DPC and Data Centers in Europe and the US
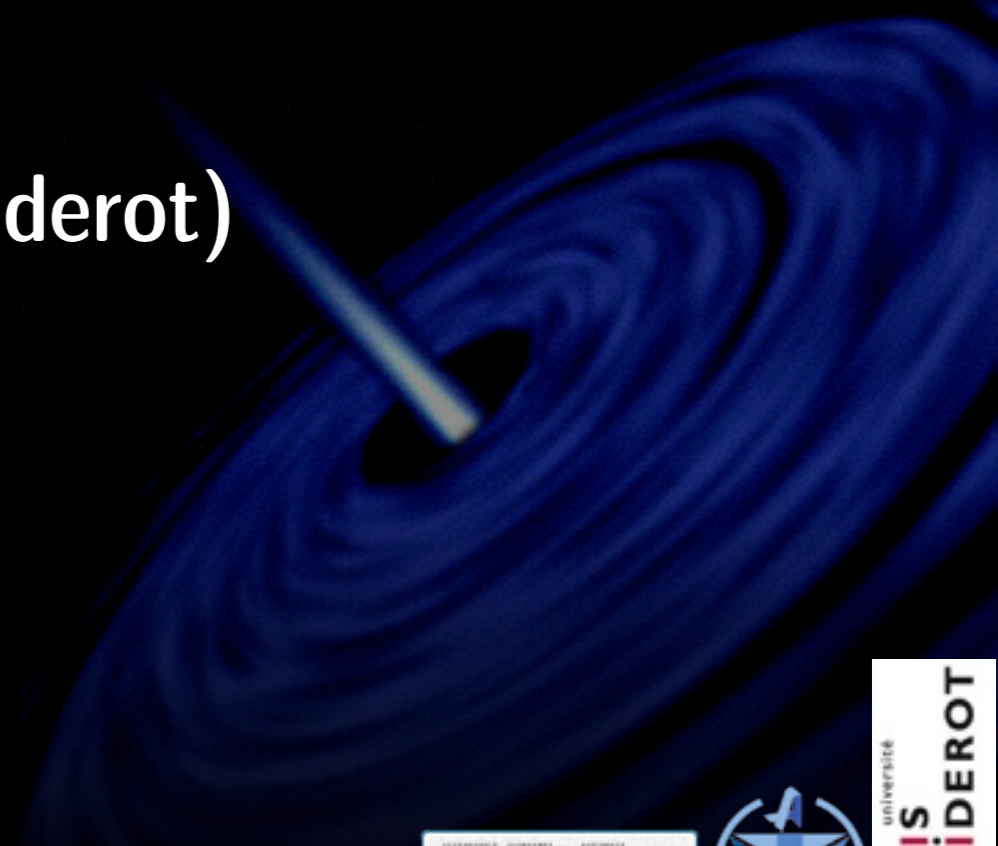
Antoine Petiteau

