RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

cea cnrs Inria

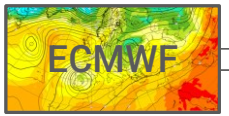
FRANCE 2030
PROGRAMME DE RECHERCHE
NUMÉRIQUE POUR L'EXASCALE

ExaDoST - Work Package 1
Exascale I/O and Data Storage

WP Leaders:
Francieli Boito (Université de Bordeaux) &
François Tessier (Inria Rennes)

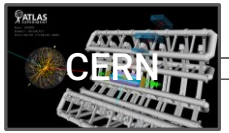
Trends

→ (New) workloads (**AI**, **data analytics**), using **more data** and using it **in new ways**.
→ More data from **sensors** and **scientific instruments**.

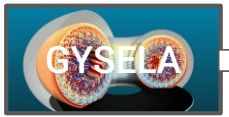


2023: **40 TiB / day**
"Shortly": **180 TiB / day**
"Near future": **700 TiB / day**
20x

SDP:
Input: **10 Tbps**
Output: **700 PB / year**



2021: **240 Gb/s** storage bw
2023: **> 1 EiB** of storage
2027: **2.4 Tb/s** storage bw
~350 PB / year (raw data)
10x



2023: **10s PB** managed per simu
2023: **13.2 TB** of data distributed **simultaneously** into **24'576 files** per checkpoint

2022: **2 PB** dataset processed
2023: **80 PB** generated by a **single job**
2023: **700 PB** storage system on **Frontier** has only a **90 days retention policy**

Trends

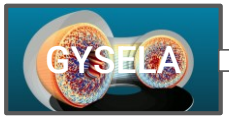
→ (New) workloads (**AI**, **data analytics**), using **more data** and using it **in new ways**.
→ More data from **sensors** and **scientific instruments**.

2023: **40 TiB / day**
"Shortly": **180 TiB / day**
"Near future": **700 TiB / day**
20x

2021: **240 Gb/s** storage bw
2023: **> 1 EiB** of storage
2027: **2.4 Tb/s** storage bw
~350 PB / year (raw data)
10x

2022: **2 PB** dataset processed
2023: **80 PB** generated by a **single job**
2023: **700 PB** storage system on **Frontier** has only a **90 days retention policy**

SDP:
Input: **10 Tbps**
Output: **700 PB / year**



2023: **10s PB** managed per simu
2023: **13.2 TB** of data distributed **simultaneously** into **24'576 files** per checkpoint

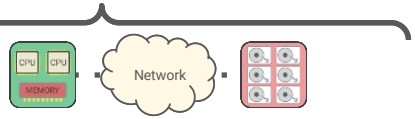
NumPEx demonstrators

Compute-centric to data-centric shift

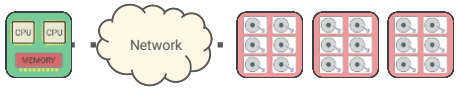
↗ I/O pressure on large-scale storage systems

