Storage Allocation over Hybrid HPC/Cloud Infrastructures

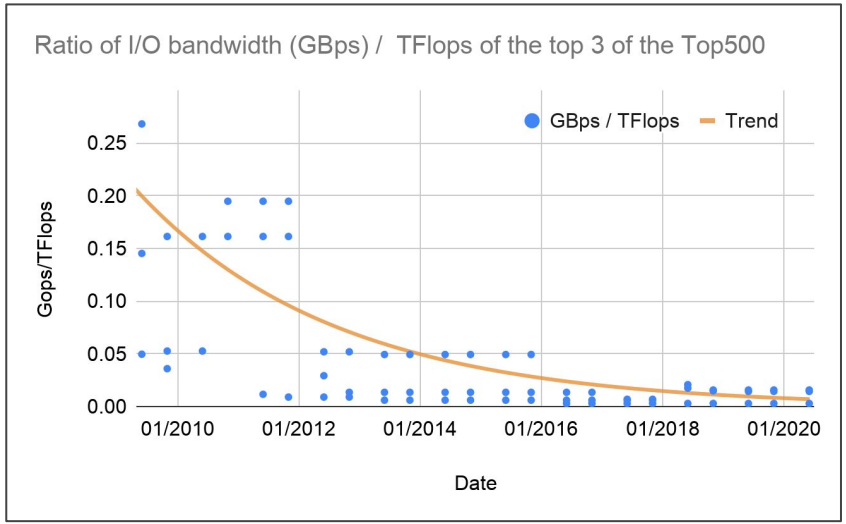
François Tessier, Gabriel Antoniu, Matthieu Robert
KerData *- Inria Rennes*

12th JLESC Workshop
Online

The data movement problem

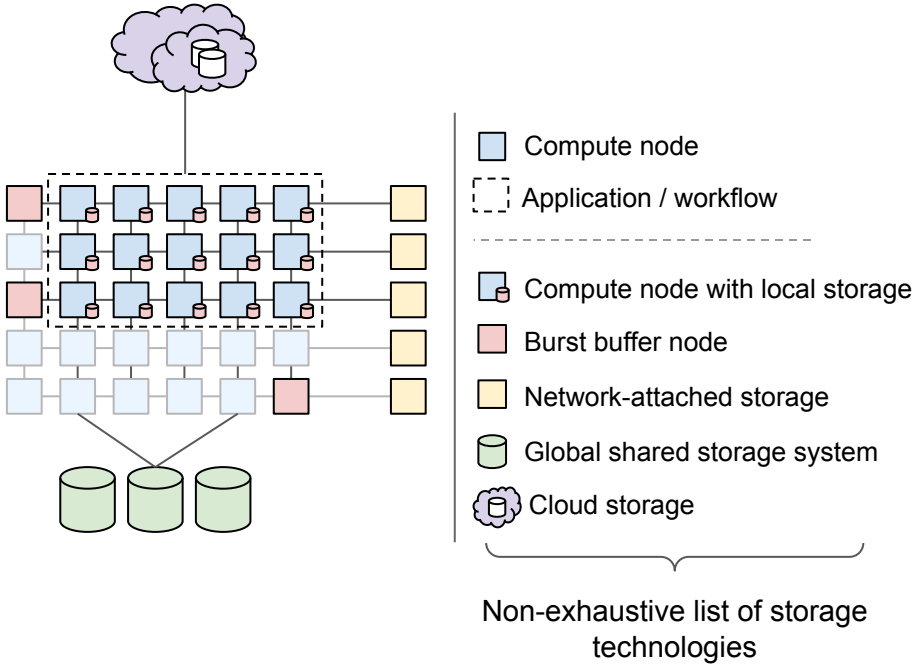
> "A supercomputer is a device for turning compute-bound problems into I/O-bounds problems."
>
> *[Kenneth E. Batcher, Kent State Univ.]*

