# PEPR NumPEx and Cloud: 7-year Perspectives of Academic Research on Large-Scale Data Management

François Tessier - Team KerData@INRIA - Rennes, France

Teratec Forum - 06/2023

# 01

## PEPRs

# PEPR: Priority Research Programs and Equipment

**What are PEPRs?**

- Action of the “France 2030” plan dedicated to the financing of national fundamental research
- **Objective**: to fund strategic research areas
  - Material science, climate, robotics, biology, agronomics, …
- 3B€ released
  - 1B€ dedicated to exploratory PEPRs
    - Selection by call for programs
    - International evaluation
  - 2B€ dedicated to strategic PEPRs
    - Support ongoing research

# PEPR: Priority Research Programs and Equipment

**What are PEPRs?**

- Action of the “France 2030” plan dedicated to the financing of national fundamental research
- **Objective**: to fund strategic research areas
  - Material science, climate, robotics, biology, agronomics, …
- 3B€ released
  - 1B€ dedicated to exploratory PEPRs
    - Selection by call for programs
    - International evaluation
  - 2B€ dedicated to strategic PEPRs
    - Support ongoing research

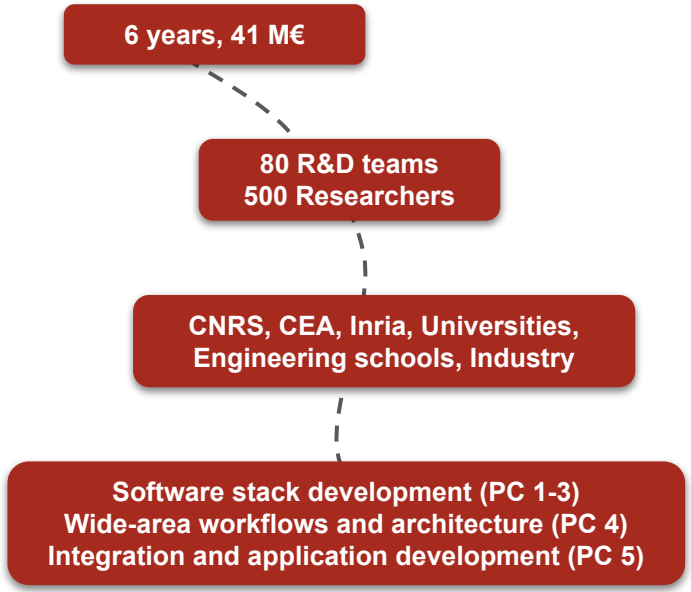
**NumPEx**

**Cloud**

# NumPEx Program

- **Goal:** Co-design the exascale software stack and prepare applications for the exascale era
- **Co-directors:** Dr. J. Bodin (CEA), Pr. M. Dayade (CRNS), Dr. J-Y Berthou (Inria)

- 5 Projects:
  - **ExaMa**: Methods and Algorithms
    *Pr. C. Prudhomme, Univ. Strasbourg*
  - **ExaSoft**: HPC software and tools
    *Pr. R. Namyst, Inria/Univ. Bordeaux*
  - **ExaDoST**: Data-oriented software and tools
    *Dr. G. Antoniu, Inria*
  - **ExaAtoW**: Architectures and Tools for Large-Scale Workflows
    *Pr. F. Bodin, Univ. Rennes*
  - **ExaDIP**: Development and Integration Project
    *Dr. J-P Vilotte, CNRS*



# NumPEx Program

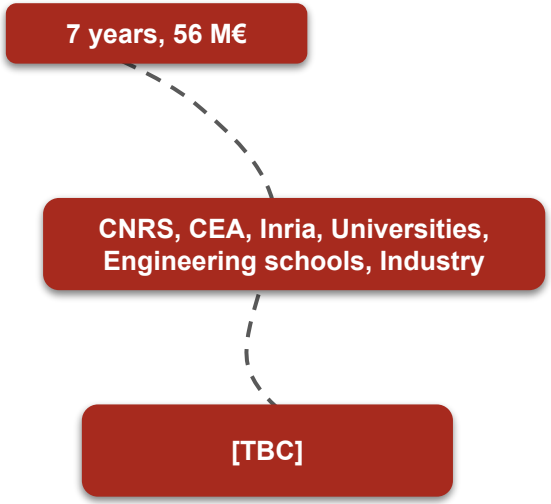
- **Goal:** Co-design the exascale software stack and prepare applications for the exascale era
- **Co-directors:** Dr. J. Bodin (CEA), Pr. M. Dayade (CRNS), Dr. J-Y Berthou (Inria)

- 5 Projects:
  - **ExaMa**: Methods and Algorithms
    *Pr. C. Prudhomme, Univ. Strasbourg*
  - **ExaSoft**: HPC software and tools
    *Pr. R. Namyst, Inria/Univ. Bordeaux*
  - **ExaDoST**: Data-oriented software and tools
    *Dr. G. Antoniu, Inria*
  - **ExaAtoW**: Architectures and Tools for Large-Scale Workflows
    *Pr. F. Bodin, Univ. Rennes*
  - **ExaDIP**: Development and Integration Project
    *Dr. J-P Vilotte, CNRS*

