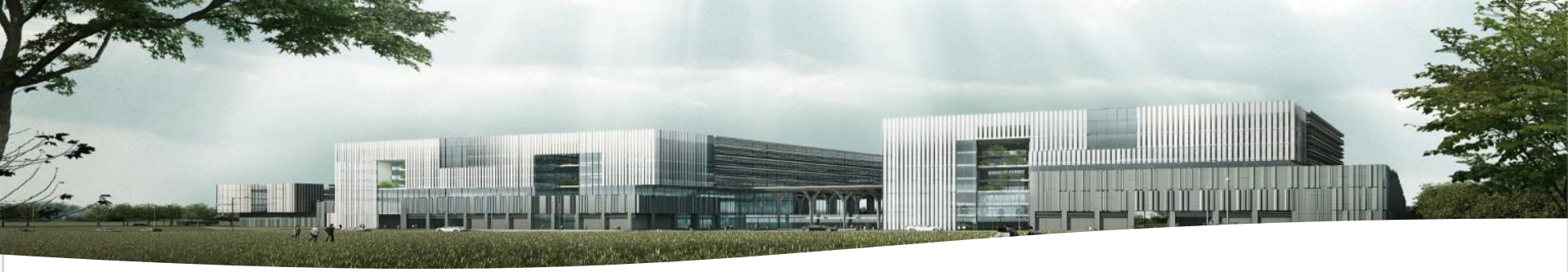


China Lustre User Group (China LUG), Beijing, October 15, 2019

Providing Architecture Support for On-Demand Services on Large-Scale HPC Platform

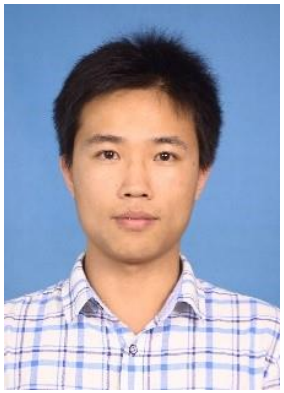
Lingfang Zeng (曾令仿)

Wuhan National Laboratory for Optoelectronics (WNLO)

Huazhong University of Science and Technology (HUST)

With Contributions from

- Wen Cheng @ **HUST**
- Xi Li @ **DDN / Whamcloud**
- Andreas Dilger @ **DDN / Whamcloud**
- André Brinkmann @ **JGU**

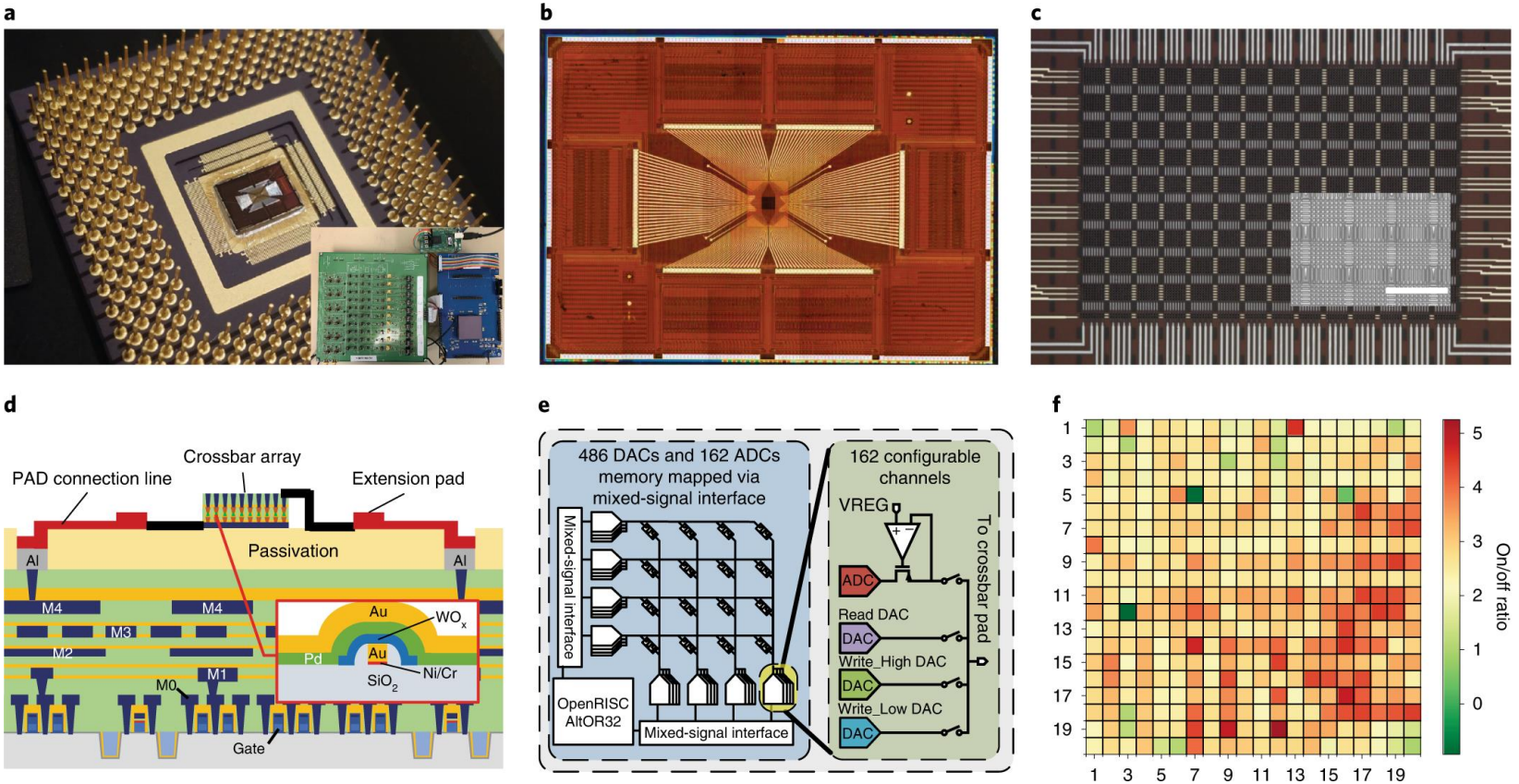



ACM A.M. Turing Award 2017: John L. Hennessy and David A. Patterson, A New Golden Age for Computer Architecture, Communications of the ACM, Volume 62 Issue 2, pp.48-60, February 2019.

Feature	Version	Remark
Progressive File Layout (PFL)	Lustre 2.10	"Fast"
Data on MDT (DoM)	Lustre 2.11	
Persistent Client Cache (PCC)	Lustre 2.13	
Hierarchical Storage Management (HSM)	Lustre 2.5	Reliability
File Level Redundancy (FLR)	Lustre 2.11	

Lustre architecture as an example

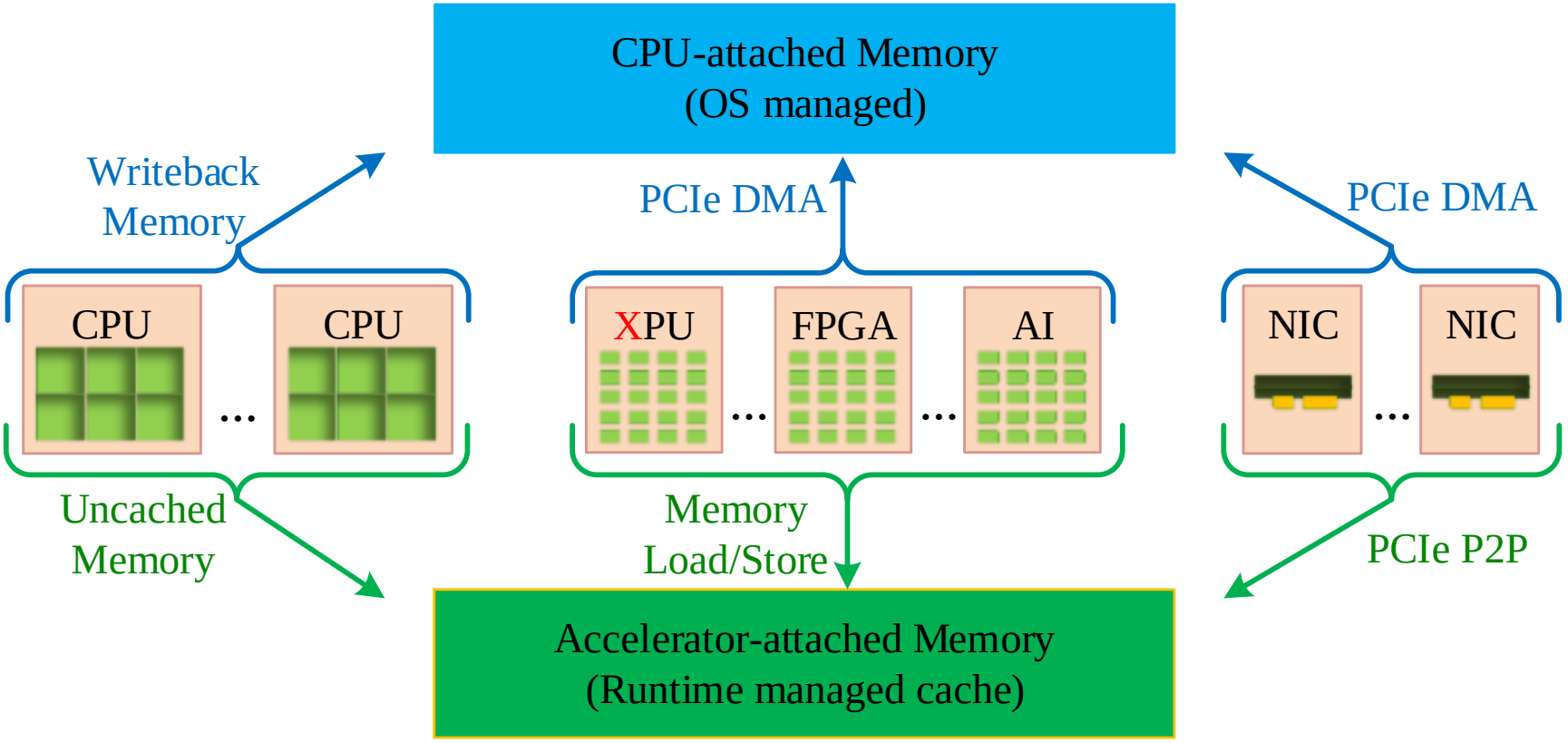
Emerging Memory Technologies



Source: *Nature Electronics*, vol.2, pp.290–299 (2019). A fully integrated reprogrammable memristor–CMOS system for efficient multiply–accumulate operations, by Fuxi Cai et. al.

Emerging Transfer Technologies

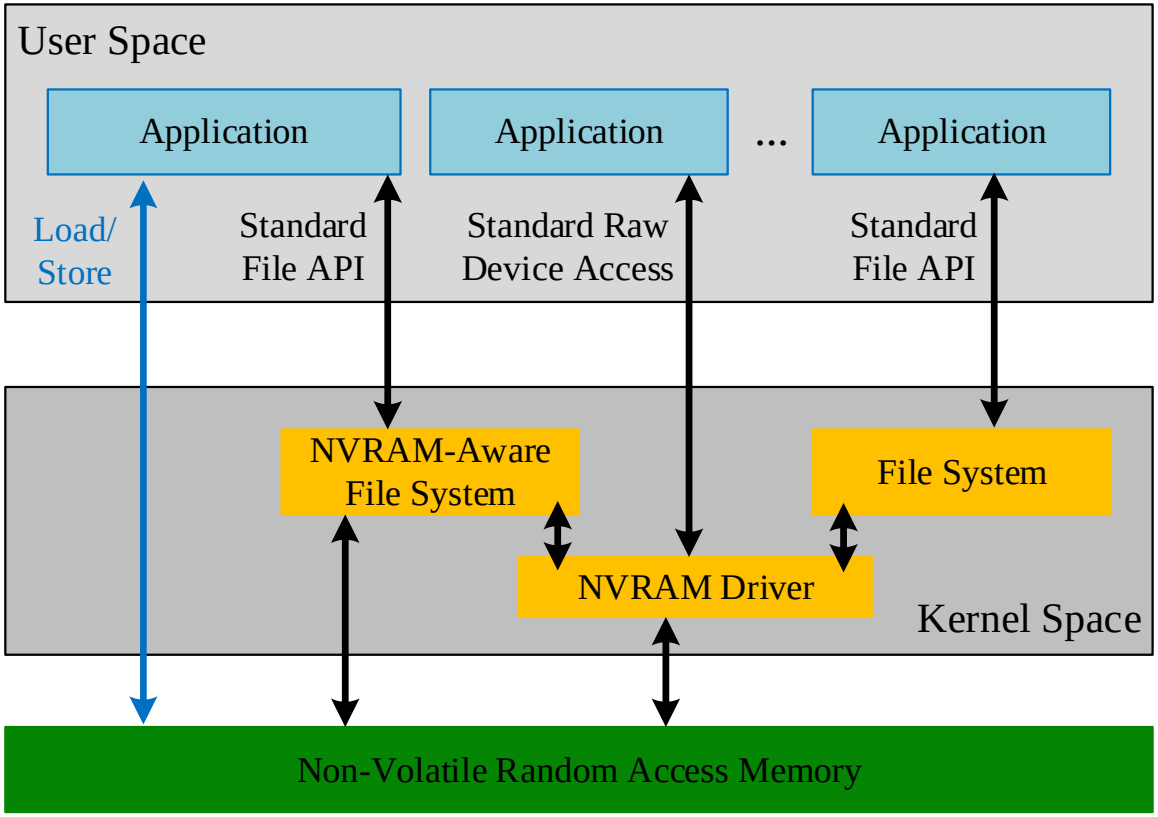
- Compute Express Link (CXL)