- 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**



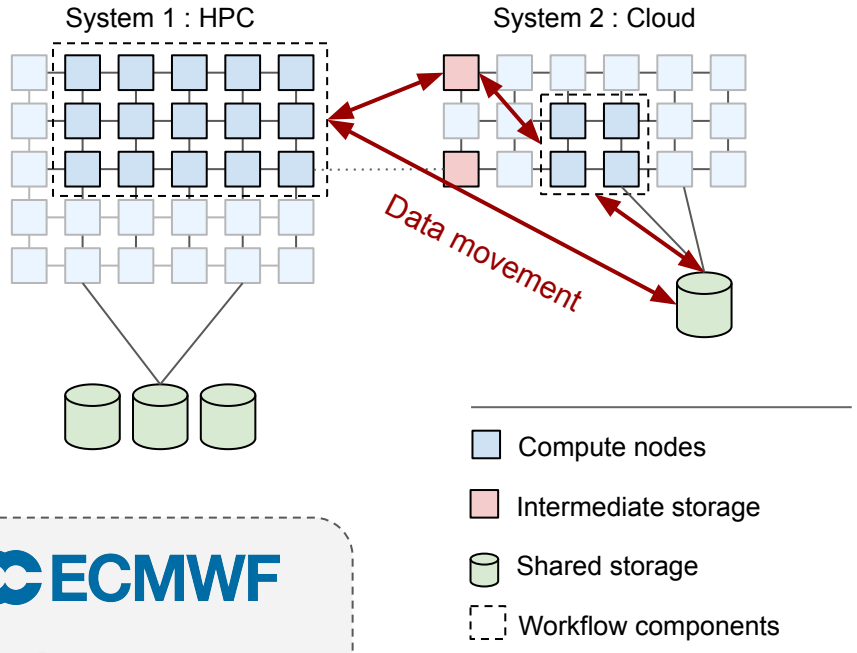
The data movement problem

- Mitigating the I/O bottleneck from an hardware perspective leads to an increasing complexity and a diversity of the architectures
  - Node-local storage (PCIe, SATA)
  - Burst buffers like Cray DataWarp, DDN Infinite Memory Engine
  - Network-attached storage (NVMeoF)
  - Cloud storage
- Strong need for flexible storage resources
  - Data-centric (hybrid) workflows
    - Weather forecast
    - Precision agriculture
  - Data services



The data movement problem

- Mitigating the I/O bottleneck from an hardware perspective leads to an increasing complexity and a diversity of the architectures
  - Node-local storage (PCIe, SATA)
  - Burst buffers like Cray DataWarp, DDN Infinite Memory Engine
  - Network-attached storage (NVMeoF)
  - Cloud storage
- Strong need for flexible storage resources
  - Data-centric (hybrid) workflows
    - Weather forecast
    - Precision agriculture
  - Data services