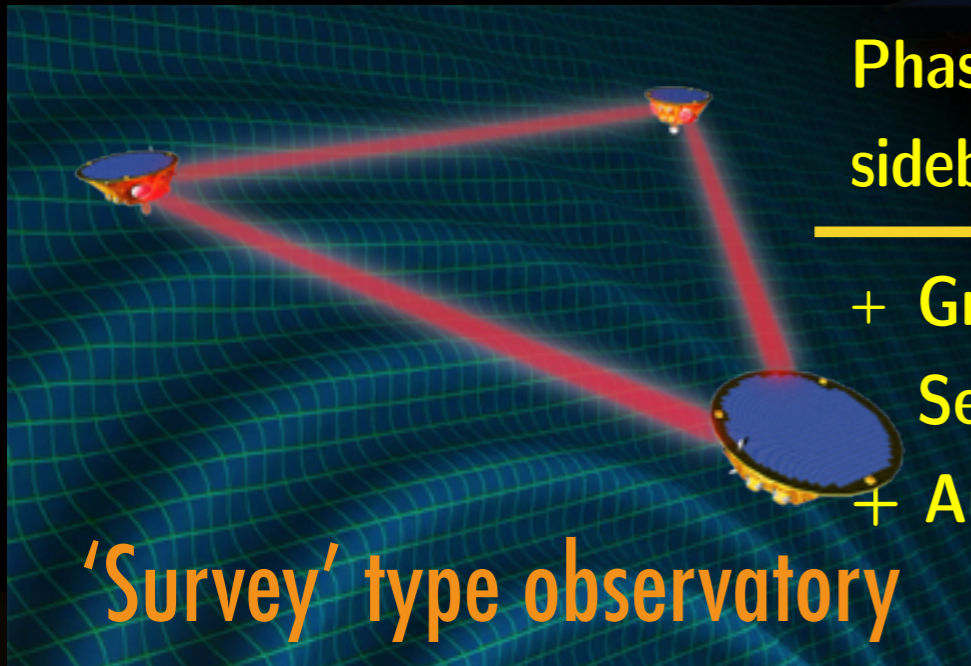**(APC - Université Paris - Diderot)**

SST #3 / NLST

Baltimore

28-29th August 2018

1

LISA data flow

Phasemeters (carrier, sidebands, distance)

+ Gravitational Reference Sensor

+ Auxiliary channels

'Survey' type observatory

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

Calibration corrections

Resynchronisation (clock)

Time-Delay Interferometry reduction of laser noise

3 TDI channels with 2 "~independent"

Data Analysis of GWs

Catalogs of GWs sources with their waveforms

