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

MAESTRO
DATA ORCHESTRATION

Dynamically Provisioning Cray DataWarp Storage

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

Cray User Group Meeting 2019
Montréal, Canada

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)

MAESTRO DATA ORCHESTRATION

Context

- But I/O performance is decreasing!

Criteria	2007	2017	Relative Inc./Dec.
Name, Location	BlueGene/L, USA	Sunway TaihuLight, China	N/A
Theoretical perf.	596 TFlops	125,436 TFlops	× 210
#Cores	212,992	10,649,600	× 50
I/O bw	128 GBps	288 GBps	× 2.25
I/O bw/core	600 kBps	27 kBps	÷ 22.2
I/O bw/TFlop	214 MBps	2.30 MBps	÷ 93.0

- 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

Context

- But I/O performance is decreasing!

System Specs	TITAN	SUMMIT	FRONTIER
Peak Performance	27 PF	200 PF	>1.5 EF (**✕ 7.5**)
Storage	32 PB, 1 TB/s Lustre file-system	250 PB, 2.5 TB/s GPFS	**2-4x** performance and capacity of Summit's I/O subsystem. Frontier will have near node storage like Summit.

NEW!

Source: https://www.olcf.ornl.gov/frontier/

- 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

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

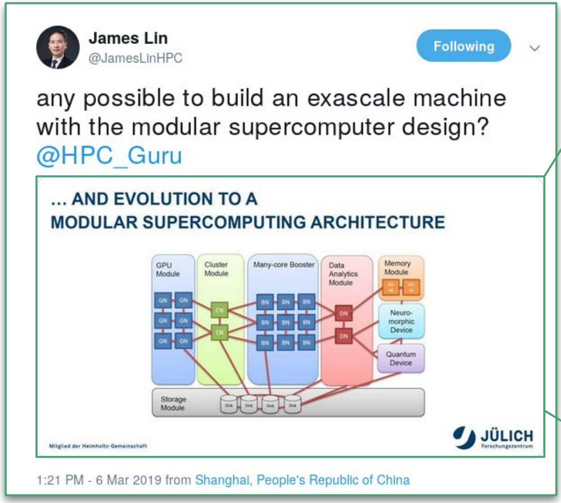
Limited support and reduced capacity

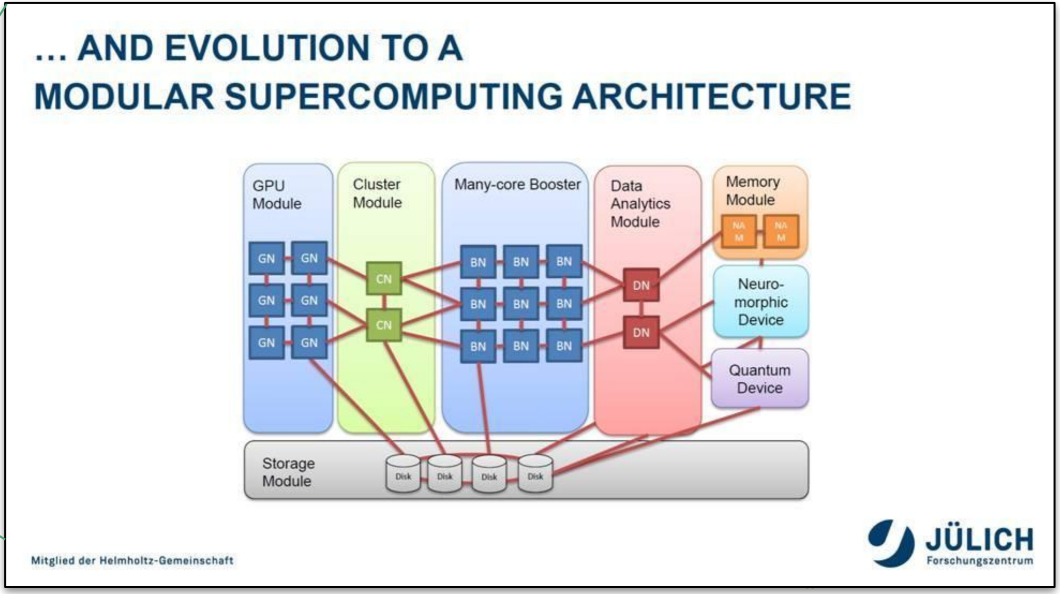
OR

Specialized and expensive

Context

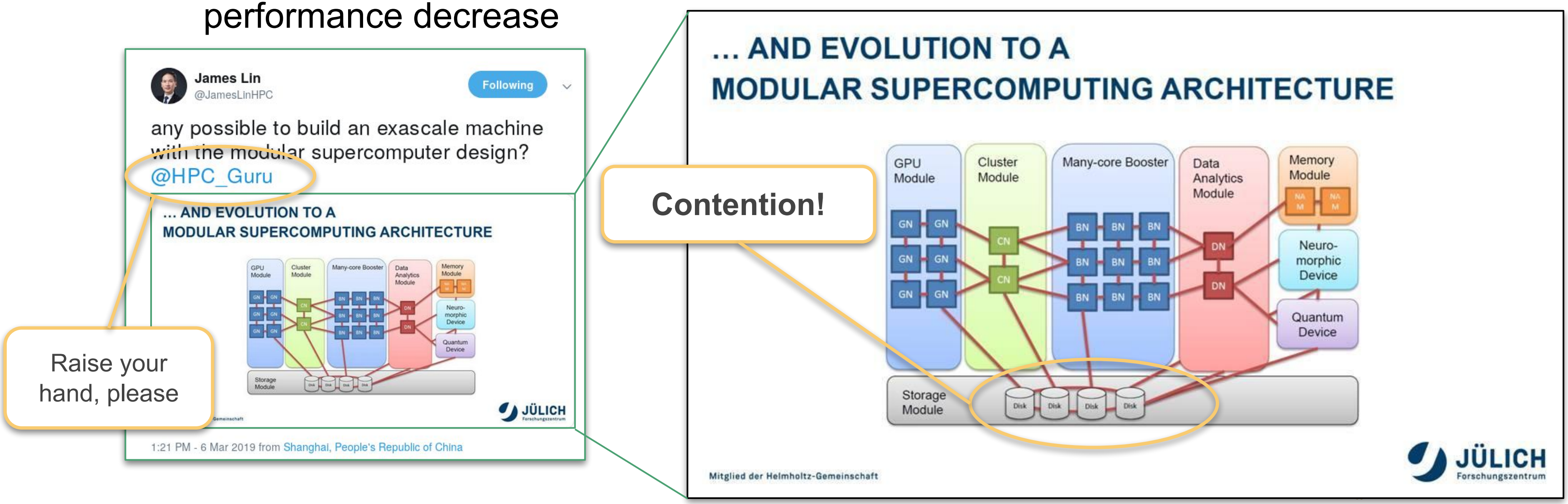
- On the HPC center side, not feasible to support a large variety of data management systems
- … and hard to provide dedicated storage resources
    - Usually, data resources are shared while compute resources are exclusive
    - Shared storage resources are subject to contention and high unexpected performance decrease





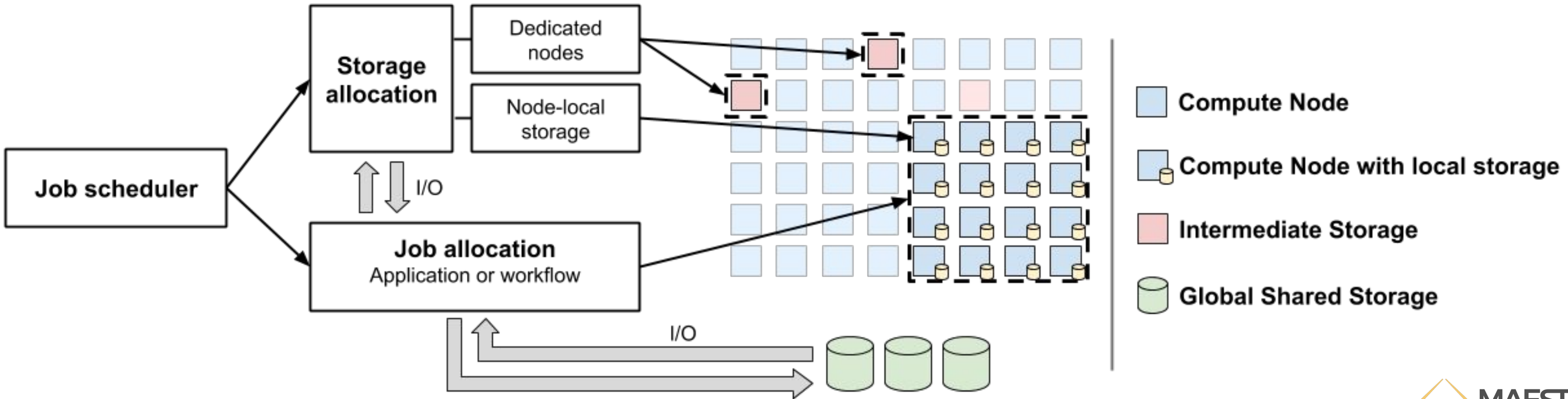
Context

- On the HPC center side, not feasible to support a large variety of data management systems
- … and hard to provide dedicated storage resources
  - Usually, data resources are shared while compute resources are exclusive
  - Shared storage resources are subject to contention and high unexpected performance decrease