# Cloud Program

- **Goal:** Support the development of software and hardware layers that will help design tomorrow's highly distributed, secure infrastructures with the smallest possible energy footprint.
- **Co-directors:** Frédéric Deprez (Inria), Adrien Lèbre (IMT Atlantique)

- 7 Projects, including:
  - **STEEL**: **S**ecure and efficient da**T**a storag**E** and proc**E**ssing on c**L**oud-based infrastructures
    *Dr. G. Antoniu, Inria*
  - [TBD]



# Cloud Program

- **Goal:** Support the development of software and hardware layers that will help design tomorrow's highly distributed, secure infrastructures with the smallest possible energy footprint.
- **Co-directors:** Frédéric Deprez (Inria), Adrien Lèbre (IMT Atlantique)

- 7 Projects, including:
  - **STEEL**: **S**ecure and efficient da**T**a storag**E** and proc**E**ssing on c**L**oud-based infrastructures
    *Dr. G. Antoniu, Inria*
  - [TBD]

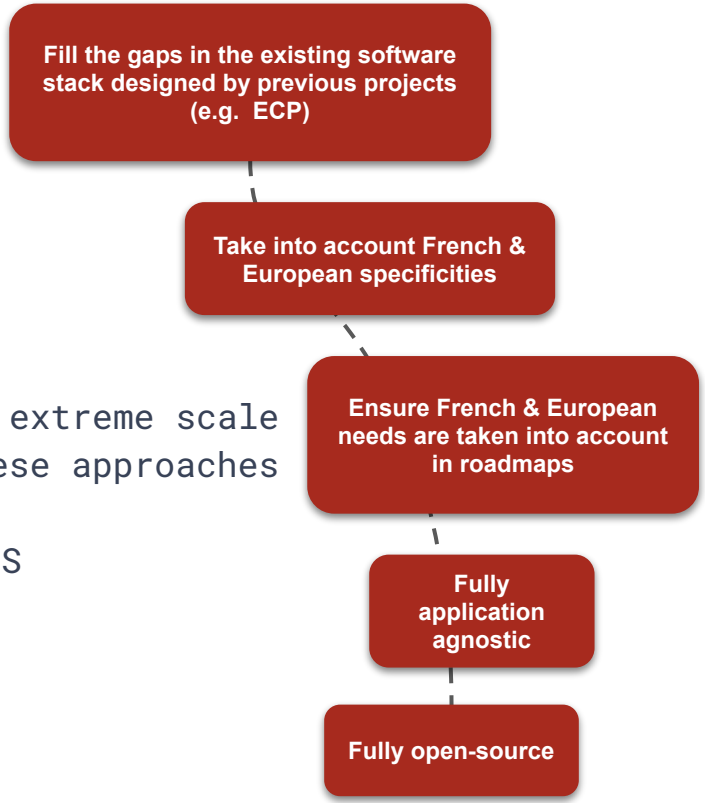
# Our Ambition

Approach:

- **Research** on data-oriented tools for HPC
- That leads to transverse, **re-usable tools**
- Usable **in production** at exascale

ExaDoST and STEEL will produce:

- **New approaches** to handle the data challenge at extreme scale
- Transverse **libraries & tools** that implement these approaches
- Validated on illustrators at **full scale**
  - Small/Mid-scale: Teralab, Grid'5000, SLICES
  - Large-scale: **GENCI**

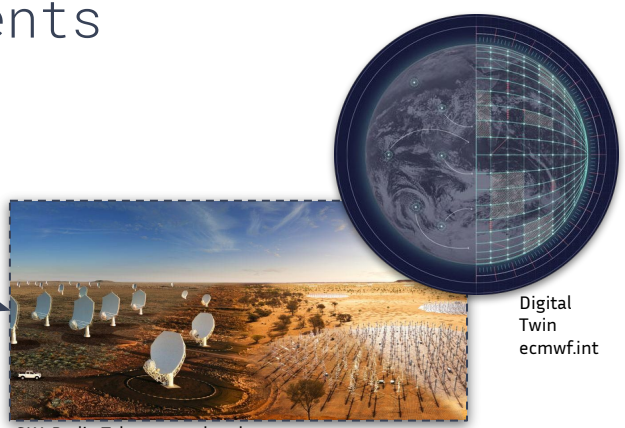
# 02

# Scientific Context

# Growing I/O requirements

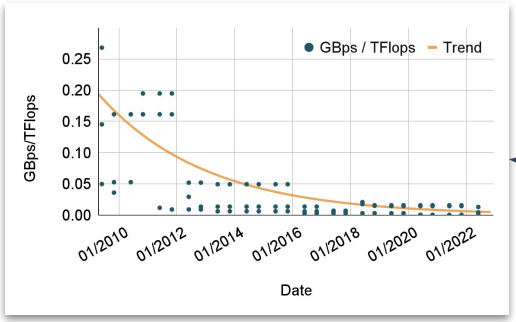
**Data deluge** from new large-scale scientific workflows

