease the challenging data analysis and simulation tasks of LISA:
  - Hardware (CPU and disk) usage not a major concern
  - Data Analysis itself is challenging: lot of unknowns, complex noises and pre-processing

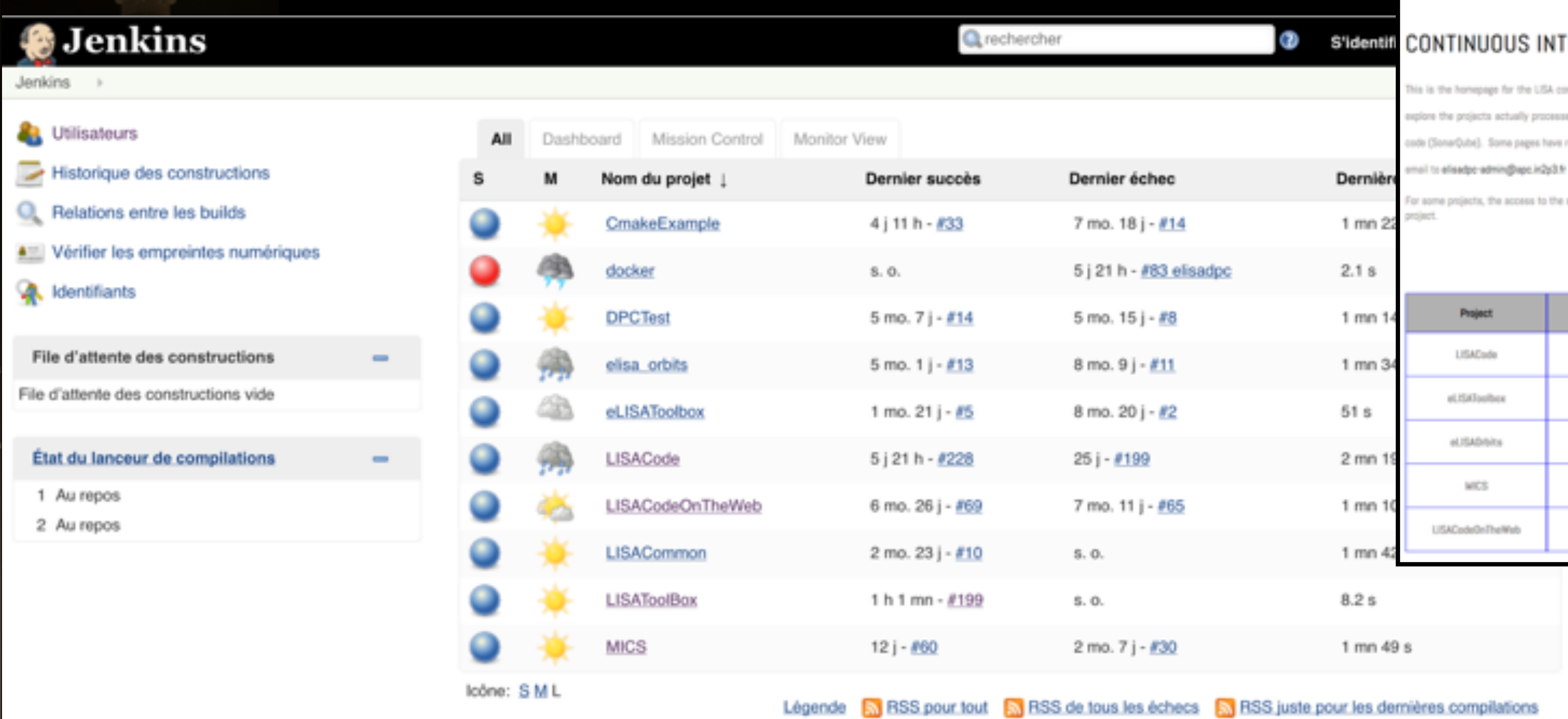
  => Keep a simple and easy to use DPC infrastructure.
  - How IT will look like in 10 years ? Will virtualization be the next standard ?
- Our guideline: The DPC has to be easy-to-use, simple, flexible and easily upgradeable until the end of the mission.

Proto-DPC: basics

- Development environment: in production
  - Goals
    - Ease the collaborative work: reason why it's already started
    - During the operation: guarantee reproducibility of a rapidly evolving and composite DA pipeline
    - In fine: keep control of performance, precision, readability, etc
  - Use existing standard tool
    - Control version system to keep track of code revision history, manage teams and workflows.
    - Continous integration (like in Euclid, LSST): suite of non-regression tests automatically run after each commit
    - Docker image: a way to encapsulate source code + its execution environment (in a single readable text file) → smooth prototyping to operation transition

Proto-DPC: basics

- Development environment: in production
  - Done:
    - Simple install of open and standard tools: Jenkins, SonarQube, gitlab CI
    - Worked on moving from 'simple' to 'automatic' using Docker
    - More projects, more users to come.
- https://elisadpc.in2p3.fr/home



Proto-DPC: basics

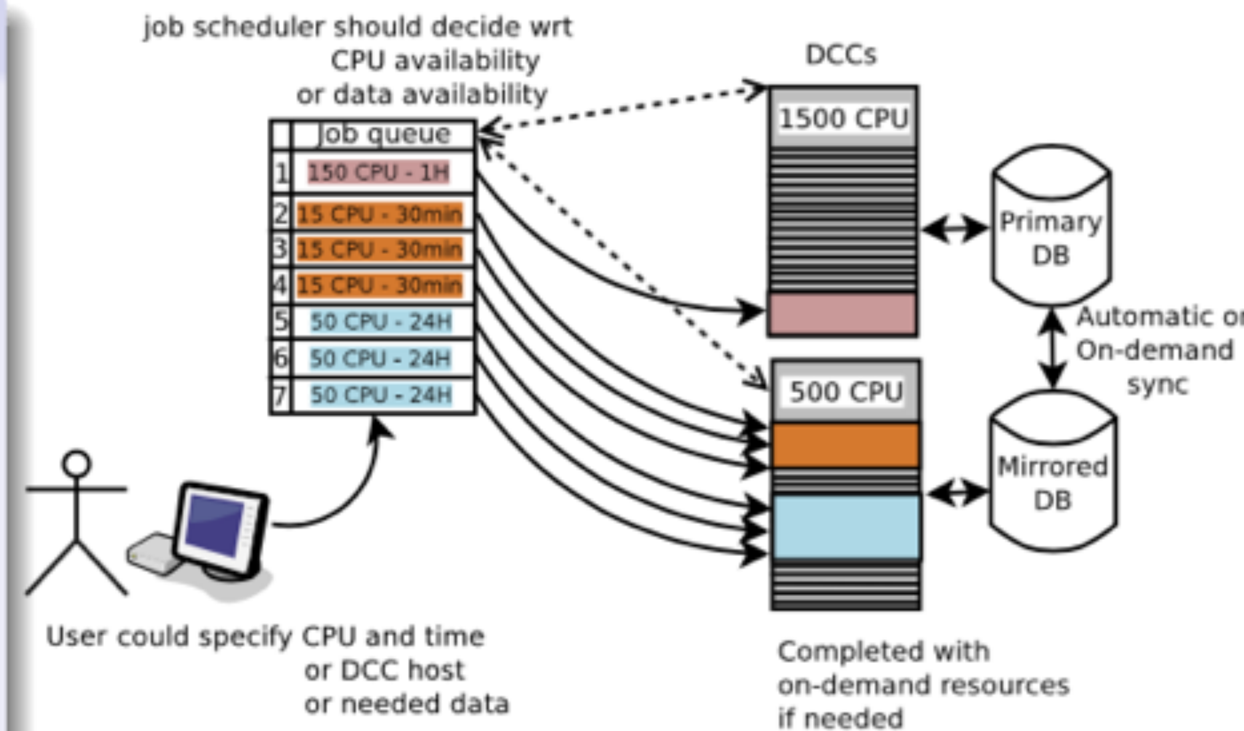
- Data basis & data model: in R&D
  - Motivations
    - Data sharing among people and computing centers
    - Mainly processed, temporary or intermediate data: need meta data management to use them
    - A lot of information: a web 2.0 (intuitive) interface is mandatory (search engine, DB request, tree view to show data dependancies, etc)
  - Context
    - Not very big LISA data volume
    - But still implies some specific developments even if using standard data format. One has to define LISA data model first ...
      - LDC, simulations, LPF data
      - Django website + its sqlite DB: first version ready

Proto-DPC: basics

- **Execution environment**: in R&D

Objectives: a composite computer center

- Pooling of CPU resources with a single scheduler for all DCCs
  - the user-friendly way to go
  - a dynamic CPU pool to adapt the resources to the actual needs (the economic way)
  - transfering data if needed
- Assumptions
  - it's easy to plug new hardware
  - it's easy to transfer data

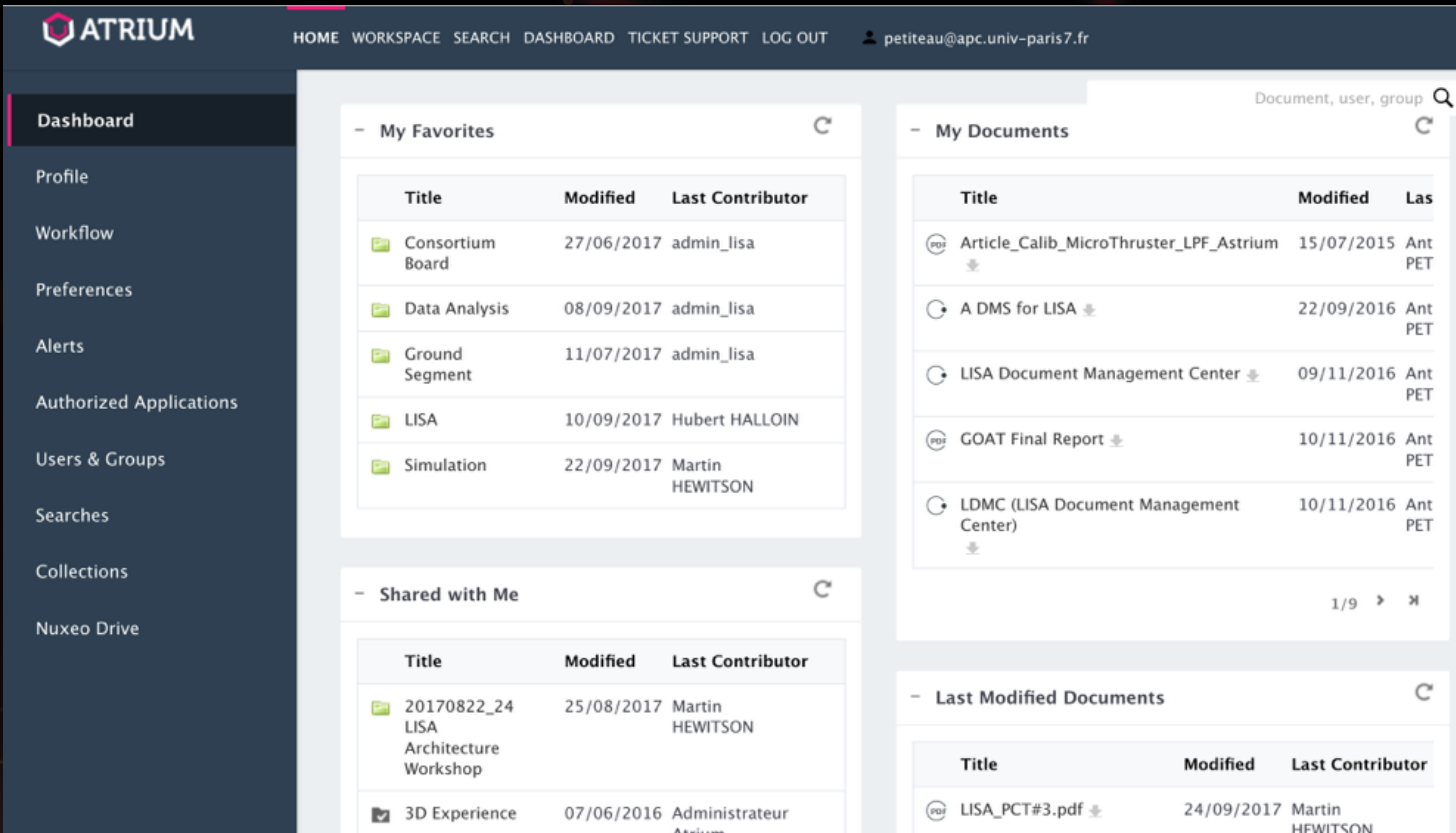


same principles than grid computing with a shorter learning phase.

R&D activities

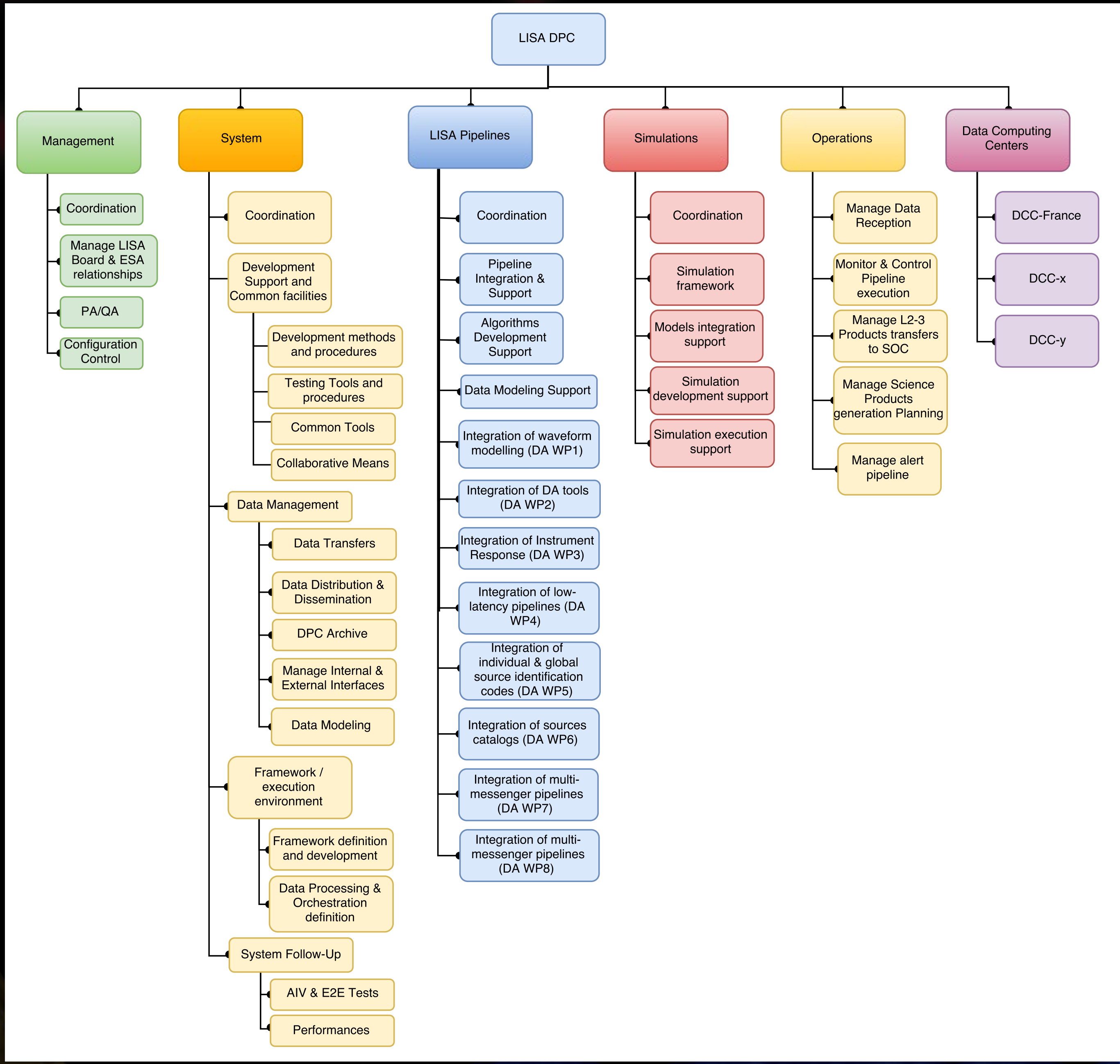
- Docker ochestrator R&T study performed by CNES
- APC involved in the French cloud network
- Doing some actual testing of cloud platform and containers orchestration (singularity).