Dynamic Resource Provisioning

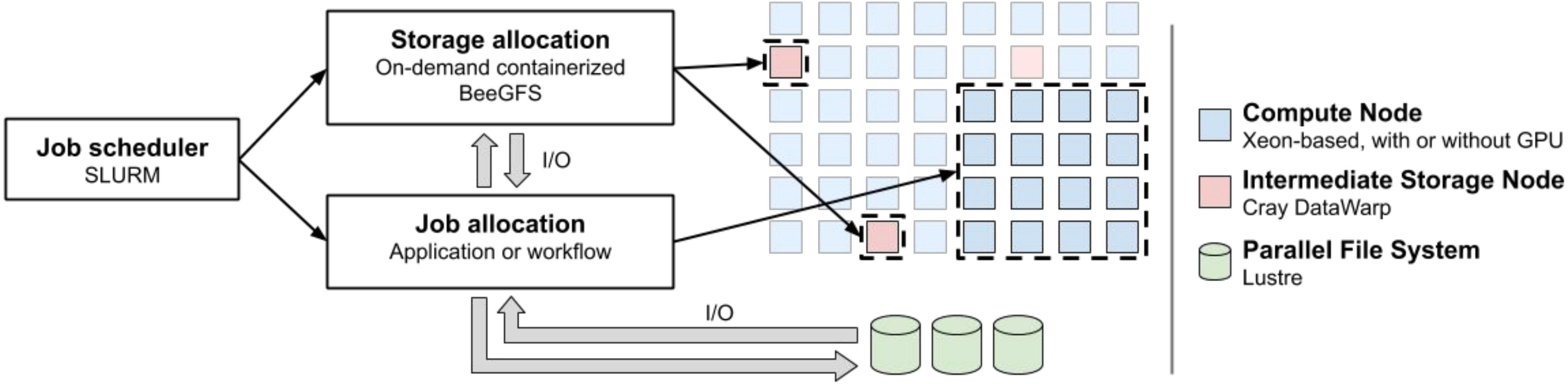
- Provisioning of storage system at job level:
  - Storage available during the job lifetime
  - Storage resources dedicated to a job (isolation)
- Dynamically supply a data management system on top of those resources
  - Several types supported: file system, object-based storage, database
  - Containerized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



MAESTRO
DATA ORCHESTRATION

Our Approach

- 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

Accessing DataWarp Nodes

Standard implementation of DataWarp

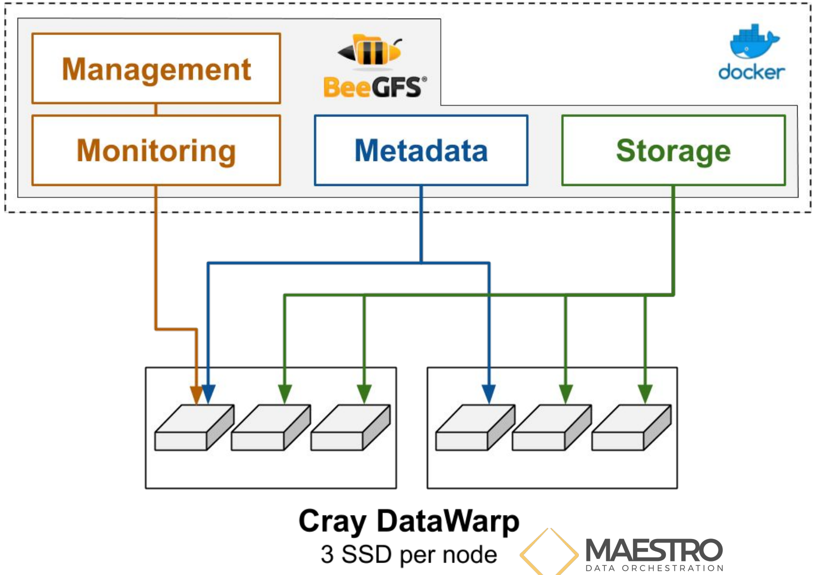
- Projection of DataWarp storage onto the compute node (through DVS)

Repurposing

- System customization to reconfigure the nodes
  - From hidden **service** nodes to standard **compute** nodes
  - Mapping of a compute node image to boot with
- Setup the local NVMe storage
  - XFS file system
  - Mount point with permissions granted to any user
- New SLURM constraint: `storage`

