I/O at the Center for Information Services and High Performance Computing

HPC-I/O in the Data Center Workshop @ ISC 2015

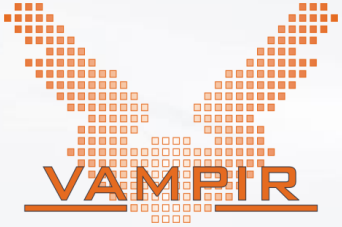
Zellescher Weg 12
Willers-Bau A 208
Tel. +49 351 - 463 – 34217

Michael Kluge (michael.kluge@tu-dresden.de)

ZIH
Center for Information Services & High Performance Computing

About us

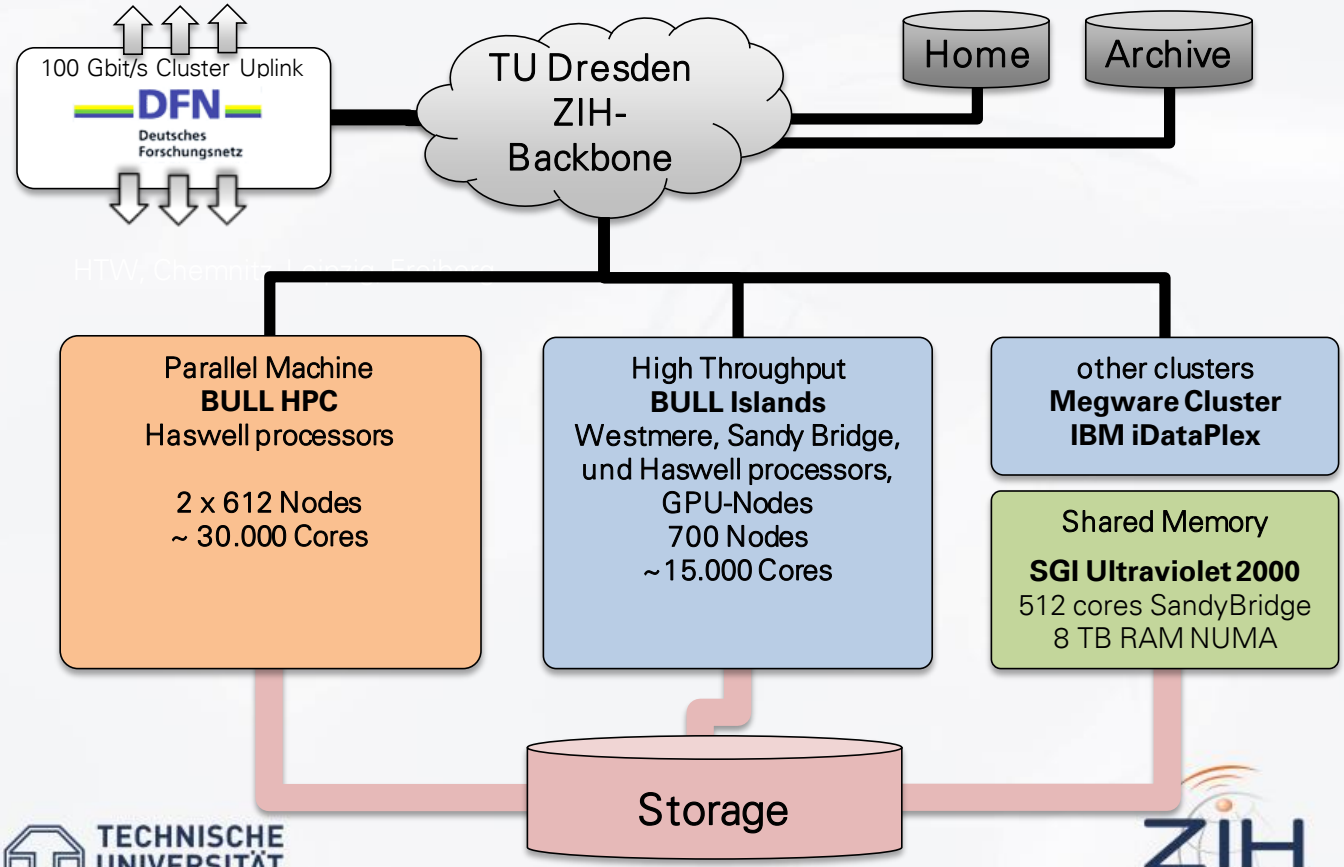
- ZIH: about us
  - HPC and service provider
  - Research Institute
  - Big Data Competence Center
- Main research areas:
  - Performance Analysis Tools
  - Energy Efficiency
  - Computer Architecture (CCoE, IPCC)
  - Standardization Efforts (OpenACC, OpenMP, OpenSHMEM, …)