Trends

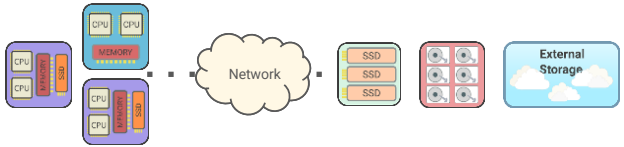
“Traditional” approach



Horizontal scaling



Vertical scaling



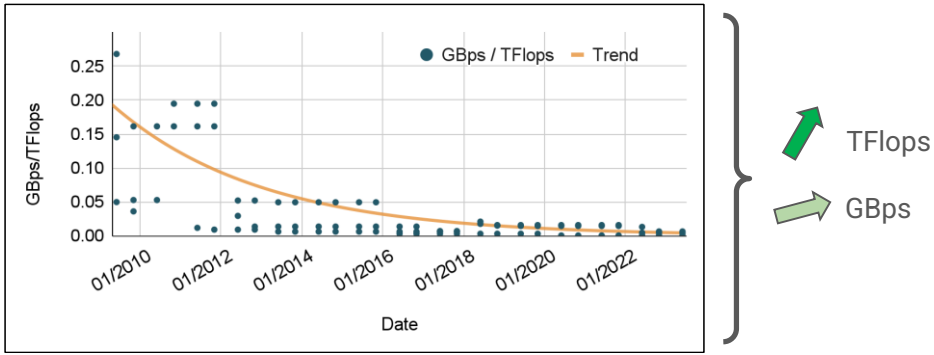
Storage **heterogeneity**

Software stack to manage all of these

(Lustre PFL & DNE, DAOS, Mochi, ADIOS2, Damaris, GekkoFS, BeeGFS, RADOS, CHFS…)

Trends

➞ What about performance then ?
➞ Extracted from the **Top500** over the past 15 years:



WP Objectives

To optimize the I/O performance of applications and workflows, and leverage emerging storage technologies

- **Support the I/O and storage requirements** of complex simulation/analytics/AI workflows running on hybrid HPC (+cloud, +edge) systems
- Promote **efficient I/O resource usage**
- Make the **I/O infrastructure adaptable to applications**' characteristics
- **Scale up modern I/O** and data storage methods and tools
- Develop and integrate **new output formats** for checkpoint/restart and for scientific analysis

Participants

Partner	Type of position	Name of participant
Inria Bordeaux	Researchers	**Francieli Boito,** Luan Teylo, Emmanuel Jeannot, Brice Goglin, Mihail Popov
	Engineers	Mahamat Abdraman
	PhD student	Alexis Bandet (former), Serge Meurrens (from Dec 2025)
	Postdocs	Maxime Gonthier (from Apr 2026), **1 open position**
	Interns	Alexandre Laffargue (2024), Axel Malmgren (2024-2025), Mahamat Abdraman (2024), Iheb Becher (2024), Laora Aimi (2025), **1 open position**
Inria Rennes	Researchers	**François Tessier**, Gabriel Antoniu, Guillaume Pallez, Silvina Caino-Lores, Jakob Luettgau
	Engineers	Julien Monniot (Jan - May 2025)
	PhD student	Théo Jolivel (+ CEA)
	Interns	Ugo Thay, Remy Chiv, **1 open position**
CEA - Maison de la simulation	Researchers/engineers	Julien Bigot, Yushan Wang, **1 open position**
CEA DAM	Researchers	Philippe Deniel, Thomas Leibovici, Arnaud Durocher, Maxime Delorme
DDN	Researchers	Jean-Thomas Acquaviva
	PhD student	Méline Trochon (+ Inria Bordeaux, + Inria Rennes)

Summary

Datasets

- I/O performance data @ Zenodo
- 1 I/O traces repository

Scientific Dissemination

- 4 conference papers: IPDPS'26 (under review), IPDPS'25, IPDPS'24, Euro-Par'24, HiPC'24
- 4 workshop papers: PDSW'24, ESSA'25 (2 papers)
- 1 pre-print @ HAL
- 3 internship reports
- Multiple talks: NHR'25, ESSA'25, COMPAS'24, JLESC, Per3S, …

External Collaborations

- ECLAT Joint-Laboratory
- MeerKAT, South Africa
- University of Honolulu, HI, USA
- University of Darmstadt, Germany
- LNCC, Brazil

Project's deliverables

- 2 reports submitted

Software production

- FIVES
- IOPS
- Interference Simulator
- MOSAIC
- TOTO

Poster

Focus

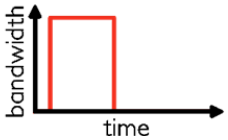
Focus 1: MOSAIC

MOSAIC: I/O Pattern Categorizer

Automatic Categorization of I/O Patterns Based on 3 Pillars:

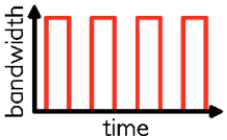
Temporality

When are I/O performed during the execution of an application?



