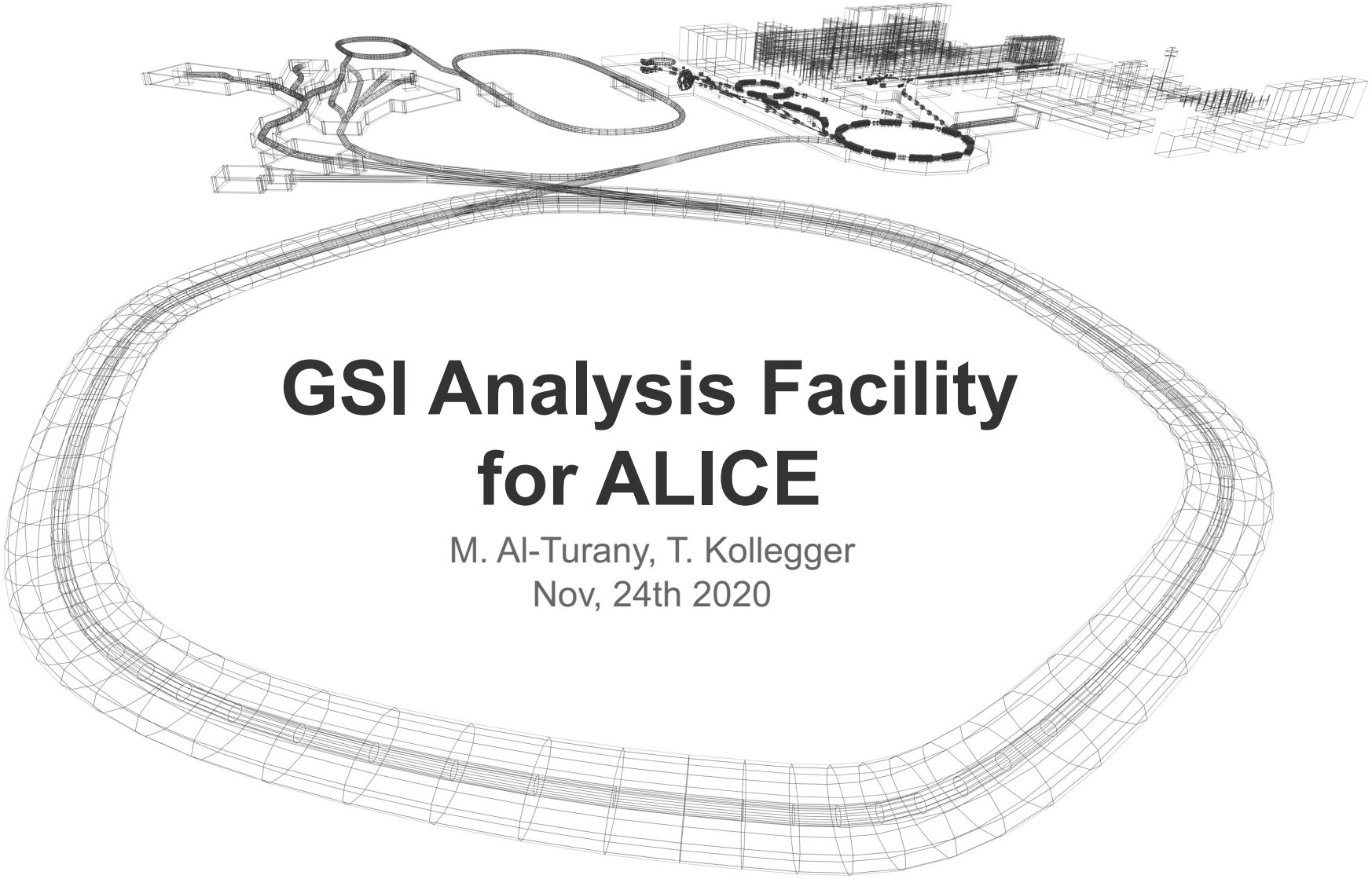
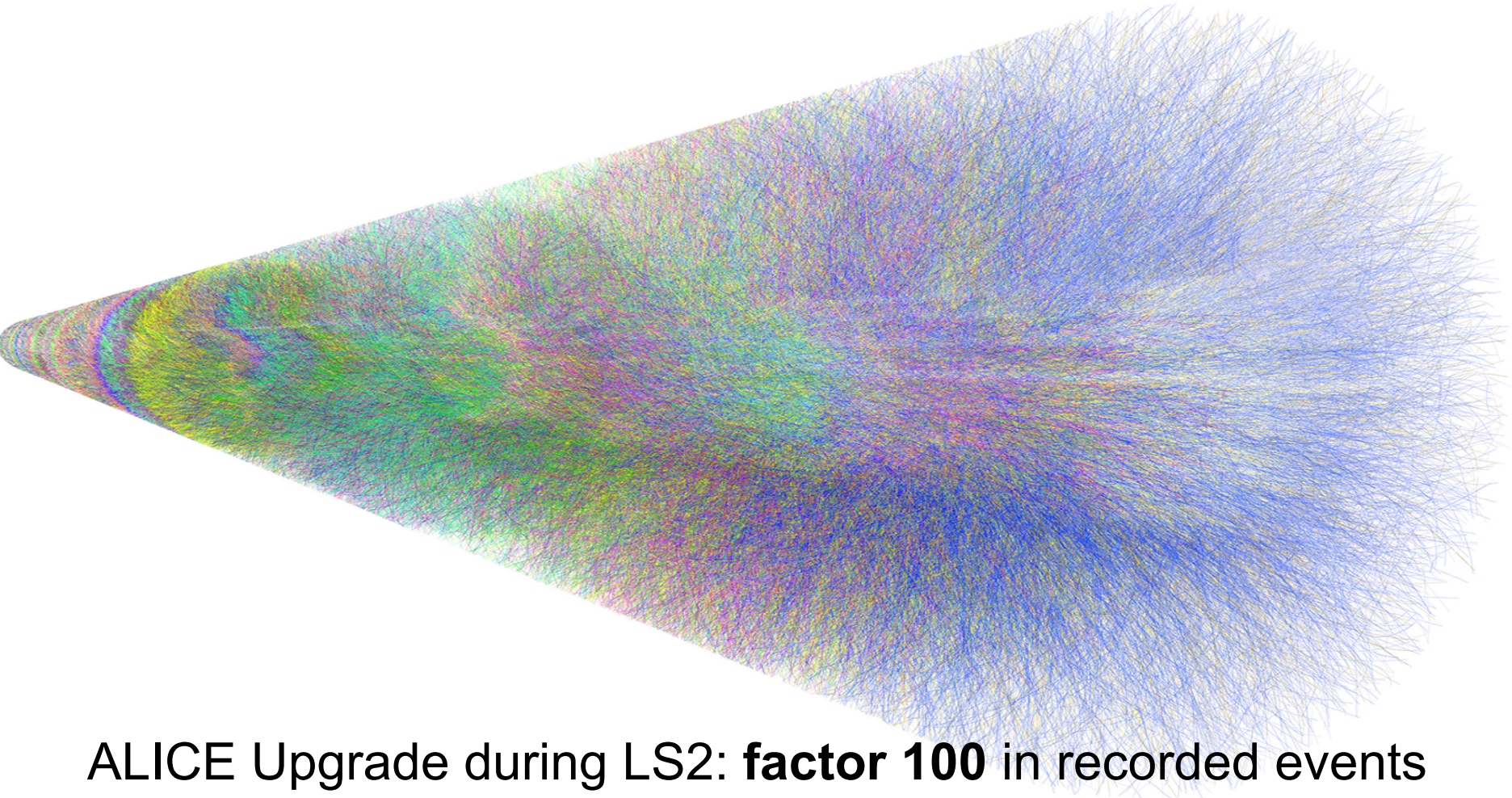