- ➔ Multiplication of data sources
- ➔ Drastic increase from scientific projects



Digital Twin ecmwf.int

SKA Radio Telescope - skatelescope.org




↗ PFlops ↘ TBps

=

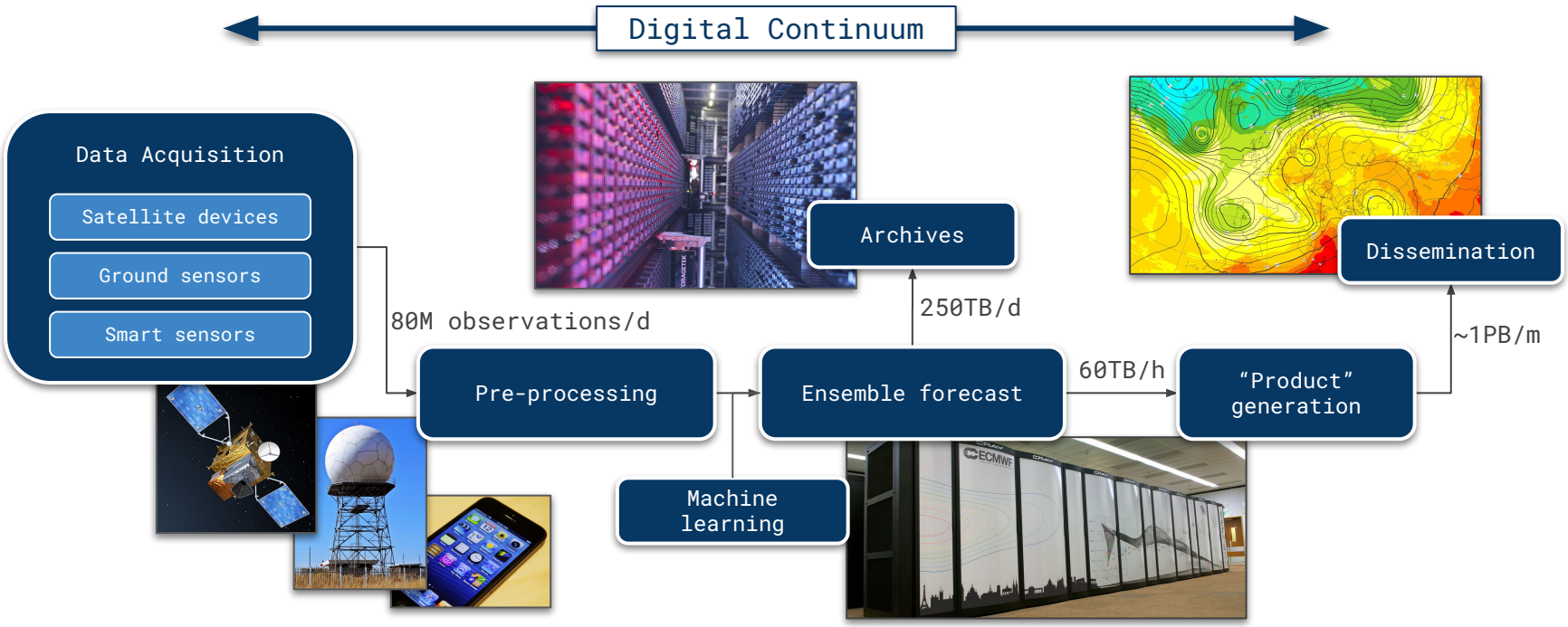
↗ gap between compute and I/O performances