Reference: https://venturebeat.com/2019/03/11/intel-and-tech-companies-create-cxl-standard-to-link-datacenter-cpus-to-accelerators/

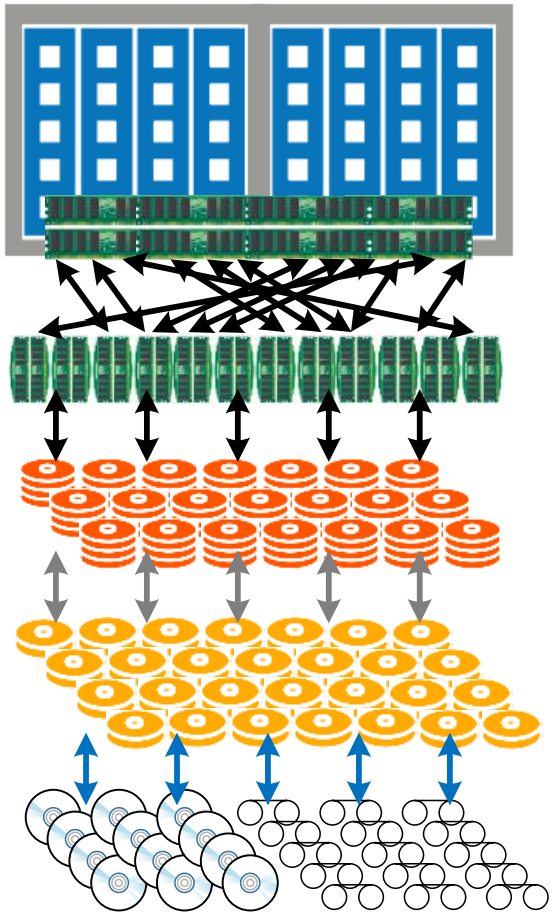
NVRAM Programming Model

- Use NVRAM like a Disk (e.g. RAMDisk)
- Use NVRAM like an SSD (no page cache, e.g. DAX)
  - **Typical context switch range: above 2-3 us**
- MMU Mappings
  - **Typical NUMA range: 0 -200 ns**



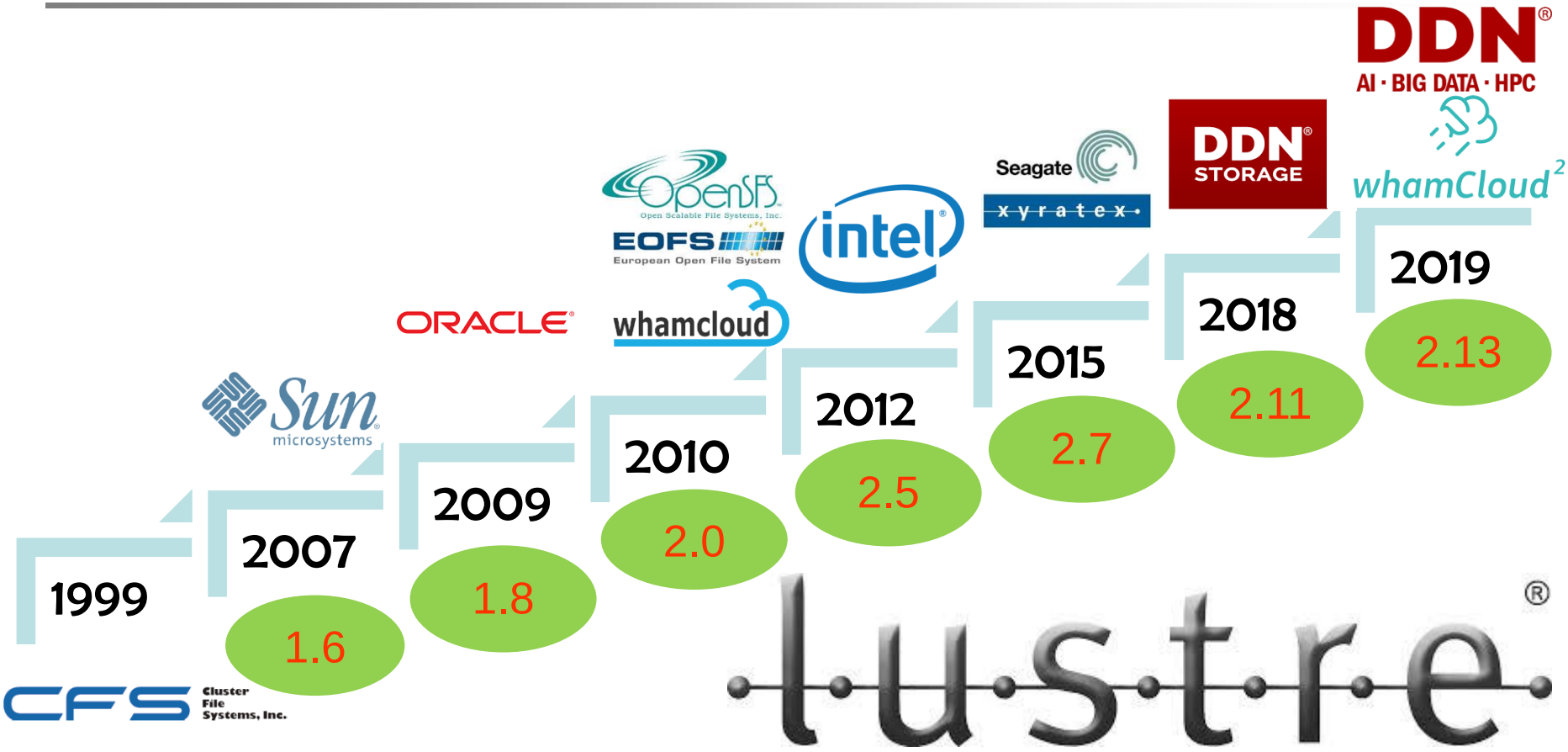
Hierarchical Storage Management Tier

- Compute servers
  - HBM
  - NVRAM/SCM
- Performance storage
  - DRAM
  - SSD
  - (performance HDD)
- Capacity storage
  - DRAM
  - Capacity HDD



Lang's Law: the more tiers, the more tears

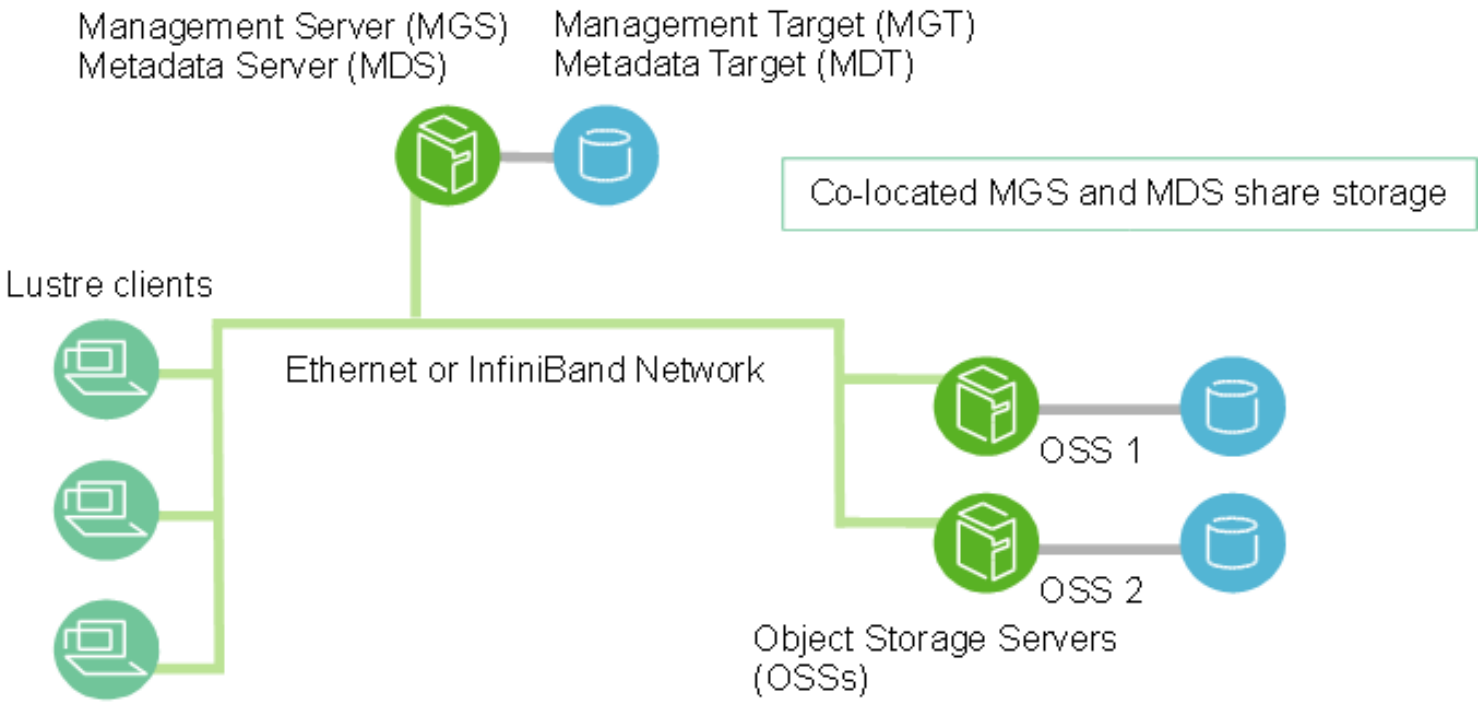
About Lustre* Journey



Reference: https://en.wikipedia.org/wiki/Lustre_(file_system)