GSI Analysis Facility for ALICE

M. Al-Turany, T. Kollegger

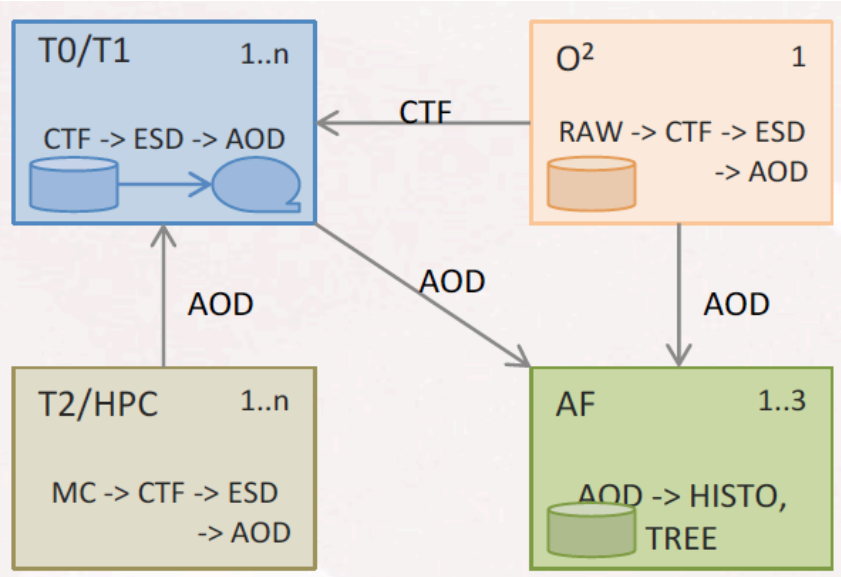
Nov, 24th 2020

ALICE Run 3



ALICE Upgrade during LS2: **factor 100** in recorded events
How do quickly and efficiently analyze this data?

ALICE Analysis Facilities

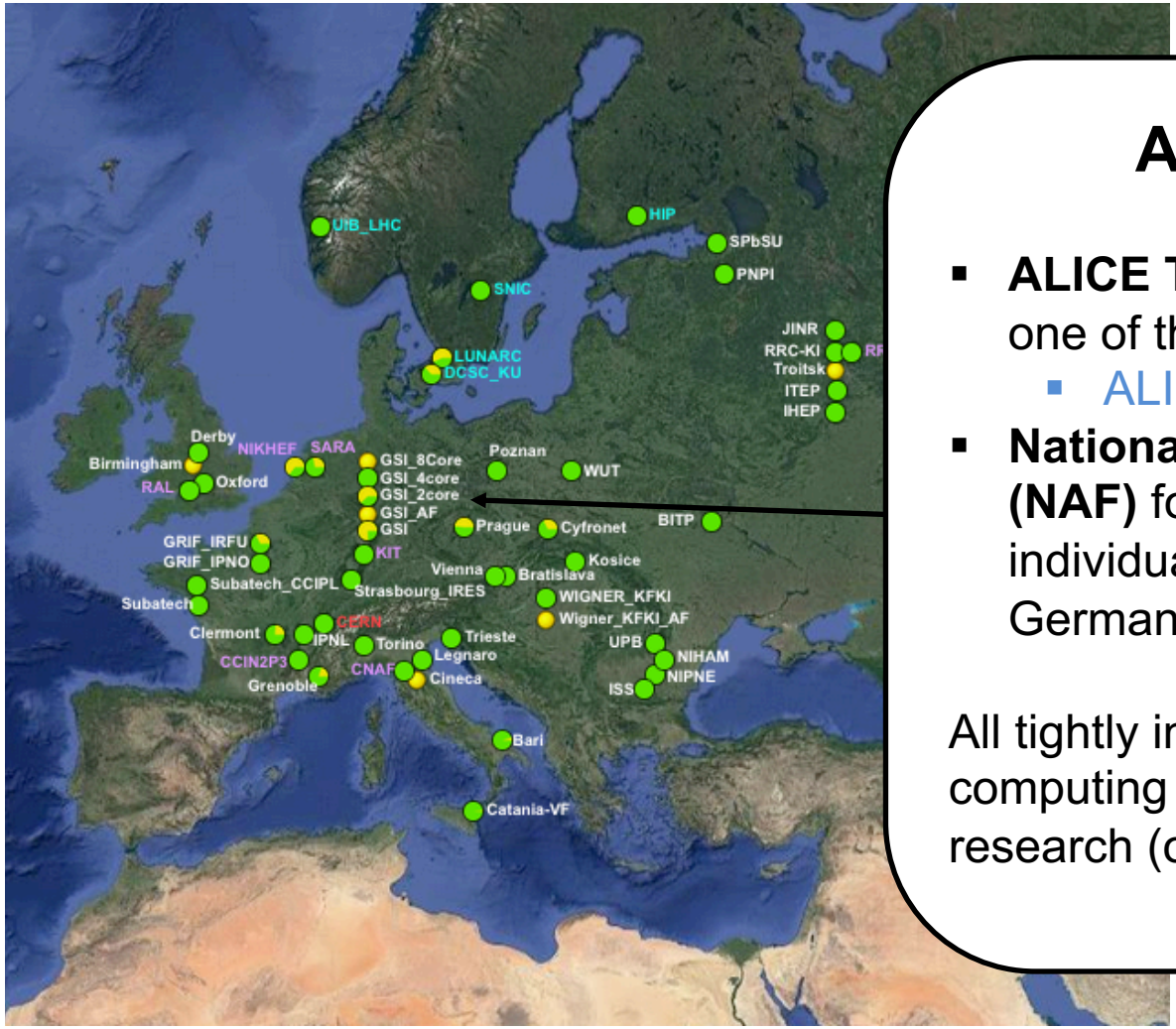


ALICE O² TDR,
P. Buncic (~2014)