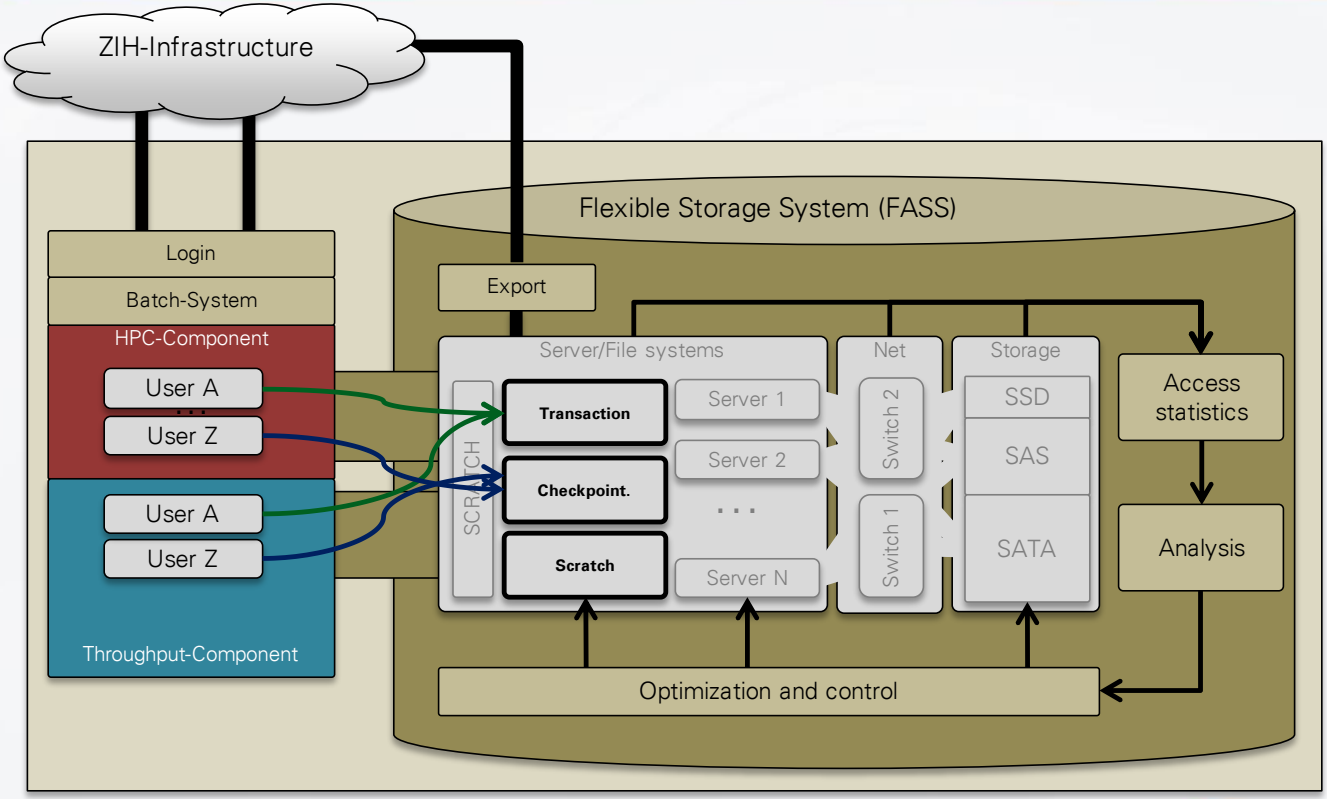
I/O related Topics

- I/O performance analysis
  - visualization of performance data
  - automated analysis
- design of I/O architectures related to
  - exascale system design
  - (big data and per user) workflows
- data management
  - dCache
  - iRods

ZIH HPC and BigData Infrastructure



Architecture of our storage concept (2010)



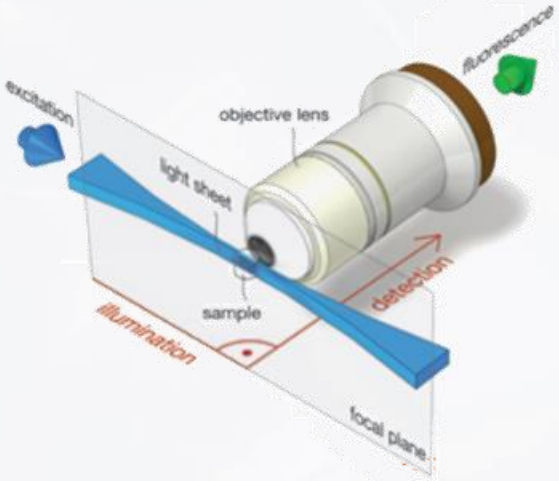
I/O intensive User Codes

- Simulations: CFD, Material Science, Climate, Electrical Engineering, Live Sciences, Physics, …
- everything from checkpoints to in-situ analysis
- the codes use HDF5, NetCDF, Silo, “bunch of images”, …
- research towards different Big Data interfaces

- we see all kinds of effects in the file system (imbalanced use of storage targets etc., one week >70% peak bandwidth)
- basically random I/O at the storage layer (lots of 1 MB requests though …)