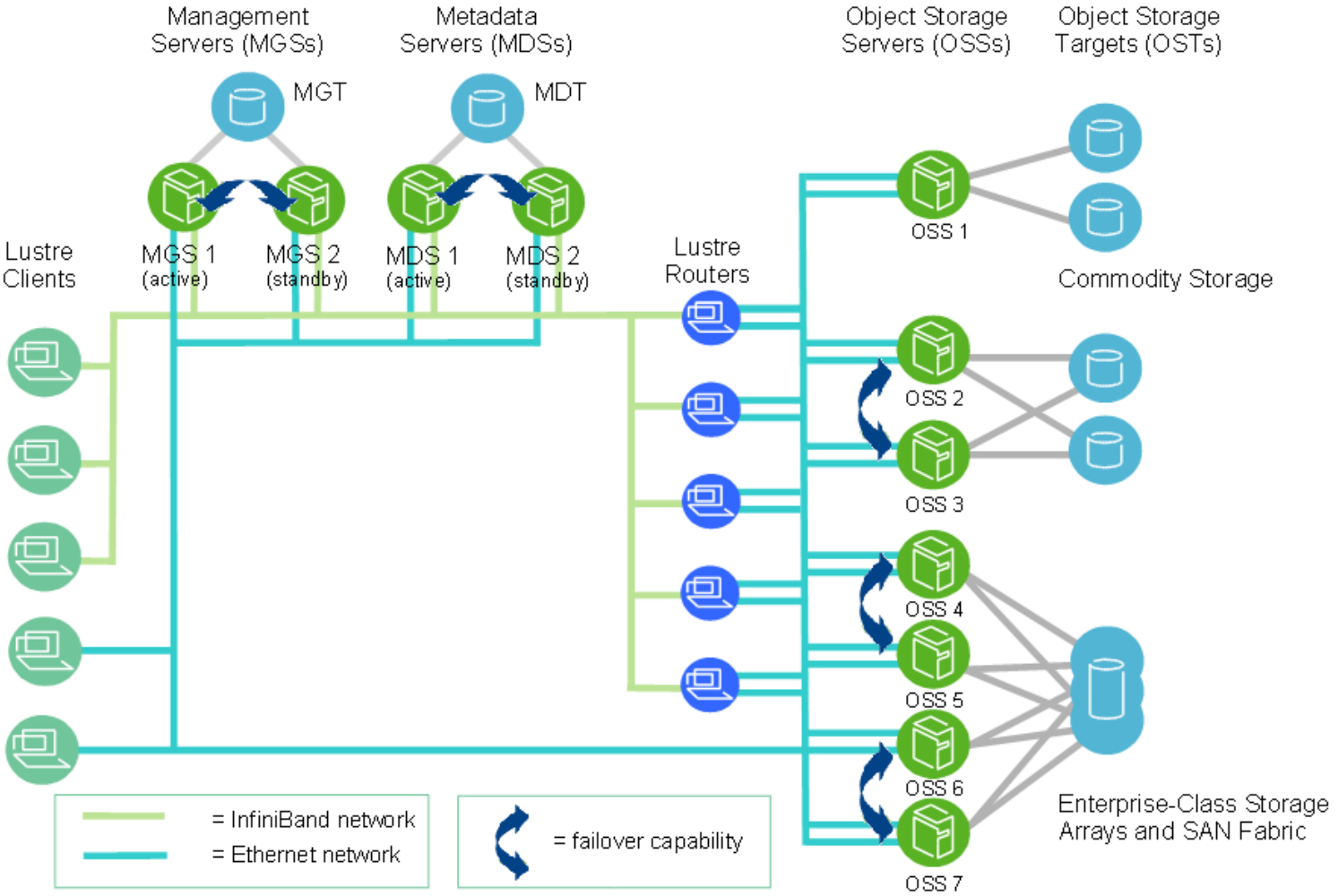
Lustre Basic Architecture

- Lustre file system components in a basic cluster



Reference: http://doc.lustre.org/lustre_manual.xhtml

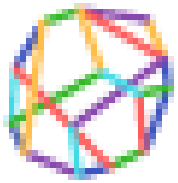
Lustre Architecture at Scale



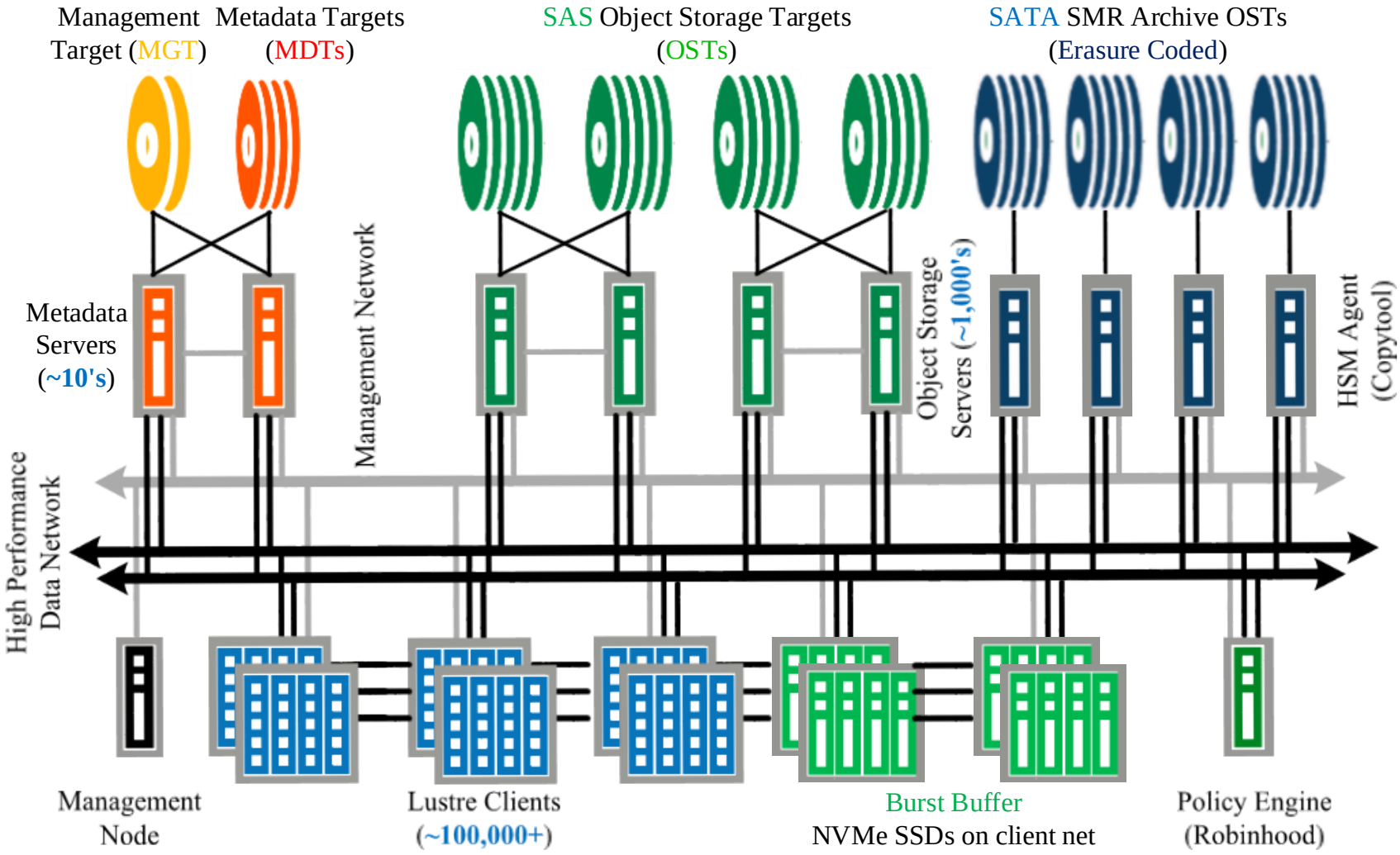
Reference: http://doc.lustre.org/lustre_manual.xhtml

"Fast" Architecture

- Progressive File Layout (PFL, Lustre 2.10)
- Data-on-MDT (DoM, Lustre 2.11)
- I/O forwarding
- Burst buffer
- Caching
  - LPCC (flash-based, CLUG'18, SC'19)
  - LPCC (NVRAM-based, LAD'19)
  - Lustre 2.13



Lustre Architecture (Burst Buffer)



Problems

- Over-provisioning
  - Forwarding nodes themselves are over-provisioning, e.g. 170:1~16:1
  - 1-1 applications are mostly insensitive to additional layer resources (Xu Ji, et. al., FAST'19)

Rank	Machine	Vendor	# C_node	# F_node	File system
3	Taihulight	NRCPC	40,960	240	Lustre
4	Tianhe-2A	NUDT	16,000	256	Lustre + H2FS
5	Frontera	Dell	8,008	60	Lustre
6	Piz Daint	Cray	6,751	54	Lustre + GPFS
7	Trinity	Cray	19,420	576	Lustre
12	Titan	Cray	18,688	432	Lustre
13	Sequoia	IBM	98,304	768	Lustre
14	Cori	Cray	12,076	130	Lustre + GPFS
16	Oakforest-PACS	Fujitsu	8,208	50	Lustre
20	K computer	Fujitsu	82,944	5184	FEFS (based on Lustre)

https://www.top500.org/list/2019/06/?page=1

Problems (cont.)

- On the critical path
  - I/O forwarding & Burst buffer
  - “failslow” phenomenon
- Data movement
  - Stage in/out, e.g. million files of 1000 bytes each
  - The most “space/time efficient” way is not to transfer it at all

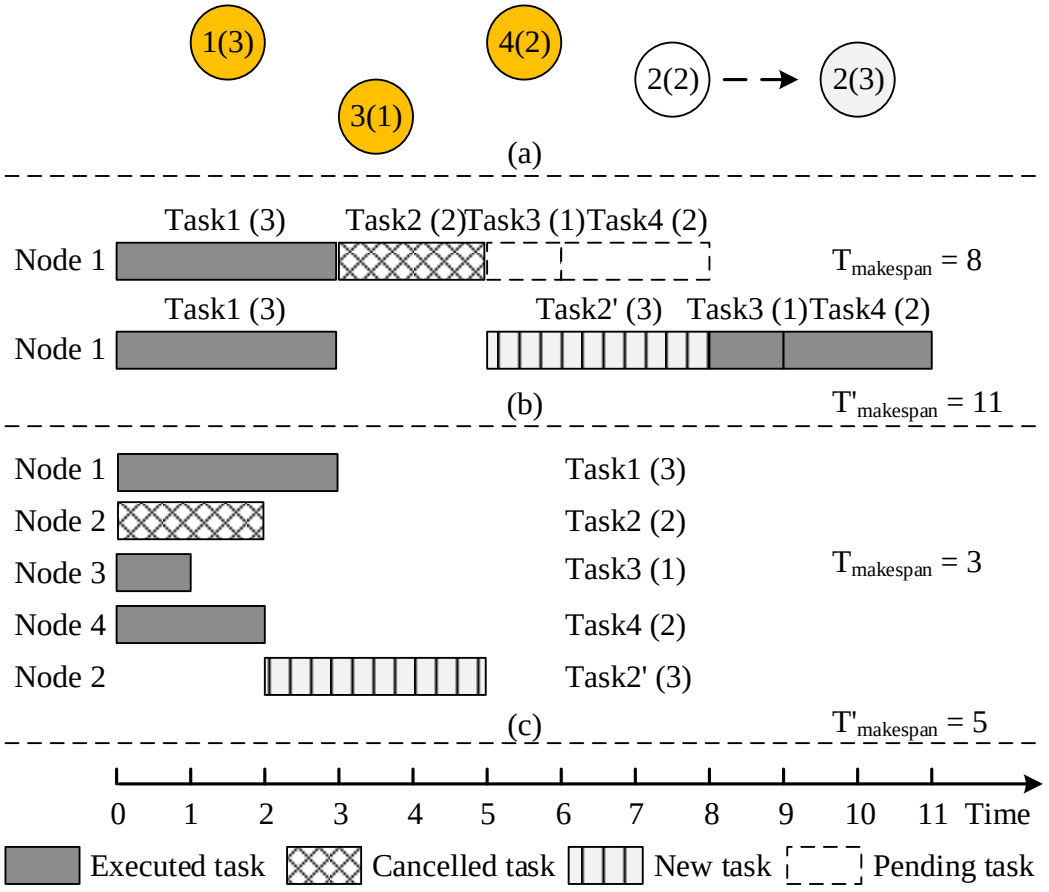
What is On-Demand Services

- Allows consumers to customize computing capabilities
  - Server time
  - Network storage
  - Automation- without requiring human interaction
- Lustre architecture has been trying to adapt to these requirements

Refer to: https://www.igi-global.com/dictionary/on-demand-services/41881

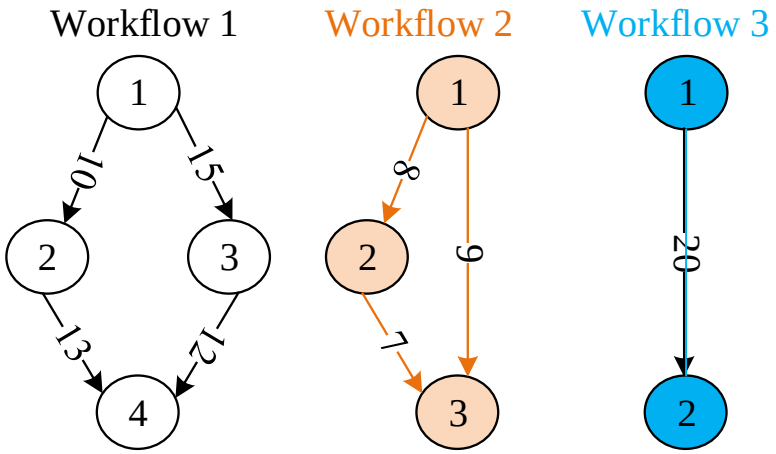
Applications (Bag-of-Task, Simulation)

- Certain vs. Uncertain



Applications (Workflow, Simulation)

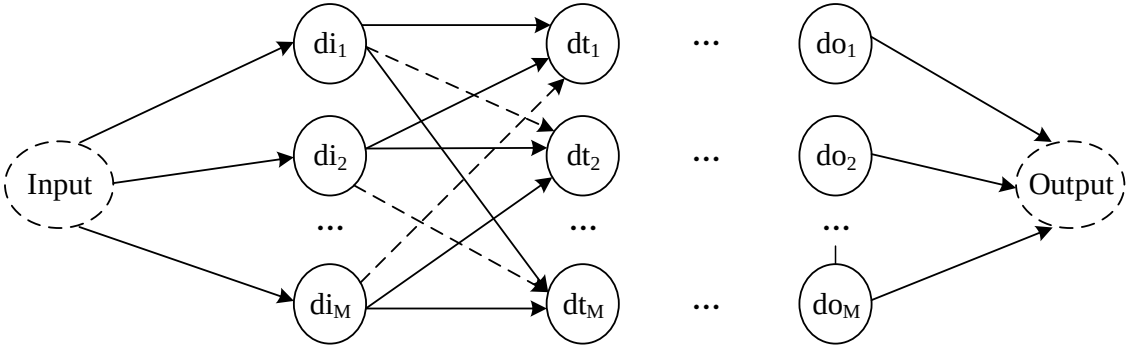
- Directed Acyclic Graph (DAG)
  - Computation cost
  - Communication cost
  - ……



