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

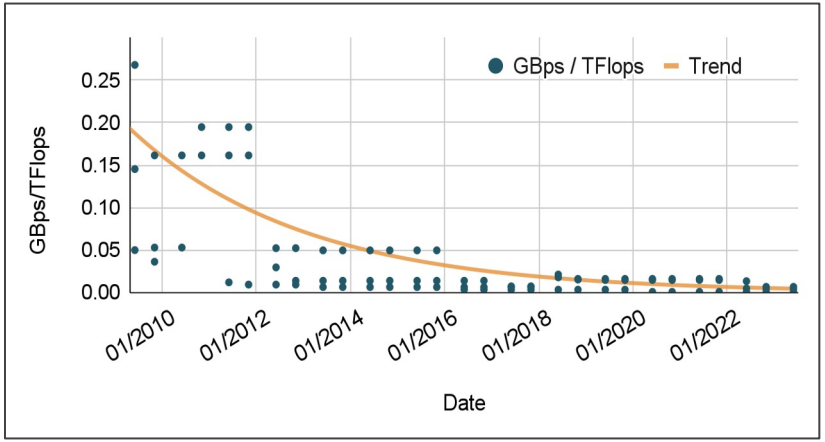
WP Leaders:
Francieli Boito (Université de Bordeaux) &
François Tessier (Inria Rennes)

1. Challenges

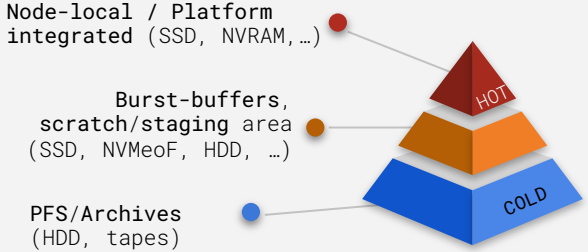
Trends

> “A supercomputer is a device for turning compute-bound problems into I/O-bounds problems.”
>
> *[Kenneth E. Batcher, Kent State Univ.]*

- Ensemble forecast (ECMWF)
  - 60TB generated per hour
  - Projection : **+40% per year**
- LHC data archives (CERN)
  - 250PB of accumulated data
  - In 2030 : **4300PB (x17)**
- Q Continuum cosmological simulation (DOE)
  - **2PB per simulation campaign**

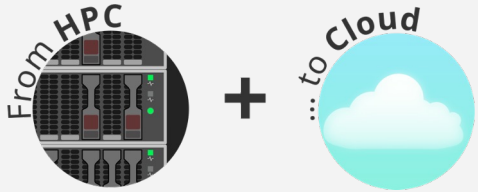


Trends



Deep storage hierarchy

New storage technologies



Hybrid infrastructures

→ Lustre FS

→ 679 PB capacity tier
- 47,700 HDD
- ~5 TBps

→ 12 PB performance tier
- 5,400 NVMe
- ~10 TBps

Vertical and horizontal scaling

WP Objectives

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

- Inria Bordeaux
  - Researchers: Francieli Boito, Luan Teylo, Emmanuel Jeannot, Brice Goglin
  - Engineers: Mahamat Abdraman
  - PhD Students: Alexis Bandet
  - Interns: Iheb Becher
  - <+ open positions: 1 PhD Student, 1 Post-doc>
- Inria Rennes
  - Researchers: François Tessier, Gabriel Antoniu, Guillaume Pallez, Silvina Caino-Lores, Jakob Luettgau
  - Engineers: Julien Monniot (to start in Jan. 2025)
  - PhD Students: Julien Monniot, Théo Jolivel (+ CEA)
  - Interns: Ugo Thay
- Maison de la Simulation
  - Researchers: Julien Bigot, Yushan Wang, <+ open position>
- CEA
  - Researchers: Philippe Deniel, Thomas Leibovici, Arnaud Durocher, Maxime Delorme
- DDN
  - Researchers: Jean-Thomas Acquaviva
  - PhD Students: Méline Trochon (+ Inria Bordeaux, + Inria Rennes)