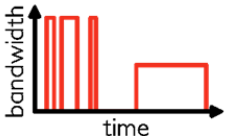
Periodicity

Are I/O operations repeated over time? Are some files frequently accessed?

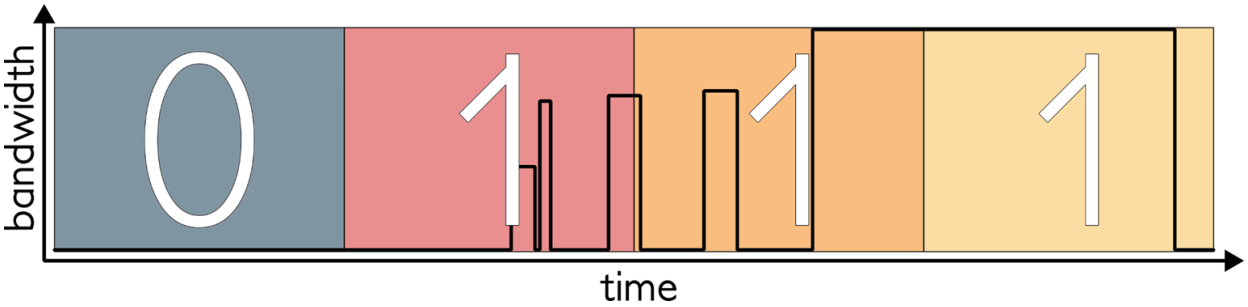


Metadata Load

What is the impact of the execution of an application on the metadata servers?



MOSAIC: Example of a Categorization

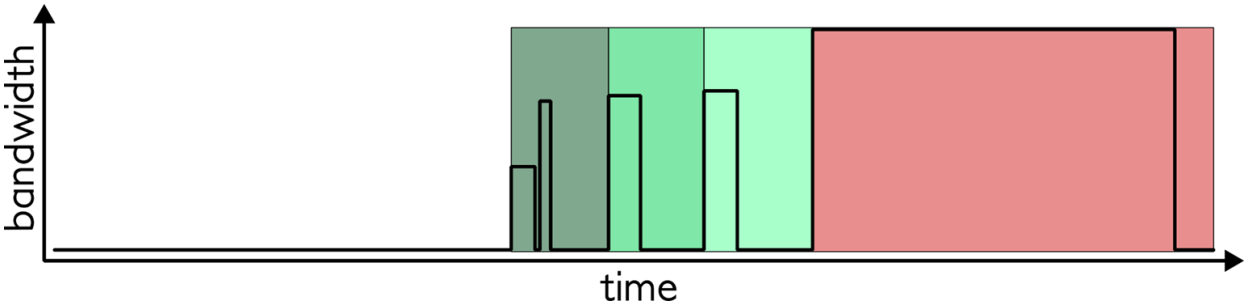


example of the I/O write activity of an hypothetical application

Assigned Classes

write_0111	➜	**no** write **activity** during the **first quarter**
write_periodic	➜	**some operations** are **repeated** over time
metadata_high_spikes	➜	executions leads to **multiple metadata** request **spikes**

