This project has received funding from the European Union's Horizon 2020 research and innovation program through grant agreement 801101.

MAESTRO
DATA ORCHESTRATION

Dynamic Provisioning of Storage Resources: A Case Study with Burst Buffers

François Tessier, Maxime Martinasso, Matteo Chesi, Mark Klein, Miguel Gila
Swiss National Supercomputing Centre, ETH Zurich, Lugano, Switzerland

High Performance Storage Workshop (IPDPSW)
May 2020

JÜLICH FORSCHUNGSZENTRUM · cea · appentra make code parallel · ETH zürich · CSCS · ECMWF · SEAGATE · CRAY

Context

Complex workflows or frameworks in various scientific domains have increasing I/O needs

Institution	Scientific domain	Workflows	Data size (real & projection)
European Centre for Medium-Range Weather Forecasts (ECMWF)	Weather Forecast	Ensemble forecasts, data assimilation,...	12PB/year
Paul Scherrer Institute (PSI)	Synchrotron imaging	X-ray spectroscopy, high resolution microscopy,...	10-20PB/year
Cherenkov Telescope Array (CTA)	Astrophysics	Gamma Rays & Cosmic Sources,...	25PB/year

- Workloads with specific needs of data movement
  - Big data analysis, machine learning, checkpointing, in-situ, co-located processes, …
  - Multiple data access pattern (model, layout, data size, frequency)

Context

Scientific domains require more and more often varied data managers (object-based storage, database, …)

- Data management inside a workflow usually relies on a global shared parallel file system
  - Unique data access semantic (POSIX)
  - Performance variability
- Workflow specific data managers are installed on a use case basis