LISA Doc. Management Syst.



DPC Work Packages

- As for DA/Science, Instrument and Simulation, work packages will be identified
- Preliminary list based on the activities described in the organigram
- Work in progress



DPC Challenges

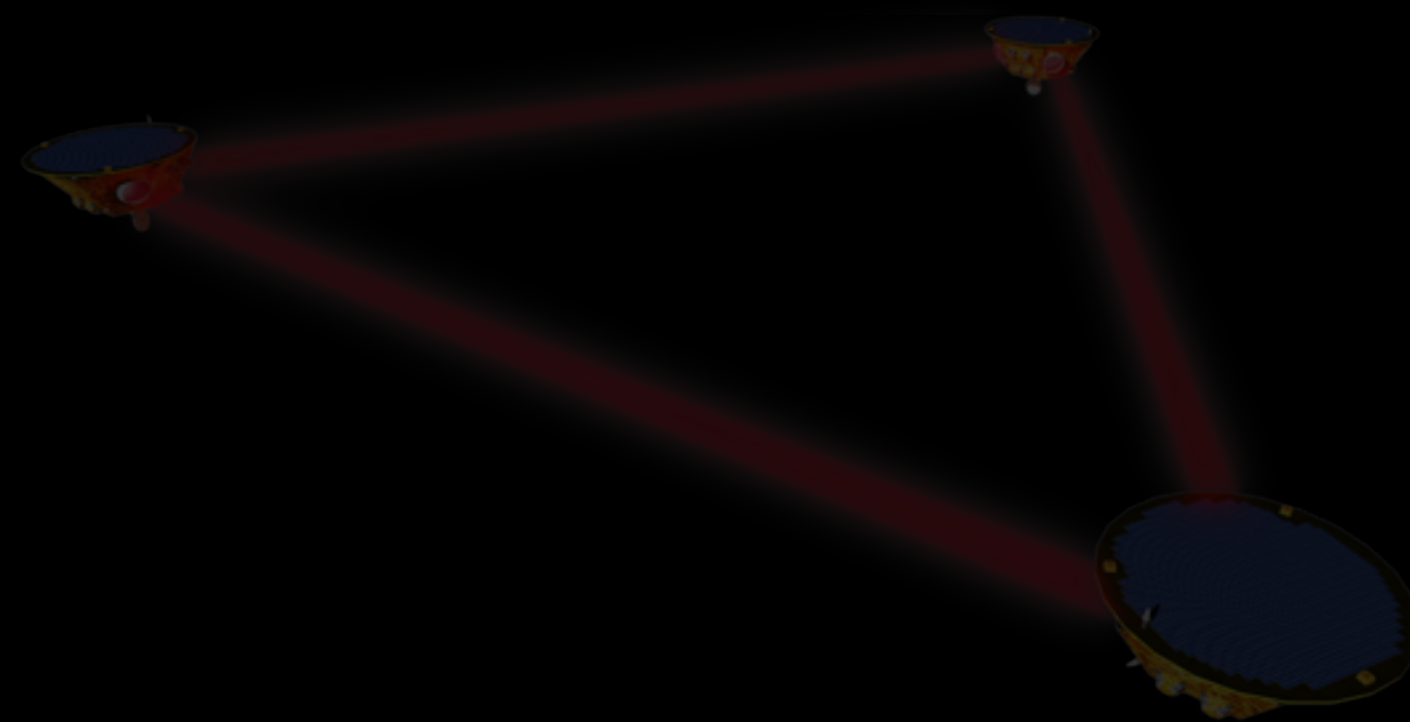
- Flexibility
- Distributed infrastructure
- Easy-to-use: scientists implement in the infrastructure with the support of DPC engineers
- Multiple pipelines
- Data Model & Data Format
- History tracking
- Storages for products and simulation
- CPU need to run pipelines, simulations and pipelines on simulated data

LISA Data Processing Group

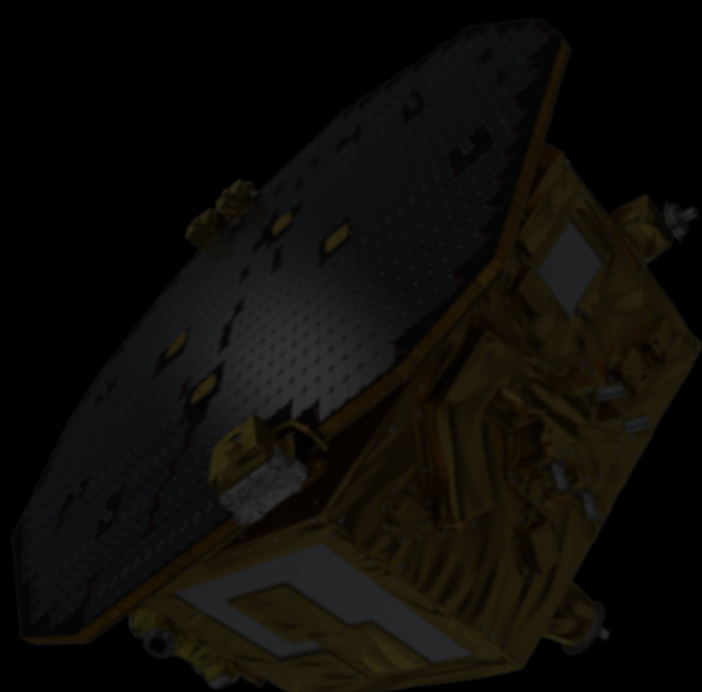
- The development of the DPC includes:
  - definition and maintenance of the pipeline and data analysis development environment;
  - design and implementation of the pipeline and analysis operations environment;
  - design and implementation of data storage facilities and databases;
  - implementation and operation of consortium IT services;
  - management and implementation of pipelines for simulation and data analysis.

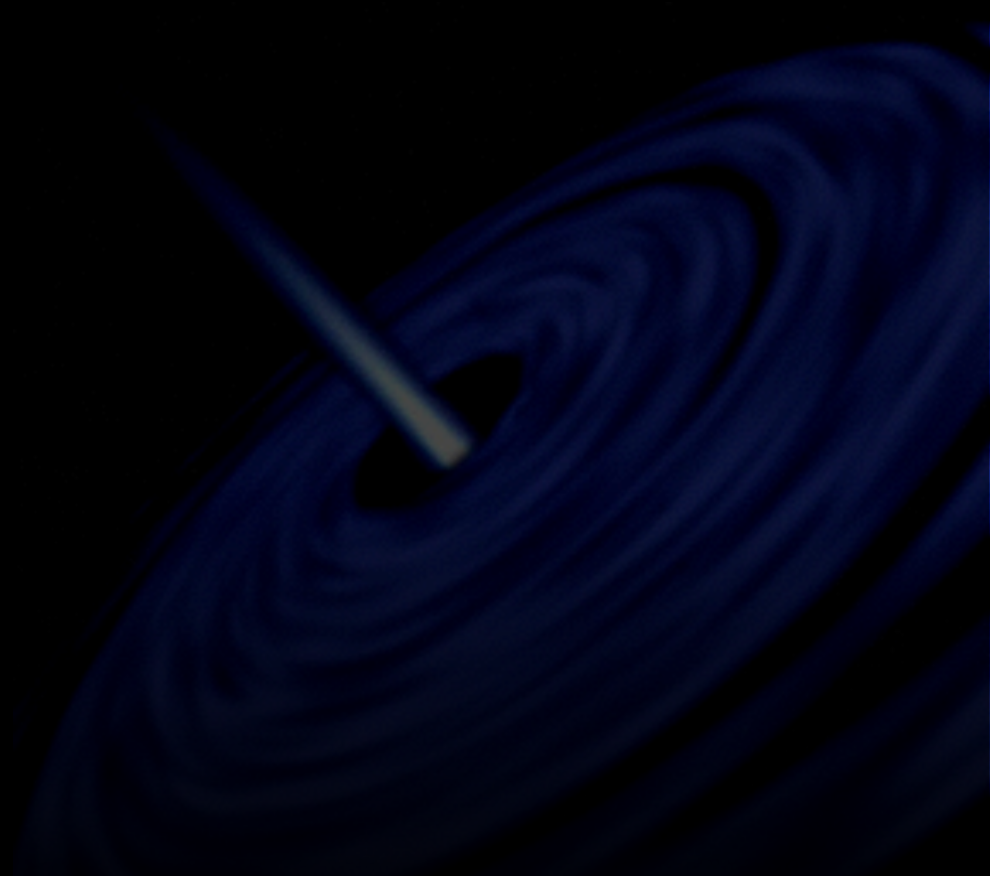
Conclusion

- LISA DPC is processing L1 (TDI) data received from the SOC to produce L2 & L3 data (catalogs) sent to the SOC.
- Particularities of LISA data: first of this kind, discovery mission => flexibility & evolution
- DPC deliverable from the consortium.
- Unique entity managing a "DPC software" running on DCCs
- Activities: Management, System, Operations, Hardware, Simulation, Pipelines
- Integration of methods and pipelines for simulation and DA but definition and prototype defined by LSG and Simulation WG
- Existing proto-DPC already used by the preliminary consortium and agencies

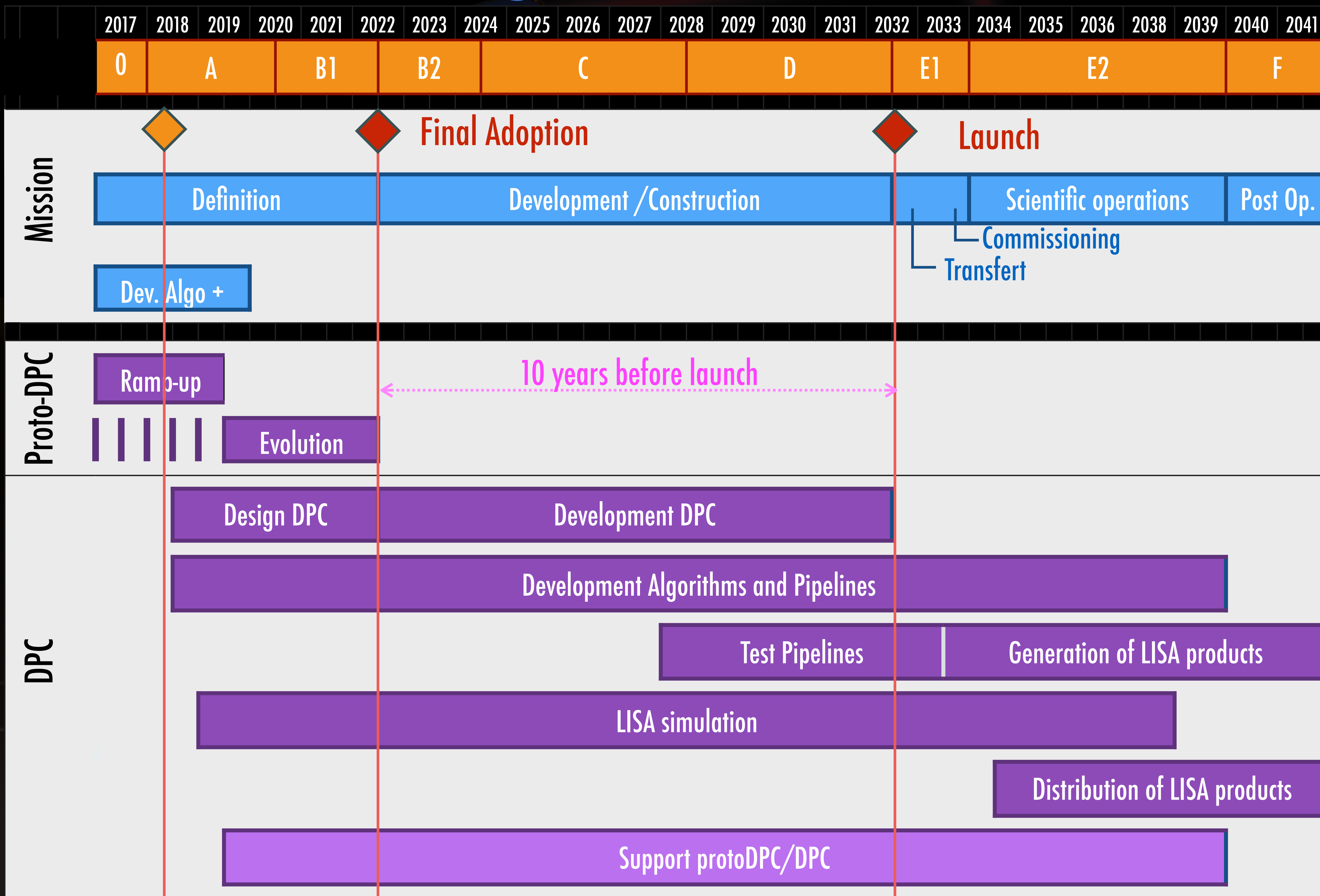


Thank you





Planning of development



Others experiments

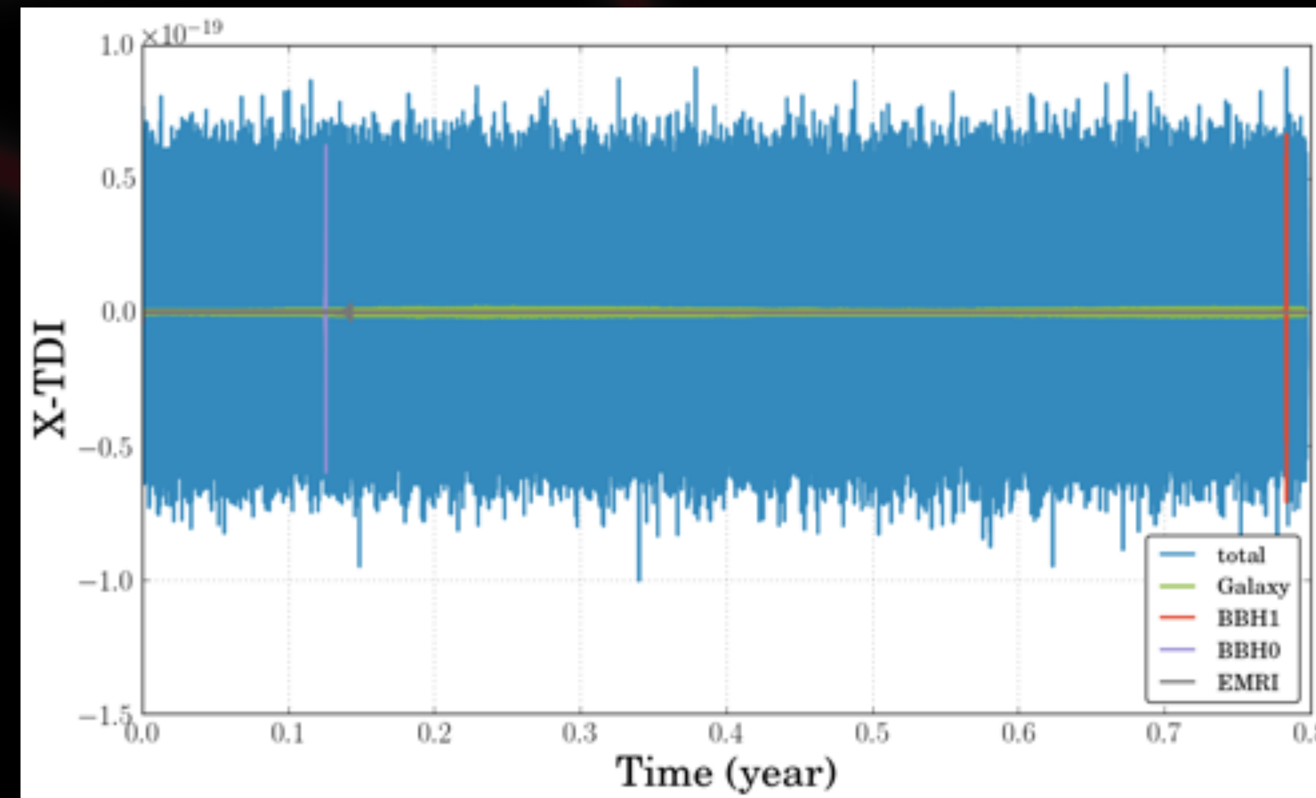
- Others GW observatories: LIGO/Virgo
  - Similarity in some type of sources (binaries, stochastic background, etc) and some type of noises
  - But difference in SNR, data volumes, available channels, etc