LISA data flow

Mission Operation Centre

Phasemeters (carrier, ...stance)

+ Gravitational Reference Sensor

+ Auxiliary channels

'Survey' type observatory

Science Operation Centre

GW sources

- 6×10^7 galactic binaries
- 10-100/year SMBHBs
- 10-1000/year EMRIs
- larae number of Stellar Oriain

Distributed Data Processing Centre

- unknown sources

L0

Calibrations corrections

Resynchronisation (clock)

Time-Delay Interferometry reduction of laser noise
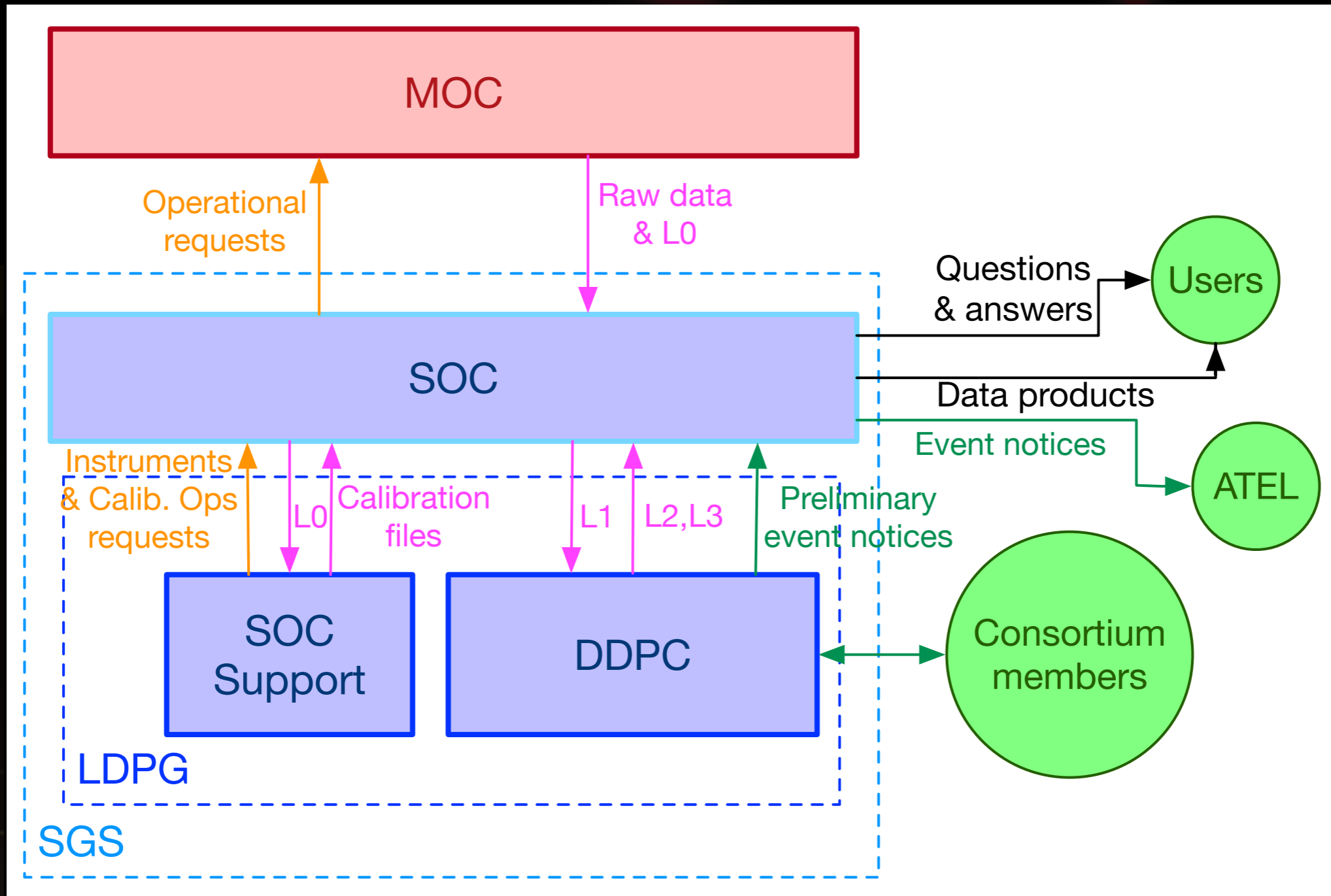
L1 — 3 TDI channels with 2 "~independents"