MOSAIC: Example of a Categorization



example of the I/O write activity of an hypothetical application

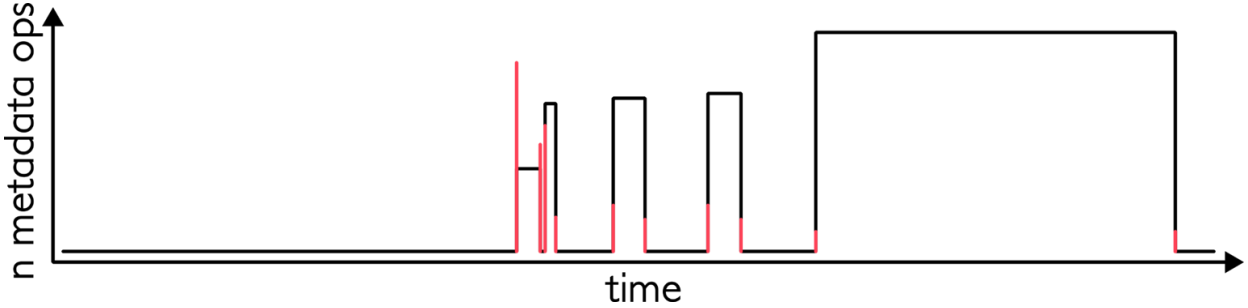
Assigned Classes

write_0111 → **no** write **activity** during the **first quarter**

write_periodic → **some operations** are **repeated** over time

metadata_high_spikes → executions leads to **multiple metadata** request **spikes**

MOSAIC: Example of a Categorization



example of the I/O write activity of an hypothetical application

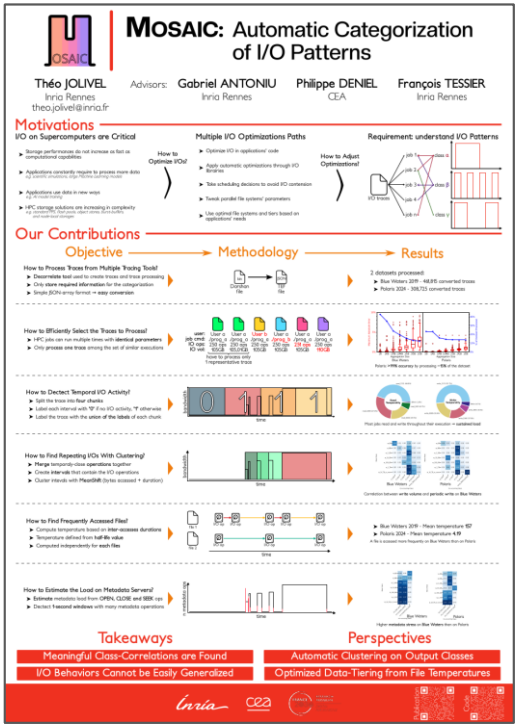
Assigned Classes

write_0111 → **no** write **activity** during the **first quarter**
write_periodic → **some operations** are **repeated** over time
metadata_high_spikes → executions leads to **multiple metadata** request **spikes**

MOSAIC: Pattern-Driven I/O Optimizations

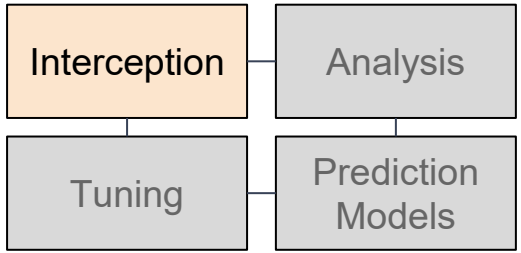
Unsupervised Clustering of Patterns	find **patterns frequently assigned together**
Study of Dataset Biases	analyze the **variations** of I/O **patterns** from **one machine to another**
Scheduling Optimizations	explore the use of **I/O patterns** to **guide scheduling decisions**
Automatic File Storage Tiering	**store files** on different **storage tiers** according to their **access frequencies** (**periodicity & file temperature**)

MOSAIC: Come See our Poster!