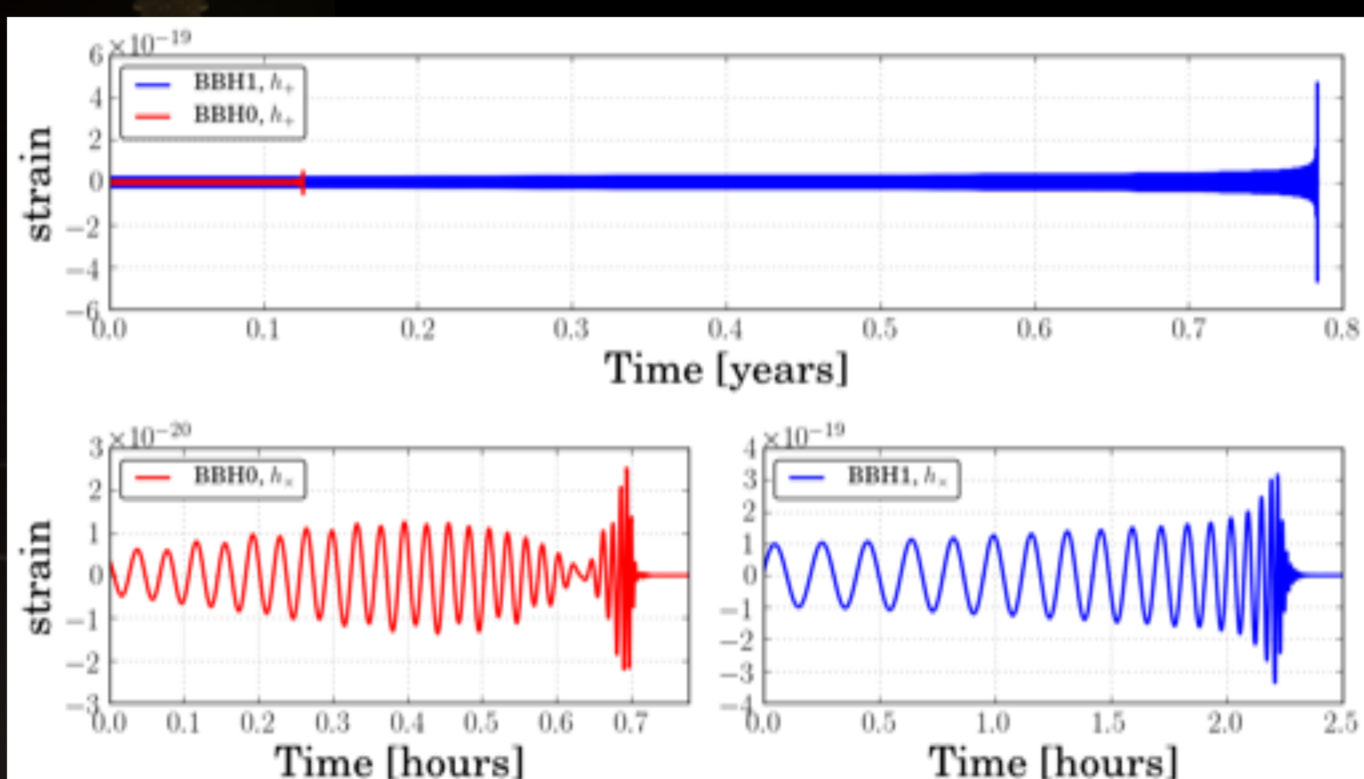
=> some data analysis methods & processing strategies can be updated for LISA

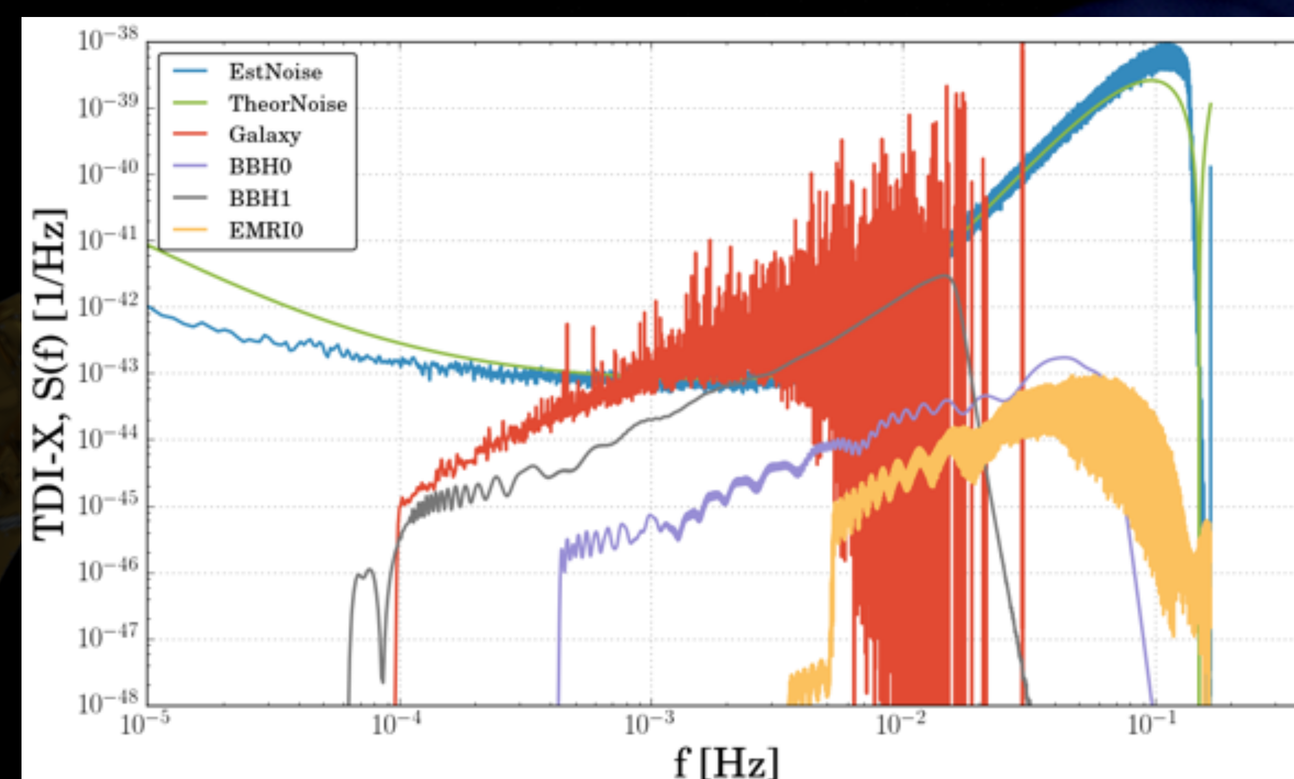
- Space-based observatories: Planck, GAIA, Euclid, ...
  - Several examples for ground segment organisation
  - Space qualified data processing for operation
  - Not possible to modify the instrument in space
  - ...

LISA Data Challenges

- Mock LDC: 2005→2011
- 2017: start of the LDC
  - Develop data analysis
  - Design the pipelines of the mission
- Example of the potential data for LDC1 (see Stas's talk)



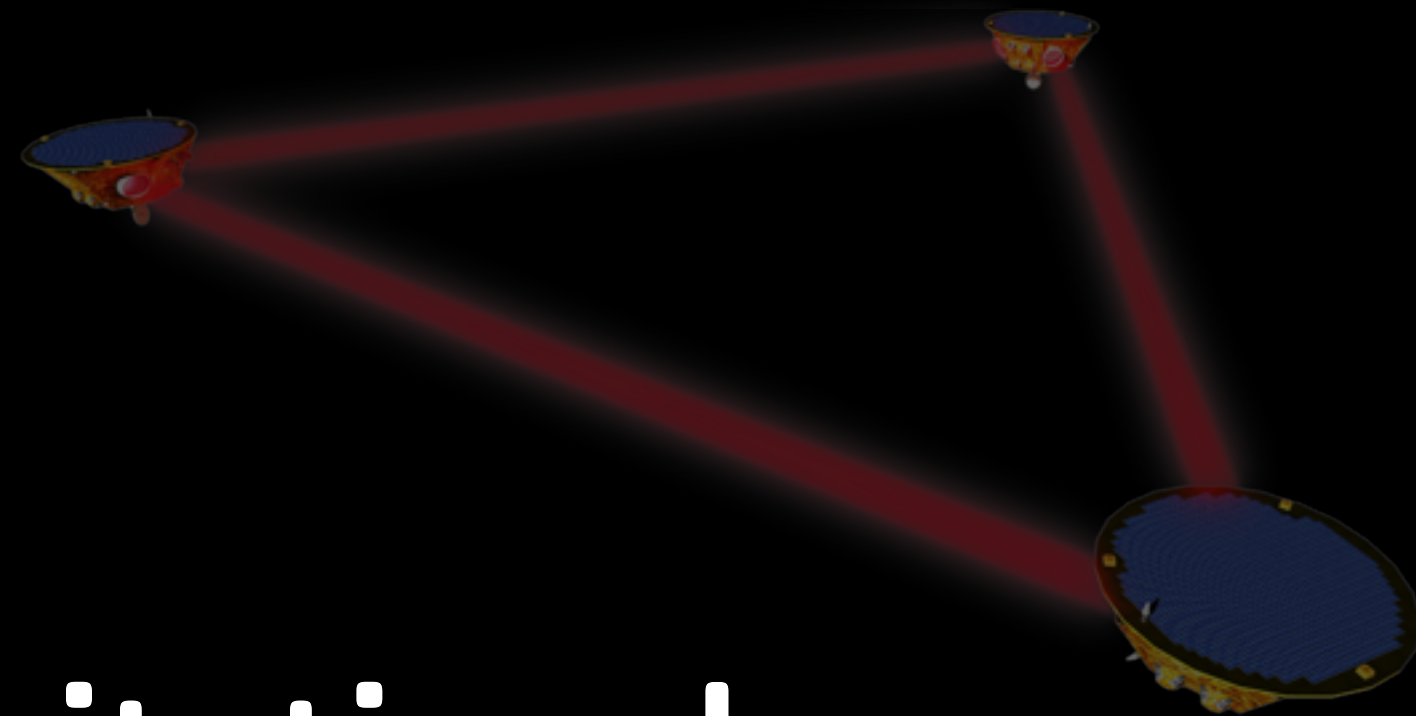




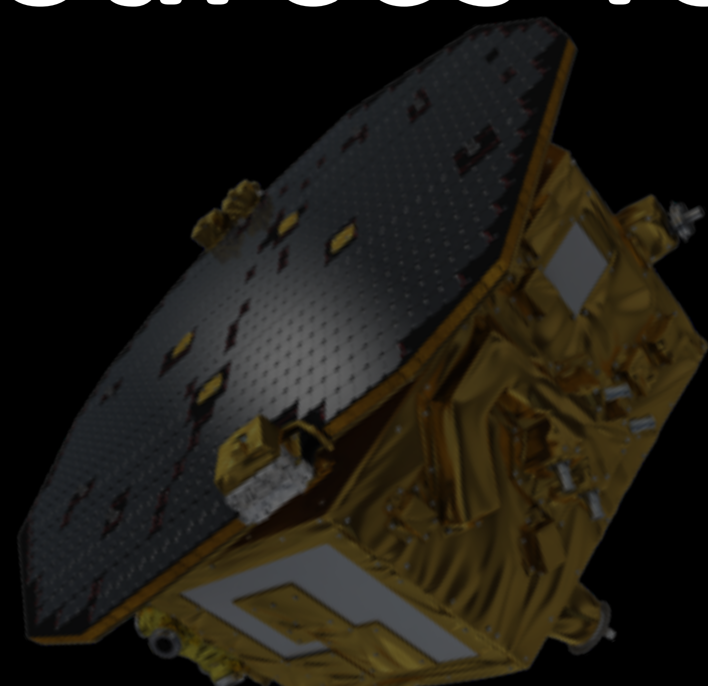
DPC in LISA proposal

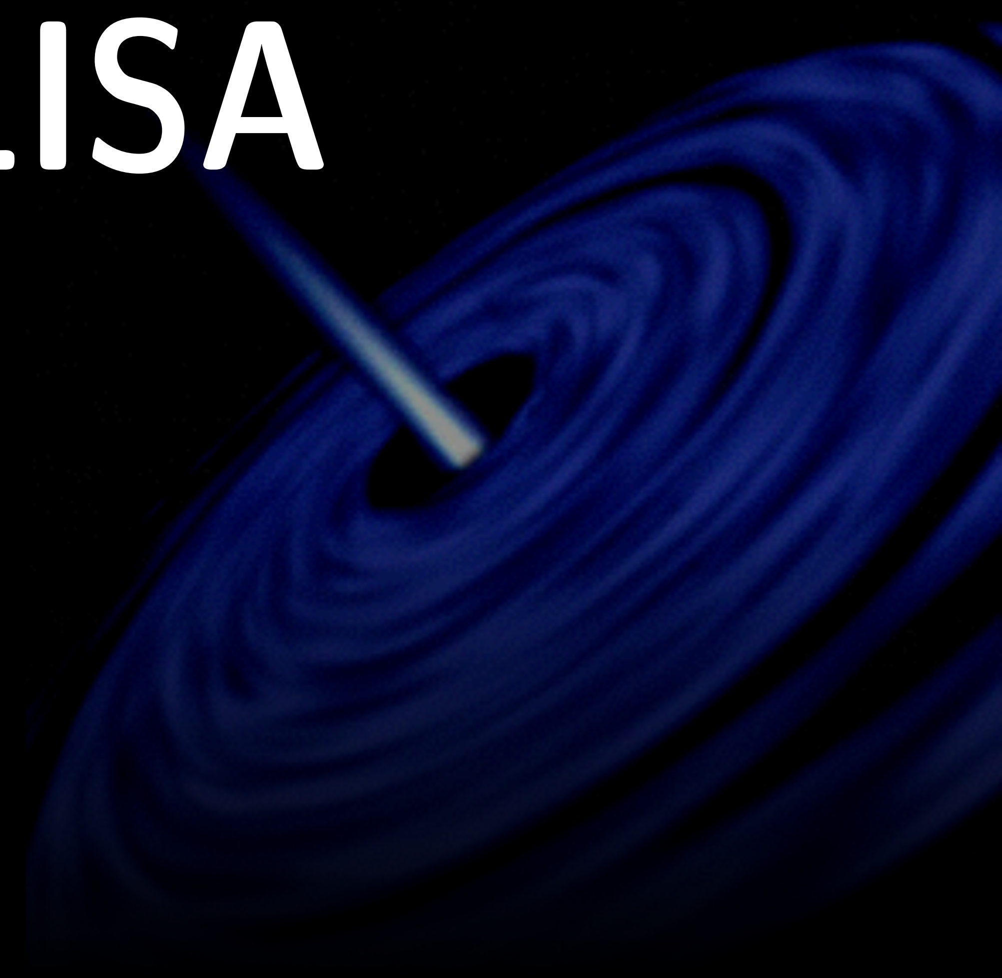
section 8: management

- <u>Ground Segment:</u> Ground Segment element includes all Mission Operations during Low Earth Operations (LEOP), and later during nominal operations and Science Operations under ESA responsibility, that is raw science data pre-processing and calibration, leading to level-1 data (TDI combinations). This task will be performed with support from France, Italy, the United Kingdom, Switzerland, Spain, Germany and the US (algorithm development) and the instrument providers (calibration during operation).

