WP2 : Exascale in situ data processing

Laurent Colombet[1]
Yushan Wang[2]

Gabriel Antoniu[3]
Julien Bigot[2]
Joshua-Charles Bowden[3]
Silvina Caino-Lores[3]
Christophe Denoual[1]
Julien Jaeger[1]

Benoît Martin[2]
François-Xavier Mordant[2]
Etienne Ndamlabin[3]
Pierre-Antoine Raclius[2]
Bruno Raffin[4]

1: CEA/DAM 2: CEA/MDLS 3: Inria Rennes 4: Inria Grenoble

Objectives

Design and implement the software building blocks required to support in-situ execution of data processing at Exascale, and to integrate these building blocks in the libraries deliverables of Exa-DoST.

Compute to bandwidth ratio is increasing exponentially

Simulations can not write data to disk freely
- need to reduce data volume
- need in-situ analytics
- need dedicated resources for asynchronous transfer and analytics

Challenge and main objectives for WP2
- For a wide range of scientific applications : enable in situ visualization, processing and analytics at **Exascale**
- Provide the infrastructure to build complex workflows at the HPC-center level coupling traditional HPC with analytics & others

Human resources funded

- Ph.D thesis at **Inria Rennes**, co-directed with **CEA/DAM** + co-advised with **MdIS**

  - Design of a dynamic management model for in situ and in transit analyses and implementation in Damaris

  - Arthur Jaquard, to start: **october 2024**. (co-advised by Gabriel, Laurent, Julien, Silvina)

- Ph.D. thesis between **MdIS** & **CEA/DAM** (*50% funded by NumPEx, 50% by CEA PTC*).

  - Elastic load balancing between exascale simulation and in situ analysis

  - Target date for thesis start : September 2025 (co-advised by Julien, Laurent, *et al.*)

- Ph.D. thesis between **MdIS** & ? (*50% funded by NumPEx, 50% by ?*).

  - Co-scheduling in-process & in-transit to limit data transfers for PDI and Deisa

  - Target start : october 2025.

- Ph.D. thesis at **INRIA Grenoble** (*100% funded by NumPEx),*

  - Thesis topic to be defined, Target start : october 2025?. (Bruno *et al.*)

Human resources funded

Research engineers (long-term contract) :

- Engineer at **MdIS**: Benoît Martin

Research engineers (fixed-term contracts) :

- Damaris engineer at **Inria Rennes**: Etienne Ndamlabin
- PDI engineer at **MdIS**:
  - Jacques Morice (to come, 10/2024)
 - Julian Auriac (to come 10/2024)

Other :

- PDI Internship at **MdIS**: François-Xavier Mordant (finished 04/2023 - 10/2023)
- Benchmark development trainee at **MdIS**: Pierre-Antoine Raclius (finished 09/2023 - 07/2024)

Work subjects proposed from the project kick-off

- PhD thesis on Triggers & GPU handling w. Damaris & Coddex (start)

- Work on in situ benchmark design combining PDI, Deisa (dev OK)

- Co-scheduling in process & in transit to limit data transfers (identified, subject ready)

- Load balancing & elasticity for analysis (identified, subject ready)

- Industrialization & Modularization (wip)

Current achievements

- **Gantt chart** for tasks' overview and collaboration coordination

- Development of an **in-situ benchmark**, successfully deployed on Adastra and Jean-Zay (w. P-A + Benoît)

- Creation of PDI team at MdlS, 2 engineers recruited

- Json plugin for PDI (w. F-X)

- GPU support for PDI (WIP, w. F-X)

- Damaris plugin for PDI (WIP, w. Josh)

- Development of framework **SPOT** (w. Benoît *et al.*)