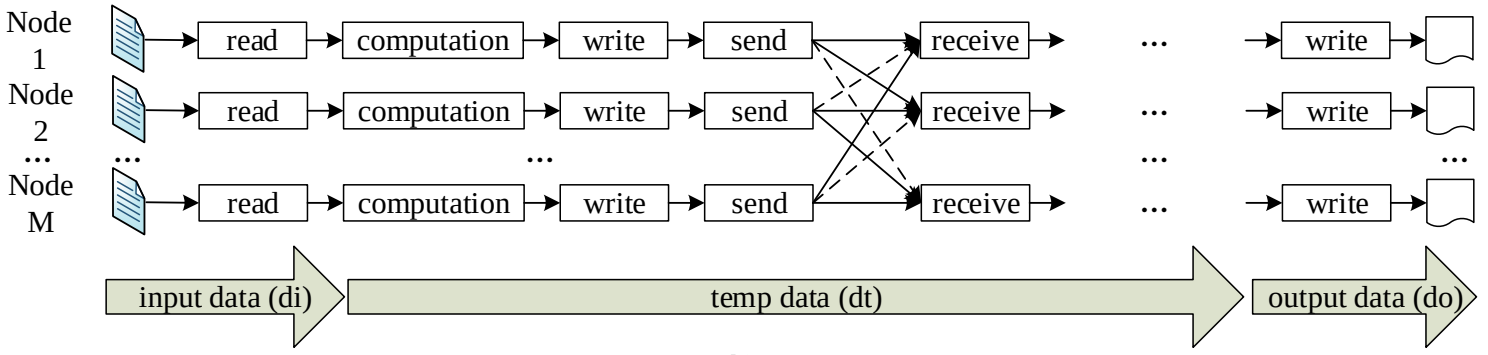
Workflow 1			Workflow 2			Workflow 3		
Task	P1	P2	Task	P1	P2	Task	P1	P2
1	8	9	1	8	13	1	8	16
2	6	4	2	14	11	2	14	18
3	10	6	3	8	6			
4	7	11						

Applications (Dataflow, Big Data)

- Model (e.g. Map/Reduce)



(a)

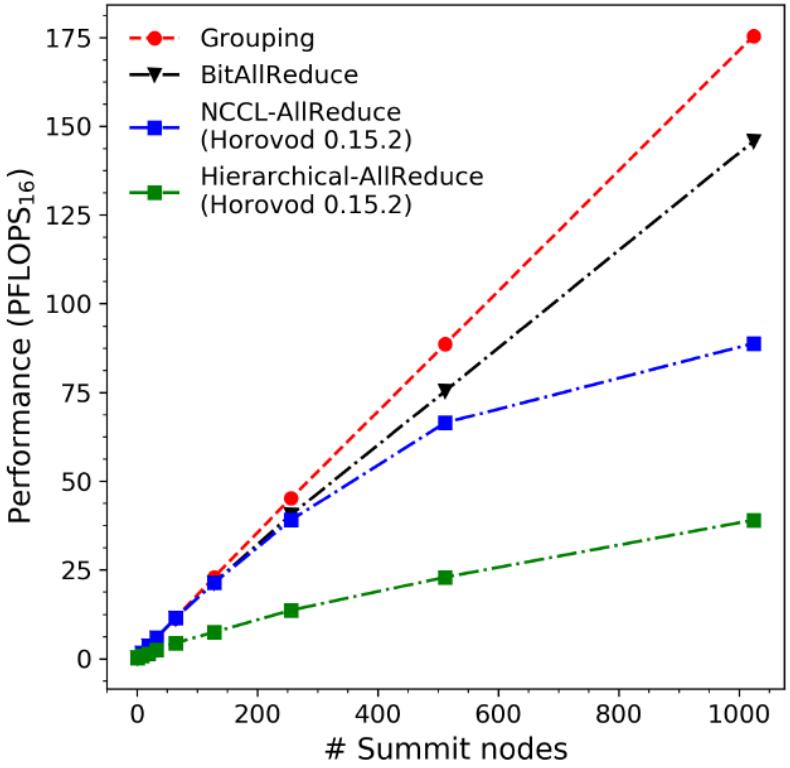


(b)

Lingfang Zeng, et. al., SABA: A security-aware and budget-aware workflow scheduling strategy in clouds, https://www.sciencedirect.com/science/article/pii/S0743731514001658

Applications (Deep Learning, AI)

- 27,600 V100 GPUs on the Summit Supercomputer
- Exascale Deep Learning (peak perf. of 2.15 $EFlops_{16}$)



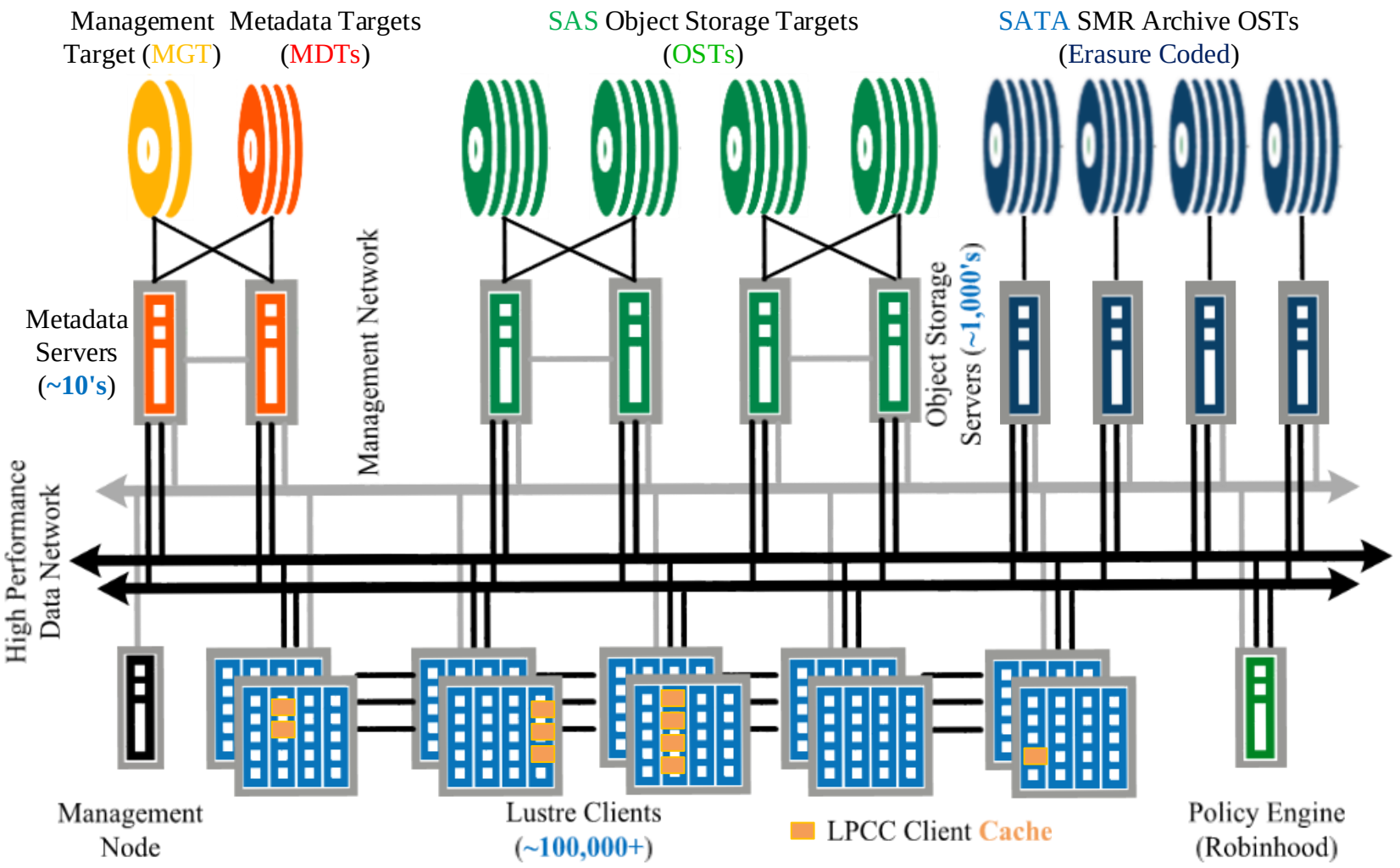
Refer to: https://arxiv.org/abs/1909.11150, 24 Sep 2019

Lustre Persistent Client Caching(LPCC)

- Large-scale testing @ DDN & JGU
- Lustre User Group (LUG) China 2018 (Keynote)
- SC'19 (CCF A)
- Lustre 2.13

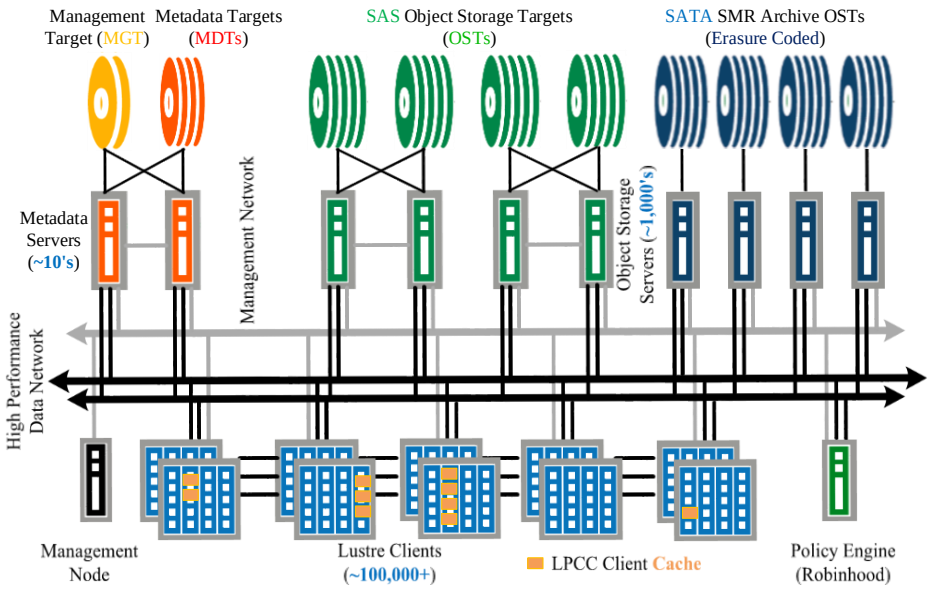
Reference: https://www.eofs.eu/events/lad19
Reference: http://lustrefs.cn/clug2018/
Reference: https://sc19.supercomputing.org/presentation/?id=pap112&sess=sess159