Analysis Facilities: specialized centers integrated in the Grid

- allowing fast-turnaround organized analysis on a significant subset of (AOD) data prior to full analysis on Grid sites, enables further processing on user laptop/…/NAF
- High-efficiency by high-bandwidth storage (10PB/day)
- Multi-core queue support (<-> new ALICE analysis framework)
- Specialized capabilities, e.g. GPUs for ML training

ALICE@GSI

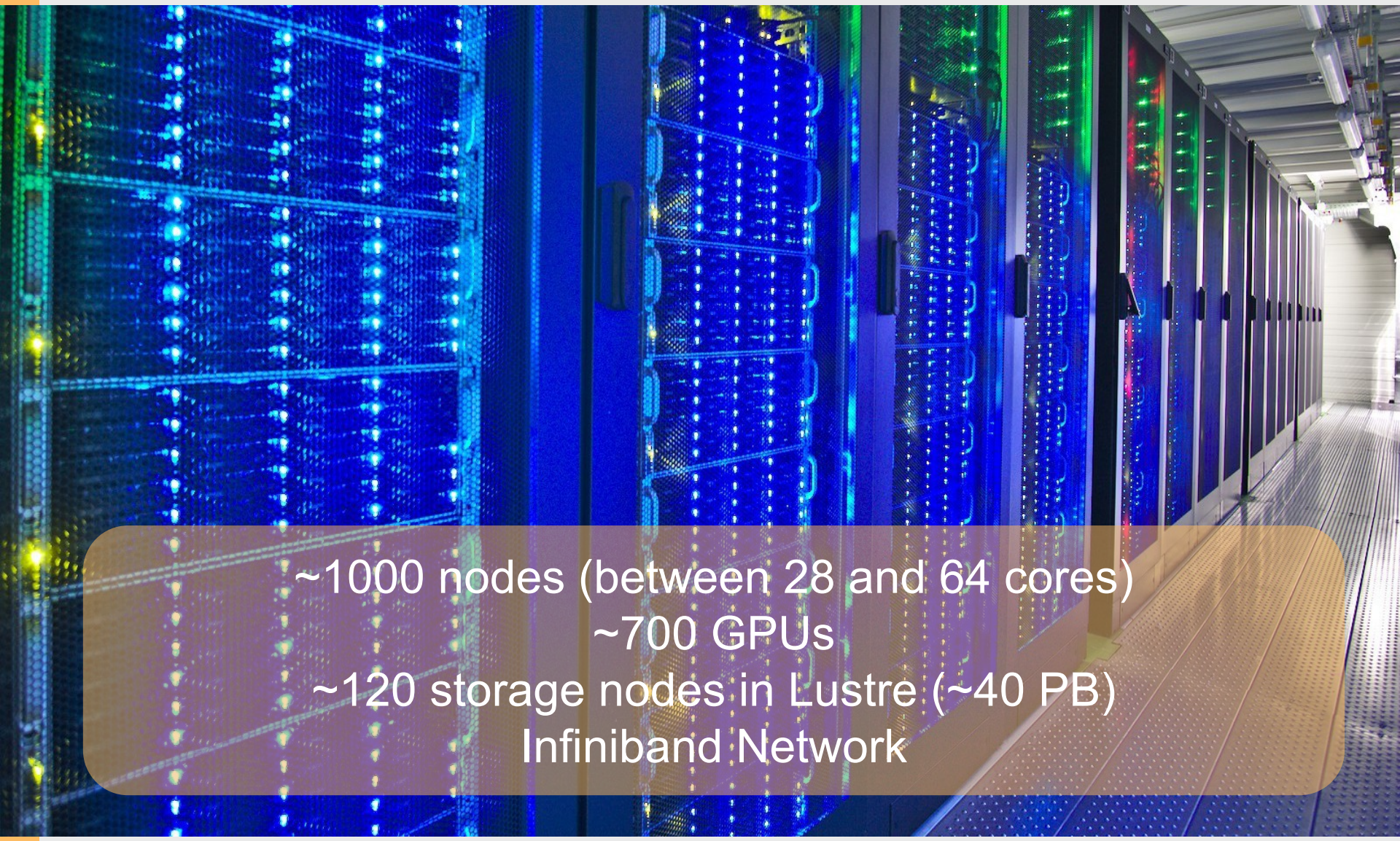


ALICE @ GSI

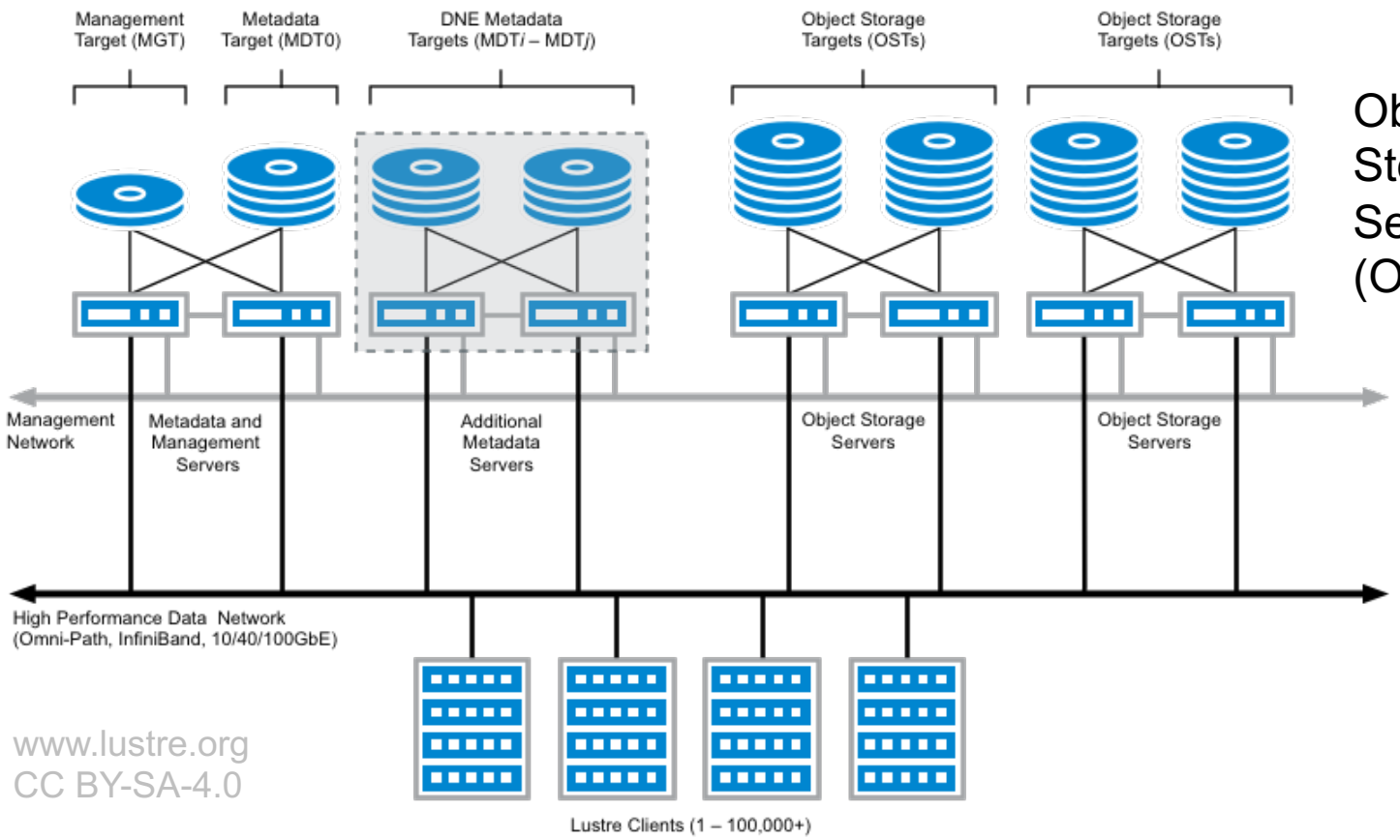
- **ALICE Tier 2** (only German, one of the largest in ALICE)
  - ALICE **Analysis Facility**
- **National Analysis Facility (NAF)** for ALICE (organized and individual user analysis for German ALICE groups)