L2 — Data Analysis of GWs

L3 — Catalogs of GWs sources with their waveform

LISA Ground Segment

LISA Data Processing

▸ **First** data and analysis of this kind + potential **unknown** sources
   => Keep **flexibility** + **continuous evolution**

▸ **Permanent** sources + **transient** sources + continuous evolution of codes, i.e. full **reprocessing** phase
   => **fluctuations** of the computational charge: **mixed infrastructure** (standard clusters + on demand, i.e. Cloud)

▸ Data analysis **challenges**: large number of mixed sources + no direct calibration of instrument
   => need to start the studies **now**!
   • Simulations
   • LISA Data Challenge

Current vision of the DDPC

‣ **DDPC**: unique entity responsable for the data processing (driving, integration of software blocks, ...)

‣ **DDPC** 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 DDPC software « suite » can run on "any" DCC.
  - Software: codes (DA & Simu.) + services (LDAP, wiki, database) + OS.

‣ First solutions:
  - Separation of hardware and software: light virtualization, ...
  - Collaborative development: continuous integration, ...
  - Fluctuations of computing load: hybrids cluster/cloud

Common system: dev./prod.

▸ **A common system:**

- **short cycle between development & production**

- **distributed hardware on DCCs (Data Computing Centres)**

- **cloud compatibility**
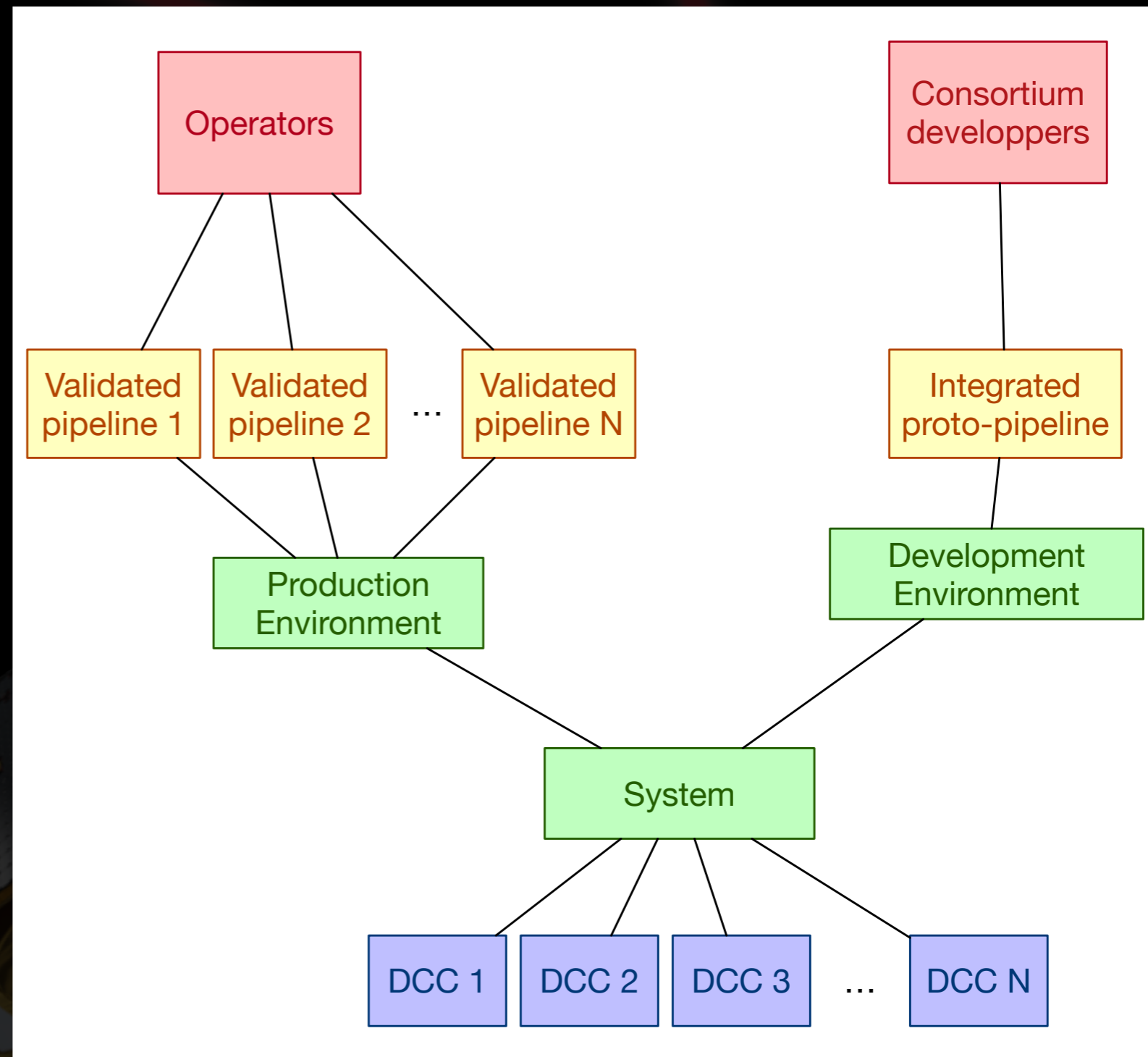
From L0 to L1

▸ Input (L0): "raw" data from the MOC

▸ Output (L1): TDI + all data "cleaned"

▸ Responsibility: SOC (ESA)

▸ With Consortium support => SOC Support group

▸ Activities / Challenges:

- Processing —————>
  - Calibration
  - Clock synchronisation
  - Ranging (estimation of delays)
  - TDI

- Hardware monitoring

- Quick-look of instrument data

- ...

From L1 to L3

- Inputs: TDI + all data "cleaned"

- Outputs: final science products (catalogs, ...)

- Responsibility: Consortium => DDPC

- Activities:

  - Data analysis pipelines and simulation:
    - Prepare, Implement, Operate;
    - Support (LSG, SimWG, LDC) design and prototyping;

  - Define, coordinate and implement software framework and management structure for data and products

  - Coordinate and operate the DCCs

  - Define, implement and maintain dev. and op. environment

Data analysis & simulations

▸ **Simulations:**

- Simulations at different scales: micro-sec to years in reasonable time
- Coherently simulate control loops, integrate discretization/ interpolation, precisions, …

▸ **Data pre-processing: clock, ranging, TDI**

▸ **Data processing: extracting science**

- For the matched filtering: optimisation of likelihood computation, variety of samplers, possibly large number of parameters, evolving number of parameters, …
- Orchestration of multiple pipelines in parallel
- Keep track of all produced data
- Incremental data: new data to integrate every day
- Fast pipeline for alerts, …