**Compute-centric** to **data-centric** shift

↗ **I/O pressure** for large-scale systems

# Growing I/O requirements

## Use-case: Digital Twin of the Earth System for Weather Forecast



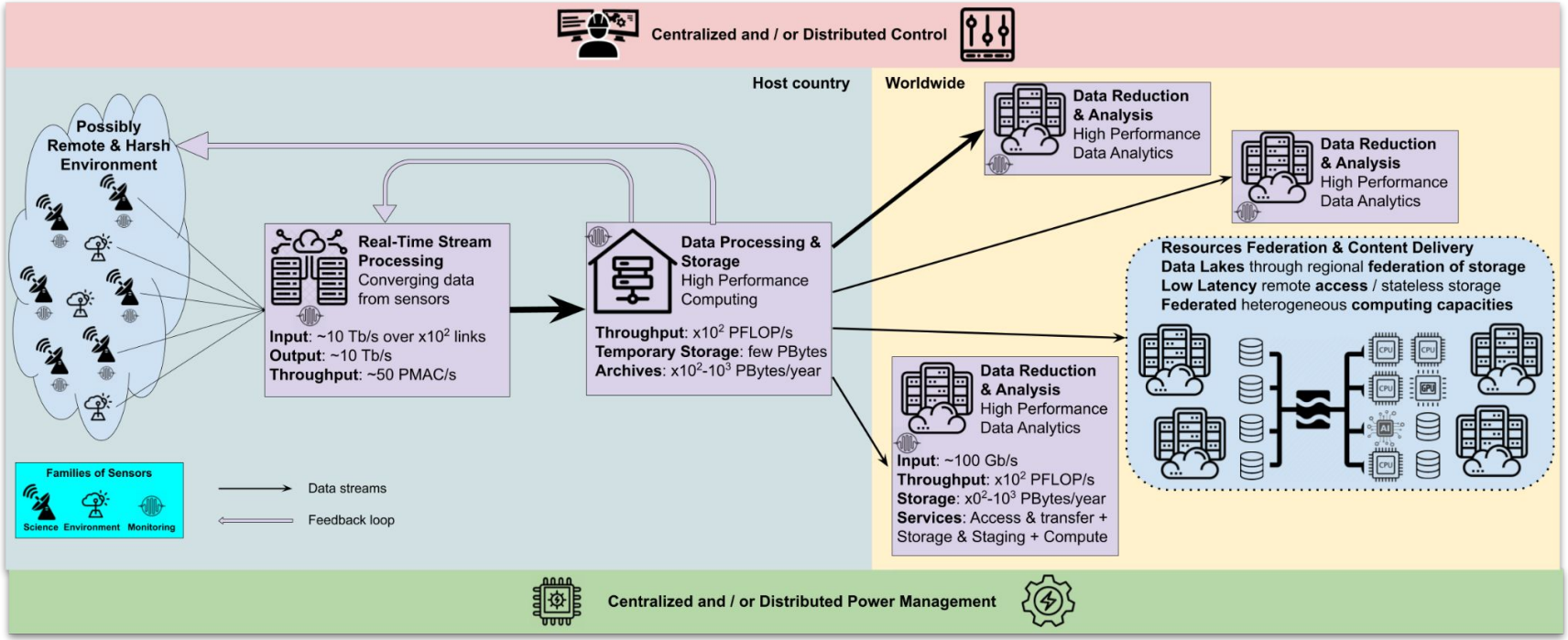

**Sources :** *ETP4HPC's SRA 4 - Strategic Research Agenda for High-Performance Computing in Europe (2020)*
*ECMWF - European Center for Medium-range Weather Forecast*
*Maestro H2020 Project*

# Growing I/O requirements

- Square Kilometer Array (SKA): Largest radio telescope in the world.
  - €1.3 B for construction, €0.7 B for the first 10 years of operation
  - More than 130k antennas (Australia) and ~200 dishes (South Africa)
  - 710 PB of science data delivered to users
  - 8 years to construct
  - 16 countries
  - Operational in 2030

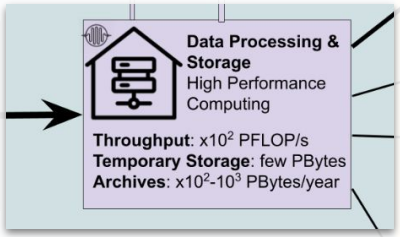

**Sources :** *SKAO website*

# Growing I/O requirements

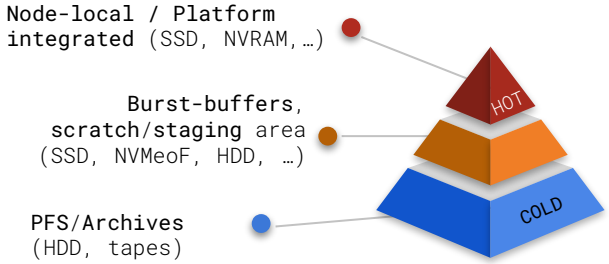


# Growing I/O requirements

- Compute:
  - ~125 Pflops peak
  - > 10 Pflops average
- IO
  - Ingestion: ~1 TB/s
  - Storage: ~40 PB
  - Data accesses: ~10 TB/s
- Budget: 35 M€

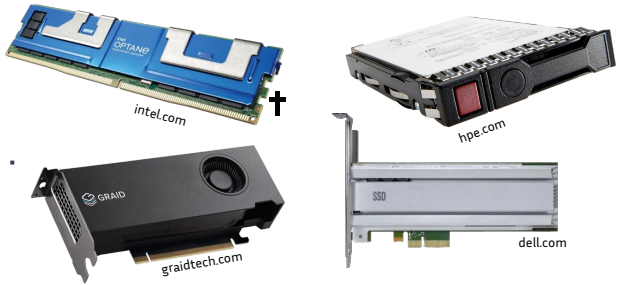


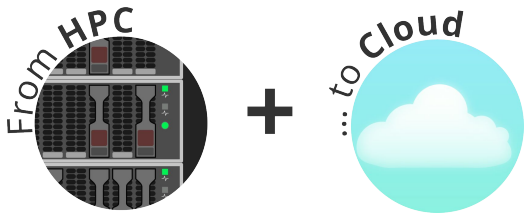
# Current trends




- Deep storage **hierarchy**

- **New** underlying **storage technologies**

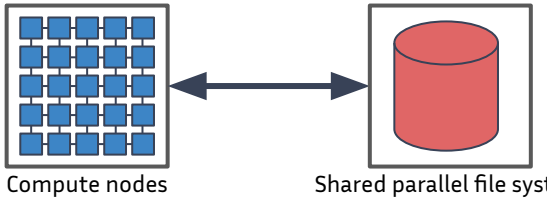


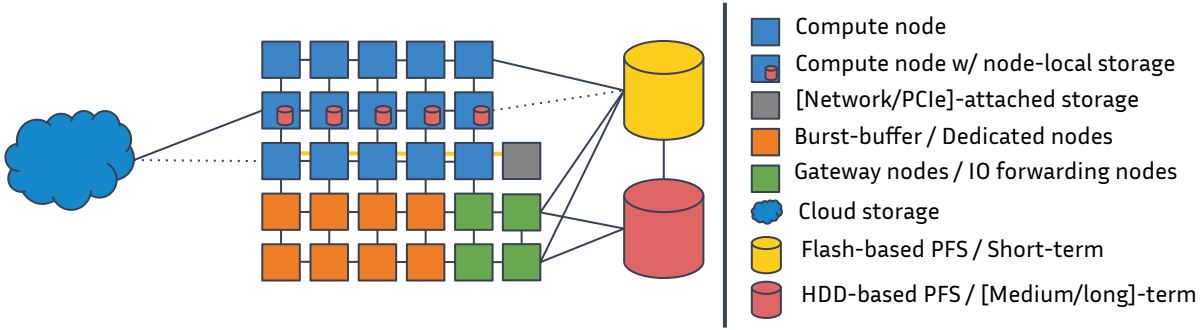

- **Migration** of HPC workflows/workloads **towards hybrid platforms**

# Finding **balance** between **performance** and **complexity**

We went from traditional storage systems…




…to more complex and hybrid resources:




↗ **Complexity, underutilization** of resources, **performance left** on the table

# Application Requirements

- **Scale up** modern I/O and data storage methods and tools
- **Support** the I/O and storage requirements of complex simulation/analytics/AI **workflows** running on hybrid HPC/cloud/edge systems
- Develop and integrate **new output formats** for checkpoint/restart and for scientific analysis, (e.g., based on the LightAMR standard)

NumPEx

- Exploit **emerging technologies** for efficient, fault-tolerant storage
- Offer efficient data storage and processing solutions on hybrid, heterogeneous infrastructures within the digital **edge-cloud-supercomputer continuum**
- Enabling **confidential storage** on clouds

Cloud

# Application Requirements

- **Scale up** modern I/O and data storage methods and tools
- **Support** the I/O and storage requirements of complex simulation/analytics/AI **workflows** running on hybrid HPC/cloud/edge systems
- Develop and integrate **new output formats** for checkpoint/restart and for scientific analysis, (e.g., based on the LightAMR standard)

NumPEx

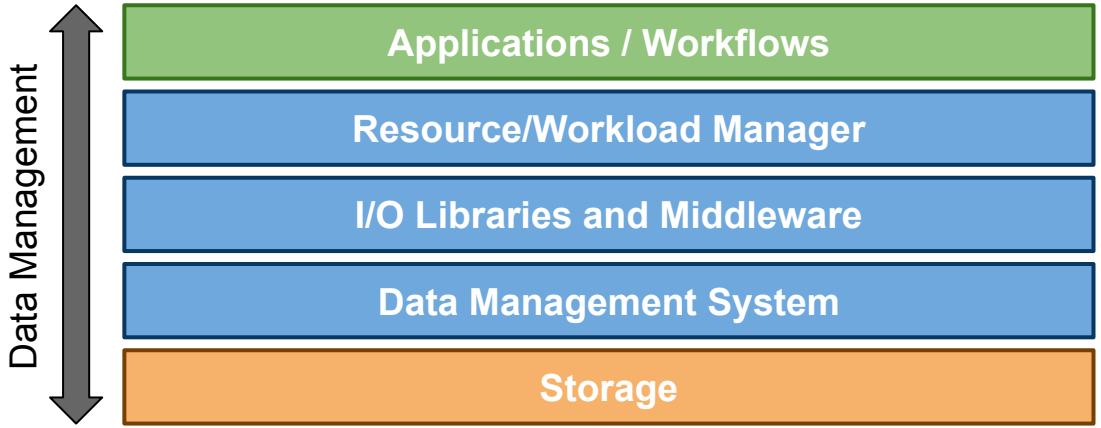
- Exploit **emerging technologies** for efficient, fault-tolerant storage
- Offer efficient data storage and processing solutions on hybrid, heterogeneous infrastructures within the digital **edge-cloud-supercomputer continuum**
- Enabling **confidential storage** on clouds