All tightly integrated into common computing systems for GSI/FAIR research (online clusters, HPC).

Compute@GSI

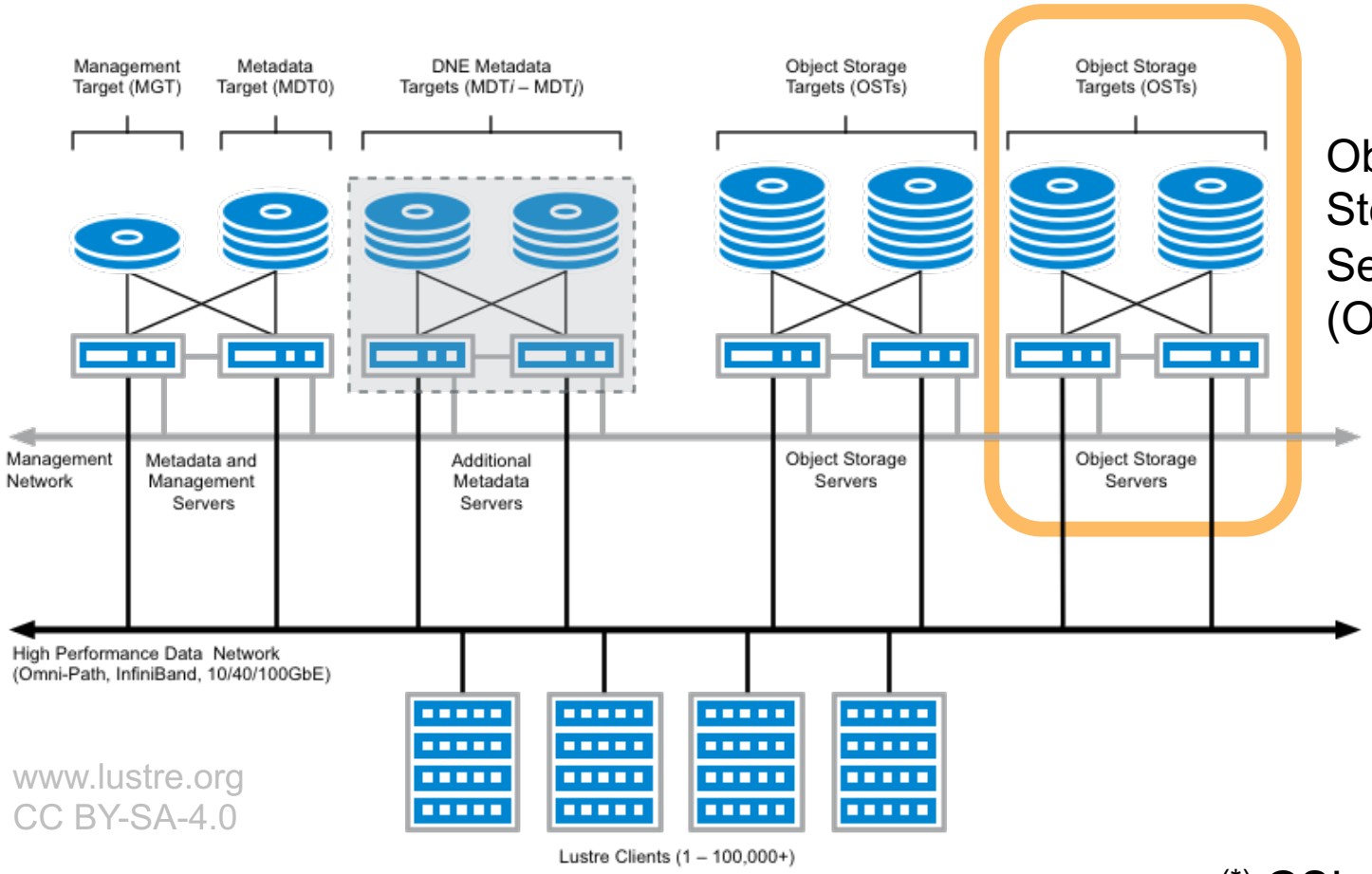

~1000 nodes (between 28 and 64 cores)
~700 GPUs
~120 storage nodes in Lustre (~40 PB)
Infiniband Network

Lustre-based Storage



Object Storage Servers (OSS)

Lustre-based Storage



Object Storage Servers (OSS) (*)