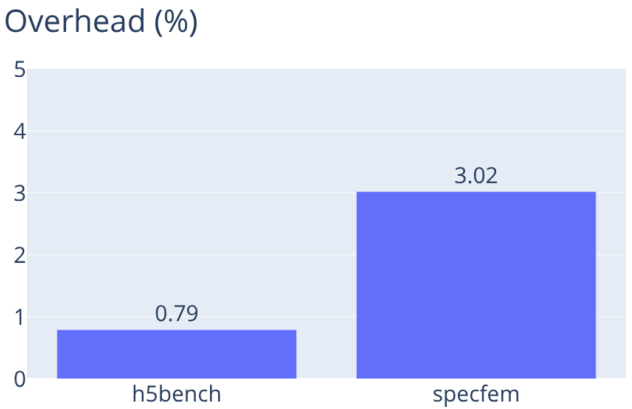
Focus 2: TOTO

Transparent and Online Tool for I/O



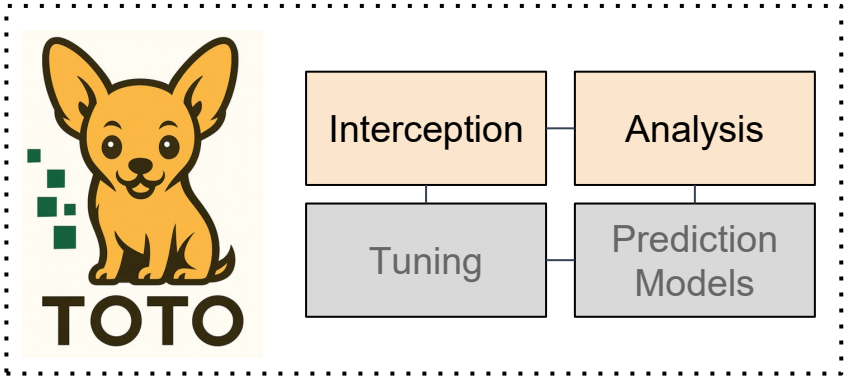
- TOTO intercepts POSIX I/O calls from each MPI rank

with LD_PRELOAD, no modification (or recompilation) required!

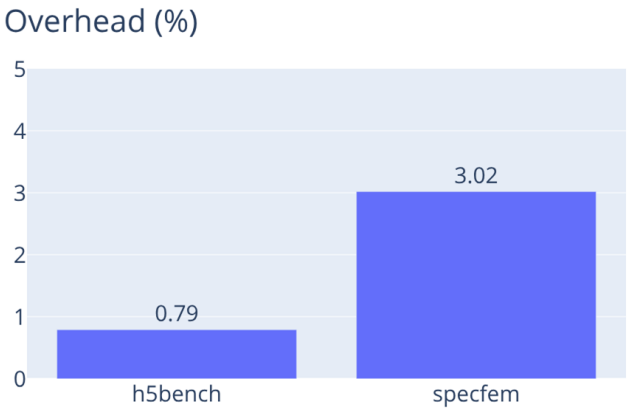


Francieli Boito, Luan Teylo, Mihail Popov, Laora Aimi, Alexis Bandet, Guillaume Pallez. ***TOTO: A Tool for Transparent I/O Tuning for HPC Applications****, to appear*

Transparent and Online Tool for I/O

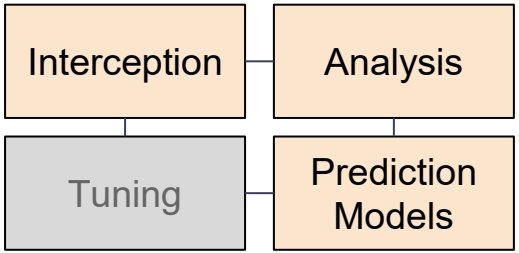


- TOTO intercepts POSIX I/O calls from each MPI rank
- Periodically, all ranks share information and characterize the current access pattern



Francieli Boito, Luan Teylo, Mihail Popov, Laora Aimi, Alexis Bandet, Guillaume Pallez. ***TOTO: A Tool for Transparent I/O Tuning for HPC Applications****, to appear*