LISA Data Processing Group
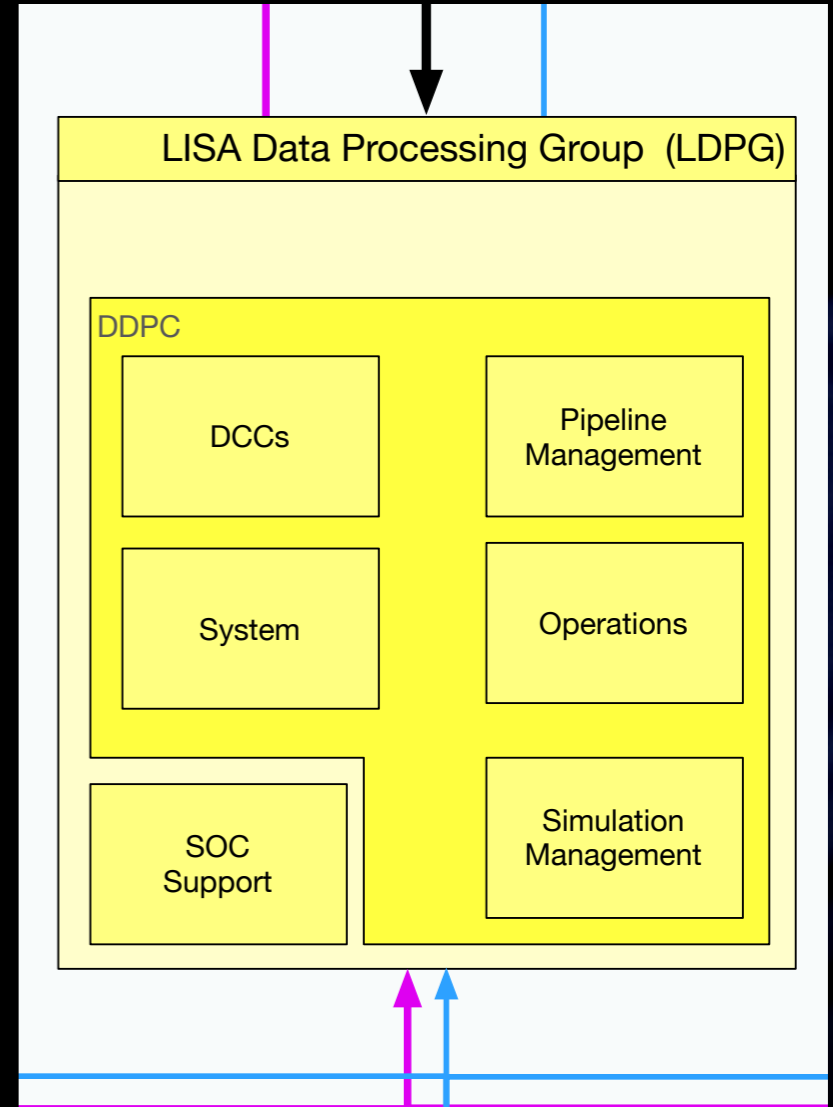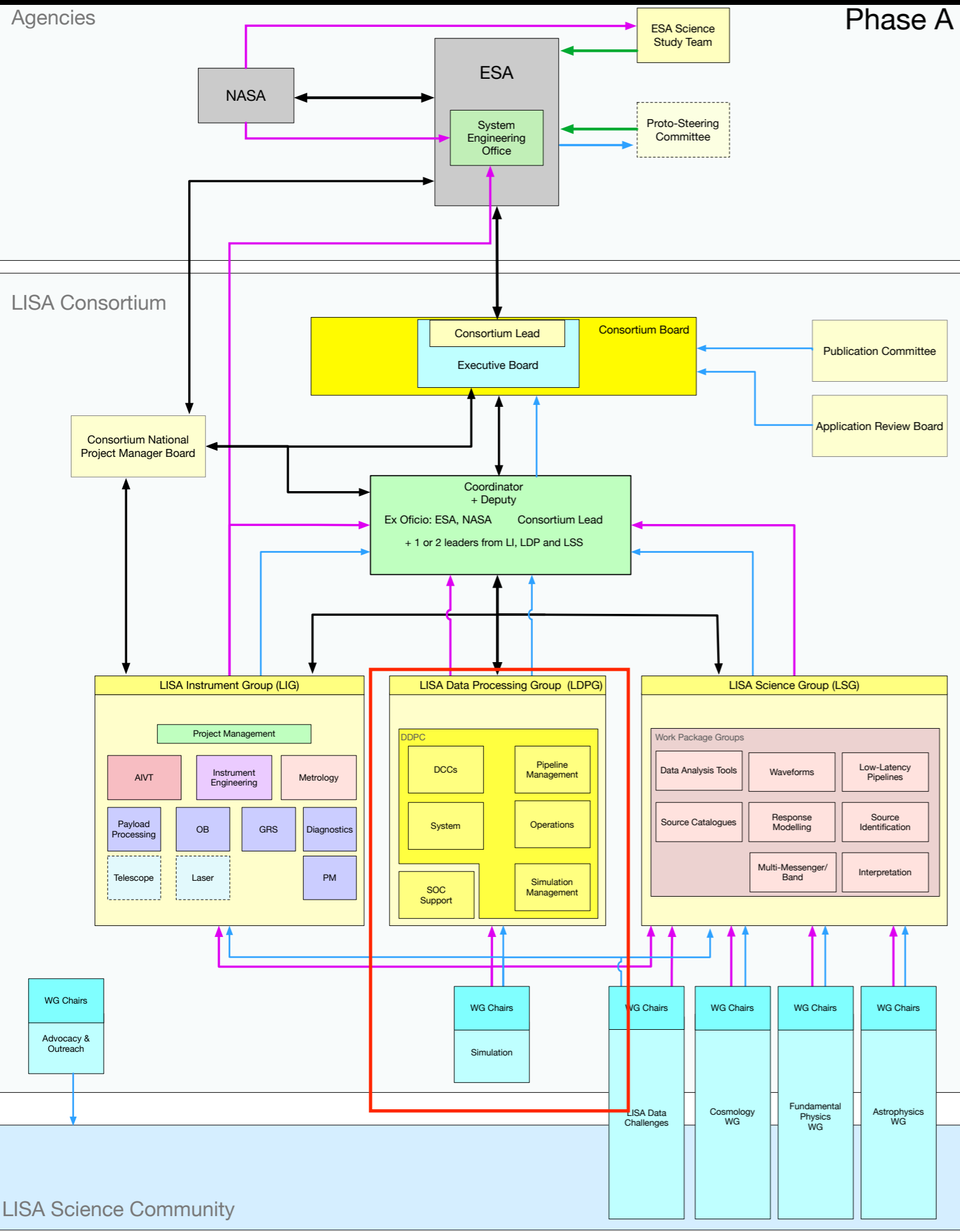