2. Scientific Contributions

Summary

Datasets

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

Scientific Dissemination

- 3 conference papers: IPDPS'24, Euro-Par'24, HiPC'24
- 1 workshop paper: PDSW'24
- 1 pre-print @ HAL
- 3 internship reports
- Multiple talks: COMPAS'24, JLESC, Per3S, …

External Collaborations

- MeerKAT, South Africa
- The LAB, Bordeaux, France
- University of Honolulu, HI, USA
- University of Darmstadt, Germany

Project's deliverables

- 1 report submitted

Software production

- FIVES
- IOPS
- MOSAIC

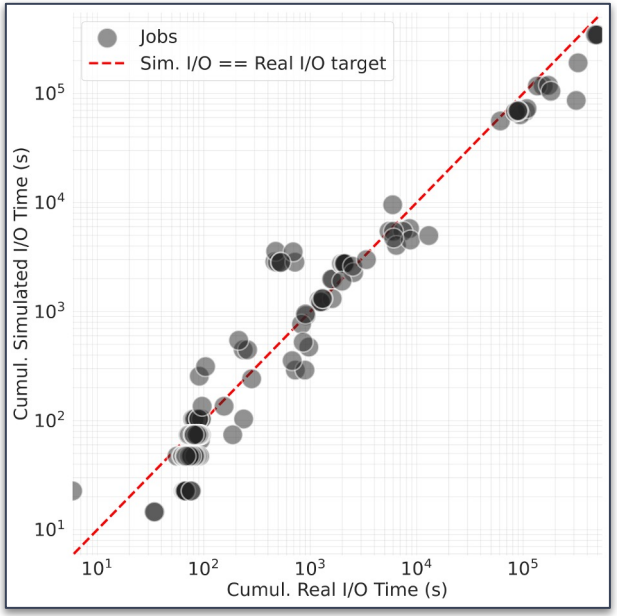
Focus 1: Simulation of Large-Scale HPC Storage Systems

FIVES is a WRENCH-based Simulator of Scheduling on Storage Systems at Scale (5S)

- Based on WRENCH/SimGrid (time-based DES)
- **Batch scheduler implementation**
- New built-in **distributed storage service**

FIVES can:

- Replay aggregated **Darshan I/O traces**…
- … on a **modeled supercomputer**…
- … including its **parallel file-system**…
- … and provides the user with **accurate results**…
- … thanks to a **bayesian optimization calibration**.