On-demand containerized BeeGFS

- BeeGFS: POSIX-compliant parallel file system based on a client-server architecture
  - Server-side: management, monitoring, metadata, storage
  - Client-side: kernel-space client, monitoring visualization
- Servers bundled in a Docker container and deployed with Sarus, a container runtime system
  - 1 metadata and 2 storage servers per DataWarp node
- Mount point on clients (compute nodes)
  - Kernel module required
  - Special privileges to mount BeeGFS



MAESTRO
DATA ORCHESTRATION

Limitations

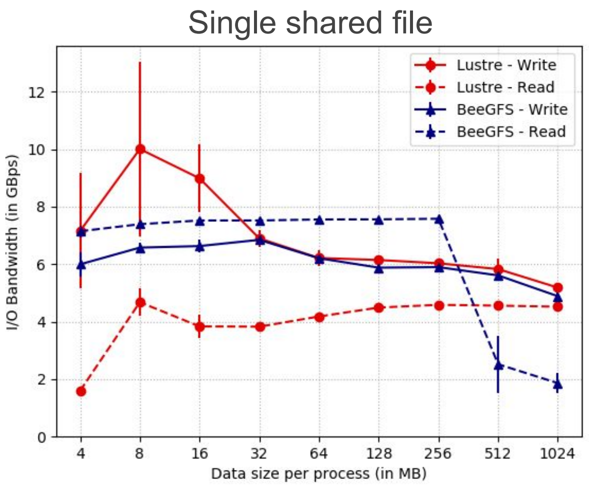
- Kernel-space file system such as BeeGFS implies special privileges
  - Load/unload kernel module: `modprobe [-r] beegfs`
  - Mount BeeGFS on compute nodes: `mount -t beegfs [...] $HOME/beegs` [...]
  - ➢ Module pre-installed on nodes?
  - ➢ Prolog script for file-system creation and mount point?

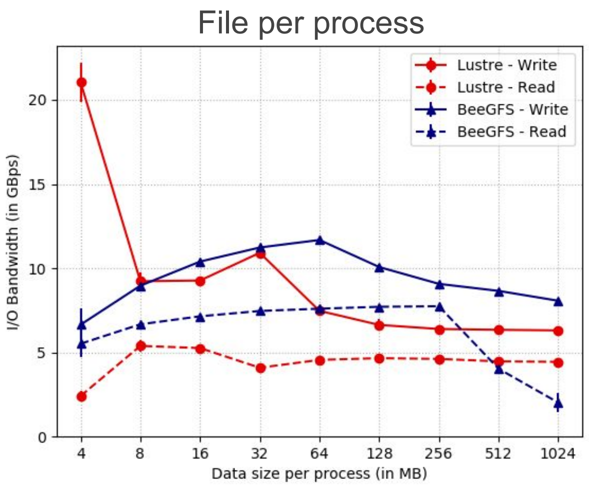
- Fresh data manager provisioned meaning no data available
  - ➢ Stage-in/stage-out phase, such as on native DataWarp?
  - ➢ Should this step be counted in the allocation time?

- Trade-off between capacity and capability
  - Better I/O bandwidth implies more disks and possibly capacity wasted