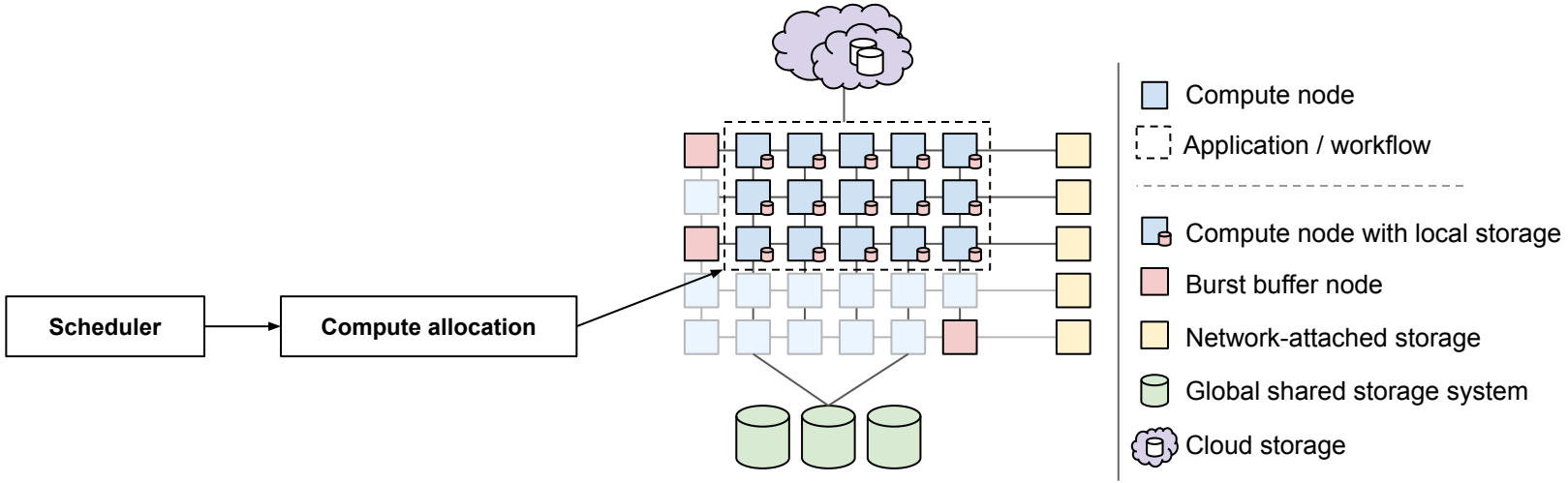


ECMWF

CybeleTech
Technologies numériques pour le monde végétal

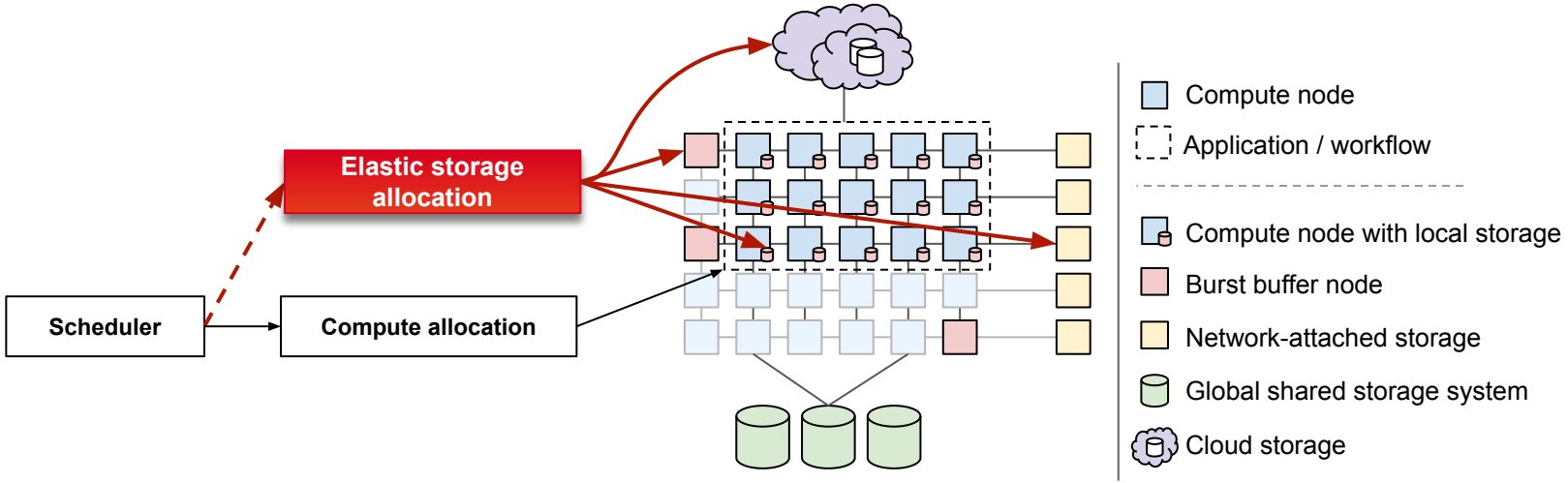
Storage allocation over hybrid infrastructures

Problem: How to provide applications and workflows with intermediate storage resources?