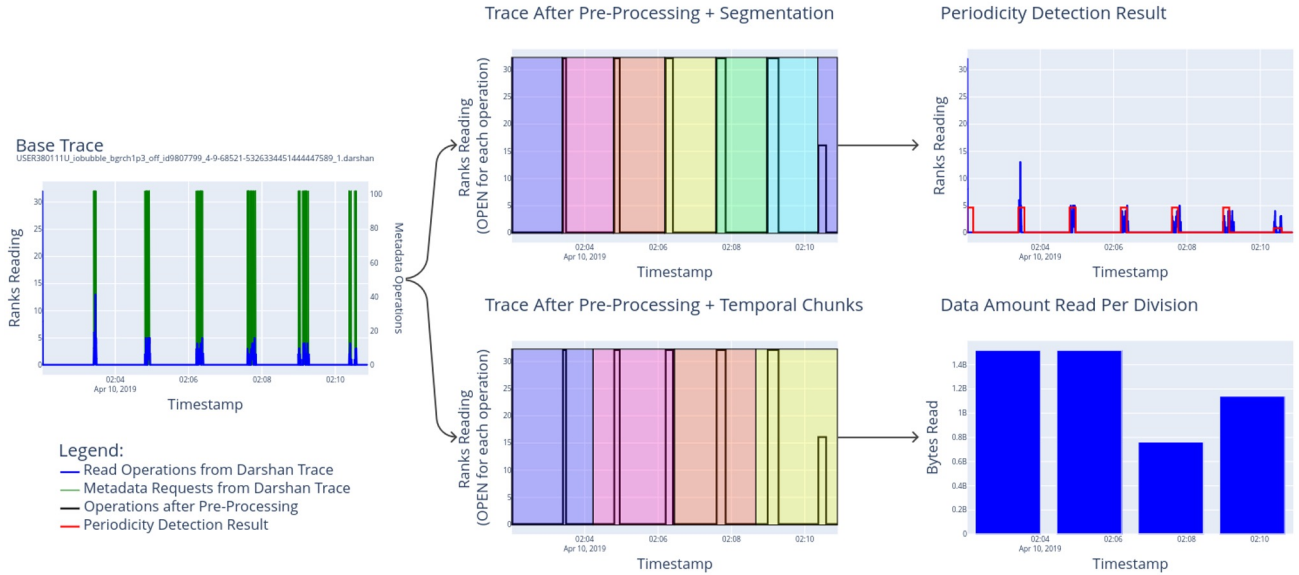


J. Monniot, F. Tessier, H. Casanova, G. Antoniu. ***Simulation of Large-Scale HPC Storage Systems: Challenges and Techniques****, HiPC 2024, Dec 2024, Bengalore, India*

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

MOSAIC

- Segmentation-based method for **detecting I/O patterns**, including periodic behavior, from **Darshan I/O traces**
- Analysis of one year of traces from the **Blue Waters** supercomputer

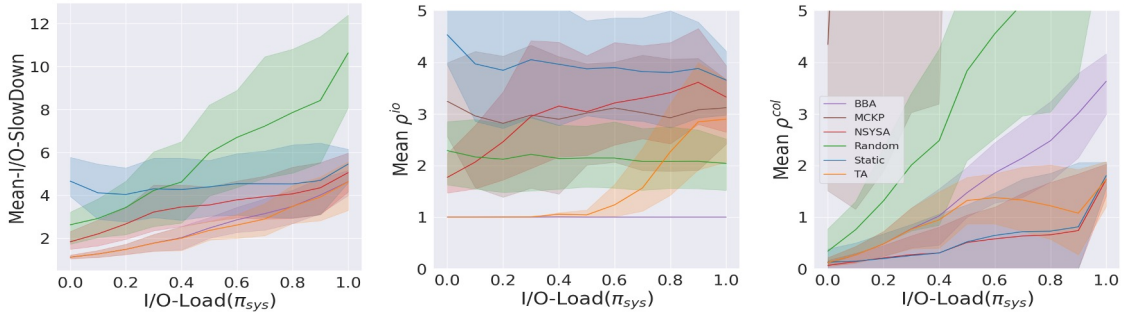


ExaDoST-funded internship and thesis (Oct. 2024)

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*

Focus 3: Scheduling Distributed I/O Resources in HPC

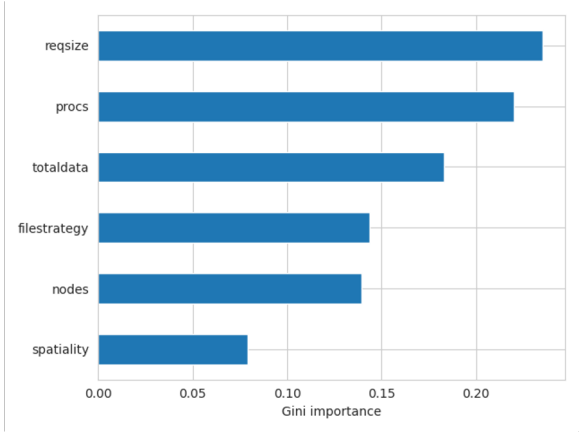
- Proposed algorithms for **allocation** and **placement** of I/O resources (OSTs, I/O nodes, etc)
- Using different application information as input
- Placement: **balancing the number of applications per resource is up to 50% better than random placement**
  - more sophisticated placement is **not** necessary
- Allocation: **BBA and TA algorithms up to 4 times better than an allocation based on the number of compute resources**
  - **BBA** requires the number of I/O resources that maximizes application performance



Alexis Bandet, Francieli Boito, Guillaume Pallez. ***Scheduling distributed I/O resources in HPC systems****. Euro-Par 2024, Madrid, Spain.*

Focus 4: Prediction of HPC I/O Resources Usage

- How can we obtain the input for the BBA allocation algorithm?
- We defined the "**best prediction"**, taking performance variability into account
  - up to **~25% better than BBA**!
- Machine learning models to **predict it from general application characteristics**
  - we can get ~80% accuracy without amount of data and spatiality (harder to obtain)
- Data set available at Zenodo: https://doi.org/10.5281/zenodo.10518127