LDPG Organisation
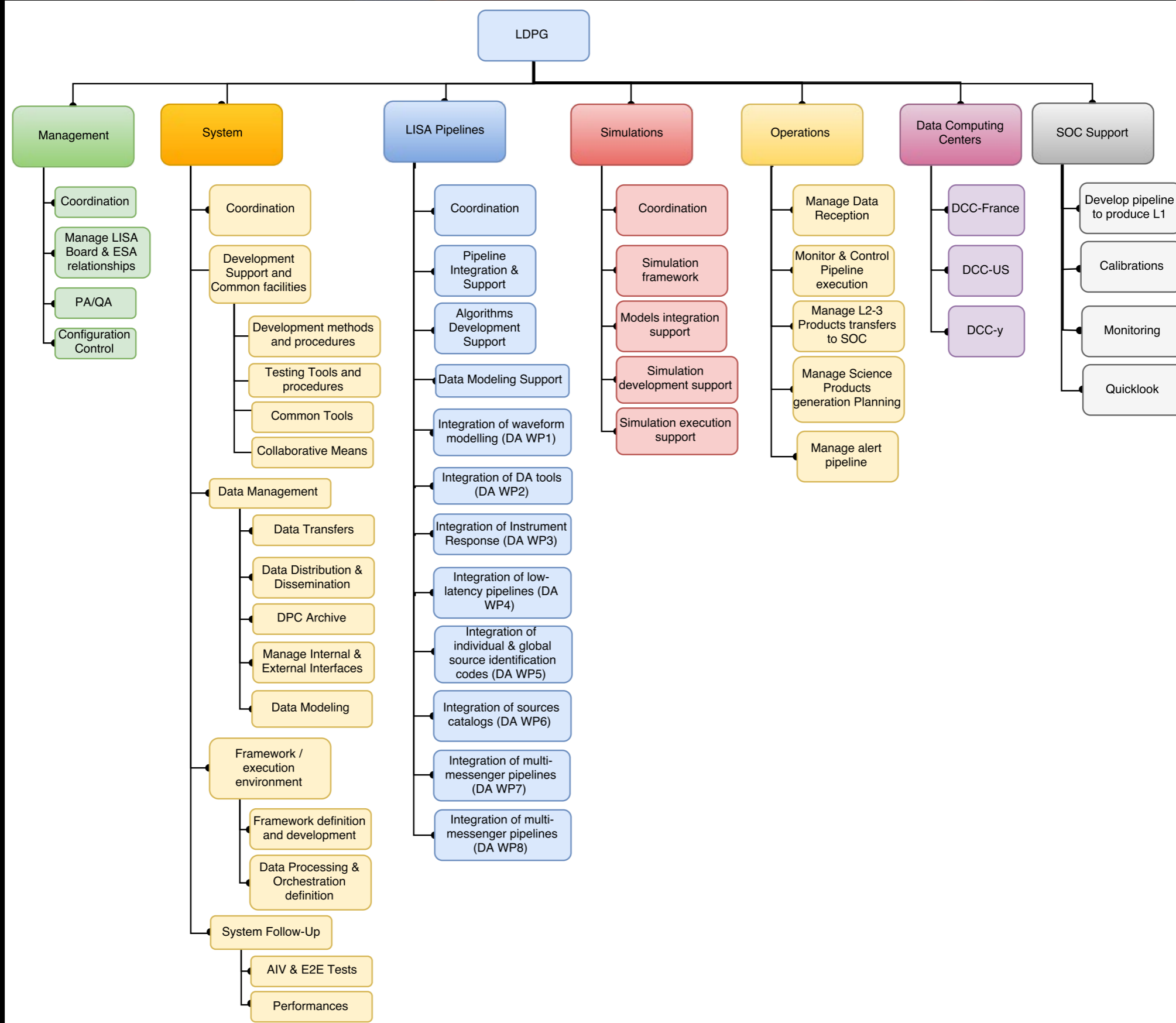
Tasks for phase A