Storage allocation over hybrid infrastructures

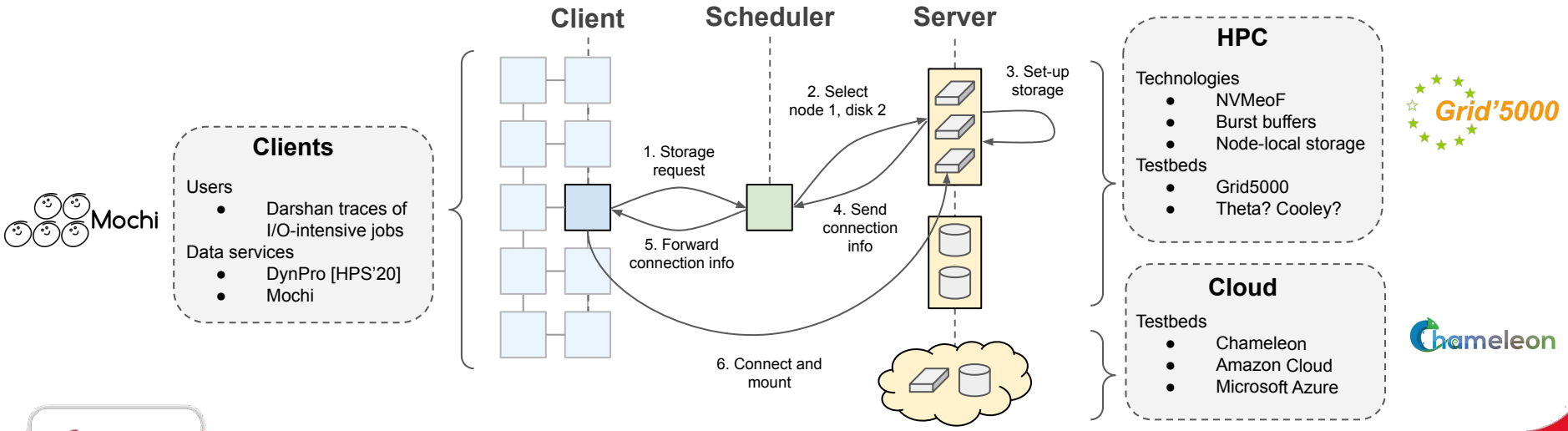
Problem: How to provide applications and workflows with intermediate storage resources?



Storage allocation over hybrid infrastructures

Goal: make storage resources allocable in the same way as computing resources on large-scale systems

- Simulator of a job scheduler for storage resources with three open research tracks:
  - **[Client]** Replay traces from Darshan logs (Theta - 2020)
  - **[Orchestrator]** Devise new storage-aware scheduling algorithms
  - **[Server]** Abstraction layer for converged HPC/Cloud storage resources