Cloud

# 03

# Identified Research Areas: focus on I/O challenges

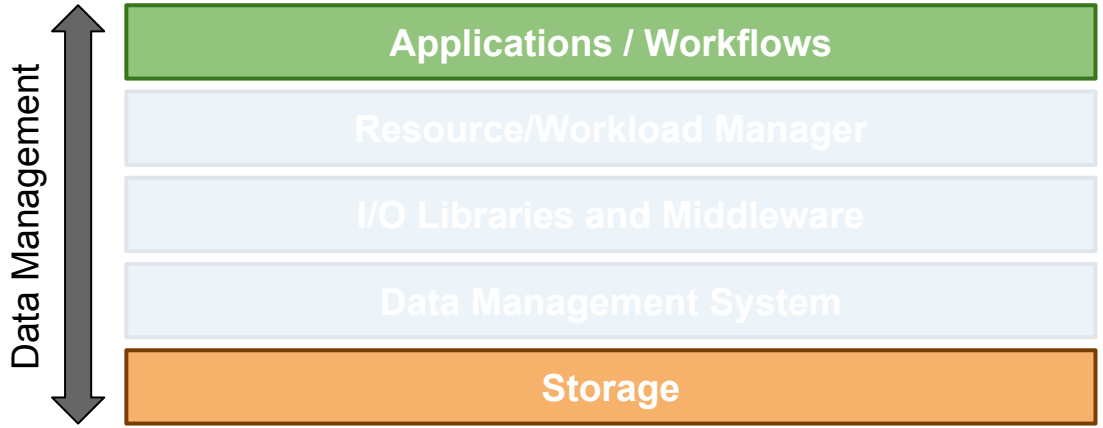
# Benchmarking and Characterization



# Benchmarking and Characterization

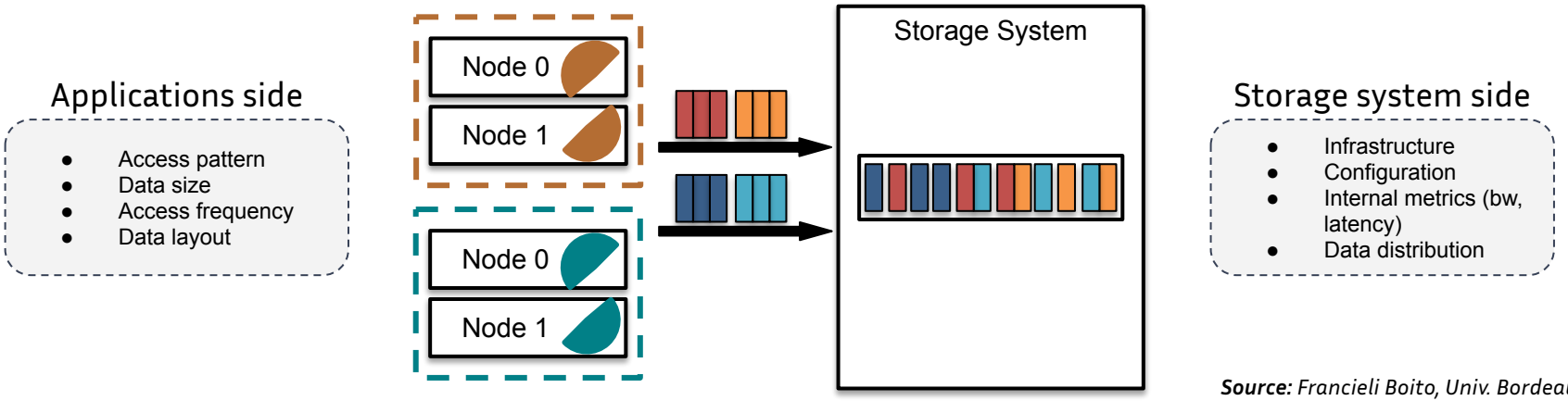
> A deep understanding of the I/O behavior of applications and storage systems is essential for developing efficient optimizations.

# Benchmarking and Characterization



# Benchmarking and Characterization

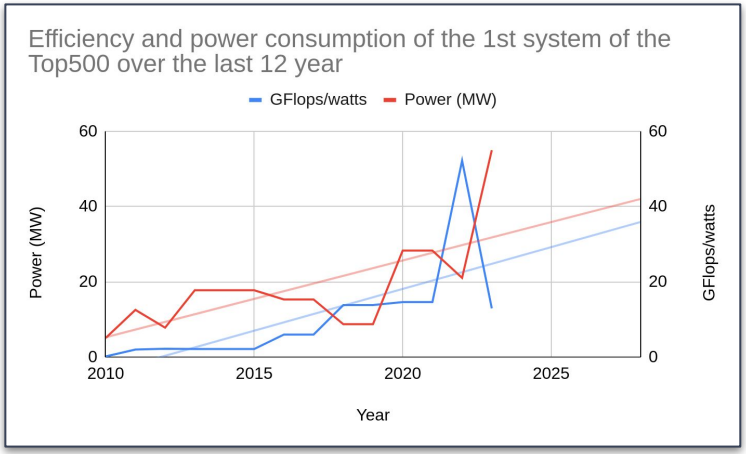
- Fine-grained benchmarking of applications' I/O behavior and testbeds' underlying storage systems
- Increase knowledge of application access patterns and storage systems performance
  - Production conditions
  - Monitoring of I/O (application and systems)
  - Benchmarking suite




***Source:*** *Francieli Boito, Univ. Bordeaux*

# Benchmarking and Characterization

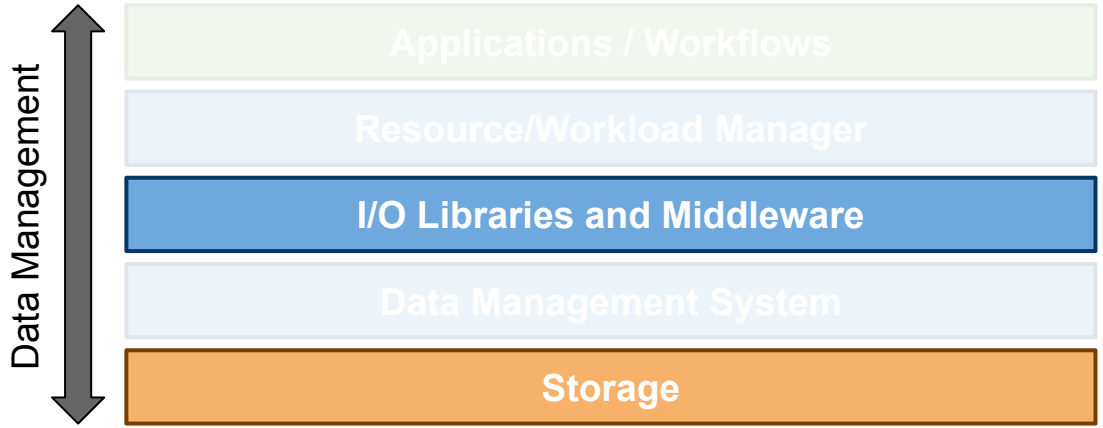
- Strategies and metrics for application characterization
  - Metrics collected from application execution
  - Predict how **performance** and **energy consumption** will be affected
  - Classification of applications based on their I/O behavior
    - Create a matching between I/O behavior and optimization techniques