(*) GSI using different OSS architecture

GSI Lustre OSS

System

- 4U Server + 4U Disk Enclosure
- 71 HDDs,10 TB
- Infiniband FDR

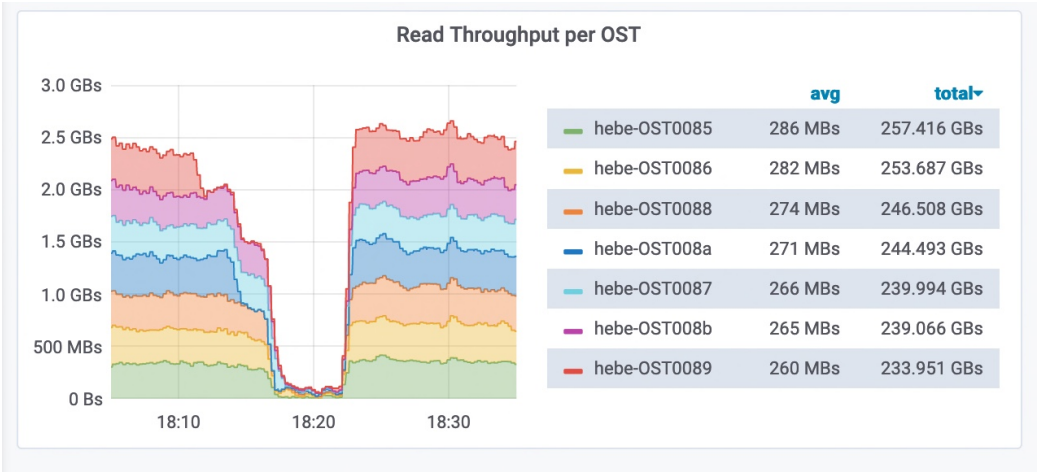
HDDs organized into 7 Object Storage Targets (OSTs)
Each OST realized as **ZFS RAIDZ2** of (8+2) HDDs (comparable to RAID 6), compression enabled,
~425 TByte usable capacity per OSS

GSI Lustre OSS

System

- 4U Server + 4U Disk Enclosure
- 71 HDDs,10 TB
- Infiniband FDR

HDDs organized into 7 Object Storage Targets (OSTs)
Each OST realized as **ZFS RAIDZ2** of (8+2) HDDs (comparable to RAID 6), compression enabled,
~425 TByte usable capacity per OSS



Performance

~350 MByte/s per OST
~2.5 GByte/s per OSS
>300 GByte/s for full system

More on Lustre…

… there is much more than just the bare OSS performance, some features being implemented (relevant for ALICE AF)

- Data-on-MDT: improve small file performance
- Lazy-Size on Metadata: improve metadata performance (stat's, e.g. "ls" on large directories)
- File-Level-Redundancy: multiple copies of simultaneously accessed files to increase performance
- Re-balancing of files: monitor access and re-distribute files between OSTs to enhance performance
- HSM: performance-tiers (Tape, HDD, SSD)
- Persistent Client Cache/ Metadata Write-Back Cache: improve performance especially for interactive-like workloads

More on Lustre…

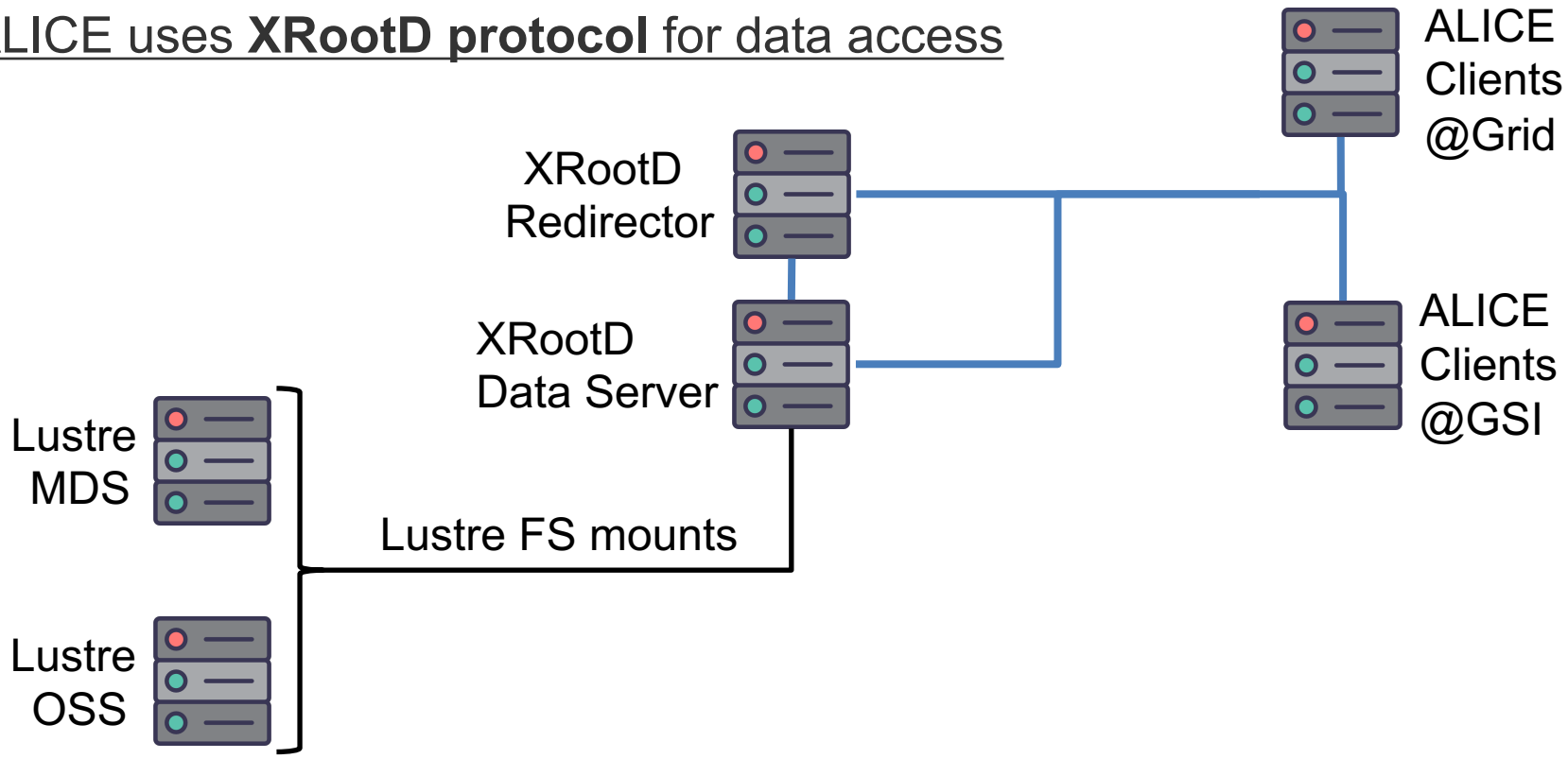
… there is much more than just the bare OSS performance, some features being implemented (relevant for ALICE AF)