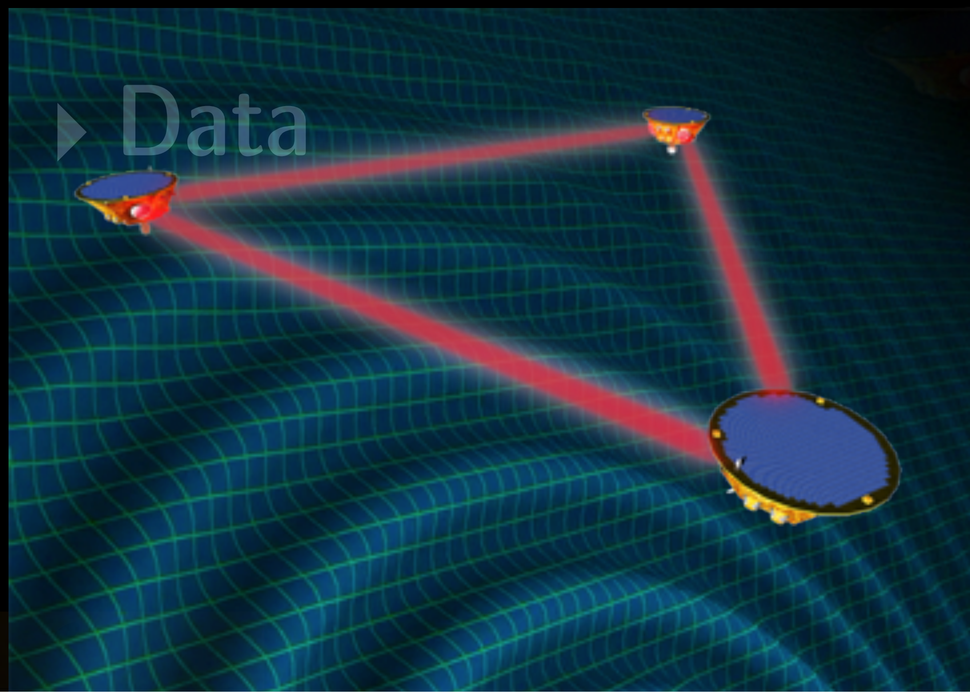


Gravitational wave sources for LISA



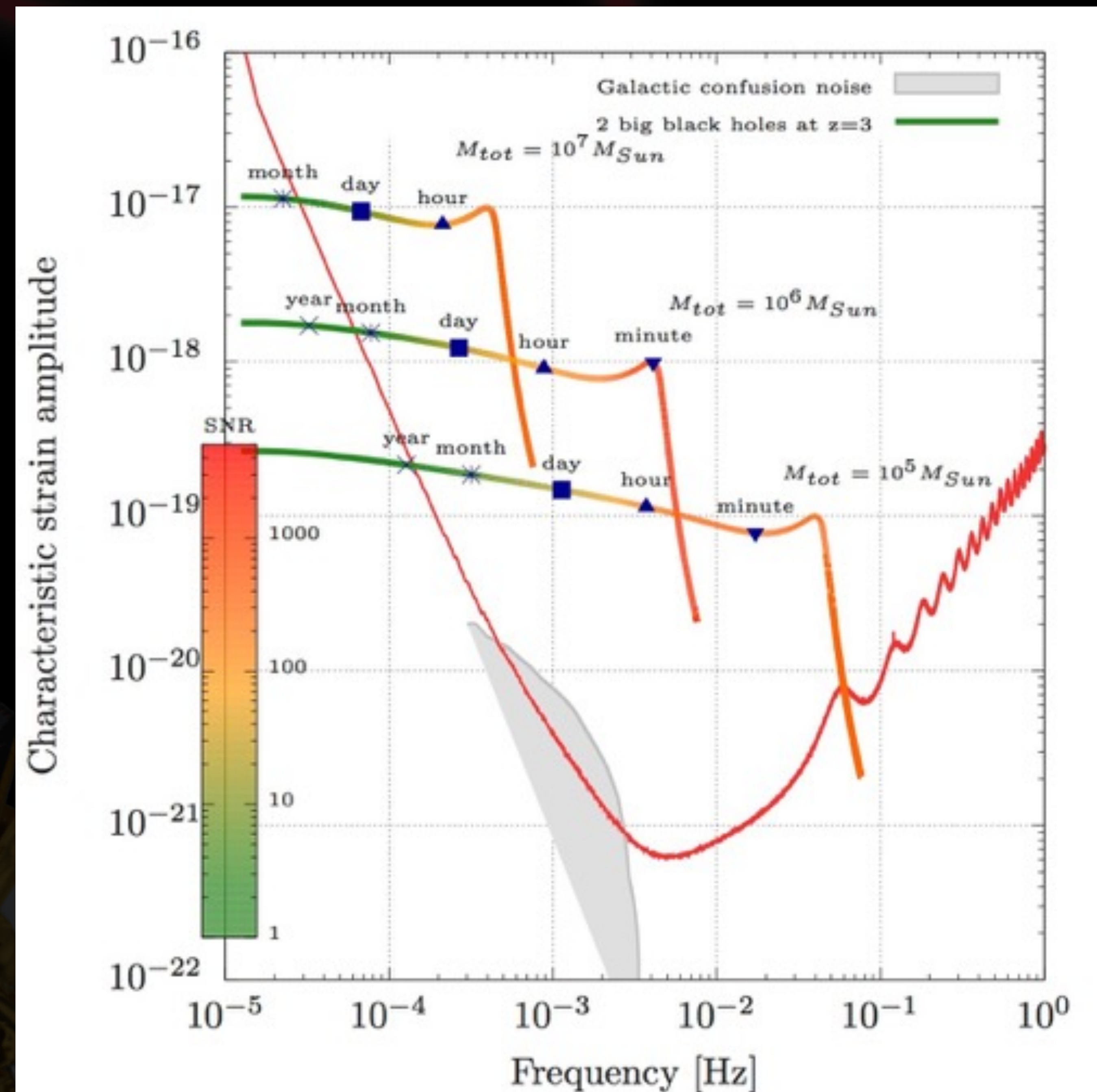


Galactic binaries



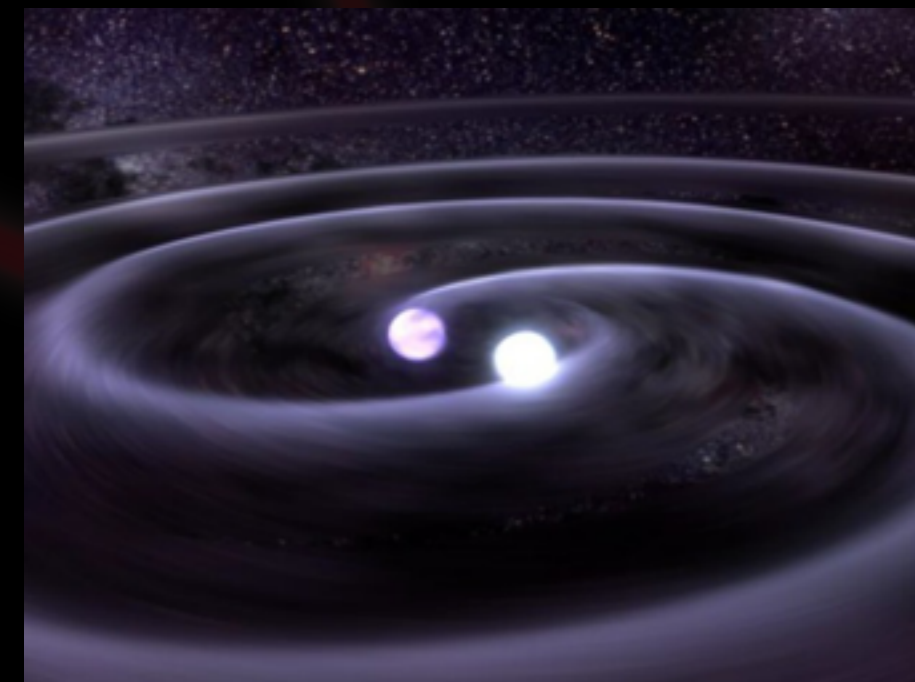
GW sources

- 6 x10^7 galactic binaries

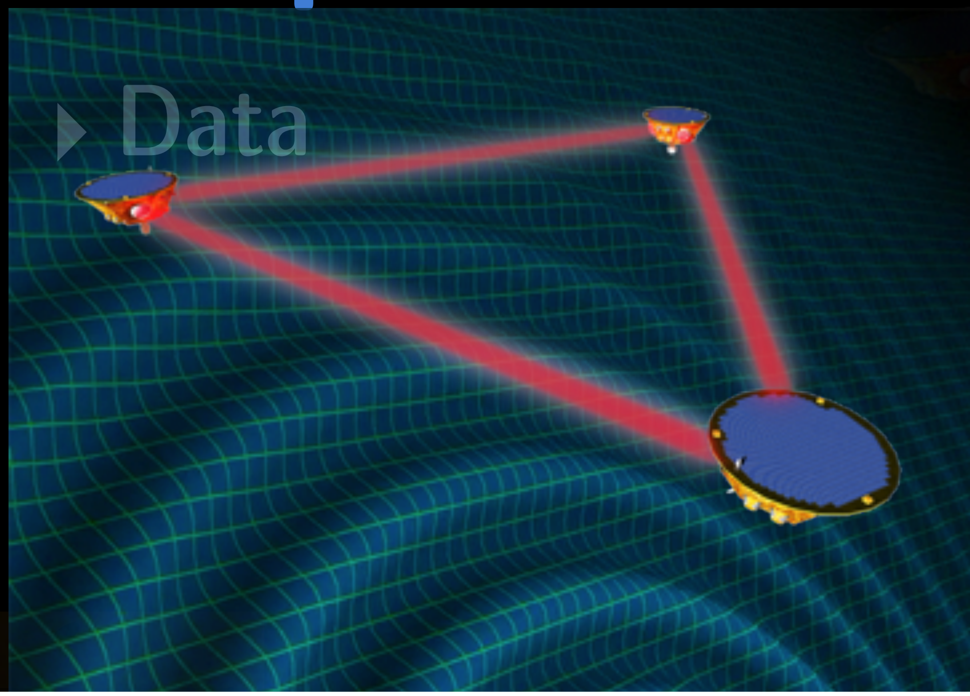


Galactic binaries

- Gravitational wave:
  - quasi monochromatic
- Duration: permanent
- Signal to noise ratio:
  - detected sources: 7 - 1000
  - confusion noise from non-detected sources
- Event rate:
  - 25 000 detected sources
  - more than 10 guarantied sources (verification binaries)