# Leverage modern storage architectures

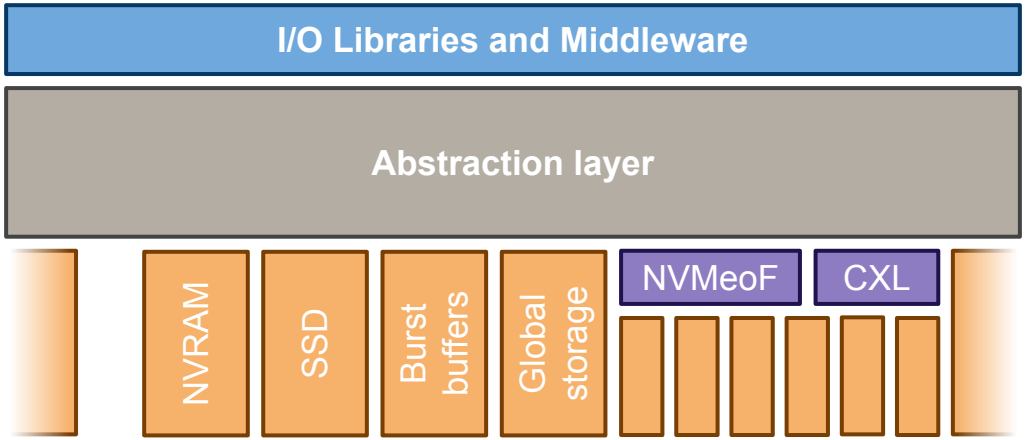
> Achieving code and performance portability across a broad variety of memory and storage tiers requires a certain level of abstraction.

# Leverage modern storage architectures



# Leverage modern storage architectures

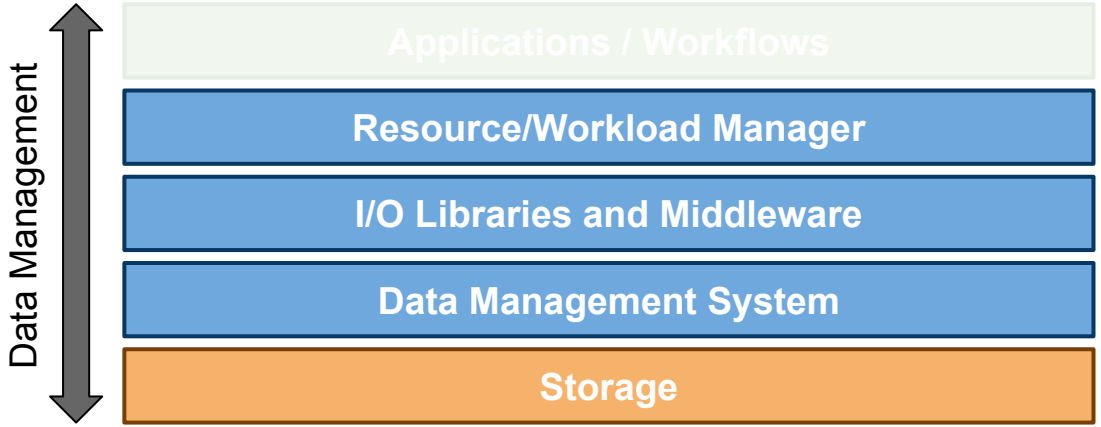
- Leverage modern storage architectures in a scalable way
  - Evaluation of existing I/O libraries and middleware
  - Support for **new intermediate storage tiers**
  - Unify the **memory/storage continuum**
    - Use-case: persistent, reliable storage of application working data, traditionally stored in DRAM



# Storage Resources Arbitration

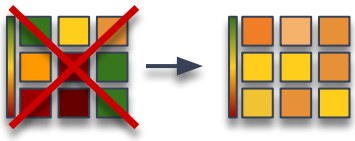
> Taking full advantage of all available resources is critical in a context where storage is central.

# Storage Resources Arbitration

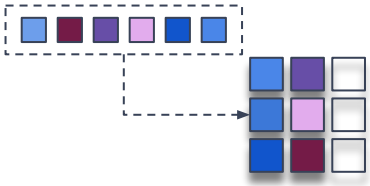


# Storage Resources Arbitration

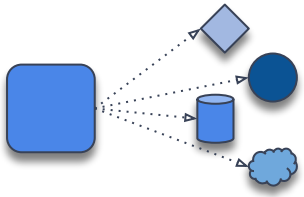
- Arbitrate storage resources between concurrent workloads
  - Scheduling of storage resources on HPC systems
    - Transpose compute resource management knowledge to storage resources
    - StorAlloc: a simulator of a storage-aware job scheduler for HPC systems [1]
      - Contention-aware provisioning of intermediate storage tiers
      - Sizing of storage infrastructures