- Data-on-MDT: improve small file performance
- Lazy-Size on Metadata: improve metadata performance (stat's, e.g. "ls" on large directories)
- File-Level-Redundancy: multiple copies of simultaneously accessed files to increase performance
- Re-balancing of files: monitor access and re-distribute files between OSTs to enhance performance
- HSM:
- Persistent Client Cache/ Metadata Write-Back Cache:

Would benefit if already taken into account during file transfer/creation
Current approach: on fs-level

Embedding into Grid

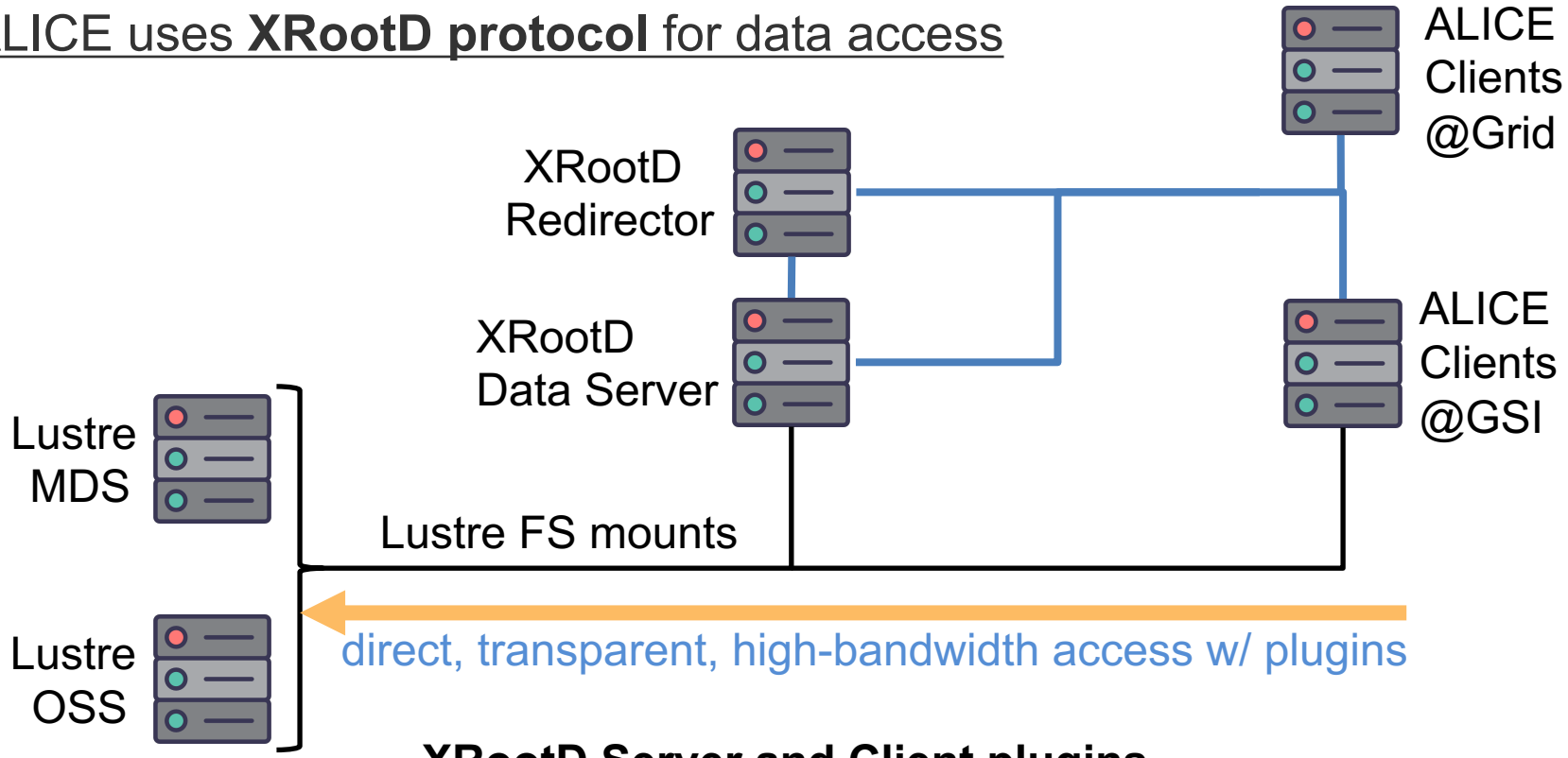
ALICE uses **XRootD protocol** for data access



'icons: Flaticon.com'. This cover has been designed using resources from Flaticon.com

Embedding into Grid

ALICE uses **XRootD protocol** for data access

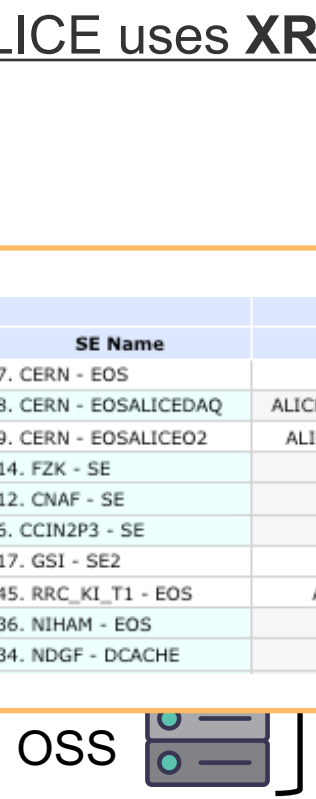


XRootD Server and Client plugins
direct FS-Level access skipping XRootD Data Servers
seamless integration into Grid and existing workflows
CHEP 2018: https://doi.org/10.1051/epjconf/201921404005

'icons: Flaticon.com'. This cover has been designed using resources from Flaticon.com