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

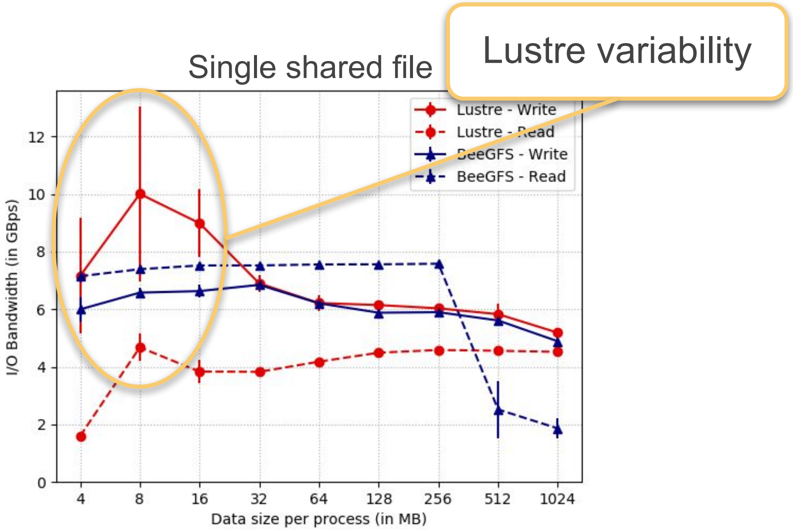


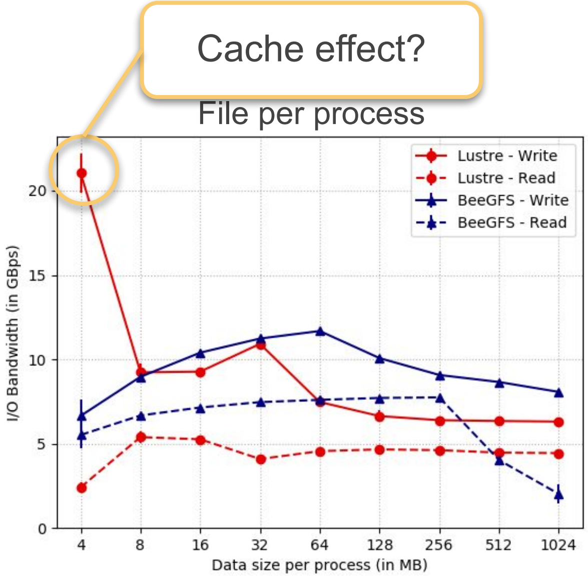


MAESTRO DATA ORCHESTRATION

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

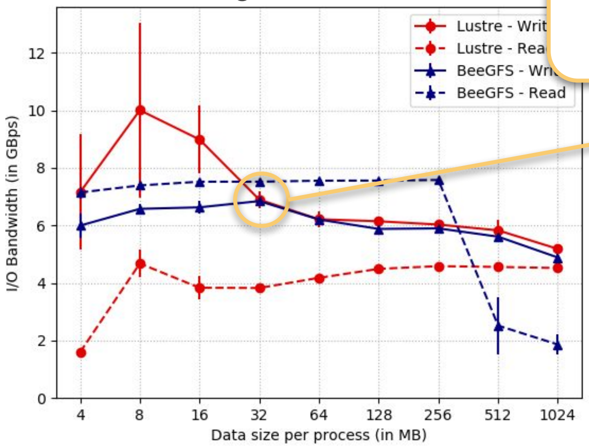




Performance Evaluation

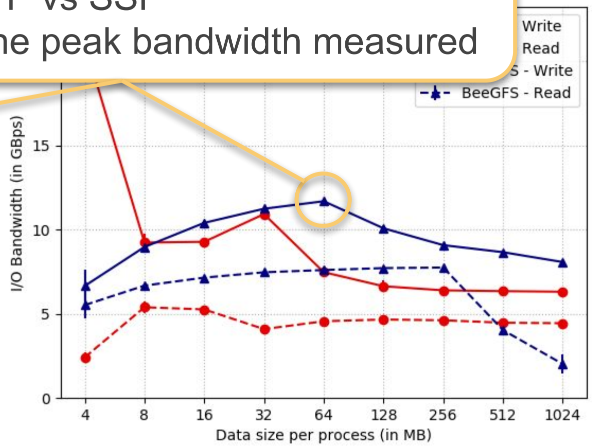
- Dom, Cray XC50 system with DataWarp at CSCS
  - Test and development system of Piz Daint (27PFlops)
  - 8 nodes with two 18-cores Intel Broadwell CPU and 64GB of DRAM
  - 4 DataWarp nodes each with three 5.9TB PCIe SSD
- On demand-BeeGFS (2 DW nodes) VS Lustre file system (Sonexion 1600, 2 OSTs)
-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

Single shared file



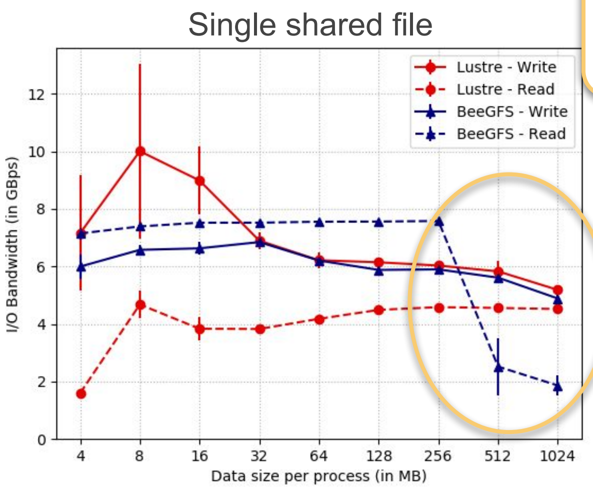
Peak write bandwidth:

- **+70%** FPP vs SSF
- **93%** of the peak bandwidth measured

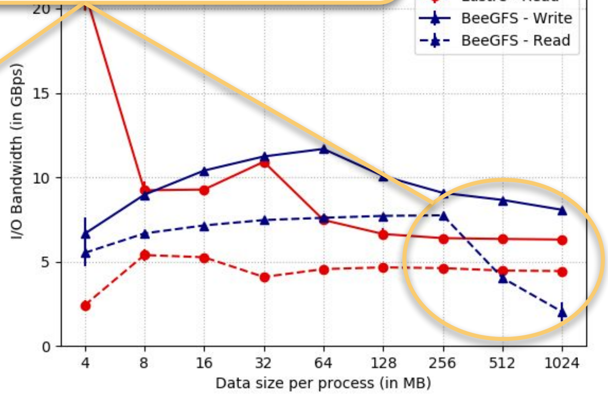


Performance Evaluation

- 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



Not enough memory on DW nodes (64GB) for the caching mechanism



MAESTRO DATA ORCHESTRATION

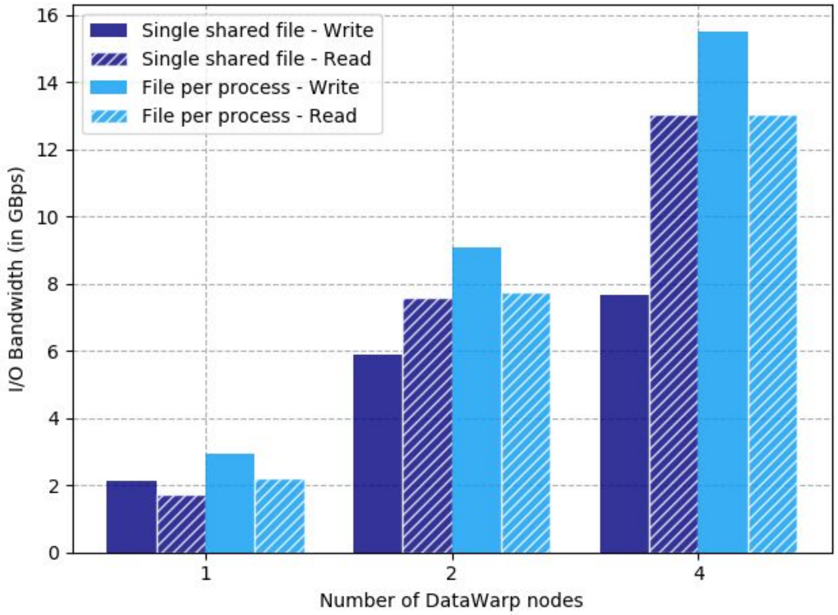
Performance Evaluation

- 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) versus global Lustre file system (2 OSTs)
- *mdtest* benchmark

		BeeGFS	Lustre	
Target	**Operation**	**Ops**		**L/B**
Directory	Creation	8276.43	37222.57	✕ 4.5
	Stat	5301788.76	182330.42	÷ 29.1
	Removal	12967.02	38732.00	✕ 3.0
File	Creation	6618.37	22916.15	✕ 3.5
	Stat	144410.46	169140.32	✕ 1.2
	Read	22541.08	45181.55	✕ 2.0
	Removal	8431.71	35985.96	✕ 4.3
Tree	Creation	2183.40	3310.42	✕ 1.5
	Removal	125.23	1298.55	✕ 10.4

Performance Evaluation

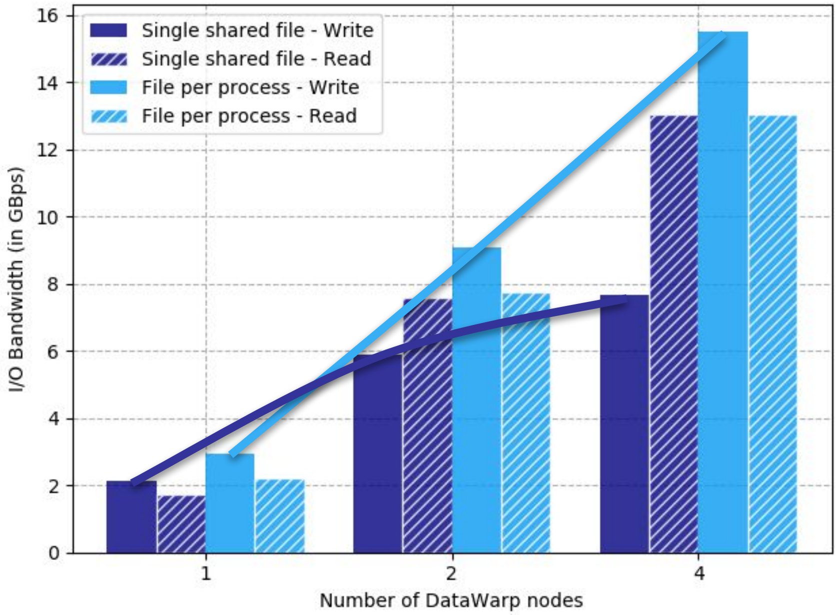
- Small-scale study of… scalability
- IOR from 8 compute nodes (36 ppn)
  - 256MB written/read per process
- Dynamically provisioned BeeGFS
  - From 1 to 4 nodes
  - Ratio metadata:storage server per node kept to 1:2
- Reasonable scalability overall
  - Except SSF - write