- Data transfer using Adios2 in Melissa (WIP w. Bruno)
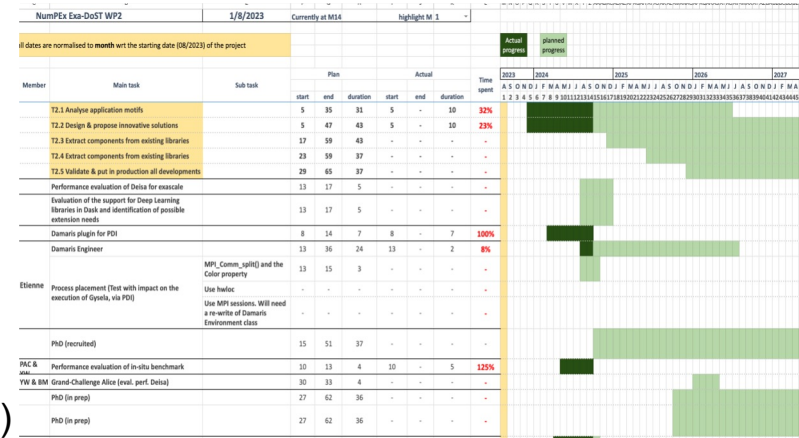
Next steps

- **PDI team @MdIS**
  - Hire one more engineer
  - GPU support on PDI, Deisa
  - Optimization, performance evaluation, and packaging of in-situ benchmark
  - Feedback mechanism in SPOT, new functionalities of PDI
  - Simple data annotation
  - New functional mode of PDI, from "push" to "pull"
  - Damaris & PDI interoperability (with KerData)

- **Arthur's thesis @KerData & CEA/DAM**
  - **Design of a dynamic management model for in situ and in transit analysis and implementation in Damaris**
  - Integration of Damaris in Coddex
  - Support programmed and trigger-based analysis in Damaris

Next steps

- **Thesis**
    - Elastic load balancing between exascale simulation and in situ analysis
    - Co-scheduling in process & in transit to limit data transfers for PDI and Deisa
    - Ideally, start the thesis with an internship