1. **Definition** and detailed **design** of the Ground Segment part delivered by the **Consortium** => start: **QUESTIONNAIRE** ... => one of the goal will be start organizing for the **SIPs**

2. **Support/Contribute** to the definition of LISA Ground Segment with **ESA** => contribution to documents (SOAD & SMP) then later SIRD and SIPs.

3. **Prototyping** (benchmarking)

4. **Support** LISA activities (Consortium & ESA): LDC, SimWG, decadal white papers, ...

5. Provide **services** to the Consortium: Doc. Management, repositories, wiki, computing facilities

Proto-DPC: basics

[Poster C. Cavet]

▸ **Development environment**: in production

- **Collaborative work**, **reproducibility** of a **rapidly evolving** & composite DA pipeline; **Keep control** of performance, precision, readability, etc

- Use existing standard tools (**version control**, **Continuous Integration**, **Docker**)

▸ **Data basis & data model**: in R&D

- **Data sharing**, a lot of information (search engine, DB request, tree view);

  - Context: Not very big data volume for data itself but large number of sub-products, simulations, ... => **LDC**, simulations, LPF data

▸ **Execution environment**: in R&D (singularity, ...)

Support LISA Consortium today

▸ **Simulation**:
- **LISACode** and **LISANode**: git with continuous integration, docker image, singularity, documentation, ...

  => realistic data used for ex for performance, pre-processing, ...

▸ **Exchange**: LDC database, Virtual Machine on demand

▸ **IT**: Repositories, Document Management System, wikis

▸ **Coming soon**:
- **Jupyter hub** available soon: share scripts
- **Singularity hub**: share image containing all LDC tools
- **Computing** facilities (prototyping DCCs)
- **Integration** of LDC DA methods submitted with responses

Contribution to the DDPC

▸ DDPC is a deliverable under **French responsibility**

▸ Will be **designed and developed by the Consortium** with contribution from all interested partners (Germany, Spain, UK, US, Romania, Italy, Switzerland, …)

▸ US is one of the main partner now: contribution as a member of the LISA Consortium

- **Take part in the LDPG**
  - Organisation
  - Design of the DPC system
  - Will be responsible of some LDPG work packages
- **Provide one or many Data Computing Center(s)**

Conclusion

▸ DPC & LDPG started :

- Proto-DPC in place and supporting Consortium activities
- Active LDPG-core & system team: questionnaire, DDPC Def., ...

▸ Basis of the concept set:

- flexibility, evolution short circle;
- fluctuations of computing, common framework running on DCCs
- Common system

▸ US:

- Fully included in the LDPG (part of Consortium)
- DCC(s)

▸ Tasks for phase A: design, support, prototyping

Thank you