Transparent and Online Tool for I/O

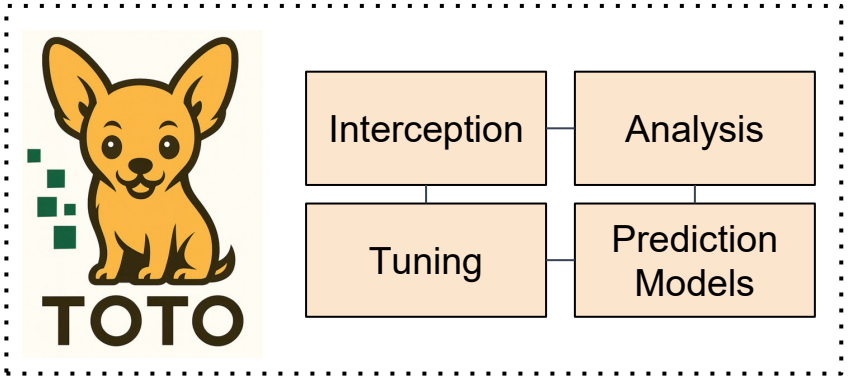


- TOTO intercepts POSIX I/O calls from each MPI rank
- Periodically, all ranks share information and characterize the current access pattern
- Prediction models can be used to support decisions
  - e.g. to choose stripe count

We were able to train a model with only 10% of the previously-required data

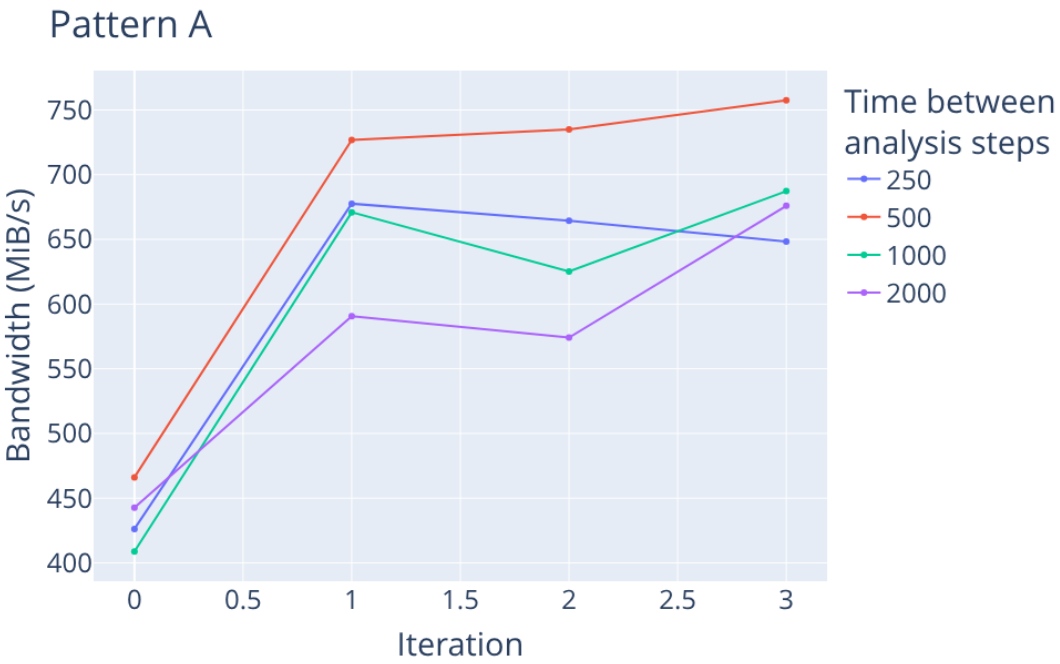
Francieli Boito, Luan Teylo, Mihail Popov, Laora Aimi, Alexis Bandet, Guillaume Pallez. ***TOTO: A Tool for Transparent I/O Tuning for HPC Applications****, to appear*

Transparent and Online Tool for I/O



- TOTO intercepts POSIX I/O calls from each MPI rank
- Periodically, all ranks share information and characterize the current access pattern
- Prediction models can be used to support decisions
  - e.g. to choose stripe count
- The master sends instructions on parameters to use
  - For now, it changes stripe count and avoids straggler OSTs

Francieli Boito, Luan Teylo, Mihail Popov, Laora Aimi, Alexis Bandet, Guillaume Pallez. ***TOTO: A Tool for Transparent I/O Tuning for HPC Applications****, to appear*



Developers: Francieli Boito & Luan Teylo
(Paper and git repository to appear soon)