- **DataMove**
    - Data transfer using Adios2 in Melissa

Next steps

- **Grand challenge Alice Recoque, test on exascale (2026)**
    - In-situ and in-transit data analysis
    - PDI, Deisa, Damaris
    - Checkpoint from GPU
    - High frequency data analysis
    - Triggering tasks based on the analysis results
    - Feedback for restarting, changing parameters
    - High frequency write of rare event data
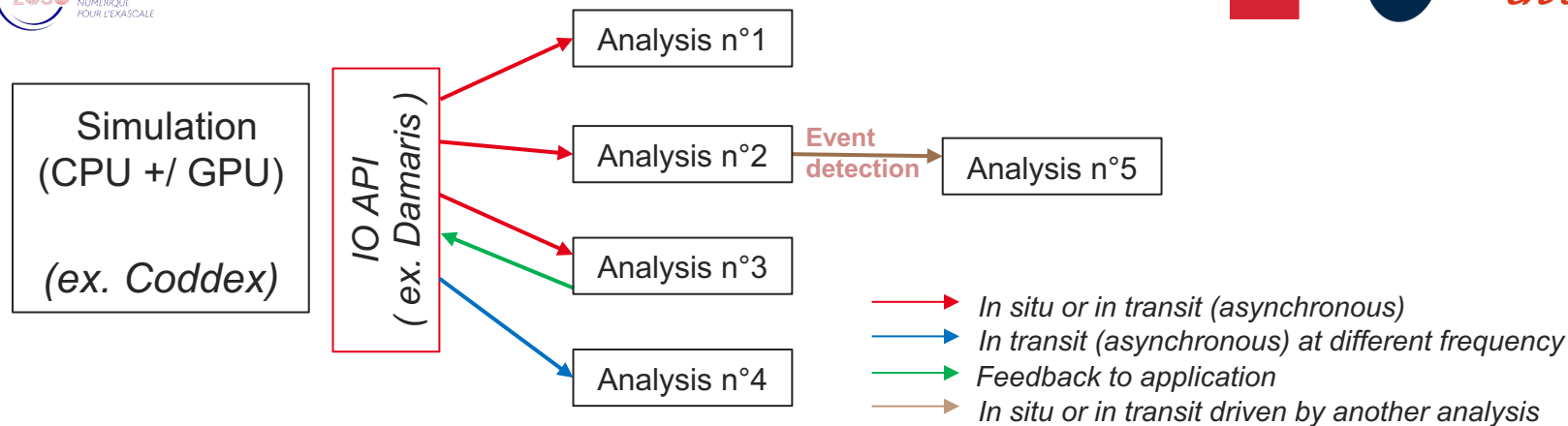
- **Publication and communication**
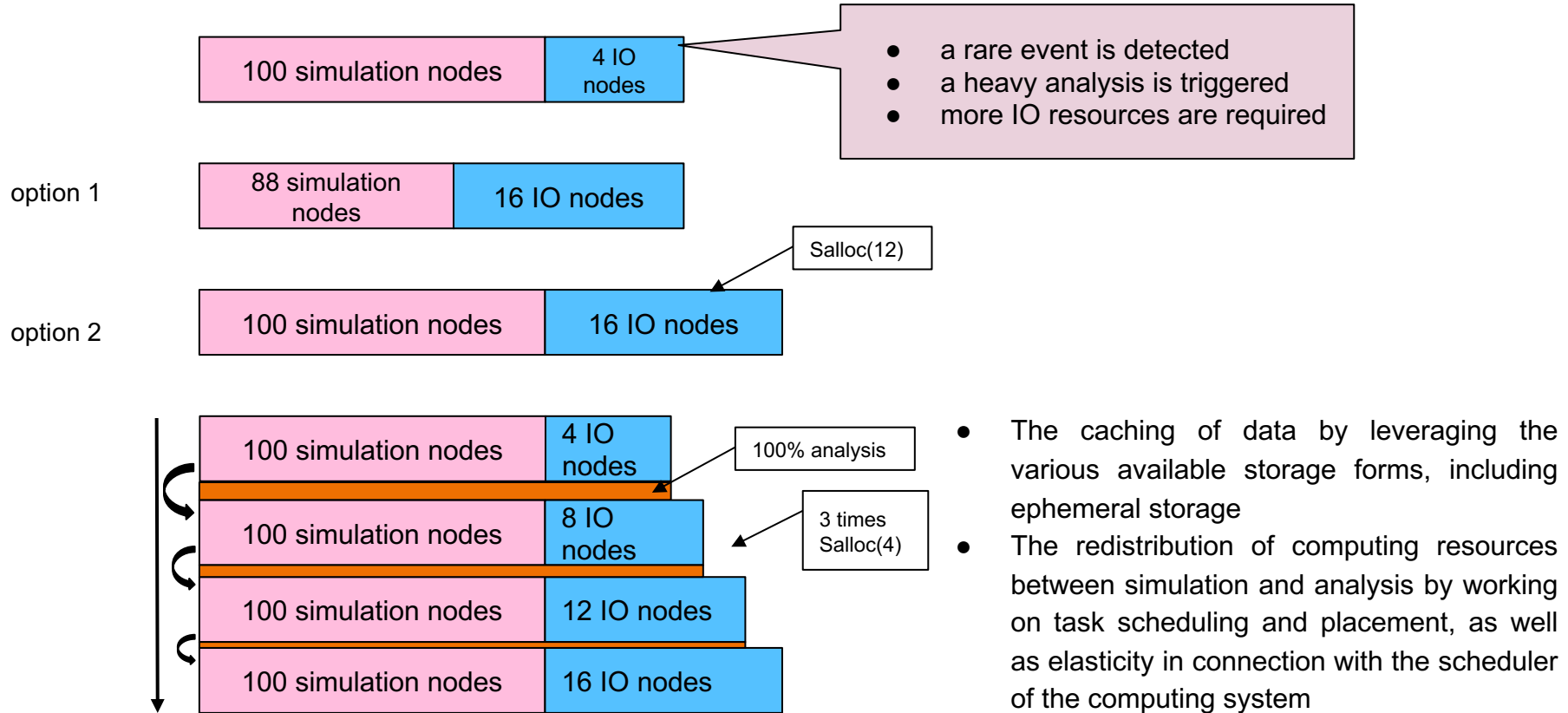
Thank you for your attention!