MAESTRO DATA ORCHESTRATION

Performance Evaluation

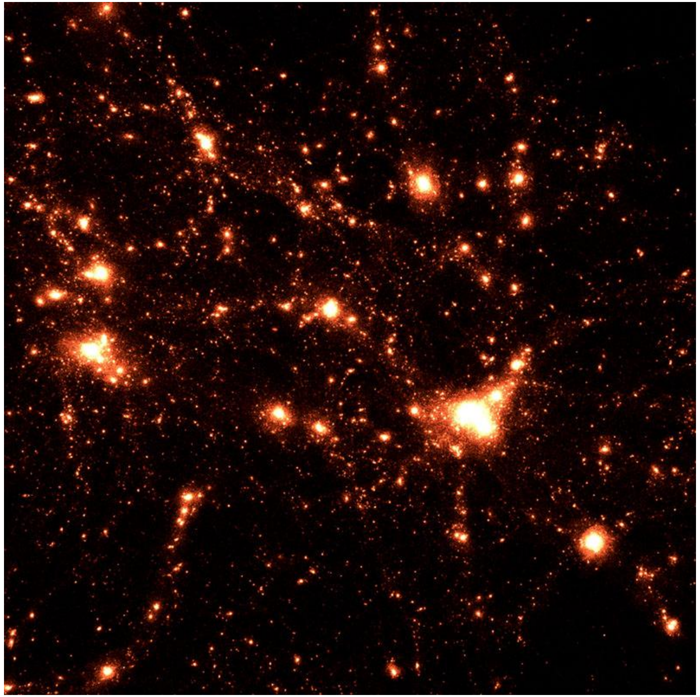
- Small-scale study of… scalability
- IOR from 8 compute nodes (36 ppn)
  - 256MB written/read per process
- Dynamically provisioned BeeGFS
  - From 1 to 4 nodes
  - Ratio metadata:storage server per node kept to 1:2
- Reasonable scalability overall
  - Except SSF - write



MAESTRO
DATA ORCHESTRATION

Performance Evaluation - HACC-IO

- I/O part of a large-scale cosmological application simulating the mass evolution of the universe with particle-mesh techniques
- Each process manages particles defined by 9 variables (38 bytes)
  - XX, YY, ZZ, VX, VY, VZ, phi, pidandmask
- Single shared checkpointing file with data in an array of structure data layout
- Average and standard deviation on 10 runs



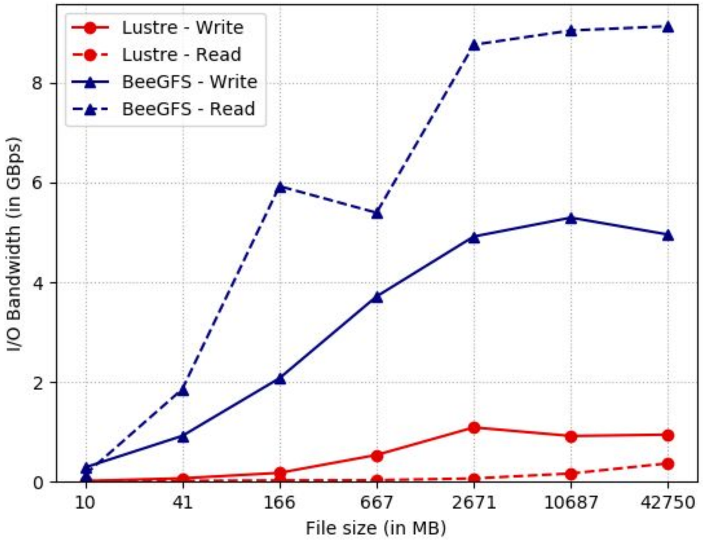
Credits: Silvio Rizzi and Joe Insley, Argonne National Laboratory

Performance Evaluation - HACC-IO

- HACC-IO from 8 compute nodes, 36 ppn
- BeeGFS (2 DW) vs Lustre (2 OSTs)