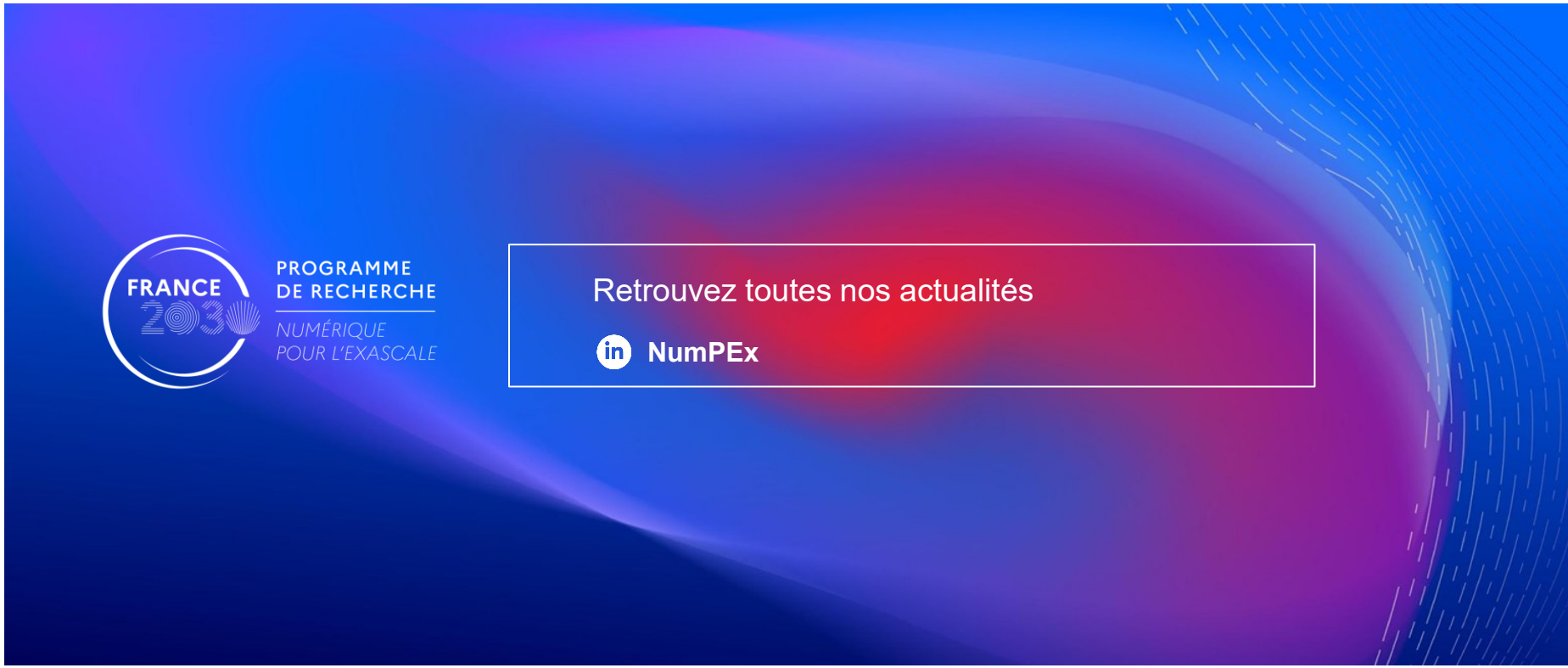
3. Next Steps

Next steps

- Continue the characterization of I/O behavior in HPC systems
  - Expand to more recent systems
  - Study the biases of I/O datasets
- Leverage the knowledge of I/O patterns
  - Develop more accurate I/O models in FIVES based on MOSAIC characterization
  - Make informed scheduling decisions - thesis starting soon (TADaaM + KerData)
- Work on the **illustrators**
  - Mitigate I/O interference for Gysela - Méline Trochon's thesis
  - Progress on profiling the DDF pipeline (SKA) - create a mockup for I/O benchmarking
  - **A big goal for this meeting!**