LPCC Architecture

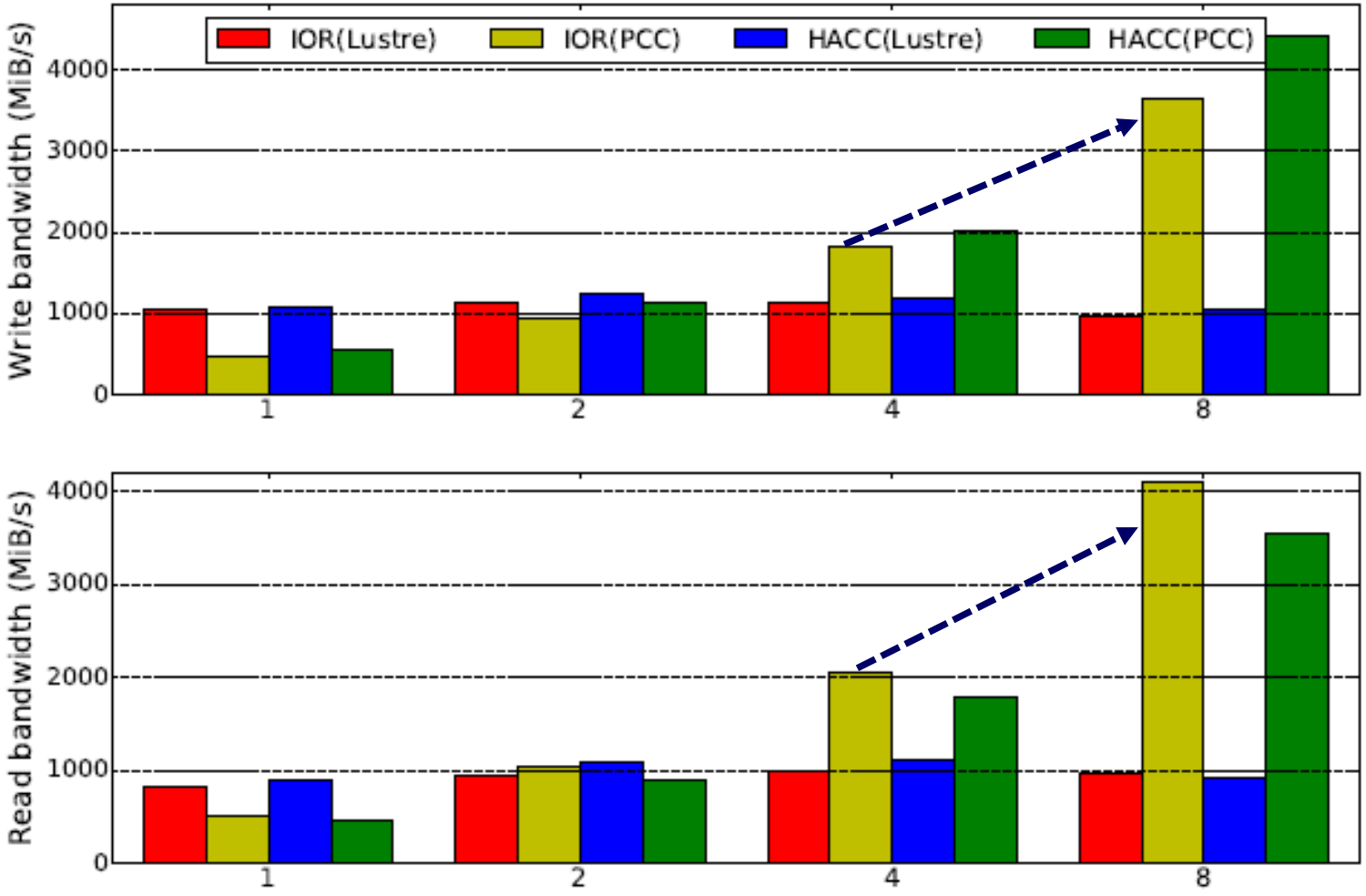


LPCC Architecture

- HDD
- Flash-based SSD
  - Evaluation on two platforms
  - Top 100 supercomputer MOGON II at the Johannes Gutenberg University Mainz (hpc.uni-mainz.de)
- NVRAM
  - Cache medium influence

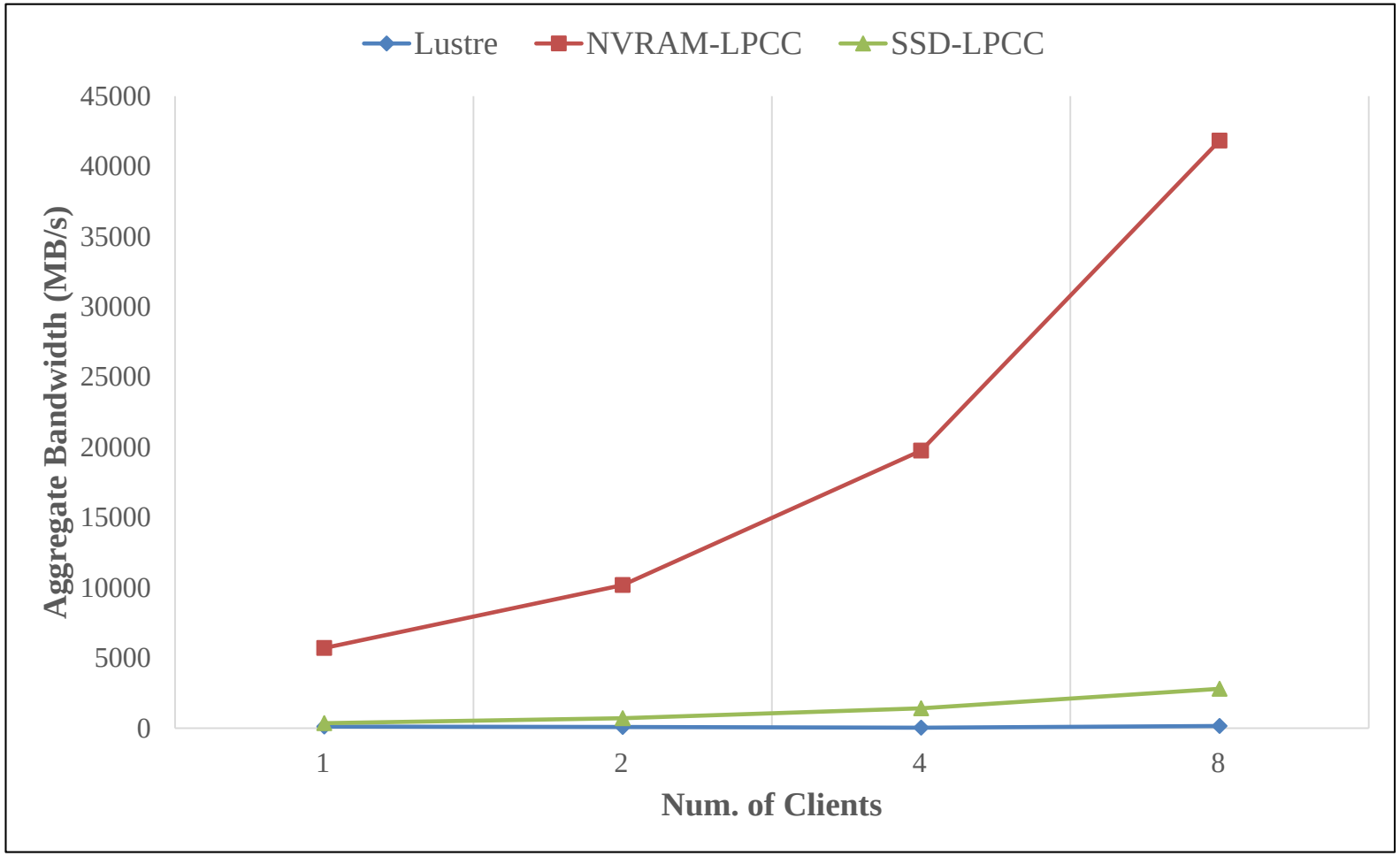


RW-PCC Scalability Evaluation



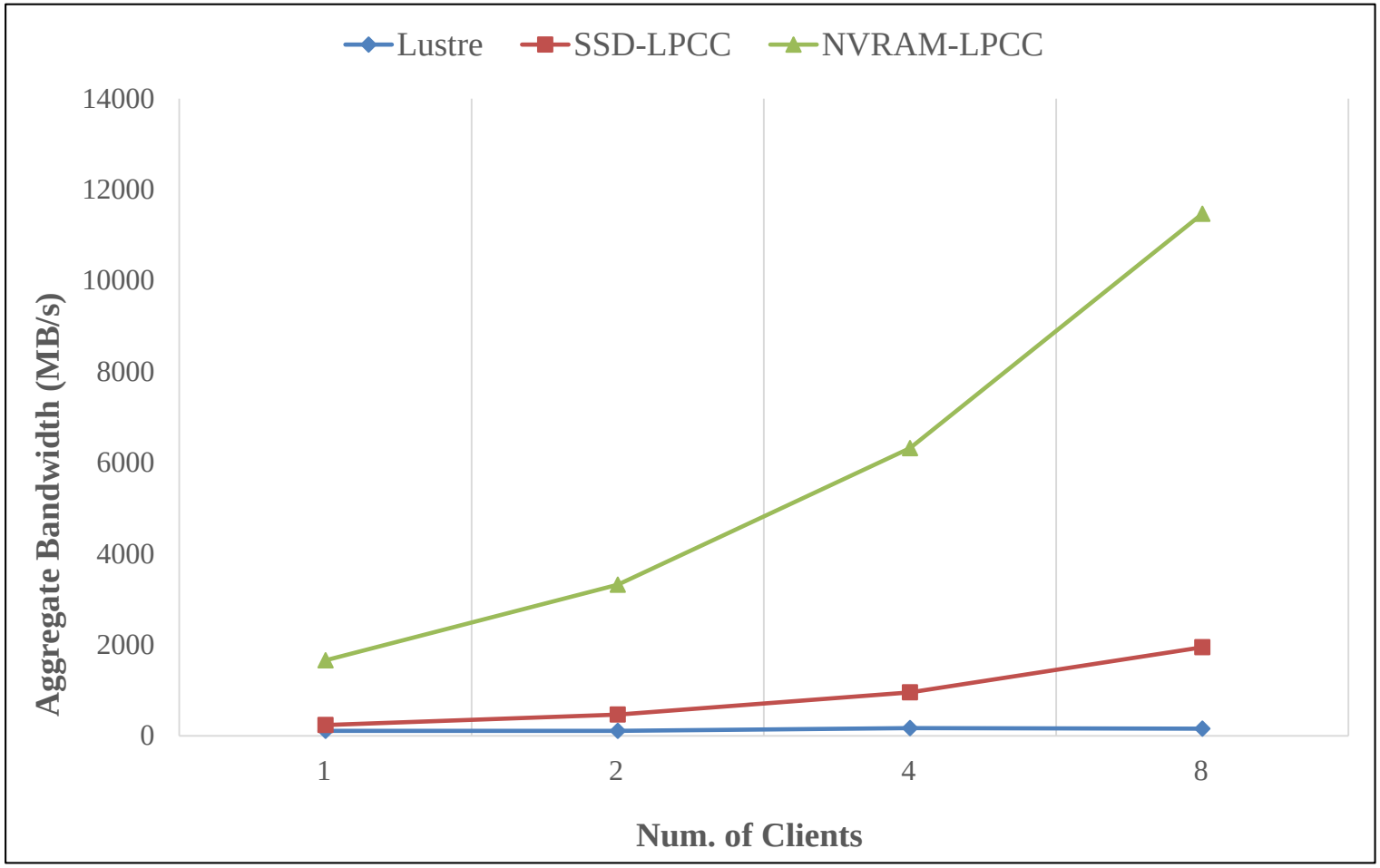
Source: Yingjin Qian,et. al., LPCC: Hierarchical Persistent Client Caching for Lustre. In Proceedings of SC19.

Read Scale Test (RW-PCC, IOR)



Source: Lingfang Zeng, Xi Li, and Wen Cheng, NVRAM-oriented Lustre Persistent Cache on Client, LAD2019

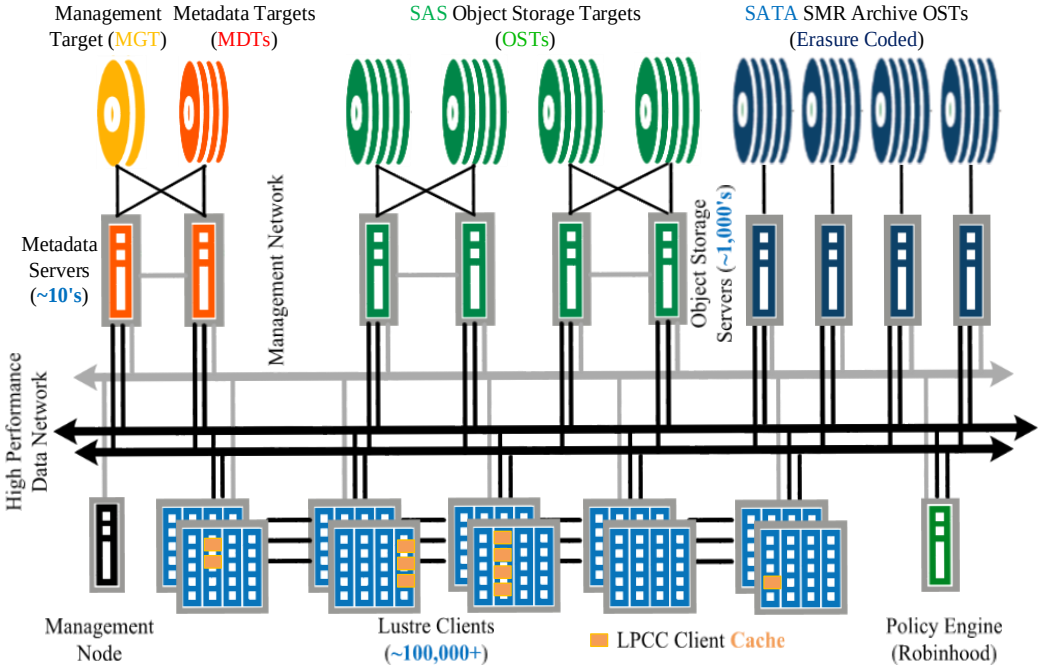
Write Scale Test (RW-PCC, IOR)