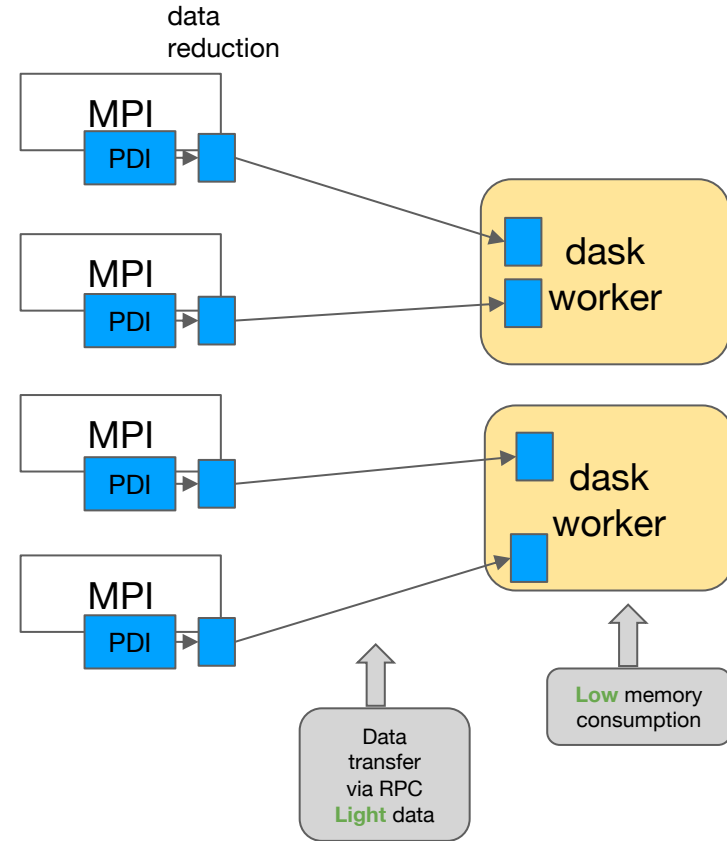in NumPEx
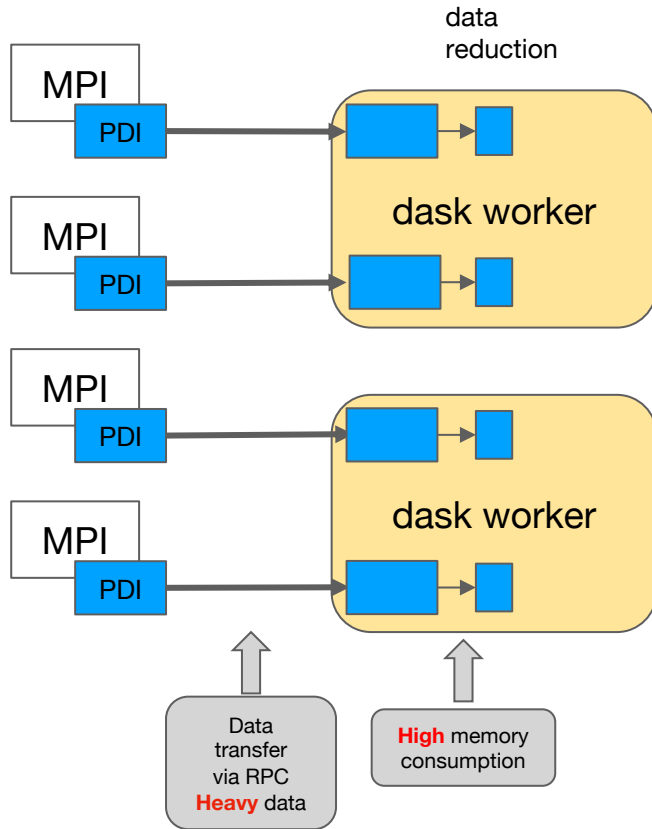
Laurent Colombet[1]
Yushan Wang[2]