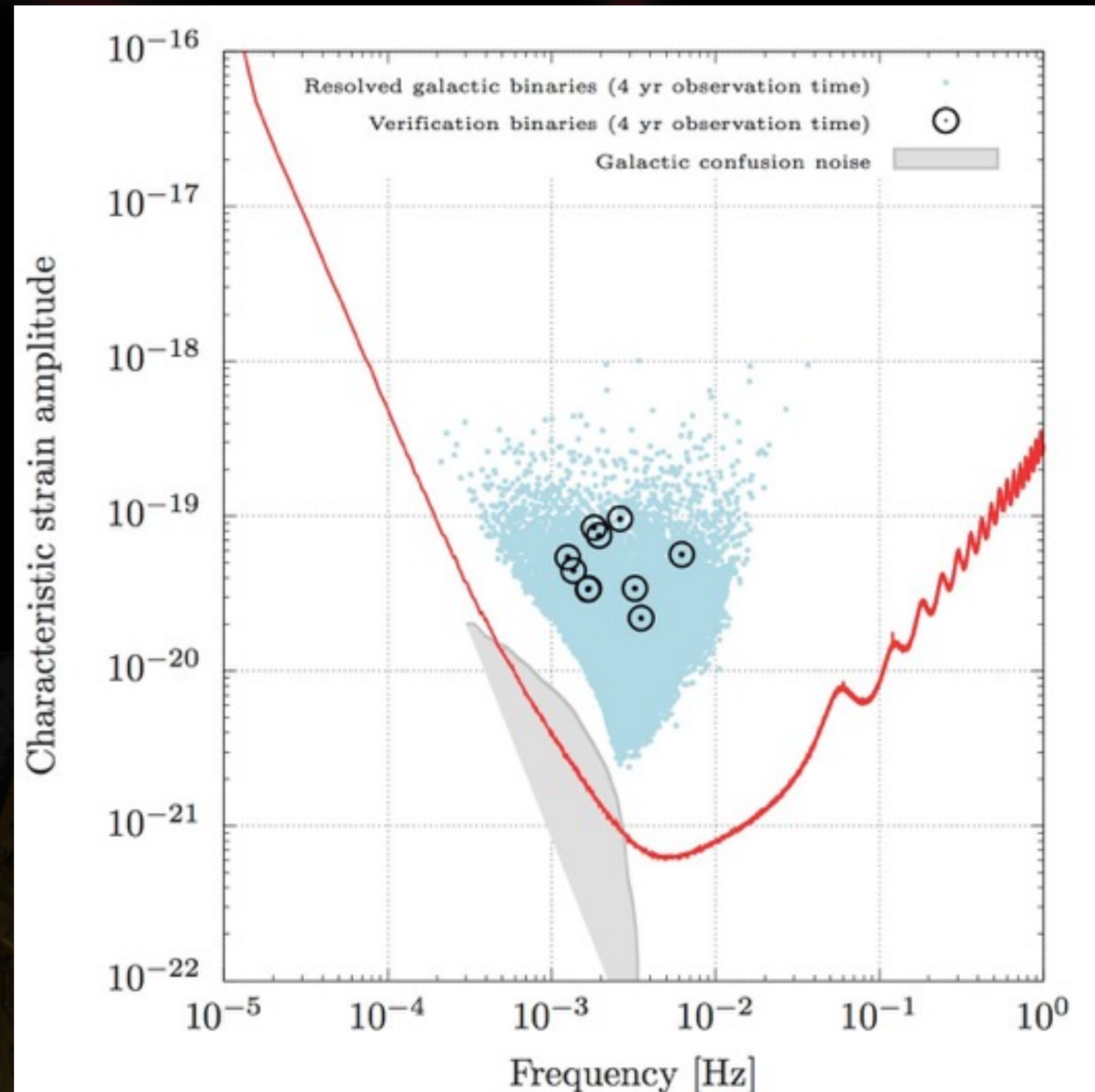


Super Massive Black Hole Binaries



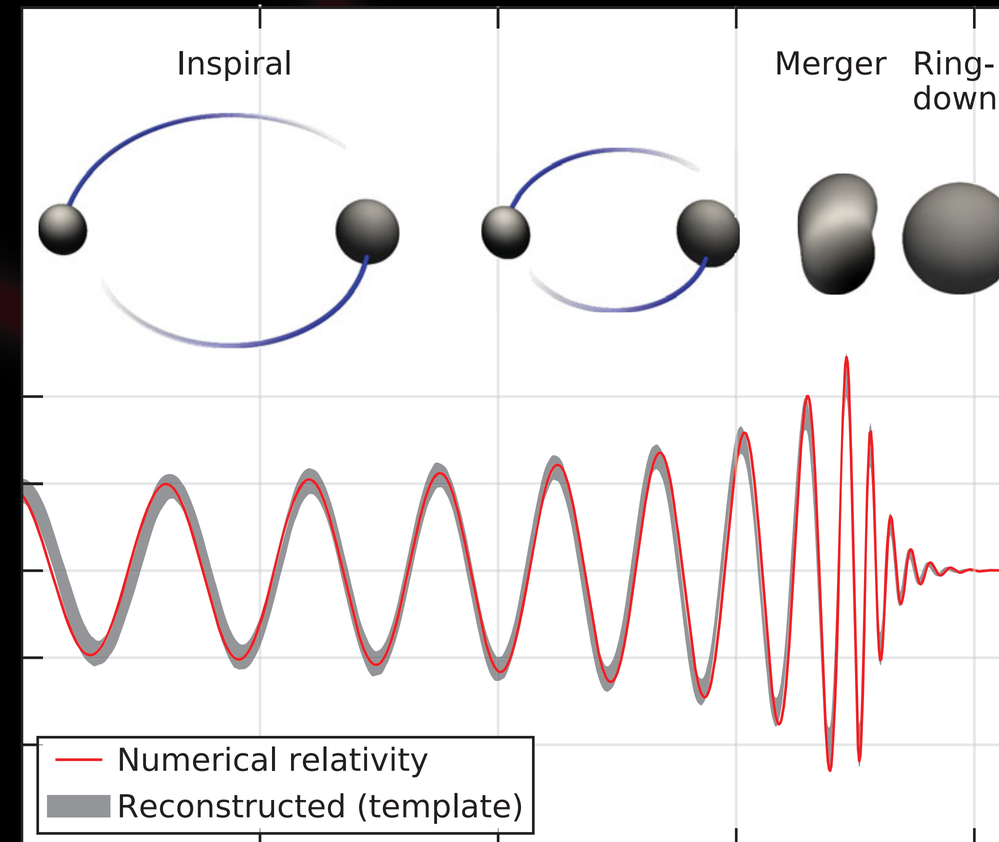
OG sources

- 6 x10^7 galactic binaries
- 10-100/year SMBHBs



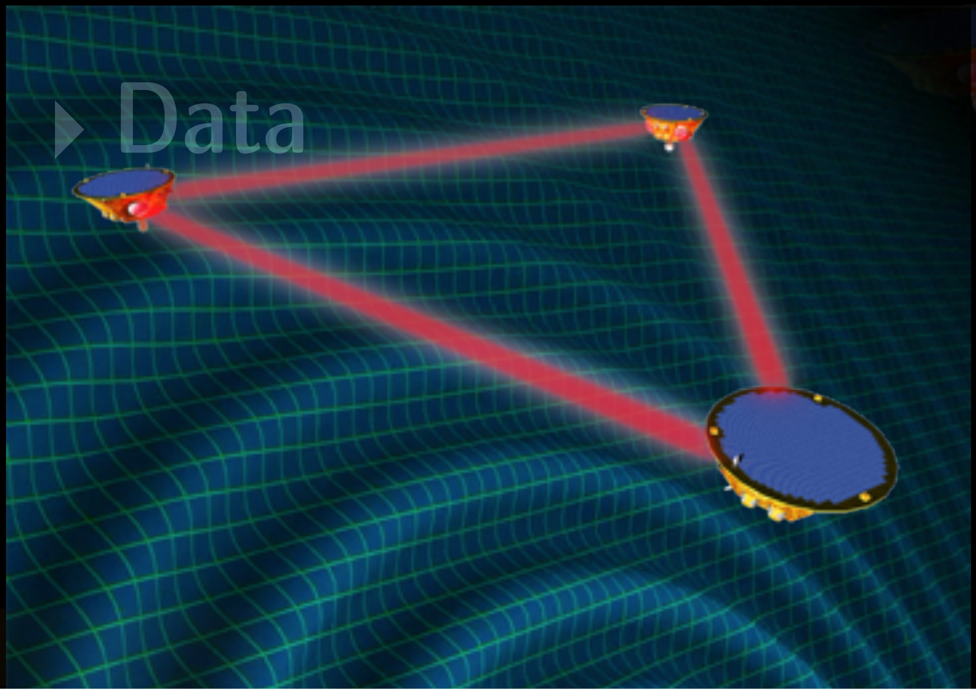
Super Massive Black Hole Binaries

- Gravitational wave:
  - Inspiral: Post-Newtonian,
  - Merger: Numerical relativity,
  - Ringdown: Oscillation of the resulting MBH.



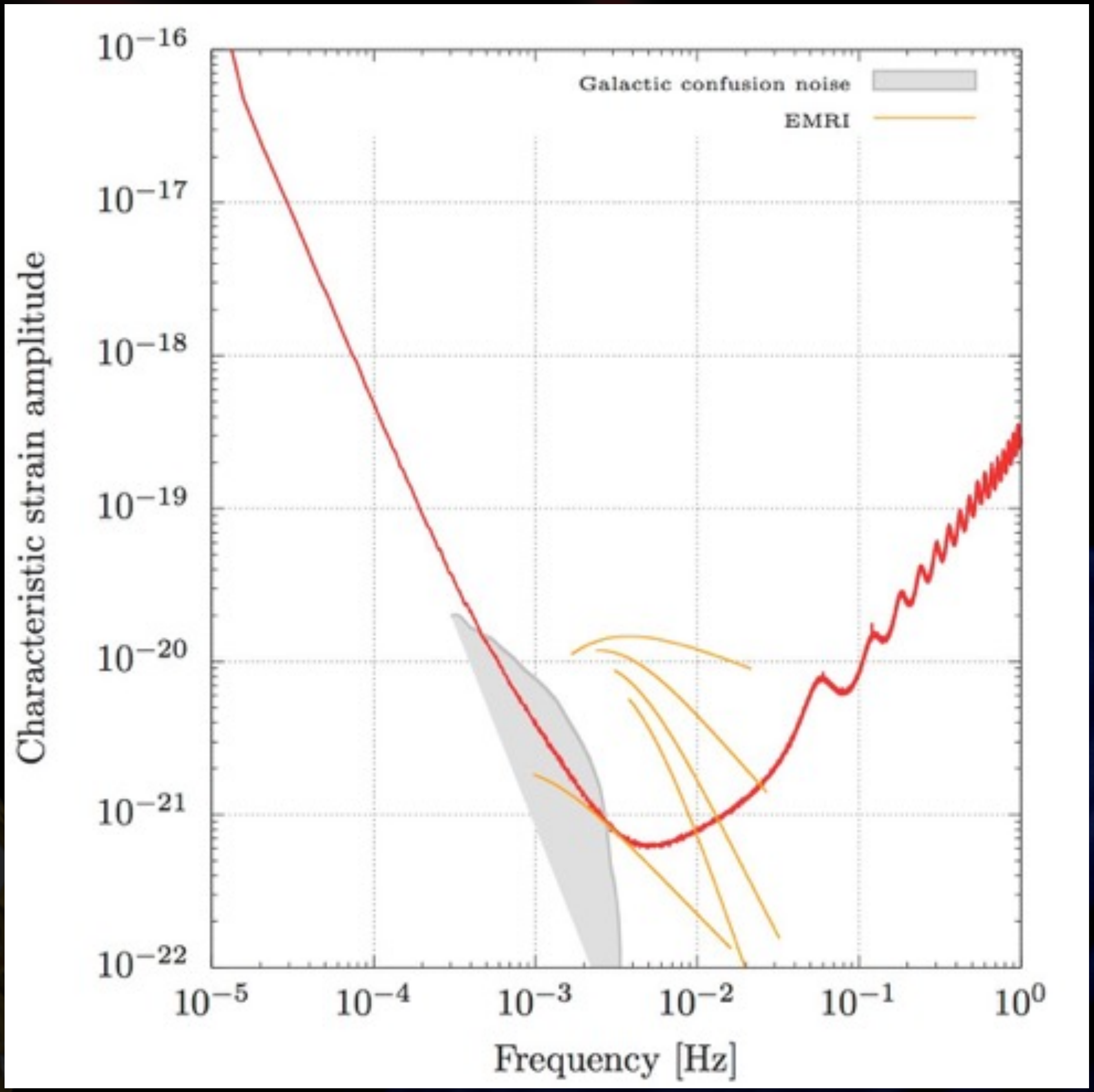
- Duration: between few hours and several months
- Signal to noise ratio: until few thousands
- Event rate: 10-100/year

EMRIs



OG sources

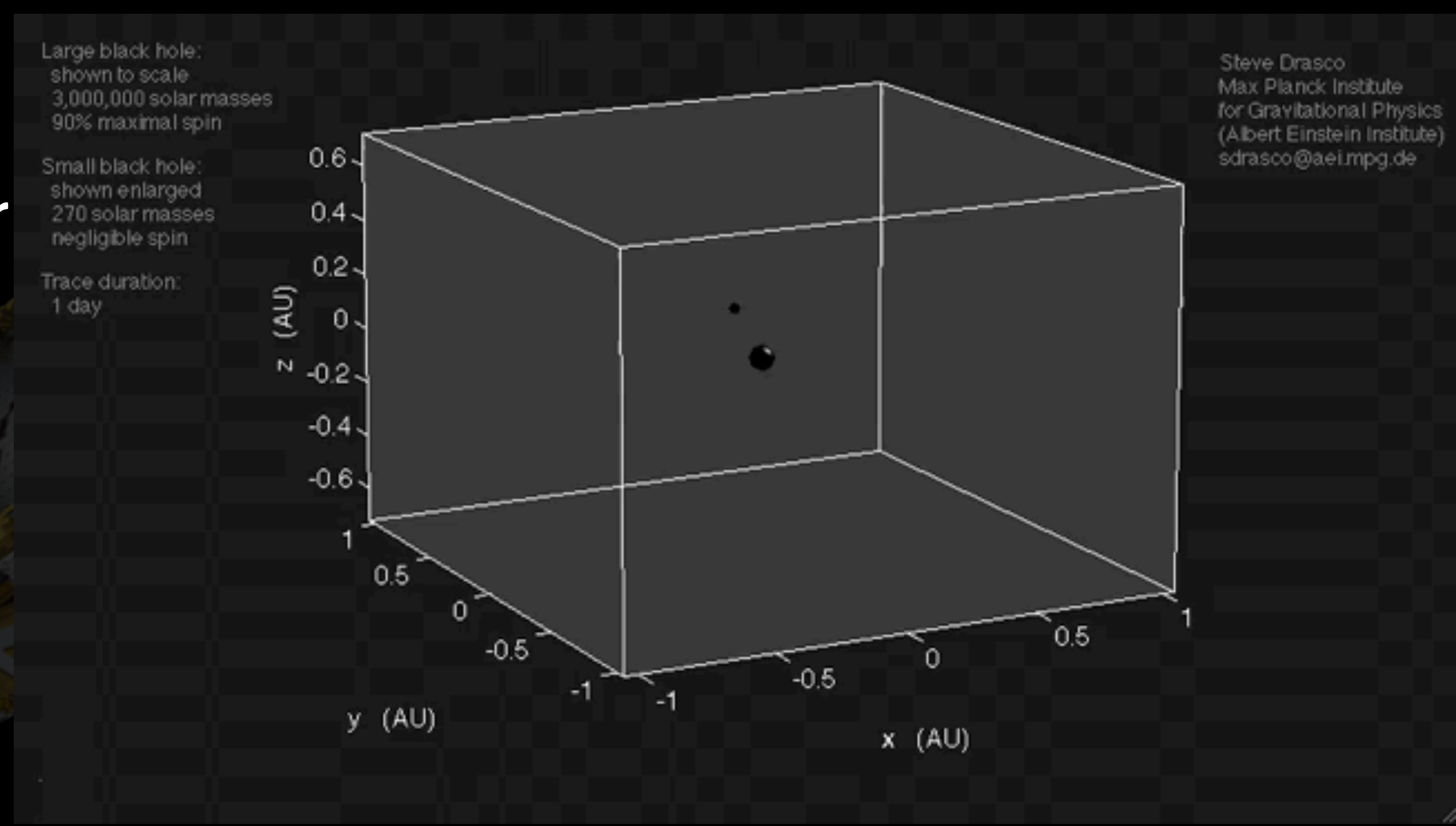
- 6×10^7 galactic binariess
- 10-100/year SMBHBs
- 10-1000/years EMRIs



EMRIs

- Gravitational wave:
  - very complex waveform
  - No precise simulation at the moment
- Duration: about 1 year
- Signal to Noise Ratio: from tens to few hundreds
- Event rate:

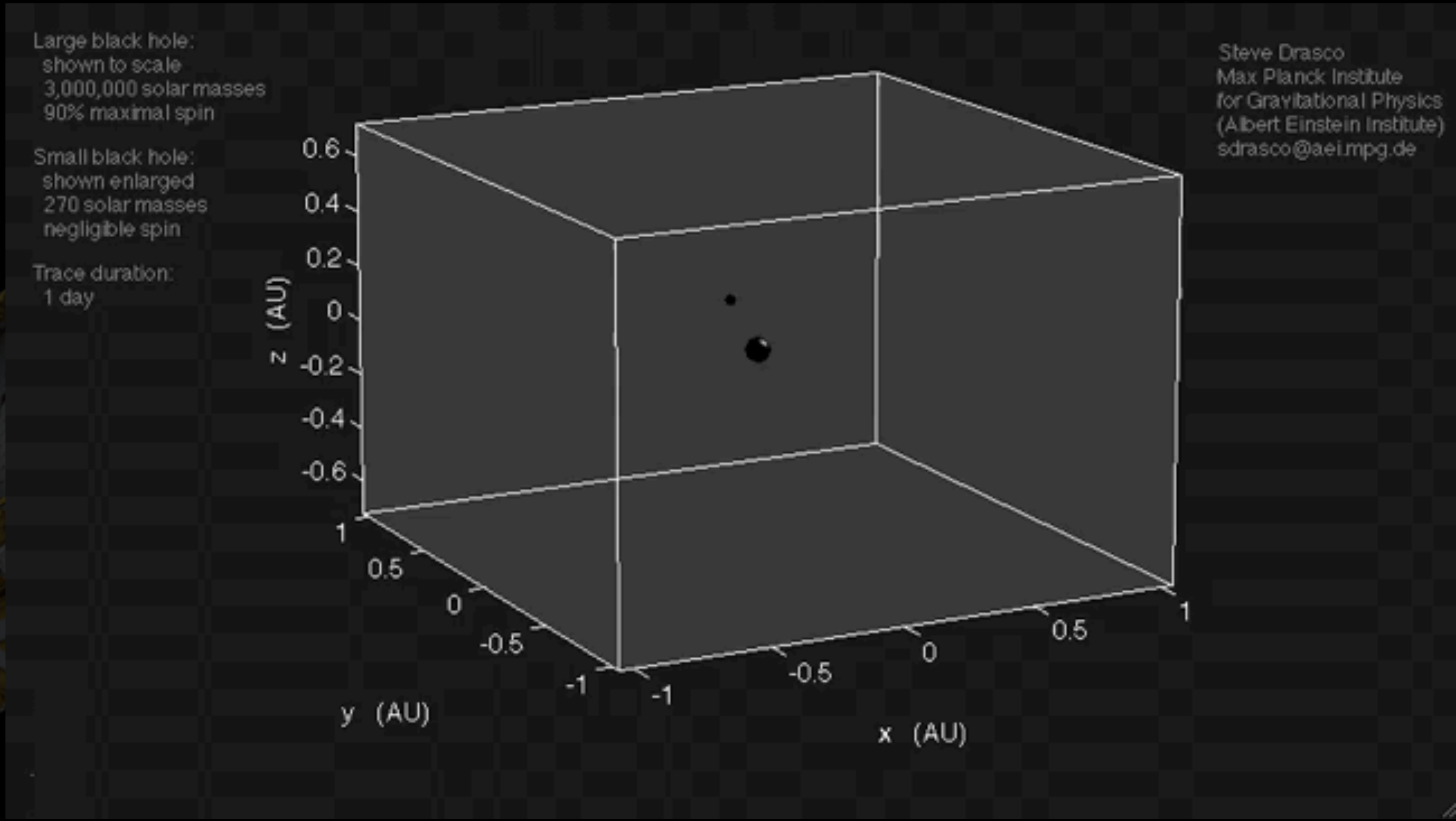
  from few events per year to few hundreds