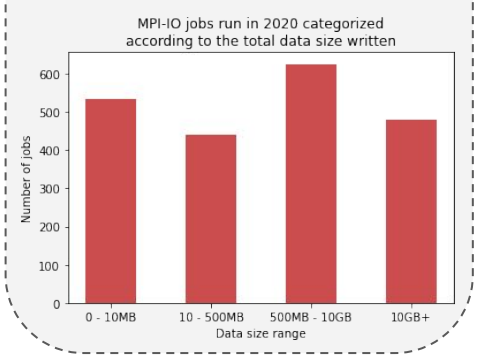


Storage allocation over hybrid infrastructures

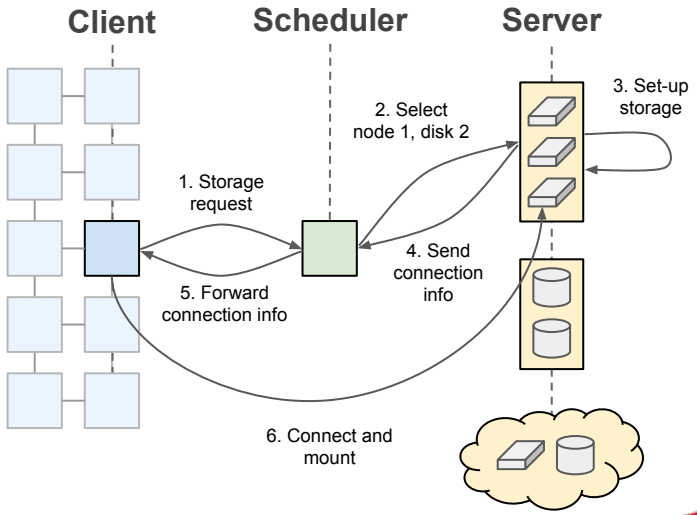
Goal: make storage resources allocable in the same way as computing resources on large-scale systems

- Simulator of a job scheduler for storage resources with three open research tracks:
  - **[Client]** Replay traces from Darshan logs (Theta - 2020)
  - **[Scheduler]** Devise new storage-aware scheduling algorithms
  - **[Server]** Abstraction layer for converged HPC/Cloud storage resources

Darshan logs
- Year: 2020
- Machine: Theta @ ANL



Data model
- Starting time
- Ending time
- Total data size read
- Total data size written
- I/O / compute time
- Data access pattern
- Job distribution (#cores, #nodes)
- ...



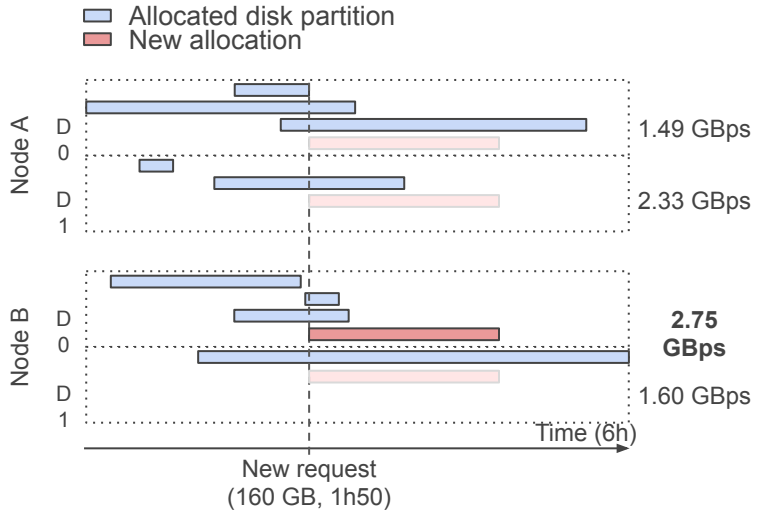
Storage allocation over hybrid infrastructures

Goal: make storage resources allocable in the same way as computing resources on large-scale systems

- Simulator of a job scheduler for storage resources with three open research tracks:
  - **[Client]** Replay traces from Darshan logs (Theta - 2020)
  - **[Scheduler]** Devise new storage-aware scheduling algorithms
  - **[Server]** Abstraction layer for converged HPC/Cloud storage resources

---

- Preliminary algorithm for storage allocation
  - Hierarchical load-balancing (nodes -> disks)
  - Best average I/O bandwidth in the worst case
    - Worst case = concurrent jobs at full capacity
  - Next steps:
    - Optimization criteria (disk usage, waiting time, max bandwidth, …)
    - Evaluation criteria



Open questions and collaboration opportunities

- Input data / clients: Darshan logs to replay traces, on-demand data services
- Scheduling algorithm minimizing I/O interference between 'storage jobs'
- Provision Cloud storage on-demand through the storage-aware scheduler (Chameleon?)

francois.tessier@inria.fr