Embedding into Grid

ALICE uses **XRootD protocol** for data access



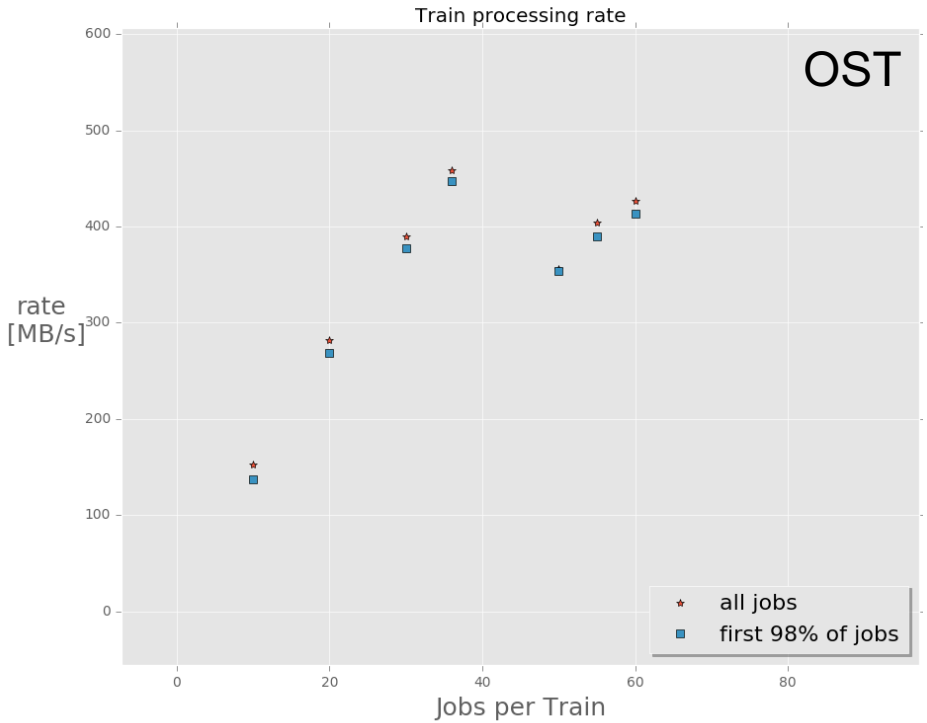
	AliEn SE		Catalogue statistics						Storage-provided info			
SE Name	AliEn name	Tier	Size ▲	Used	Free	Usage	No. of files	Type	Size	Used	Free	Usage
7. CERN - EOS	ALICE::CERN::EOS	0	29.72 PB	26.3 PB	3.419 PB	88.5%	902,599,531	FILE	29.72 PB	26.54 PB	3.173 PB	89.32%
8. CERN - EOSALICEDAQ	ALICE::CERN::EOSALICEDAQ	0	10.66 PB	265.8 GB	10.66 PB	0.002%	160	FILE	10.66 PB	353 TB	10.31 PB	3.234%
9. CERN - EOSALICEO2	ALICE::CERN::EOSALICEO2	0	10.22 PB	0.226 KB	10.22 PB	-	1	TEST	10.22 PB	2.689 PB	7.53 PB	26.31%
14. FZK - SE	ALICE::FZK::SE	1	9.046 PB	8.344 PB	719.3 TB	92.23%	211,870,207	FILE	9.197 PB	8.754 PB	453.6 TB	95.18%
12. CNAF - SE	ALICE::CNAF::SE	1	6.174 PB	4.913 PB	1.26 PB	79.59%	174,761,340	FILE	6.174 PB	4.936 PB	1.237 PB	79.96%
6. CCIN2P3 - SE	ALICE::CCIN2P3::SE	1	6.065 PB	4.91 PB	1.155 PB	80.96%	174,431,808	FILE	5.313 PB	4.913 PB	409.9 TB	92.47%
17. GSI - SE2	ALICE::GSI::SE2	2	4.3 PB	4.022 PB	284.5 TB	93.54%	159,971,351	FILE	4.3 PB	3.63 PB	685.6 TB	84.43%
45. RRC_KI_T1 - EOS	ALICE::RRC_KI_T1::EOS	1	4.152 PB	3.848 PB	311.4 TB	92.68%	122,505,785	FILE	4.152 PB	3.897 PB	261.2 TB	93.86%
36. NIHAM - EOS	ALICE::NIHAM::EOS	2	3.4 PB	3.081 PB	326.6 TB	90.62%	102,191,643	EOS	3.397 PB	3.094 PB	310.3 TB	91.08%
34. NDGF - DCACHE	ALICE::NDGF::DCACHE	1	3.27 PB	1.726 PB	1.544 PB	52.79%	35,951,666	SRM	-	-	-	-

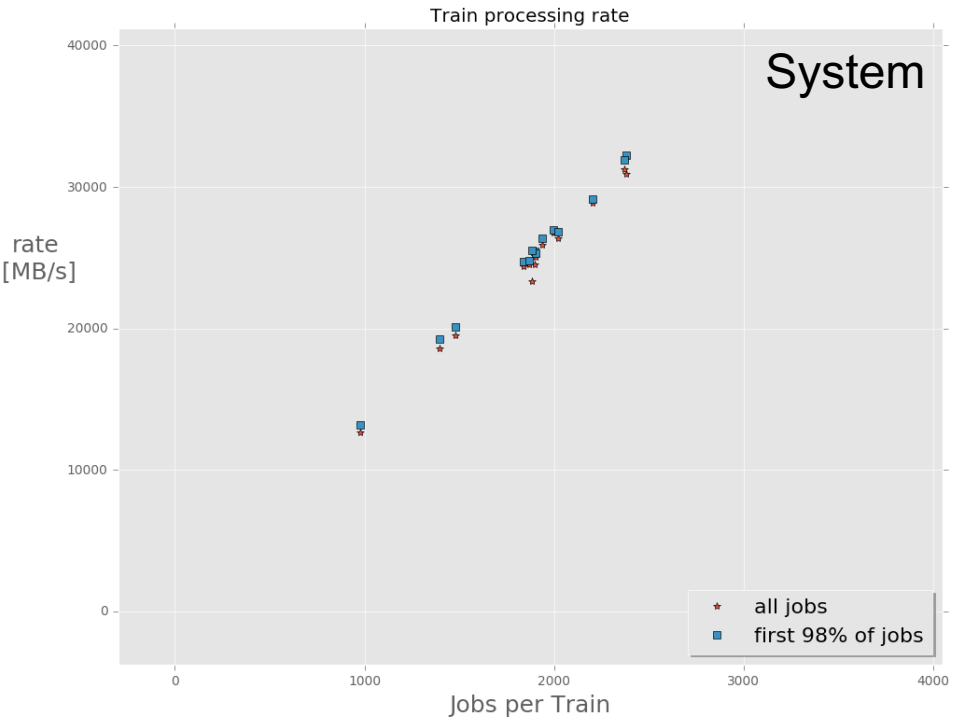
XRootD Server and Client plugins
direct FS-Level access skipping XRootD Data Servers
seamless integration into Grid and existing workflows
CHEP 2018: https://doi.org/10.1051/epjconf/201921404005

'icons: Flaticon.com'. This cover has been designed using resources from Flaticon.com

Performance

First performance tests done with the
old ALICE analysis framework on the facility in 2017/18