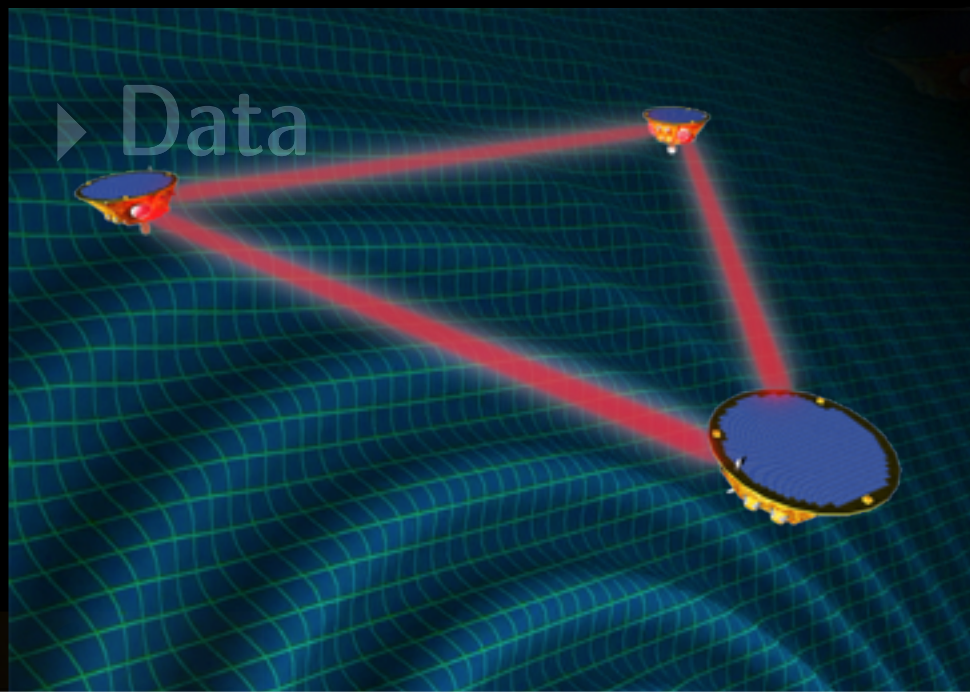
EMRIs

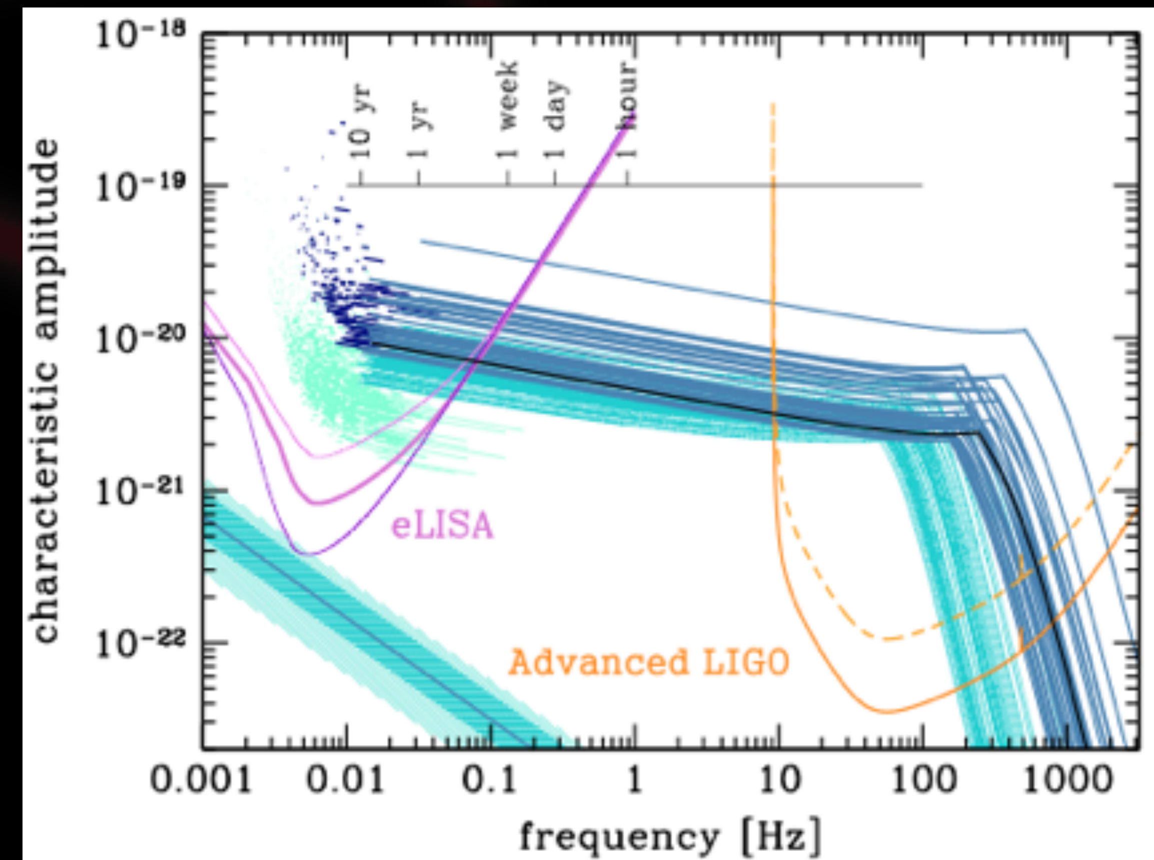
- Gravitational wave:
  - very complex waveform
  - No precise simulation at the moment
- Duration: about 1 year
- Signal to Noise Ratio: from tens to few hundreds
- Event rate:

  from few events per year to few hundreds



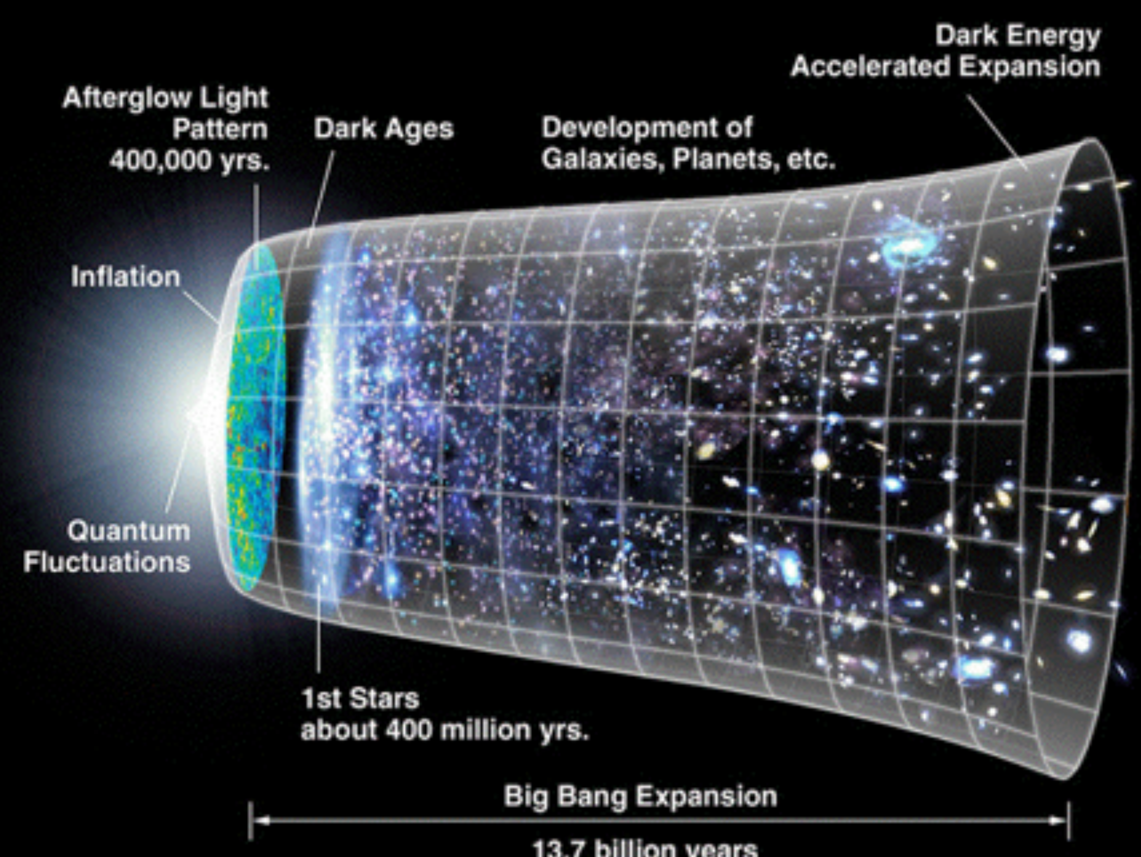
Others sources



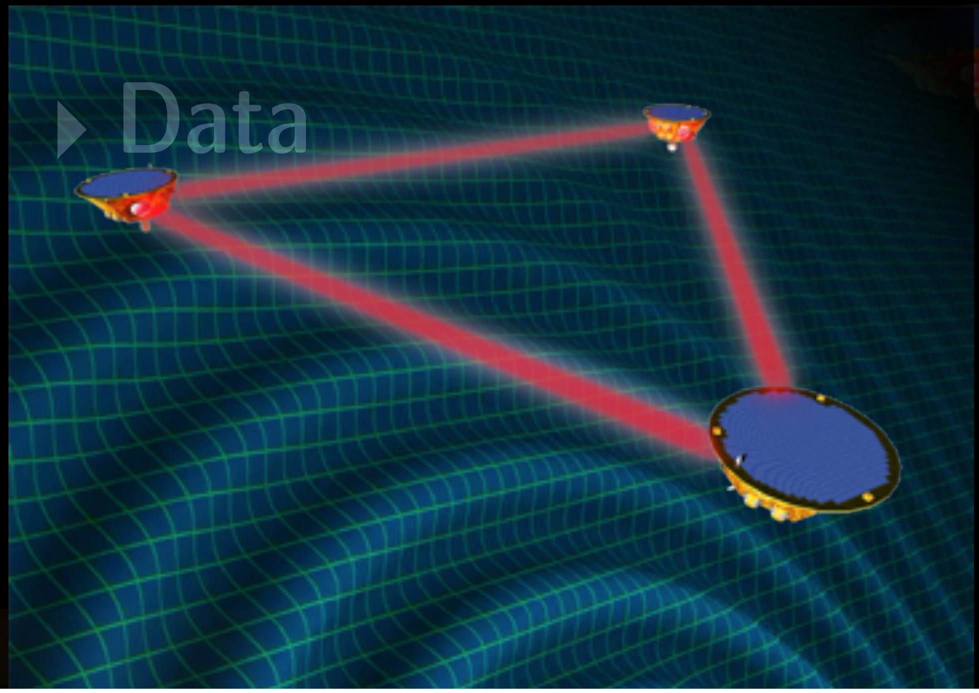


GW sources

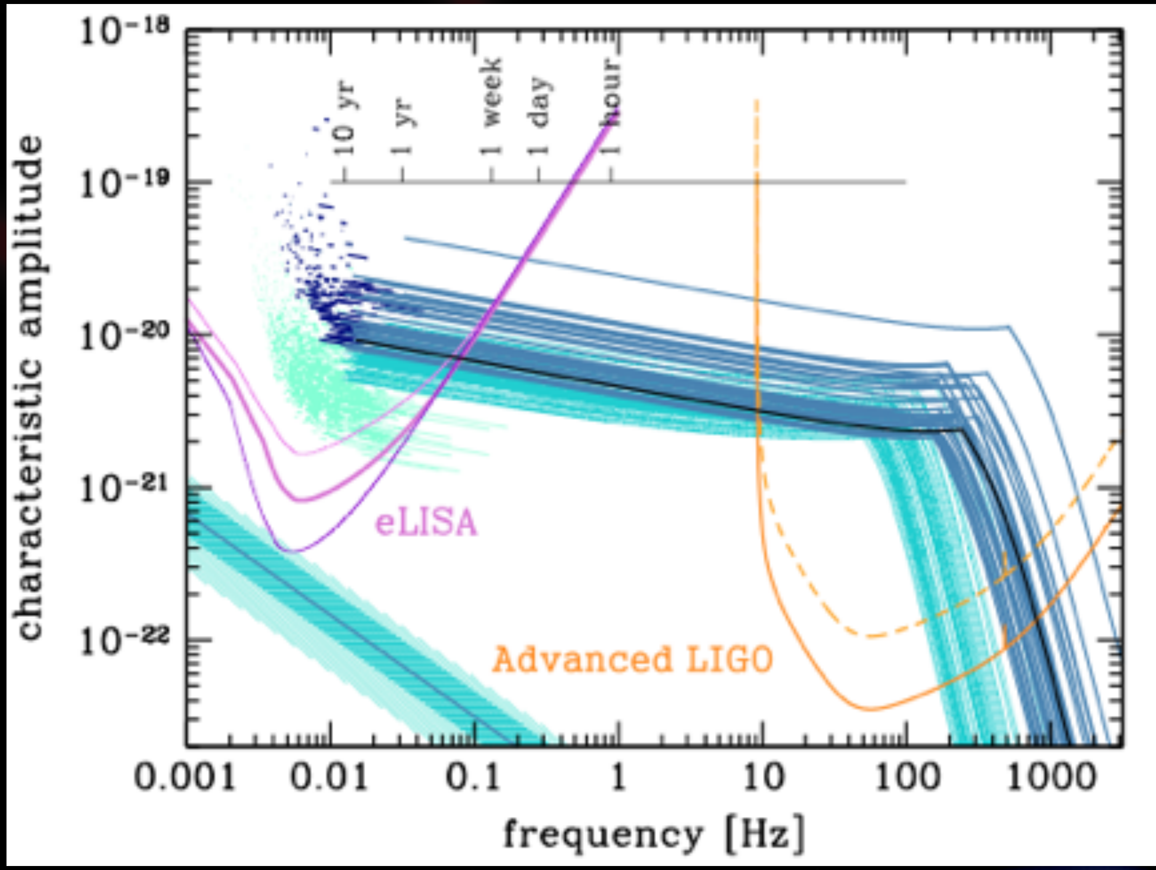
- 6 x10^7 galactic binaries
- 10-100/year SMBHBs
- 10-1000/year EMRIs
- large number of Stellar Origin BH binaries (LIGO/Virgo)
- Cosmological backgrounds
- Unknown sources



Others sources

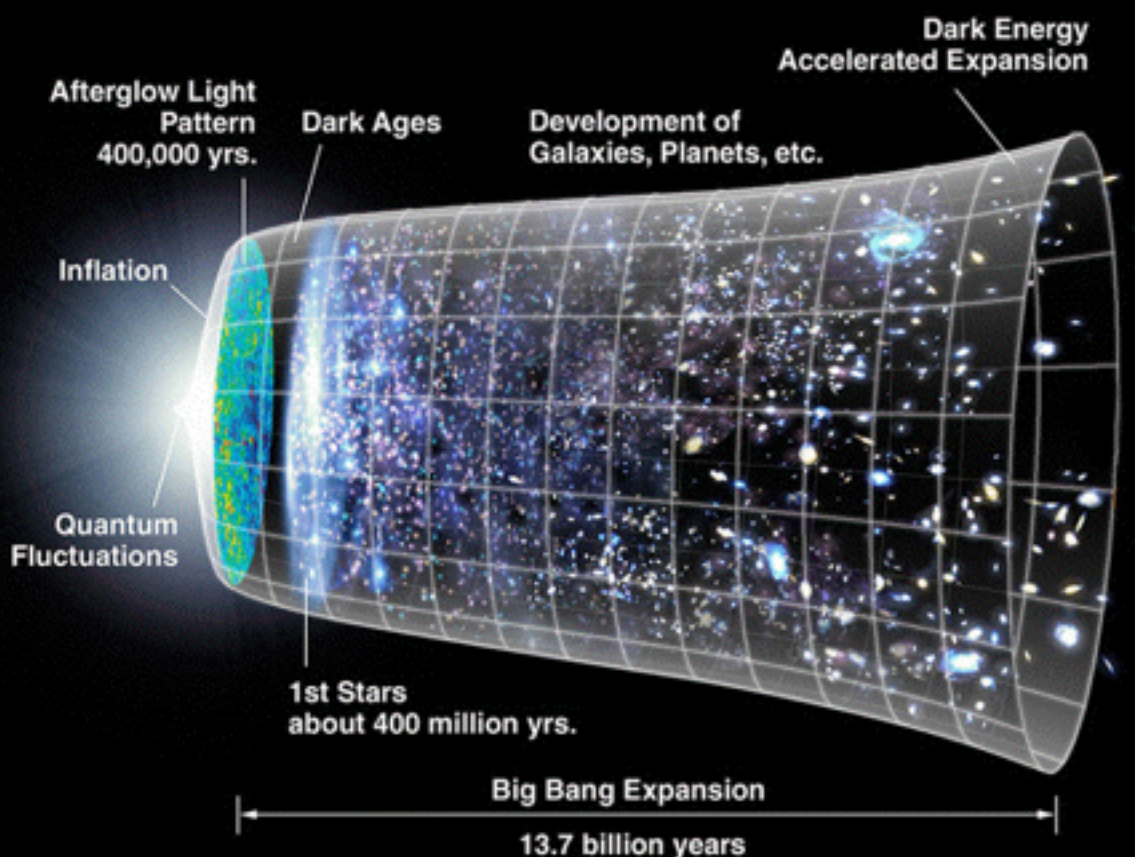


?