Make **efficient** and **fair** use of all storage resources



**Transparently allocate storage** for users and applications



Deal with **heterogeneity of hardware resources**

***[1]*** *Julien Monniot, François Tessier, Matthieu Robert, Gabriel Antoniu. StorAlloc: A Simulator for Job Scheduling on Heterogeneous Storage Resources. HeteroPar 2022, Aug 2022, Glasgow, United Kingdom.*

# Storage Resources Arbitration

- Arbitrate storage resources between concurrent workloads
  - I/O scheduling in a concurrent environment
    - Shared storage resources -> impact on performance and variability
    - I/O scheduling algorithms at the I/O library/middleware level
      - Manage concurrency to increase the bandwidth

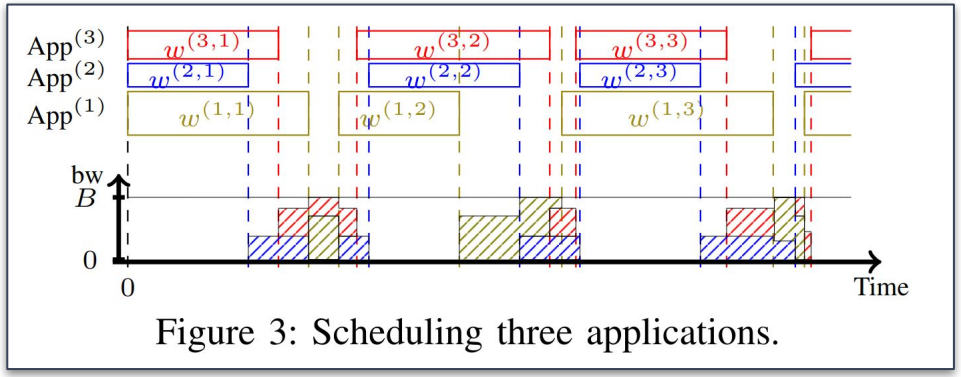
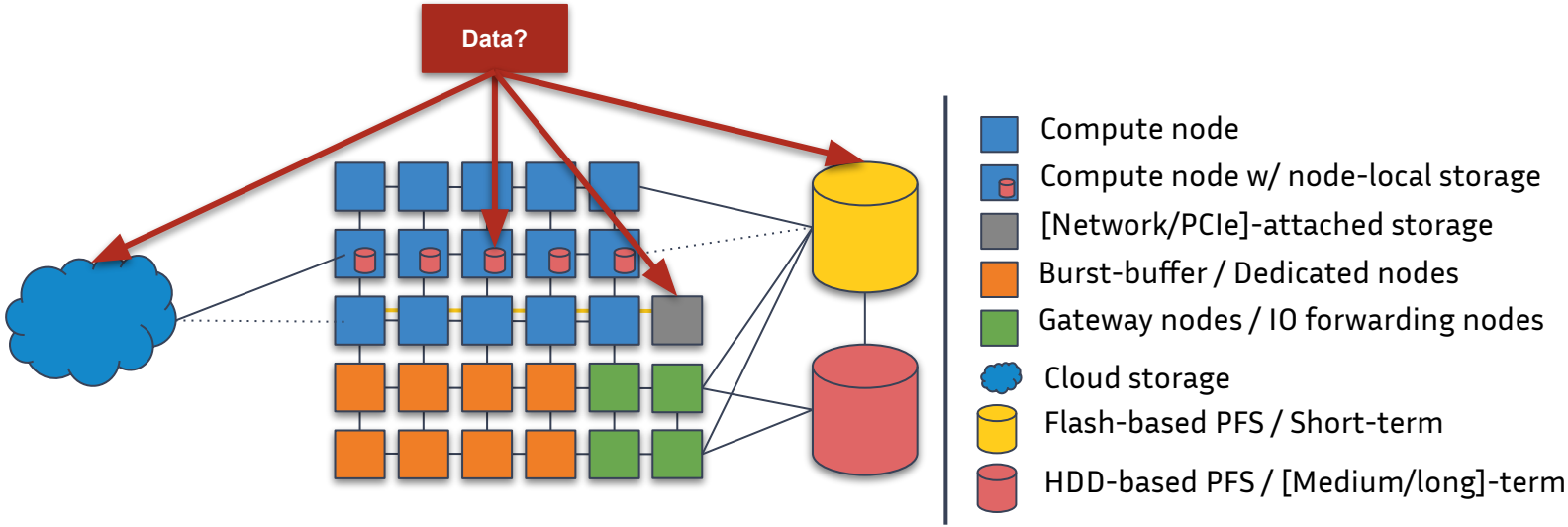


Figure 3: Scheduling three applications.

**Source:** *Guillaume Pallez, Inria*

# Storage Resources Arbitration

- Data placement on multi-tier storage infrastructures
  - For a single application or a workflow, where to place data such as:
    - Requirements are fulfilled (persistency, capacity, scope)
    - I/O performance maximized
    - Waste of resources minimized



# Conclusion

- **Be ready for Extreme Scale !**
- Two **ambitious** projects
  - An important part of the HPC community (in)directly involved…
  - … including industry!
- Choices made about the topics to cover
- Pragmatic approach for I/O and storage
  - (Almost) only what is **realistic**
  - Following the **data path**, from applications to disks

🐱 Links:
https://numpex.fr/

📧 Contacts:
francois.tessier@inria.fr