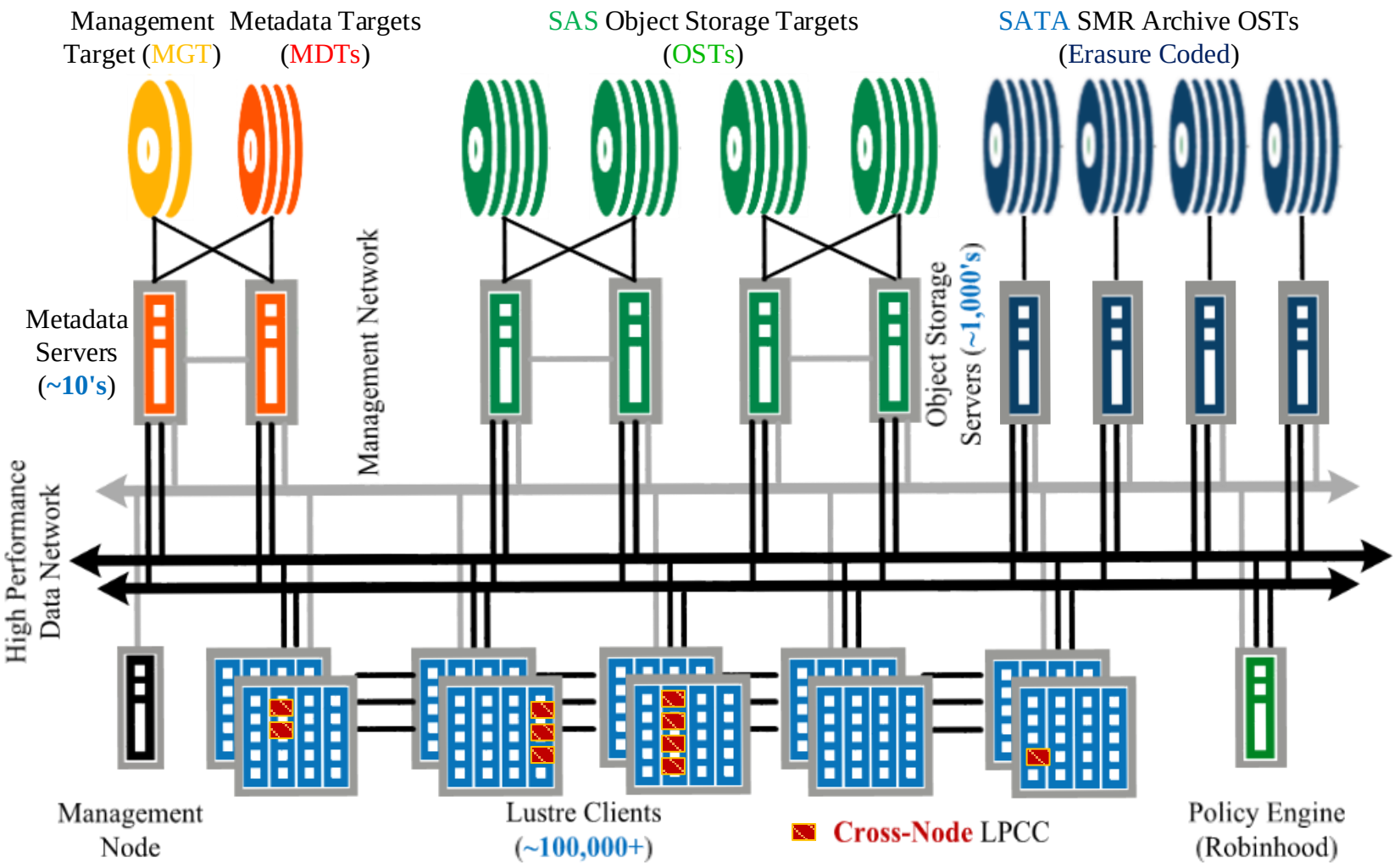
Source: Lingfang Zeng, Xi Li, and Wen Cheng, NVRAM-oriented Lustre Persistent Cache on Client, LAD2019

LPCC's Pros and Cons

- Pros 
  - Latency intensive
  - Data locality
- Cons 
  - Capacity limitation
  - Affect data layout
  - Data movement

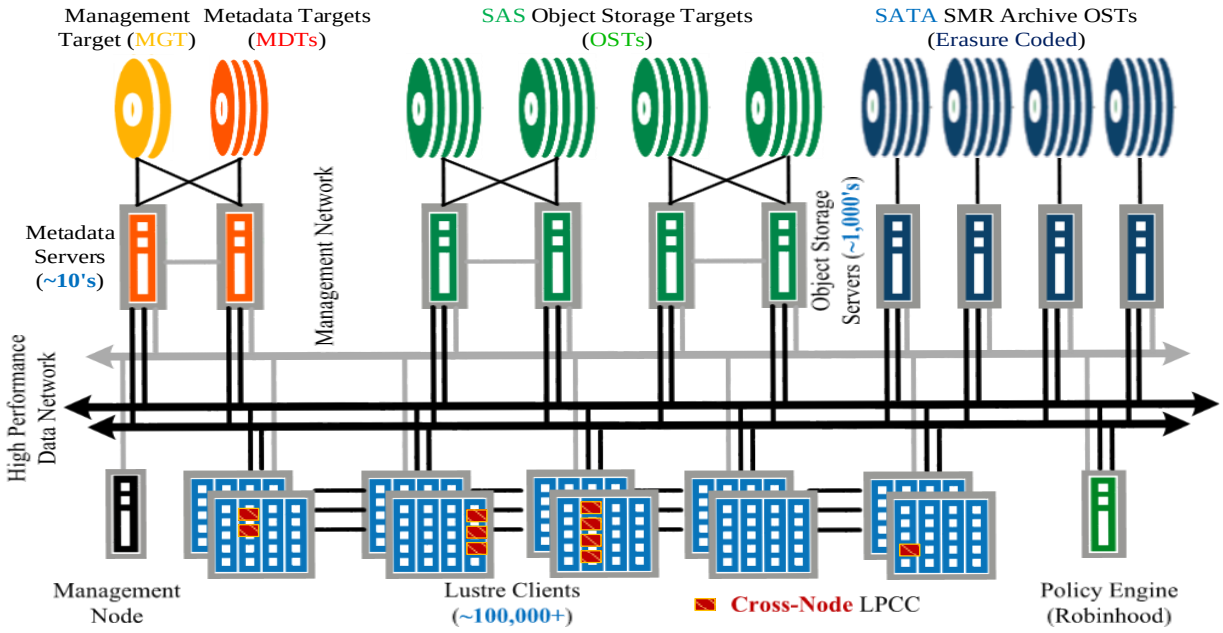


Cross-Node LPCC Architecture

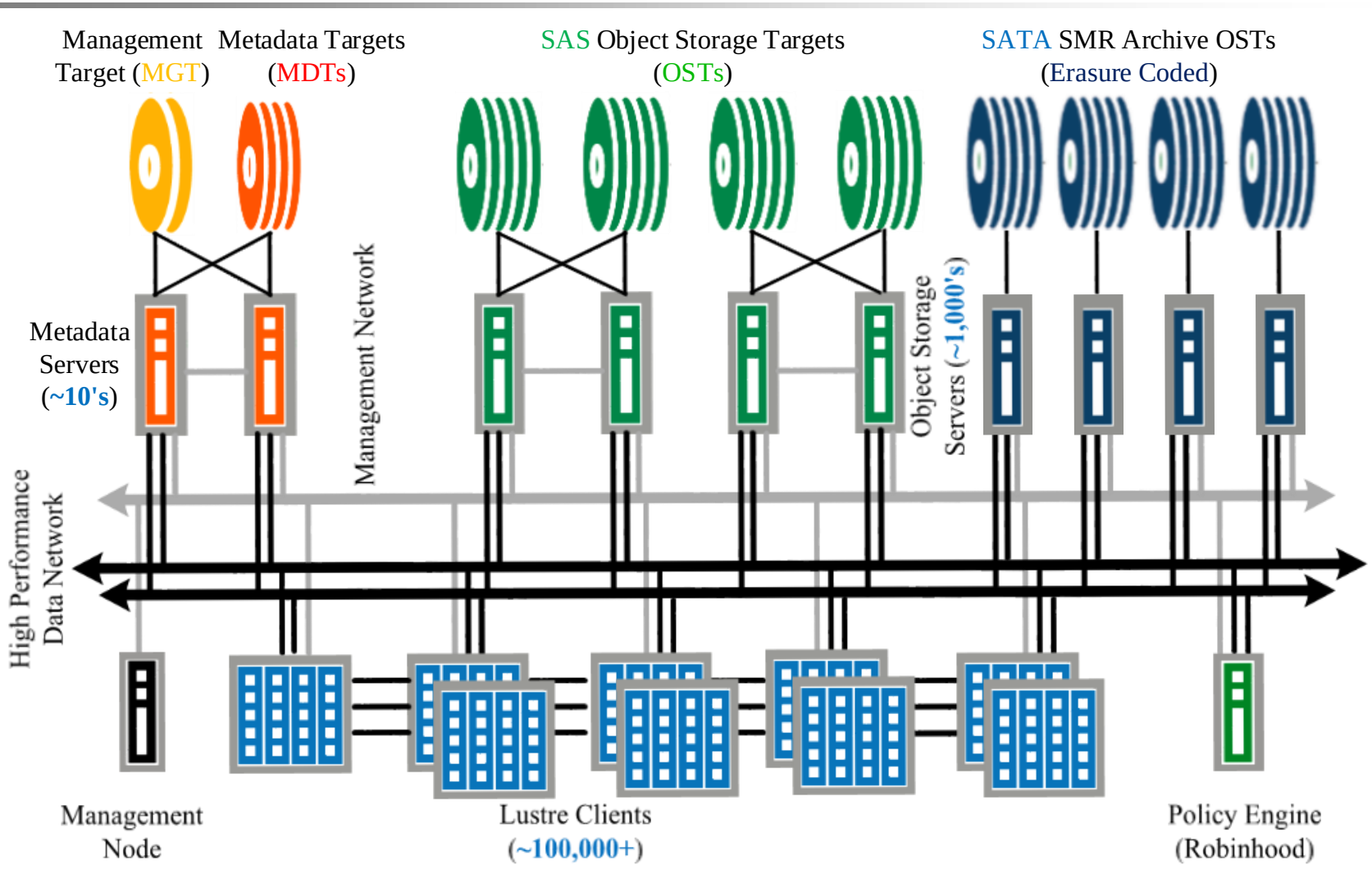


Cross-Node LPCC Architecture

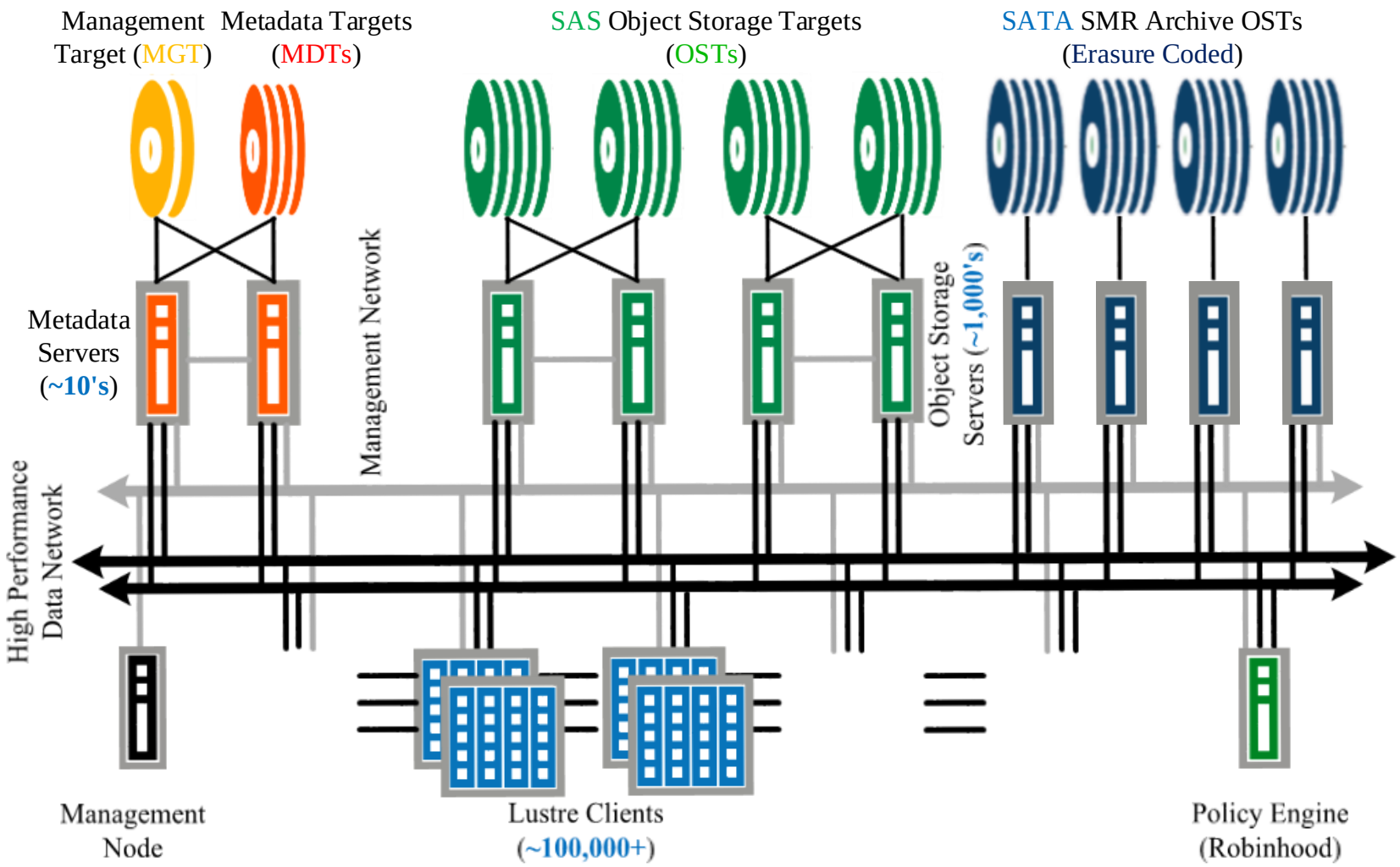
- Caching (role)
- Data can be distributed in multi clients
- Data movement