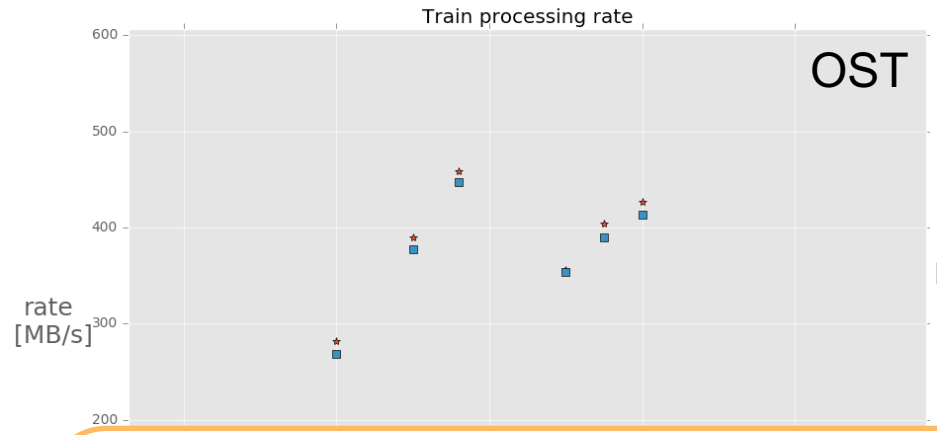


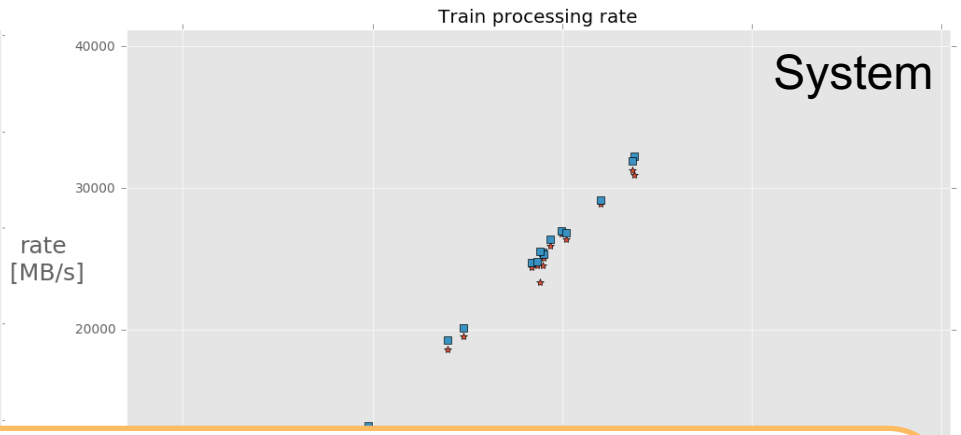


CHEP 2018: https://doi.org/10.1051/epjconf/201921404005

Performance

First performance tests done with the
old ALICE analysis framework on the facility in 2017/18



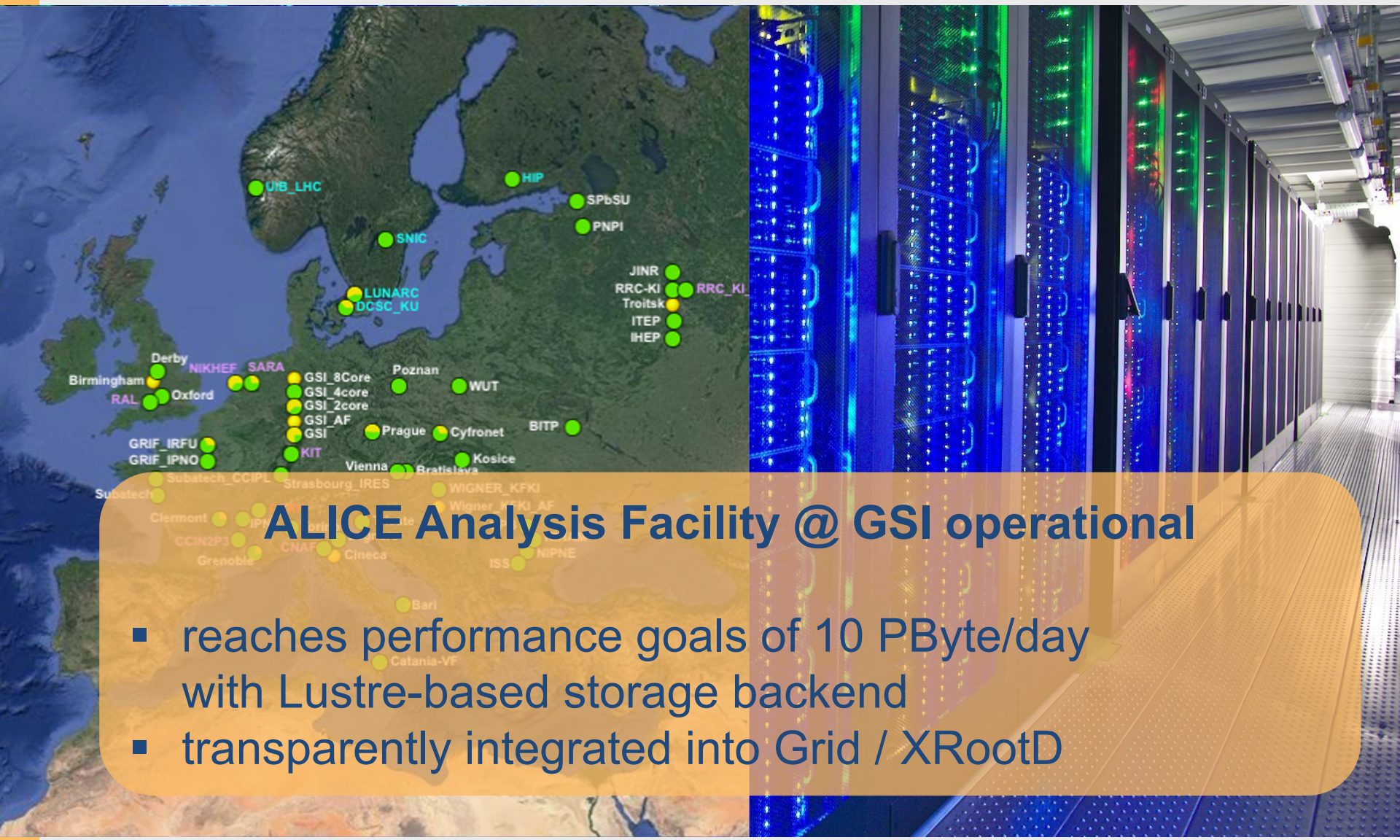


Demonstrated feasibility of 10 PByte/day goal

New analysis framework to further improve performance,
Run 3 analysis challenge currently ongoing

CHEP 2018: https://doi.org/10.1051/epjconf/201921404005

Summary



ALICE Analysis Facility @ GSI operational

- reaches performance goals of 10 PByte/day with Lustre-based storage backend
- transparently integrated into Grid / XRootD