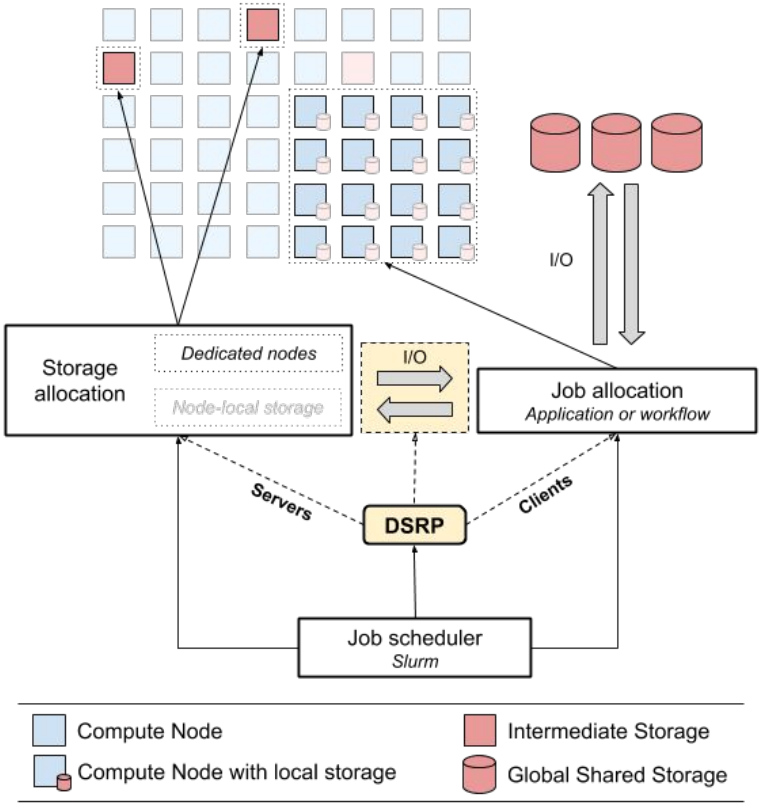
Limited support and reduced capacity

OR

Specialized and expensive

Dynamic Resource Provisioning

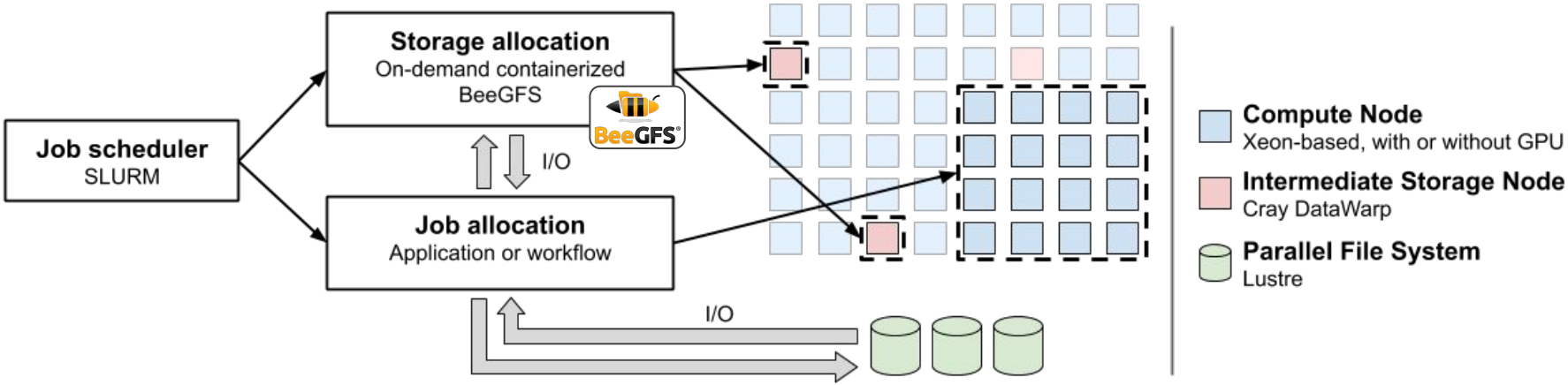
- Provisioning of storage system at job level:
  - Storage available during the job lifetime
  - Storage resources dedicated to a job (isolation)
- Dynamically supply a data management system on top of those resources
  - Several types supported: file system (current), object-based storage, database
  - Containerized data management services
  - Deployment integrated at a job scheduler level



MAESTRO DATA ORCHESTRATION

Proof of concept: BeeGFS on Cray DataWarp

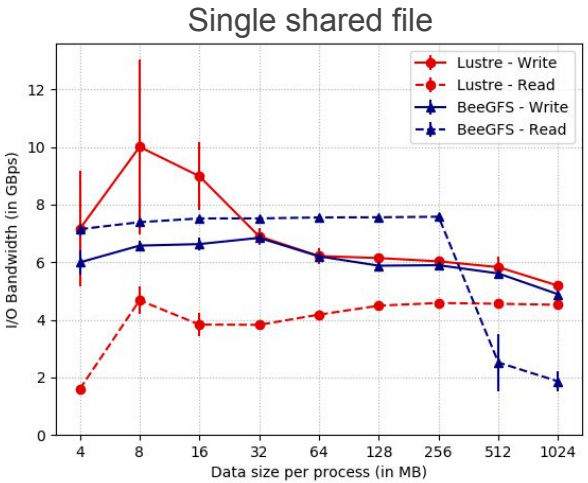
- Repurposing Cray DataWarp nodes
- Get an allocation of intermediate storage nodes along with compute nodes
- Deploy a well-sized BeeGFS across disks on DataWarp nodes
- Configure the compute nodes to act as clients of the BeeGFS instance