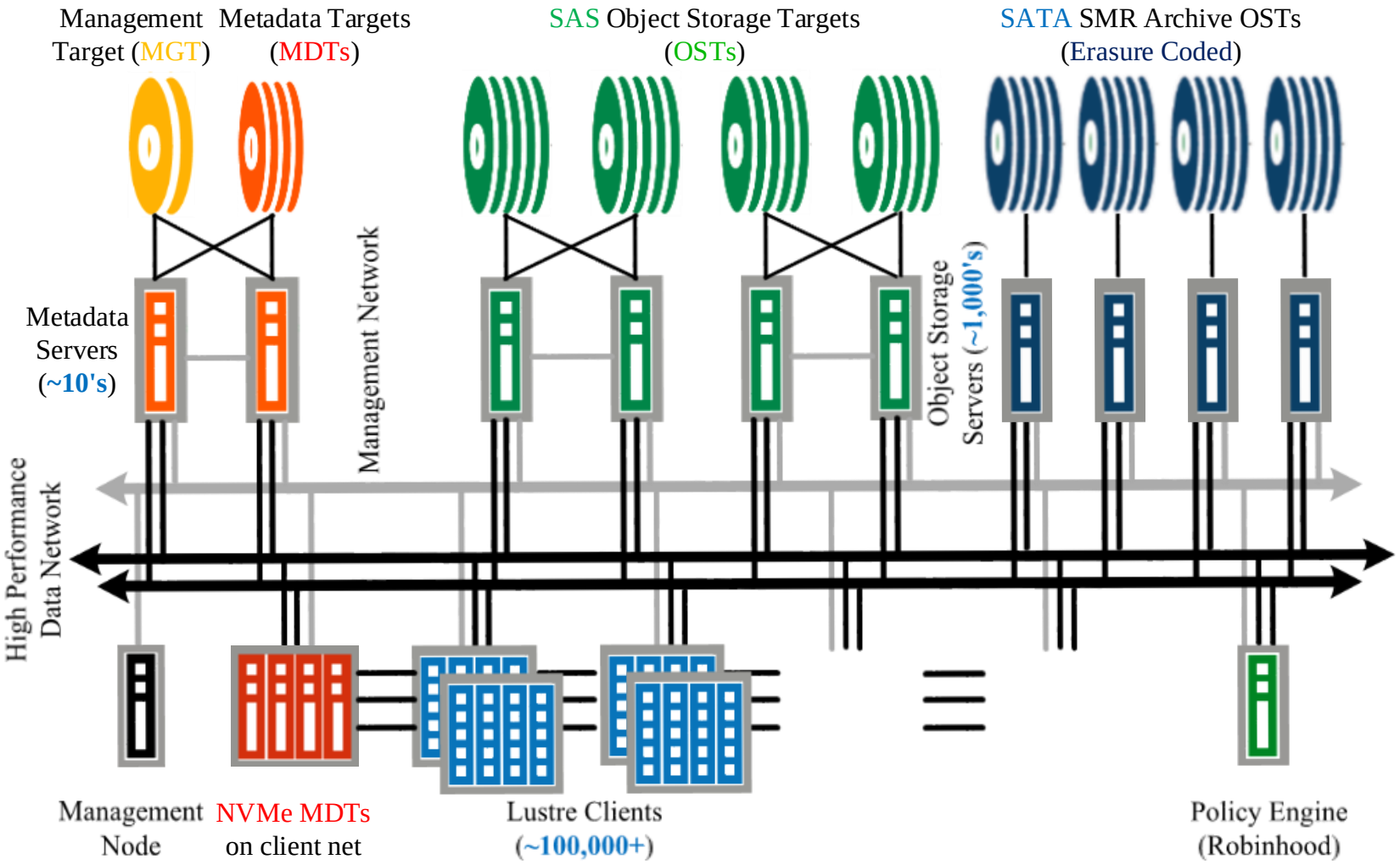
Lustre-on-Demand Architecture



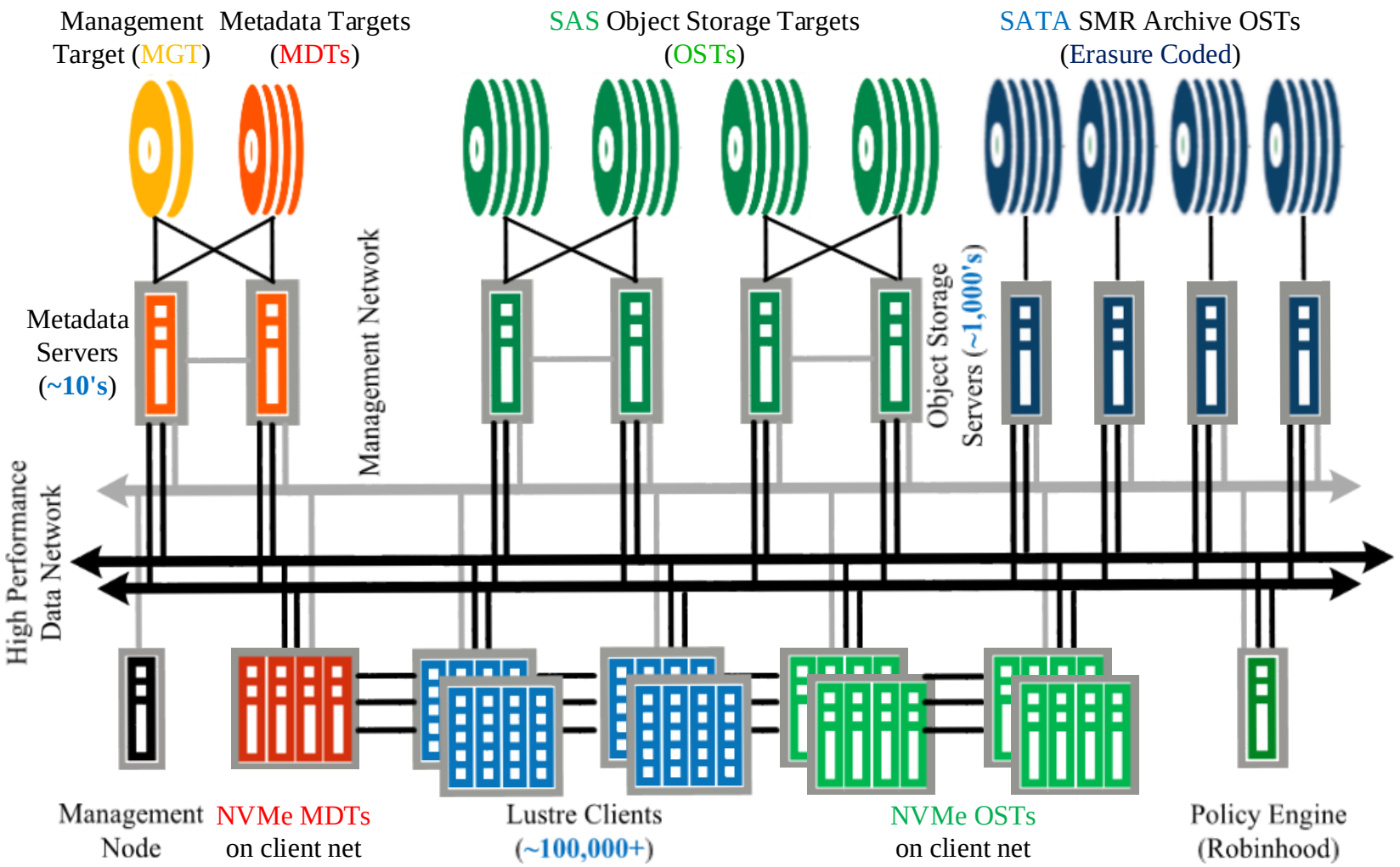
Lustre-on-Demand Architecture



Lustre-on-Demand Architecture

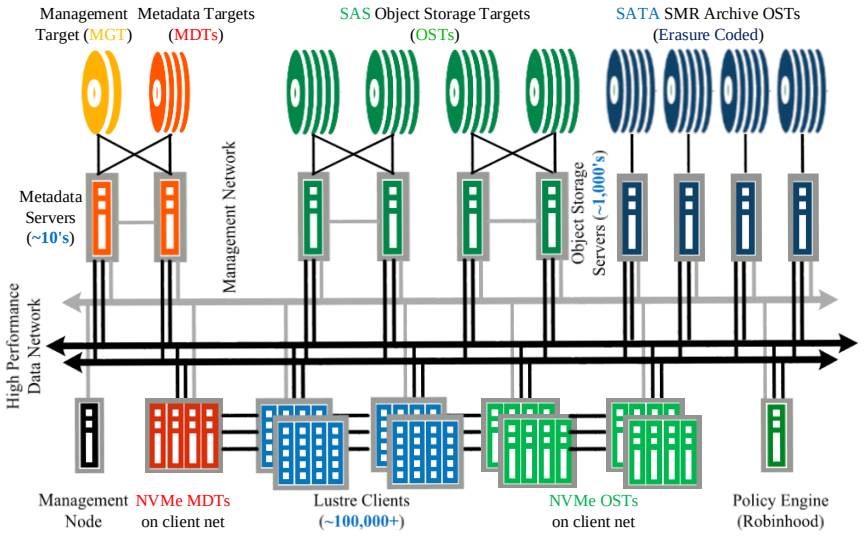


Lustre-on-Demand Architecture



Lustre-on-Demand Architecture

- Storage (also cache role?)
- Data on-demand distributed in client network
- On-demand data movement
- Heterogeneous fusion (CPU, XPU, AI, NVRAM, …)
  - Homogeneous heterogeneity
  - Flash-based (current)
  - NVRAM (near future)
  - Resource pooling/disaggregation