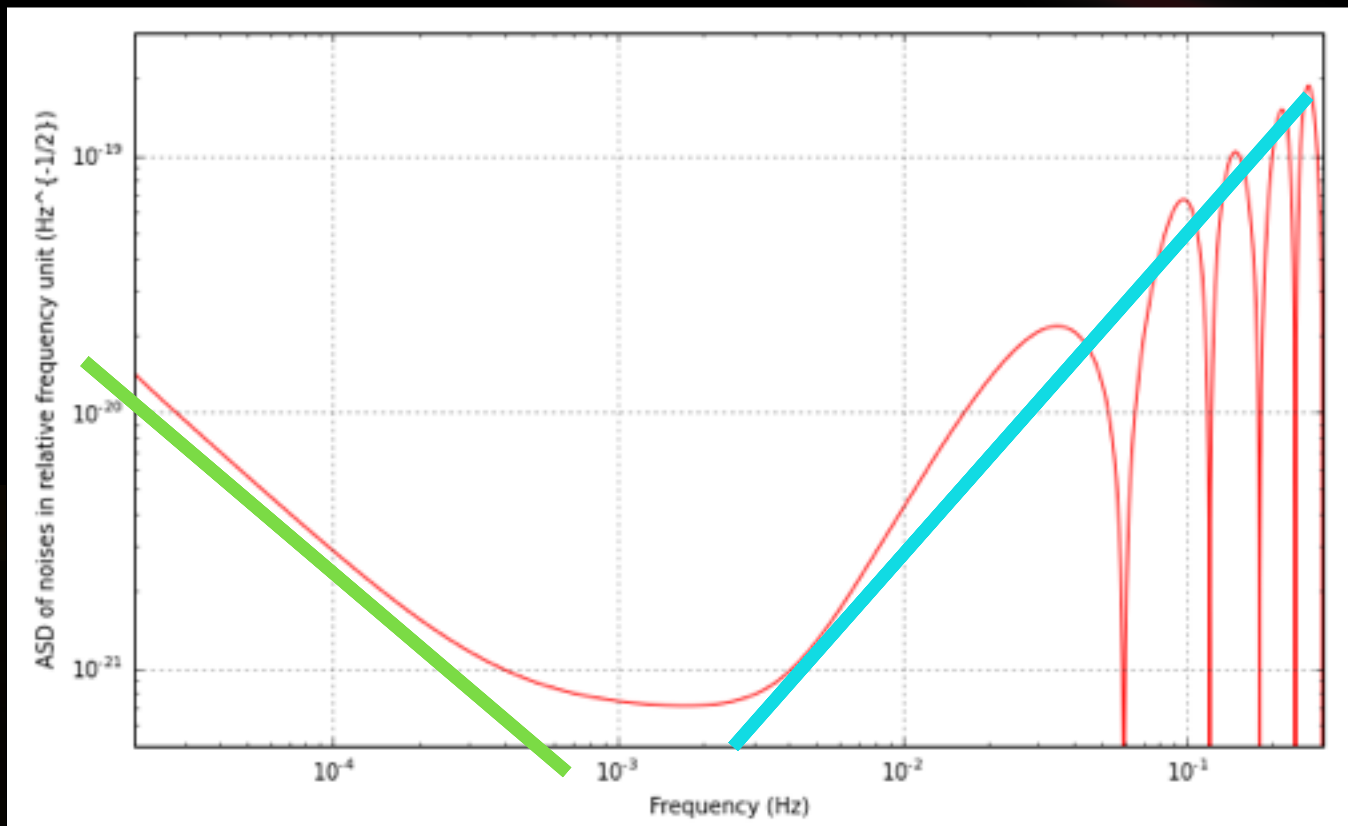
GW sources

- 6×10^7 galactic binaries
- 10-100/year SMBHBs
- 10-1000/year EMRIs
- large number of Stellar Origin BH binaries (LIGO/Virgo)
- Cosmological backgrounds
- Unknown sources