Gabriel Antoniu[3]
Julien Bigot[2]
Joshua-Charles Bowden[3]
Silvina Caino-Lores[3]
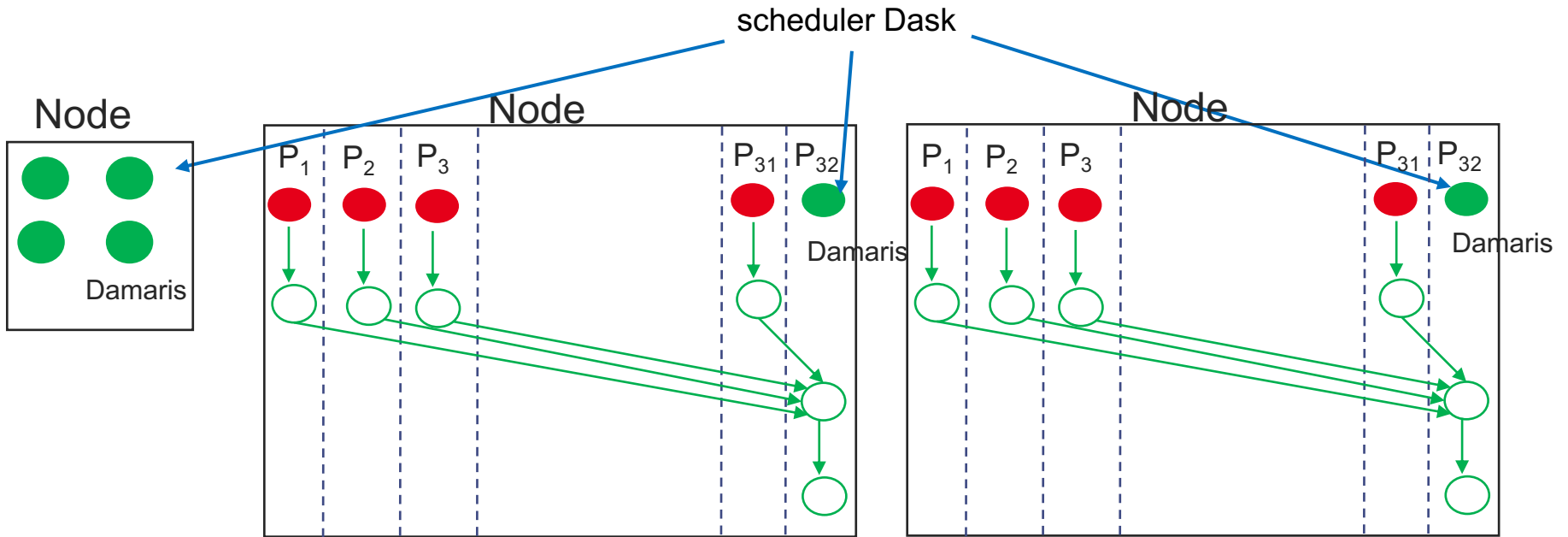Christophe Denoual[1]
Julien Jaeger[1]

Benoît Martin[2]
François-Xavier Mordant[2]
Etienne Ndamlabin[3]
Pierre-Antoine Raclius[2]
Bruno Raffin[4]

1: CEA/DAM 2: CEA/MDLS 3: Inria Rennes 4: Inria Grenoble

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

FRANCE 2030
PROGRAMME DE RECHERCHE
NUMÉRIQUE
POUR L'EXASCALE

cea   cnrs   Inria

- An analysis can either inform the code that will trigger a new analysis, or an analysis, based on a physical criterion, can trigger a new analysis by itself.
- An analysis may request an increase in the frequency of an analysis call that is already present in the contract.
- What happens if the 'sent and compacted' data needs to be modified with the arrival of analysis 5? A rollback or, if planned, a local iteration via the PDI API.
- Analysis 5 could also be a request for a rollback of the code with a different time step.

100 simulation nodes | 4 IO nodes

- a rare event is detected
- a heavy analysis is triggered
- more IO resources are required

option 1

88 simulation nodes | 16 IO nodes

Salloc(12)

option 2

100 simulation nodes | 16 IO nodes

100 simulation nodes | 4 IO nodes

100% analysis

100 simulation nodes | 8 IO nodes

3 times Salloc(4)

100 simulation nodes | 12 IO nodes

100 simulation nodes | 16 IO nodes

- The caching of data by leveraging the various available storage forms, including ephemeral storage
- The redistribution of computing resources between simulation and analysis by working on task scheduling and placement, as well as elasticity in connection with the scheduler of the computing system

scheduler Dask

Node

Node

Node

P₁ P₂ P₃ P₃₁ P₃₂

Damaris

Damaris

🔴 *Simulation task*   🟢 *Worker (for analysis)*   ⚪ *Analysis task*

SPOT: Simulation Pattern Observation Tool