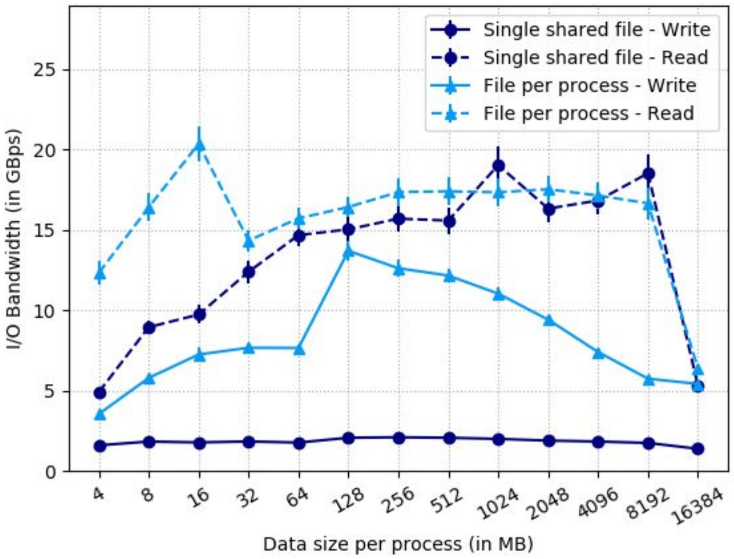
- BeeGFS peak **write** bandwidth: **5.3GBps**
  read bandwidth: **9.1GBps**

- As expected (previous work), BeeGFS highly outperforms Lustre
  - Single shared file and array of structure data layout is a bad combination on Lustre



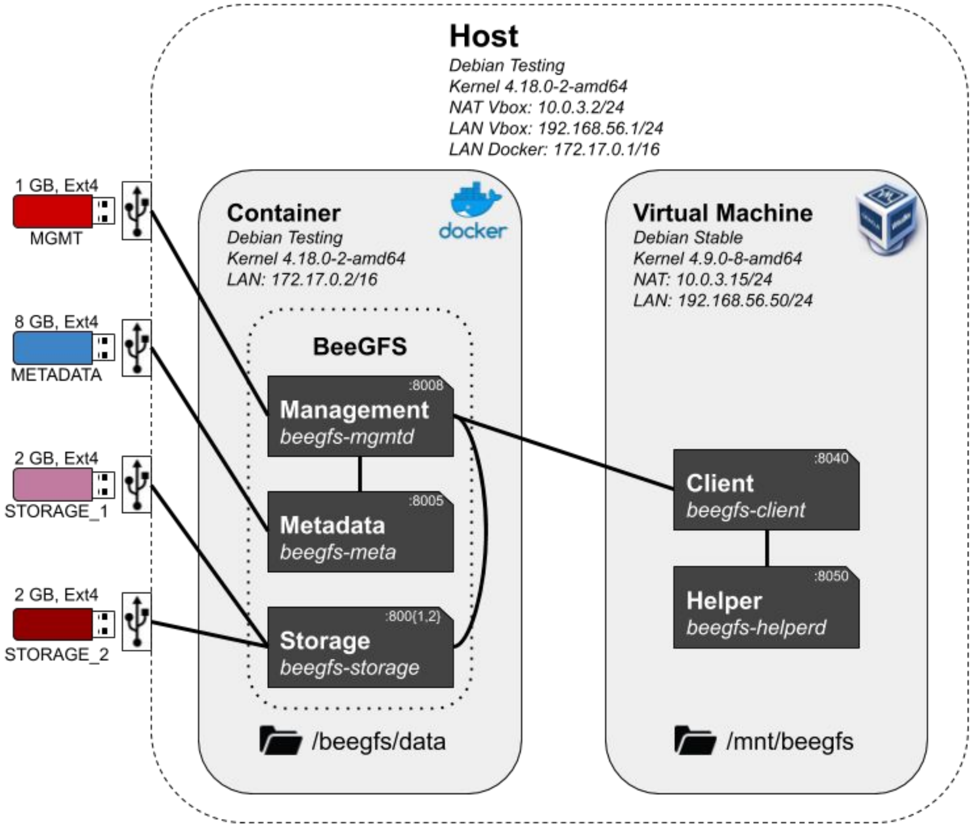
MAESTRO DATA ORCHESTRATION

Portability



- Ault, testbed platform at CSCS allowing for prototyping experimental services and platforms
  - Various types of hardware
  - Safe privileged-access level for researchers
- Ault11, compute node with a 22-core Intel Xeon Gold 6152 CPU
  - 16 3D NAND NVMe disks
- Dynamically provisioned BeeGFS
  - 1 disk for management and monitoring
  - 2 disks for metadata
  - 5 disks for storage
- Peak **read** bandwidth: **20.36GBps**
- Peak **write** bandwidth: **13.70GBps**
- In line with values communicated by the vendor

Portability For Fun



How to give a second lease of life to HPC conference USB Keys?

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

Thank you for your attention!

francois.tessier@cscs.ch

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

MAESTRO

DATA ORCHESTRATION