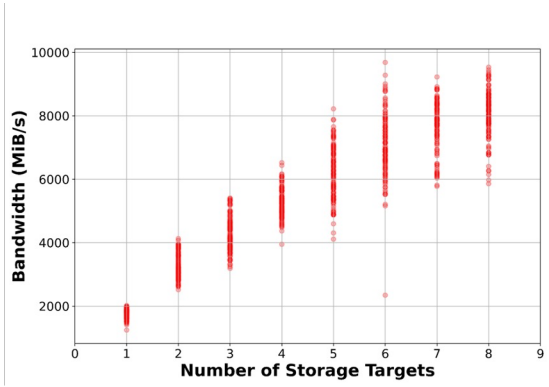


Alexis Bandet, Francieli Boito, Guillaume Pallez. ***Prediction and Interpretability of HPC I/O Resources Usage with Machine Learning****. Pre-print, 2024: https://inria.hal.science/hal-04698511*

Focus 5: I/O Performance Profiling

- **Profile parallel file system performance** by evaluating combinations of parameters: compute nodes, processes, stripe count, and size.
- **IOPS, an open-source tool**, automates parameter search
  - Designed for **ease of use** and to **minimize the number of tests**
  - Provides a **report about the results**

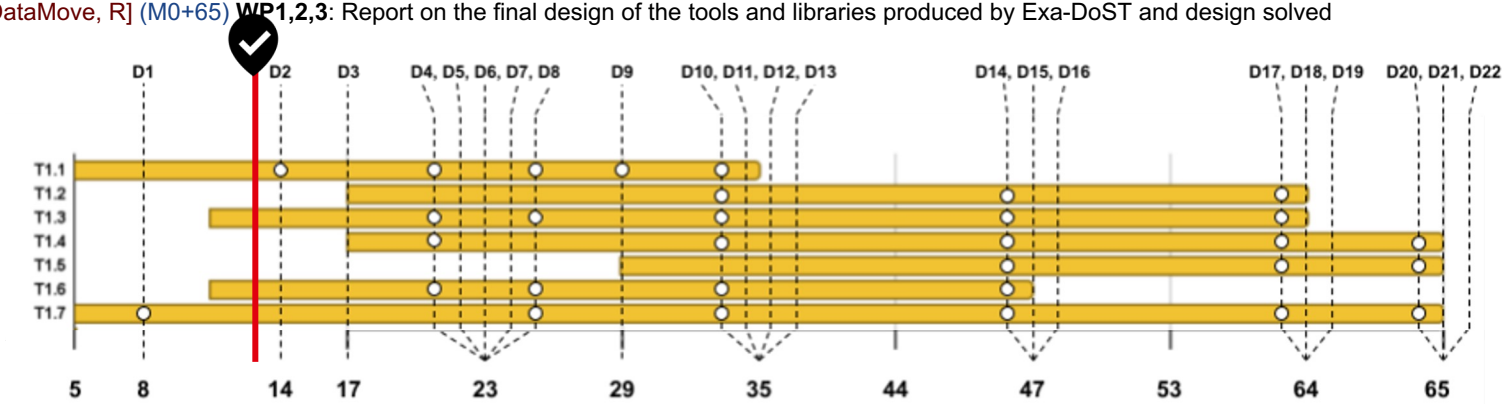
https://gitlab.inria.fr/lgouveia/iops



ℹ Exa-DoST funded internship and engineer position (Oct 2024)

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



3. Perspectives & Challenges

Perspectives & Challenges

- TADaaM + KerData currently working on **trace analysis**
  - to characterize the temporal I/O behavior of HPC applications
  - extract **common patterns**
- Need to ***work on the illustrators***
  - two internships on SKA (TADaaM+LAB, KerData)
    - more details during the WP1 session tomorrow morning
  - start of Méline Trochon's thesis soon, after an internship on Gysela
  - one of the tasks for the new recruited engineer @ TADaaM
  - a big goal of this meeting!

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

cea cnrs Inria

FRANCE 2030

PROGRAMME DE RECHERCHE

NUMÉRIQUE POUR L'EXASCALE

Retrouvez toutes nos actualités

NumPEx