Related libraries/tools

Related libraries/tools

google-gflags

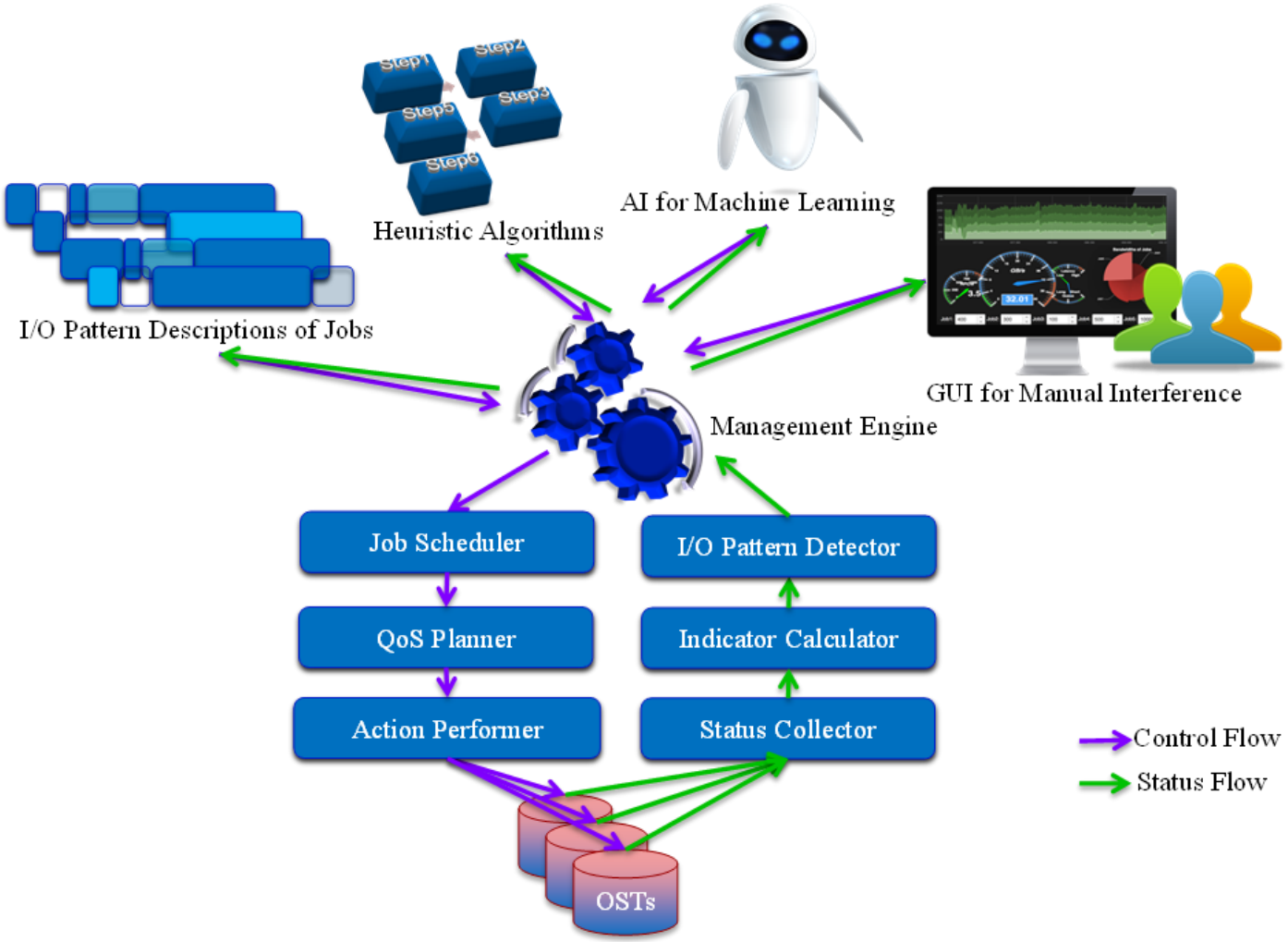
googletest
Google C++ Testing Framework

protobuf
Protocol Buffers

google-glog

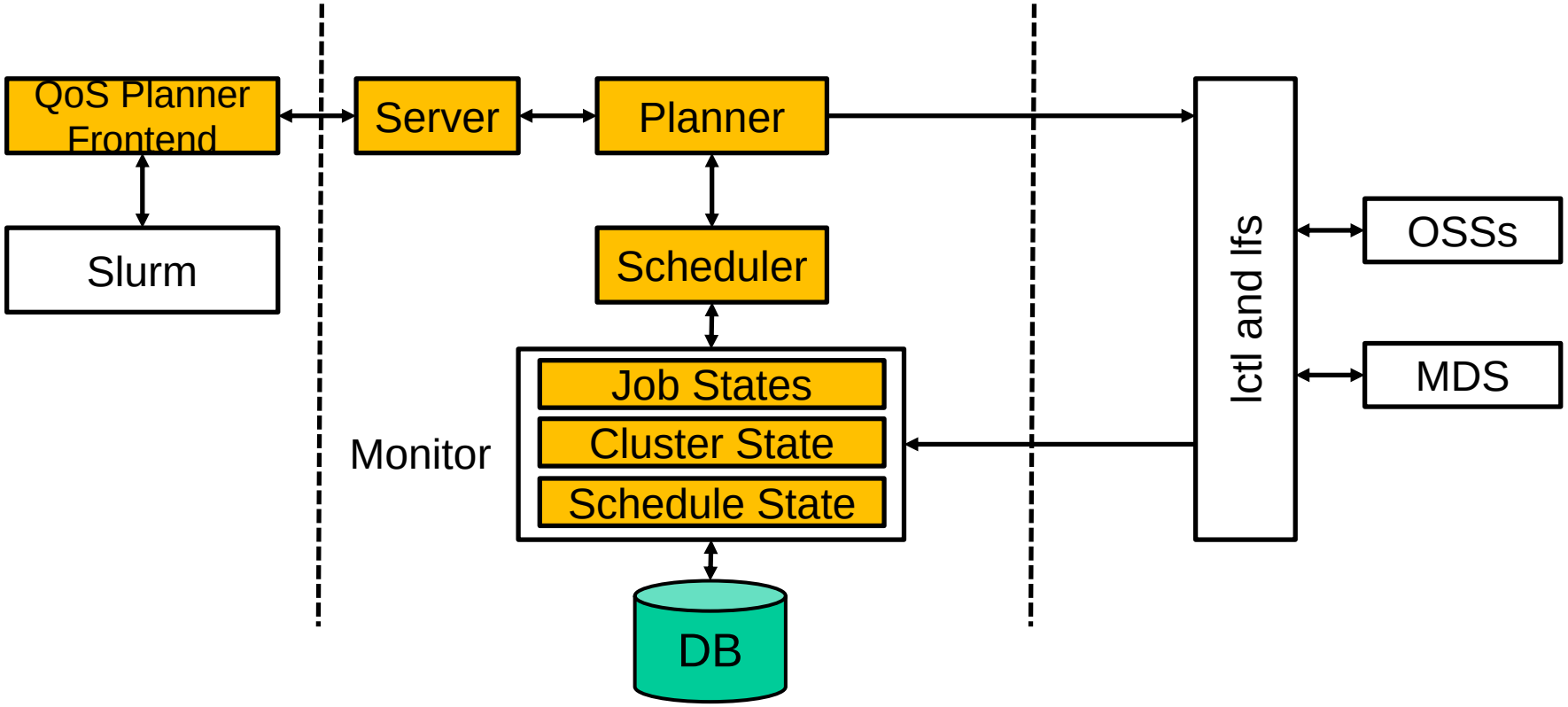
Related libraries/tools

Lustre Intelligent Management Engine



Source: Xi Li, and Lingfang Zeng, LIME: A Framework for Lustre Global QoS Management, LAD2018

QoS Planner



Source: Lingfang Zeng etc., Providing QoS-mechanisms for Lustre through centralized control applying the TBF-NRS, LUG2017

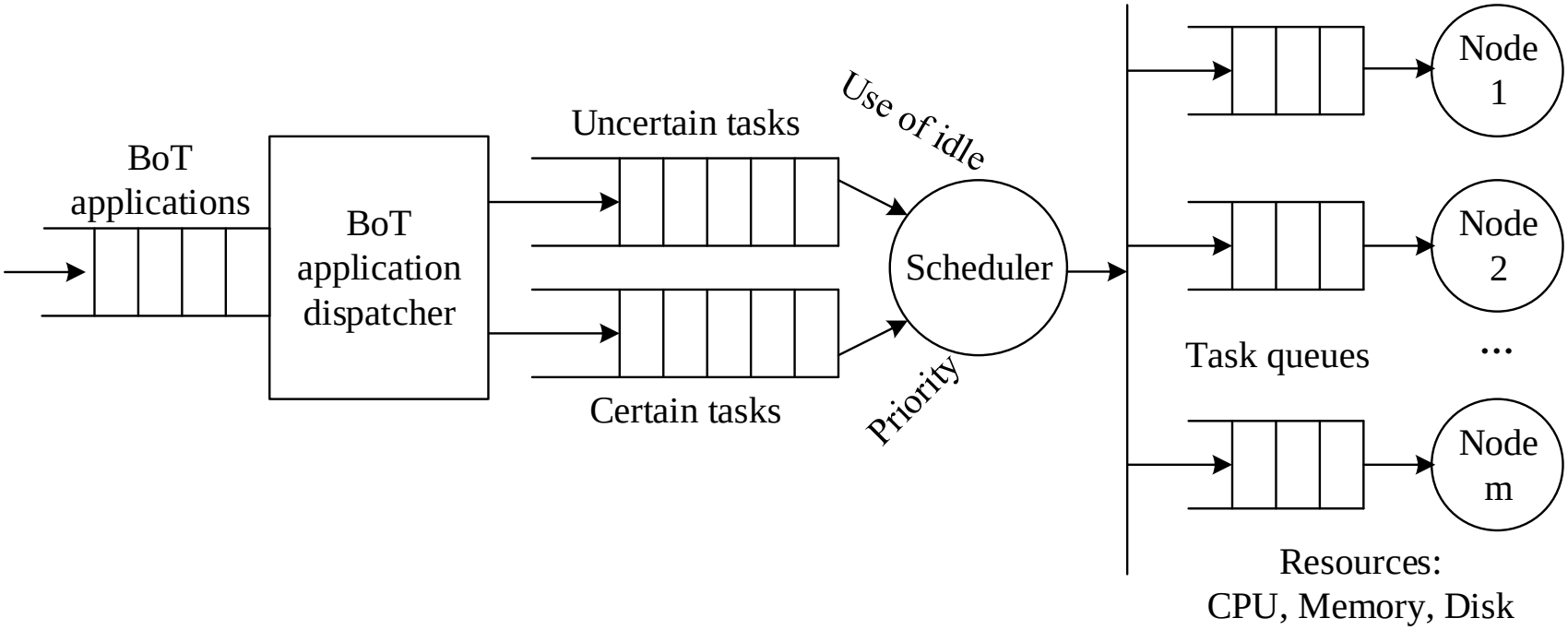
QoS Planner & Slurm Integration

- Bandwidth is defined as a global and as a local resource
- Slurm plug-in controls:
  - Globally available bandwidth - treated as *license* (one license/MB)
  - Local bandwidth - treated as *generic* resource
- Job gets rejected if one resource is not available
- Example:
  - `srun -N1 -p bigmem -A system --gres=qoslustre:100M -L lustreqos:100 sleep 5`

Source: Lingfang Zeng etc., Providing QoS-mechanisms for Lustre through centralized control applying the TBF-NRS, LUG2017

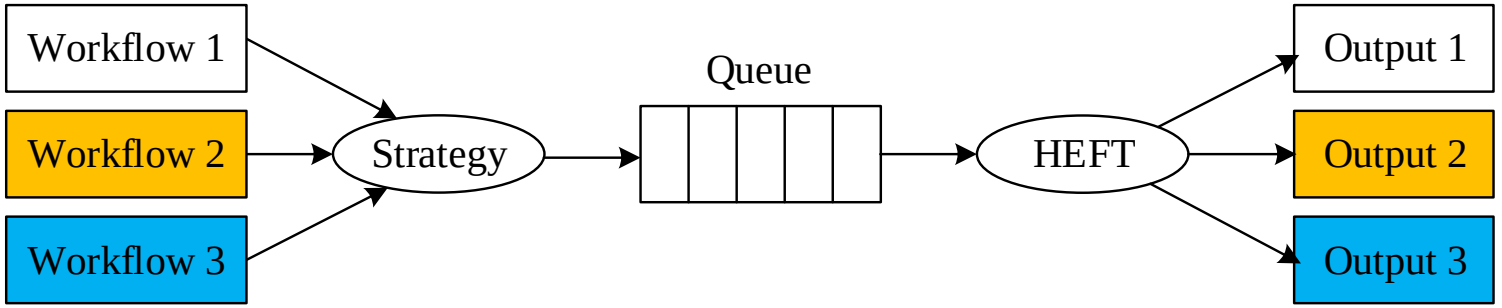
Handling of BoT (example)

- Priority Based Task Scheduling



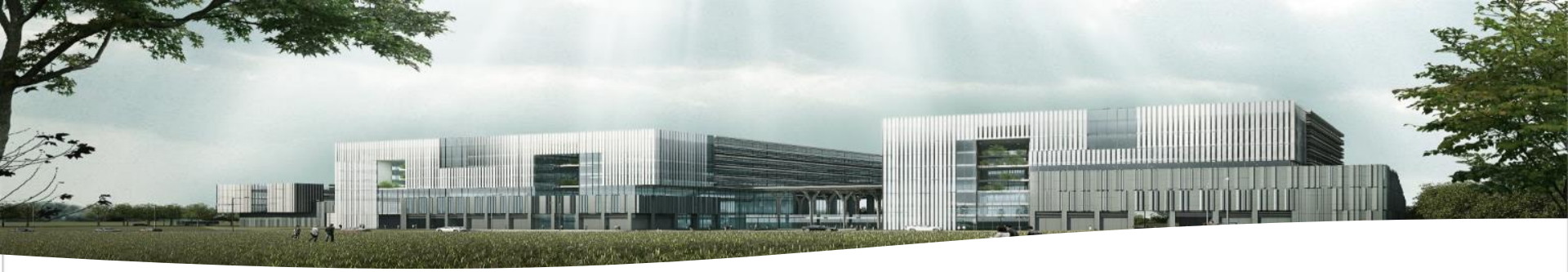
Handling of Workflow (example)

- Based on heterogeneous earliest-finish time (HEFT)



Summary

- Lustre basic architecture
- Lustre architecture evolution
  - LPCC
  - Cross-Node LPCC
- Lustre architecture trends
  - Lustre-on-Demand
  - Tools



China Lustre User Group (China LUG), Beijing, October 15, 2019

Thanks for your attention!

Providing Architecture Support for On-Demand Services on Large-Scale HPC Platform

Lingfang Zeng (曾令仿)

Wuhan National Laboratory for Optoelectronics (WNLO)

Huazhong University of Science and Technology (HUST)