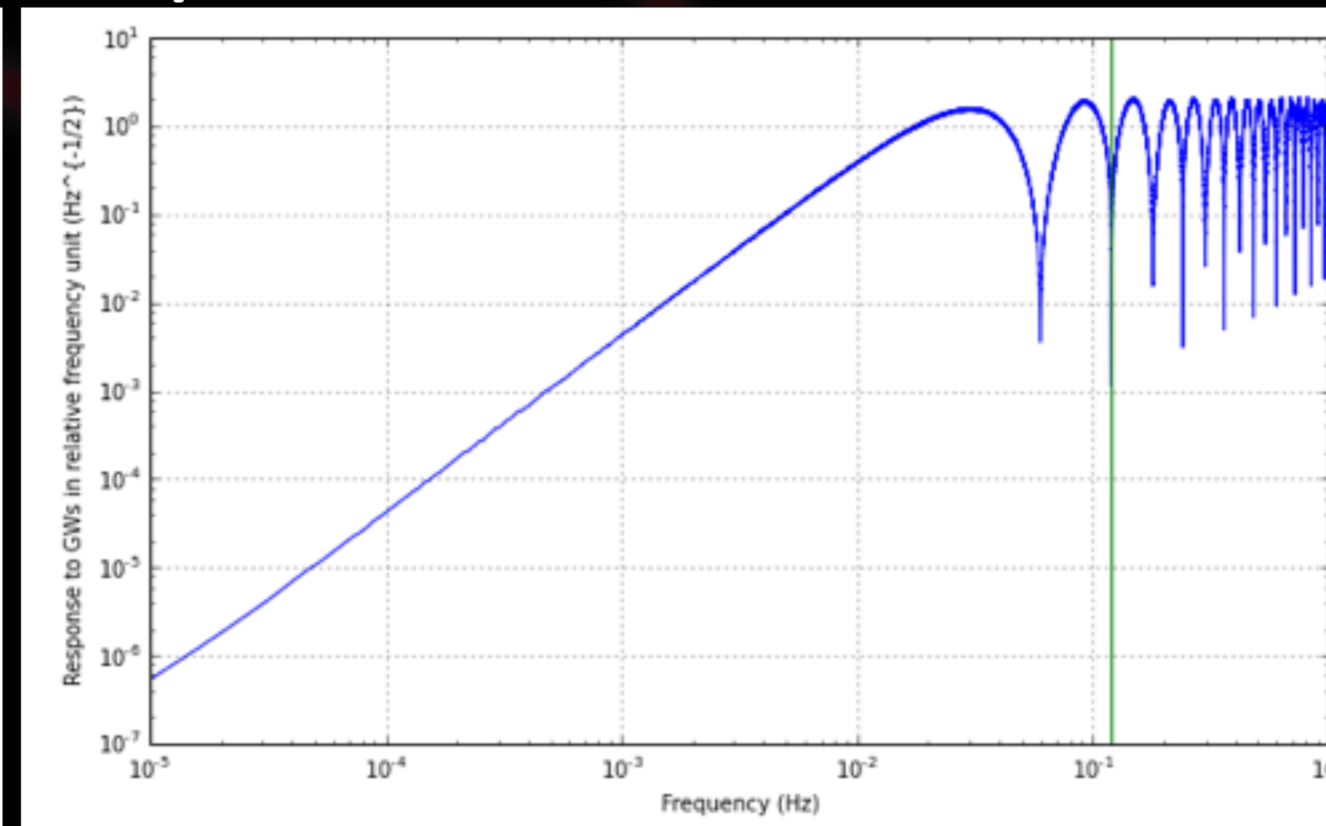


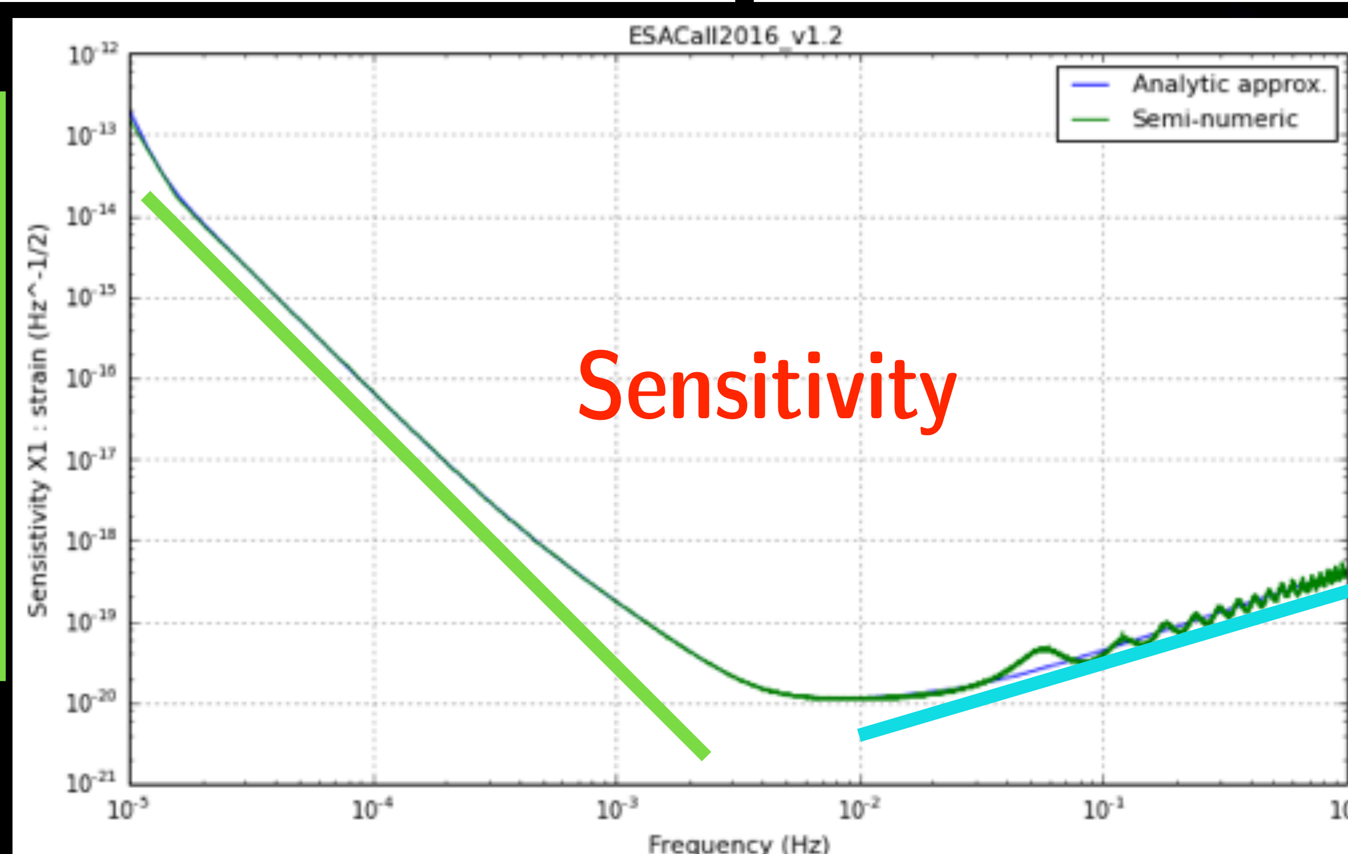
LISA

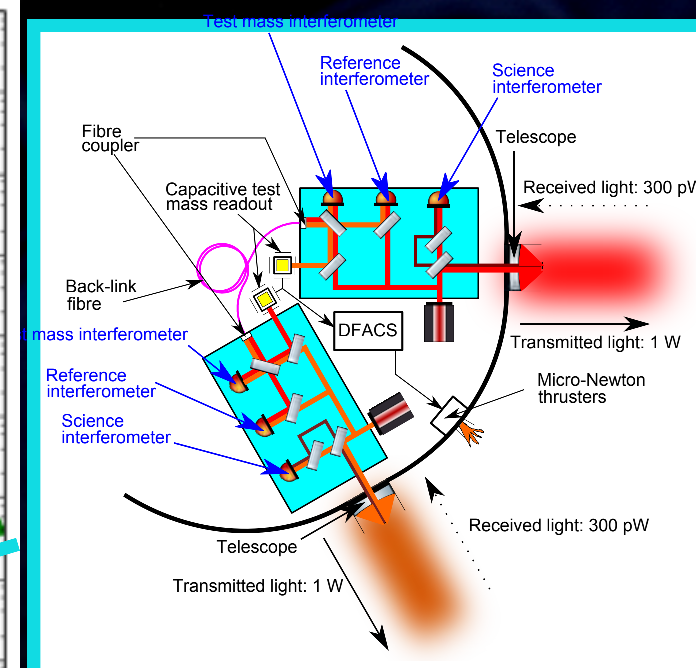
Noises



Response of the detector to GWs







Continuous integration

developer

branch develop

User

- last version
- documentation

light virtual machine with codes installed: ready to use

docker

Machine LISADPC

jenkins: compile test

documentation

performance quality

interface web

git repository

branch develop

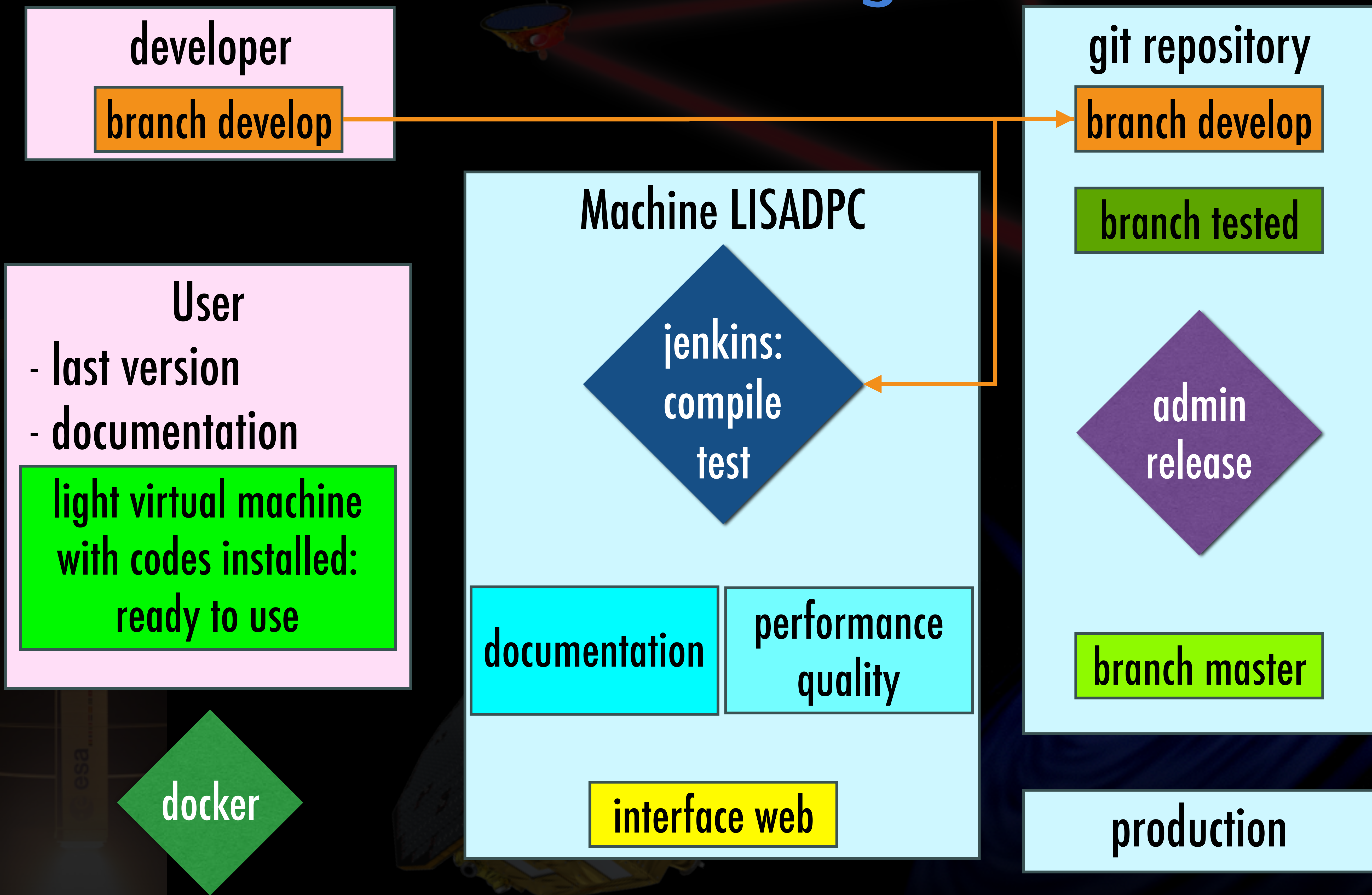
branch tested

admin release

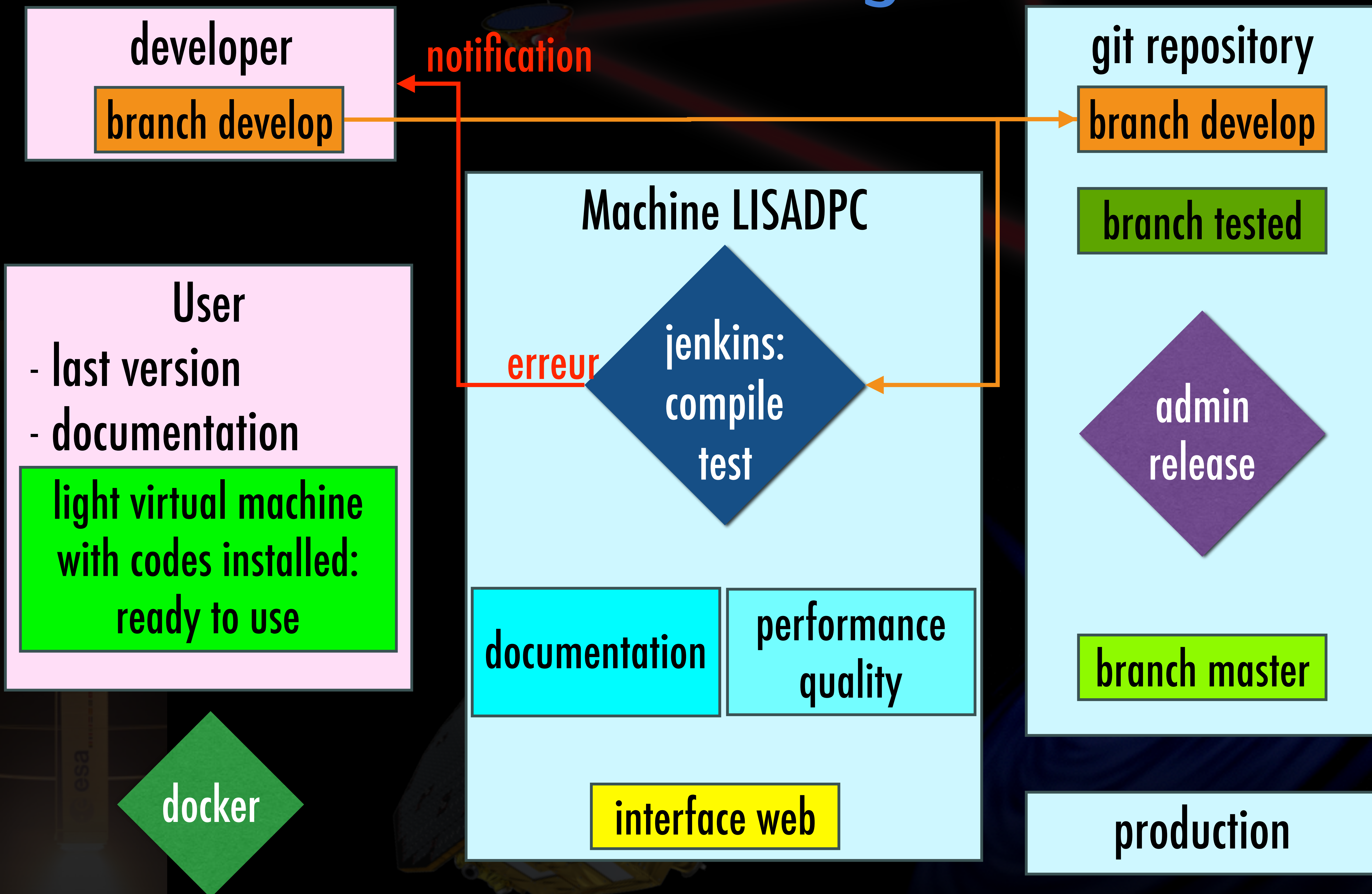
branch master

production

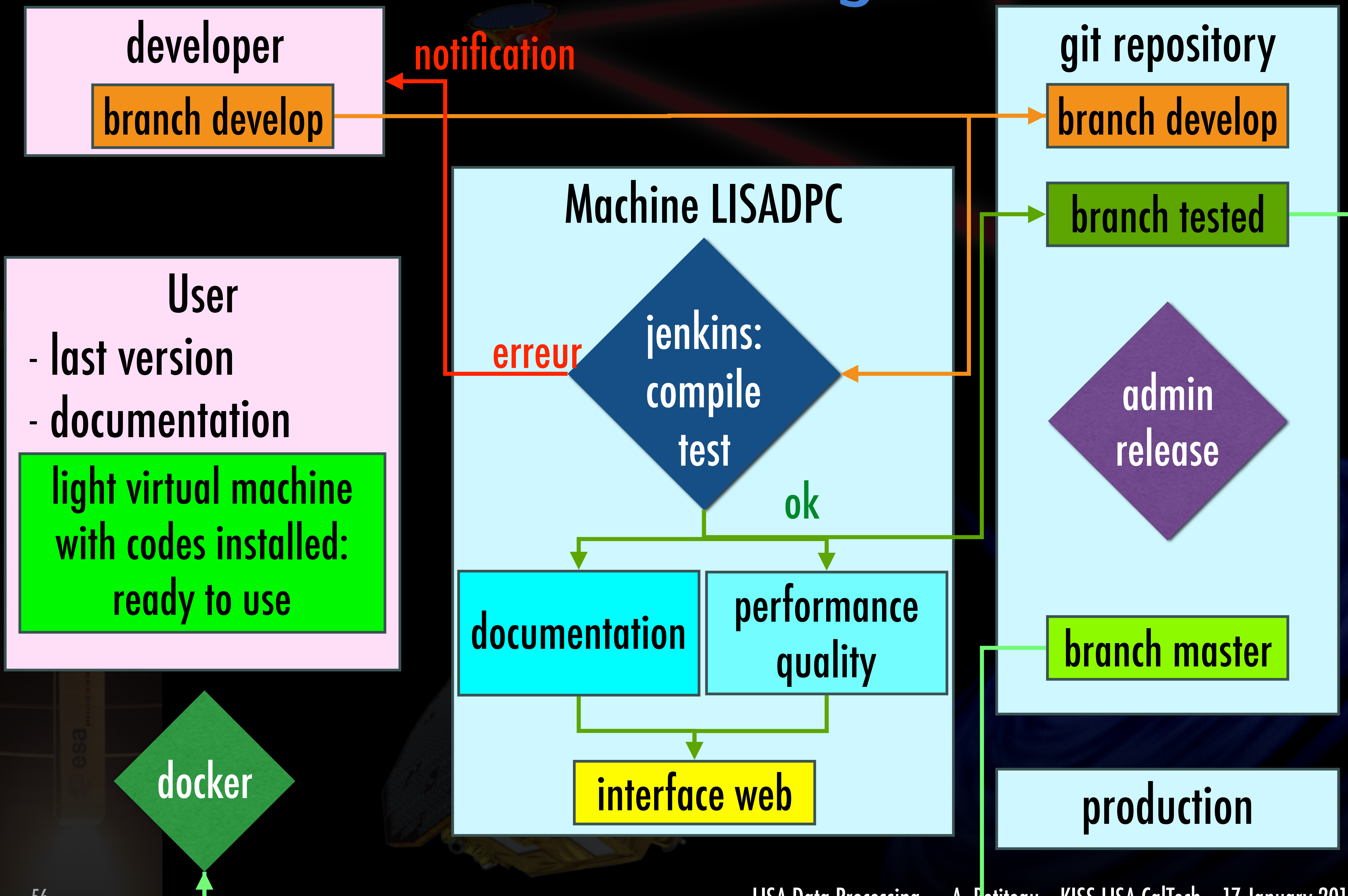
Continuous integration



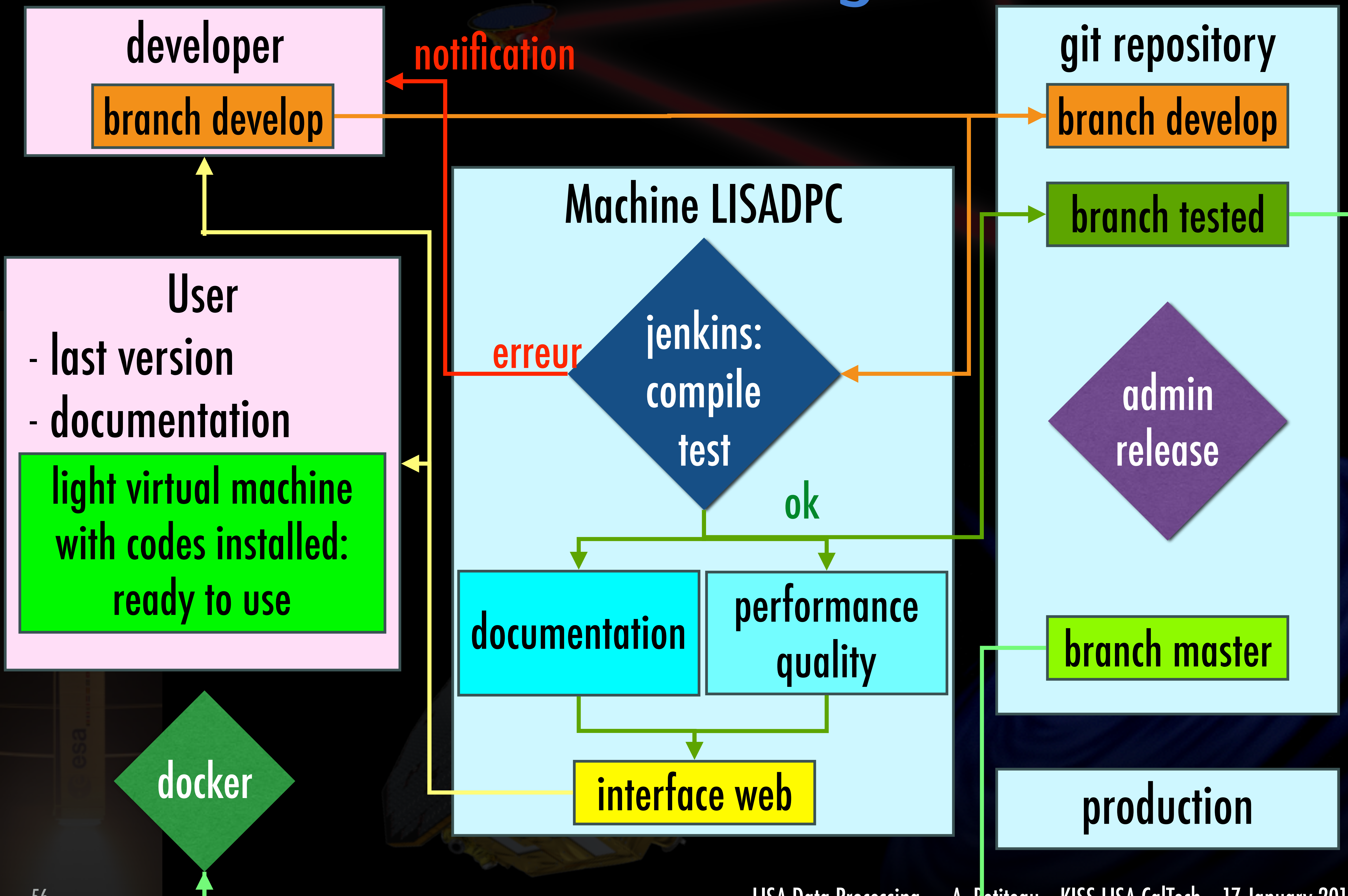
Continuous integration



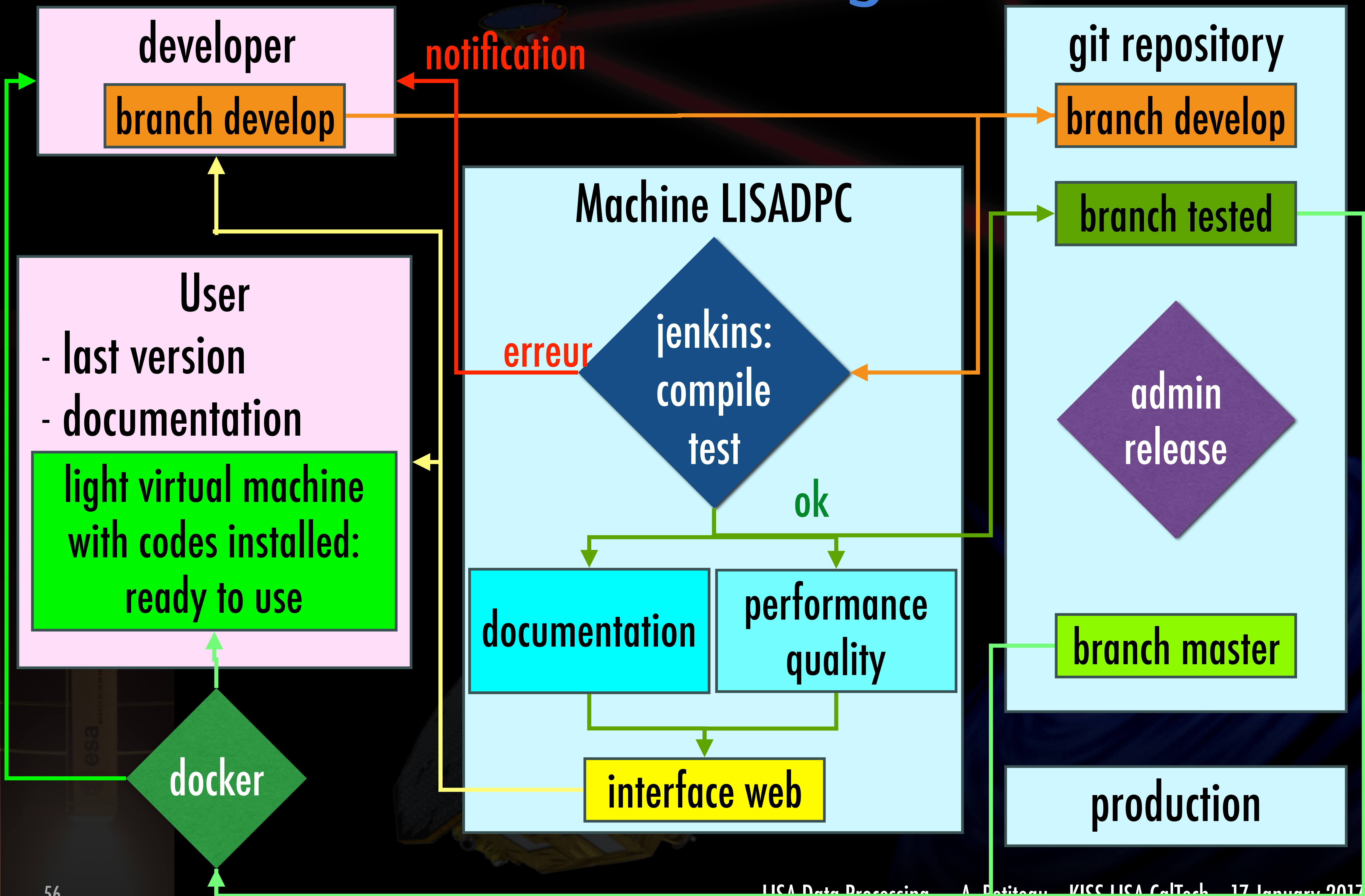
Continuous integration



Continuous integration



Continuous integration



Continuous integration