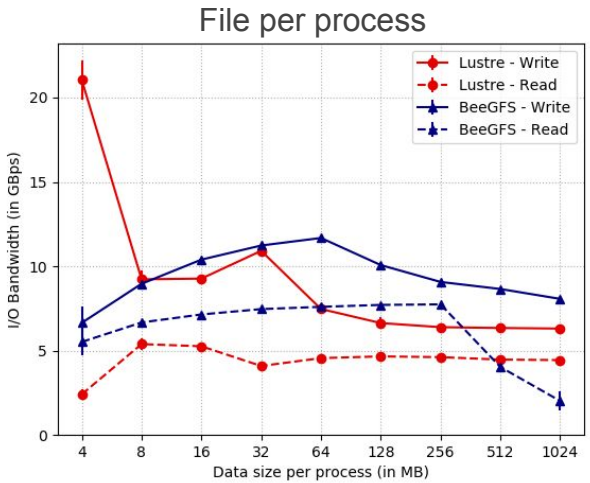


MAESTRO DATA ORCHESTRATION

Performance Evaluation

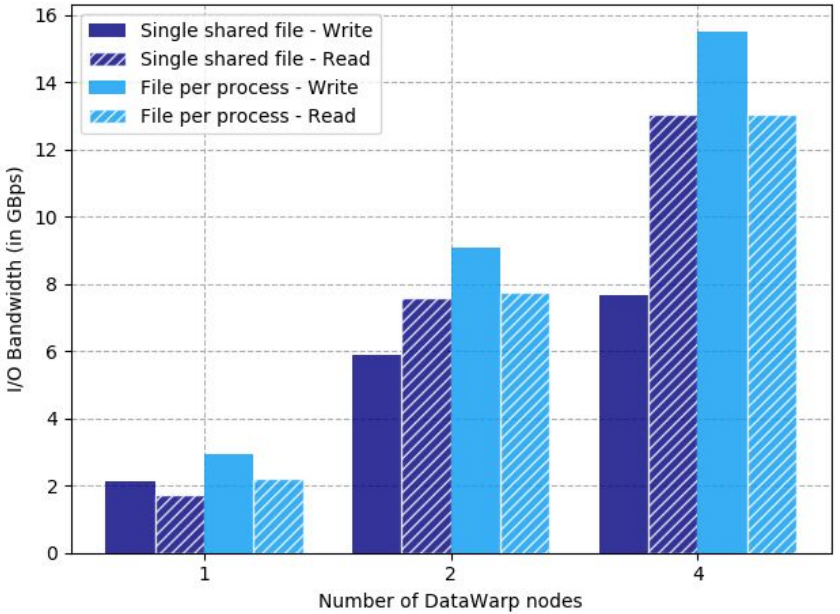
- Dom, Cray XC50 system with DataWarp at CSCS
  - Test and development system of Piz Daint (27PFlops)
  - 8 nodes with two 18-cores Intel Broadwell CPU and 64GB of DRAM
  - 4 DataWarp nodes each with three 5.9TB PCIe SSD
- On demand-BeeGFS (2 DW nodes) VS Lustre file system (Sonexion 1600, 2 OSTs)
- IOR benchmark: independent I/O, 10 runs





Performance Evaluation

- Small-scale study of… scalability
- IOR from 8 compute nodes (36 ppn)
  - 256MB written/read per process
- Dynamically provisioned BeeGFS
  - From 1 to 4 nodes
  - Ratio metadata:storage server per node kept to 1:2
- Reasonable scalability overall
  - Except SSF - write



MAESTRO DATA ORCHESTRATION

Conclusion

- Proof of concept of a mechanism to dynamically provision data managers on top of intermediate storage resources
  - Focused on containerized BeeGFS + DataWarp
- Promising performance and scalability with IOR and the I/O kernel of a real application
- Portability on different types of hardware and systems
- **Next steps**
  - Integration within the job scheduler (prolog/epilog scripts)
  - Configurable system for deployment: architecture's description, data manager-specific settings, …
  - Extends to other data managers packaged in a unique container

Acknowledgment

- This work is part of the MAESTRO EU Project
- 3-year European project, started in September 2018
- **Middleware library that automates data movement across diverse memory systems**
- https://www.maestro-data.eu/

Conclusion

Contacts

François Tessier, Maxime Martinasso, Matteo Chesi, Mark Klein, Miguel Gila
Swiss National Supercomputing Centre, ETH Zurich, Lugano, Switzerland

{firstname}.{lastname}@cscs.ch

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

MAESTRO

DATA ORCHESTRATION

JÜLICH FORSCHUNGSZENTRUM · cea · appentra make code parallel · ETH zürich · CSCS · ECMWF · SEAGATE · CRAY