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

cea cnrs Inria

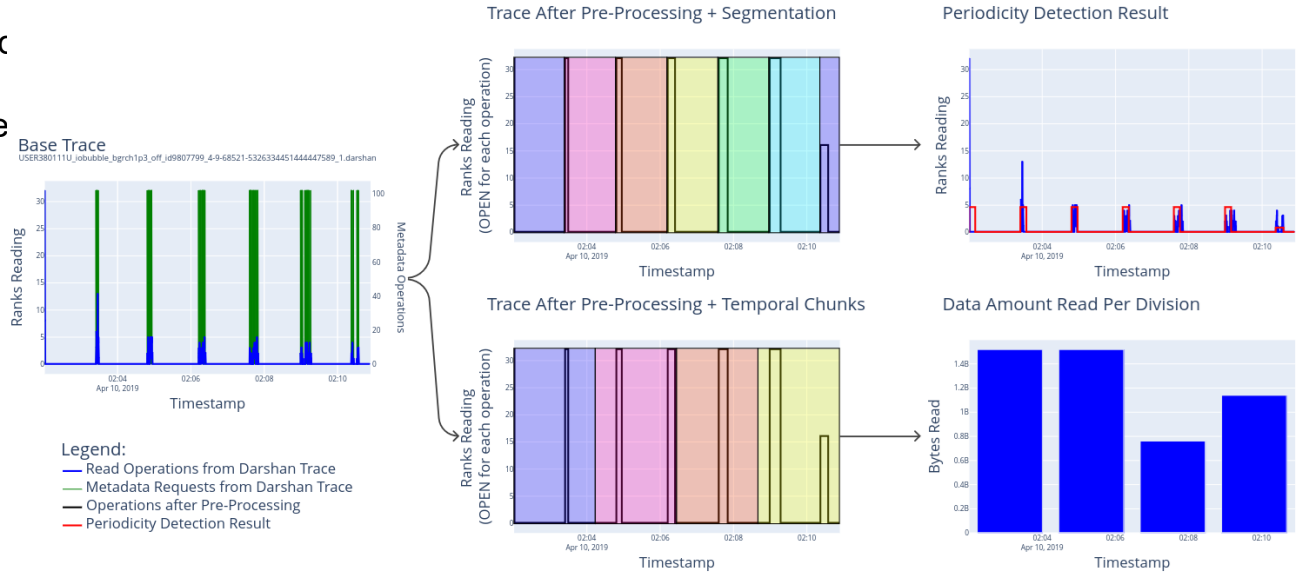


Focus 2: Detection and Categorization of I/O Patterns in HPC Applications

MOSAIC

- Segmentation-based method
  traces
- Analysis of one year of trace

internship
(Oct. 2024)



T. Jolivel, F. Tessier, J. Monniot, G. Pallez. ***MOSAIC: Detection and Categorization of I/O Patterns in HPC Applications***, PDSW 2024, Nov 2024, Atlanta, USA

Deliverables

- [MdlS, R] (M0+08) **WP1,2,3,4**: Selection of the initial release of the libraries and tools that will make up the Exa-DoST software stack.
- [TADAAM, R] (M0+23) **WP1**: Report on the solutions selected in Exa-DoST to answer the storage and IO challenges at Exascale
- [KerData, C] (M0+23) **WP1,2,3**: Intermediate coordinated release of all tools and libraries produced by Exa-DoST, including documentation
- [MdlS, C] (M0+35) **WP1,2,3**: Intermediate coordinated release of all tools and libraries produced by Exa-DoST, including documentation
- [SANL, C] (M0+47) **WP1,2,3**: Intermediate coordinated release of all tools and libraries produced by Exa-DoST, including documentation
- [DataMove, C] (M0+59) **WP1,2,3**: Final releases of all tools and libraries produced by Exa-DoST, including documentation
- [DataMove, R] (M0+65) **WP1,2,3**: Report on the final design of the tools and libraries produced by Exa-DoST and design solved