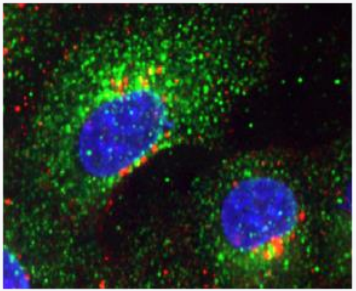
Application Example: Big Microscopy Data

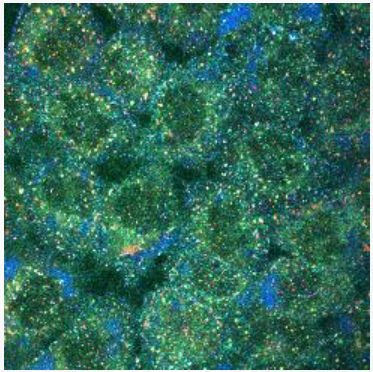
- Selective Plane Illumination Microscopy (SPIM)
- Typical configuration with one camera:
  - 0,85 GB/s and 10 files/s with goal of 24/7 operation
  - Monthly: 2 PB in 26 million files
  - More advanced types planned



Big Microscopy Data - Approaches

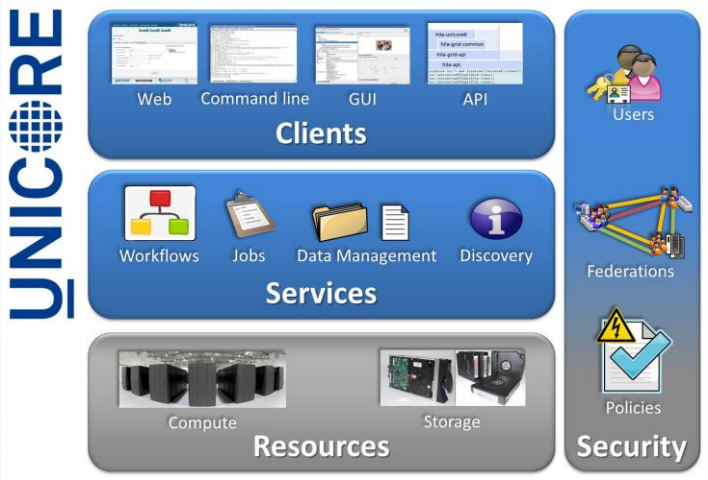
- Current workflow operational for specialized image analysis suite
- Not suitable for SPIM data anymore
- Goal of generic approach using UNICORE middleware with HPC integration
  - Data oriented processing for automating standard preprocessing tasks
  - UNICORE FTP for efficient parallel file transfers
  - Flexible HPC-integrated workflows for complex analysis graphs





Big Microscopy Data - UNICORE Middleware

- Complete middleware stack for computing and data management
- European development led by JSC
- Used for many large supercomputers in Europe (PRACE) and USA (XSEDE)
- Core of EU flagship Human Brain Project



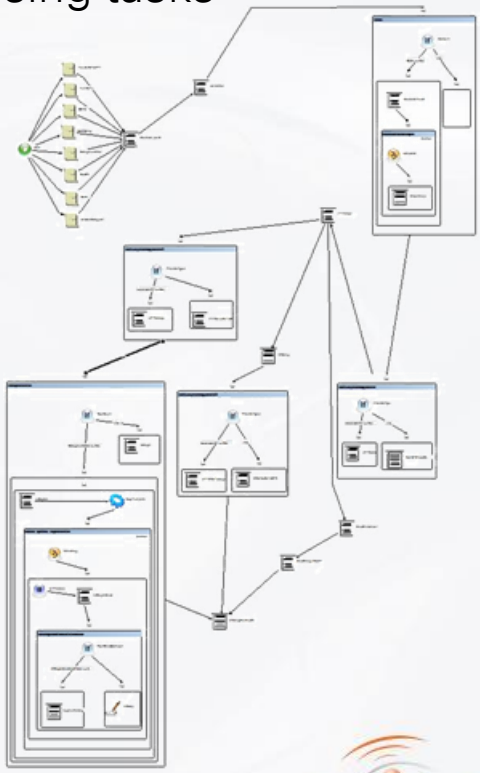
Big Microscopy Data - Data Oriented Processing

- For preprocessing of raw image data
- UNICORE monitors directory
- Pre-defined rule evaluated for new files
- Result: data directly triggers analysis

```
Name: computeMD5Sum, Match: ".*\\.pdf",
Action: {
    Type: BATCH,
    Job:  {
      Executable: "/usr/bin/md5sum",
      Arguments: ["${UC_FILE_PATH}"],
      Exports: [
        {From: "stdout",
        To: "file://${UC_BASE_DIR}/checksums/${UC_FILE_NAME}.md5"},
      ],
    }
```

Big Microscopy Data - Workflows

- easy automation of complex processing tasks
- Graphs of complex pipelines
- has conditionals and loops
- Defined once – used often
- Fosters re-usability



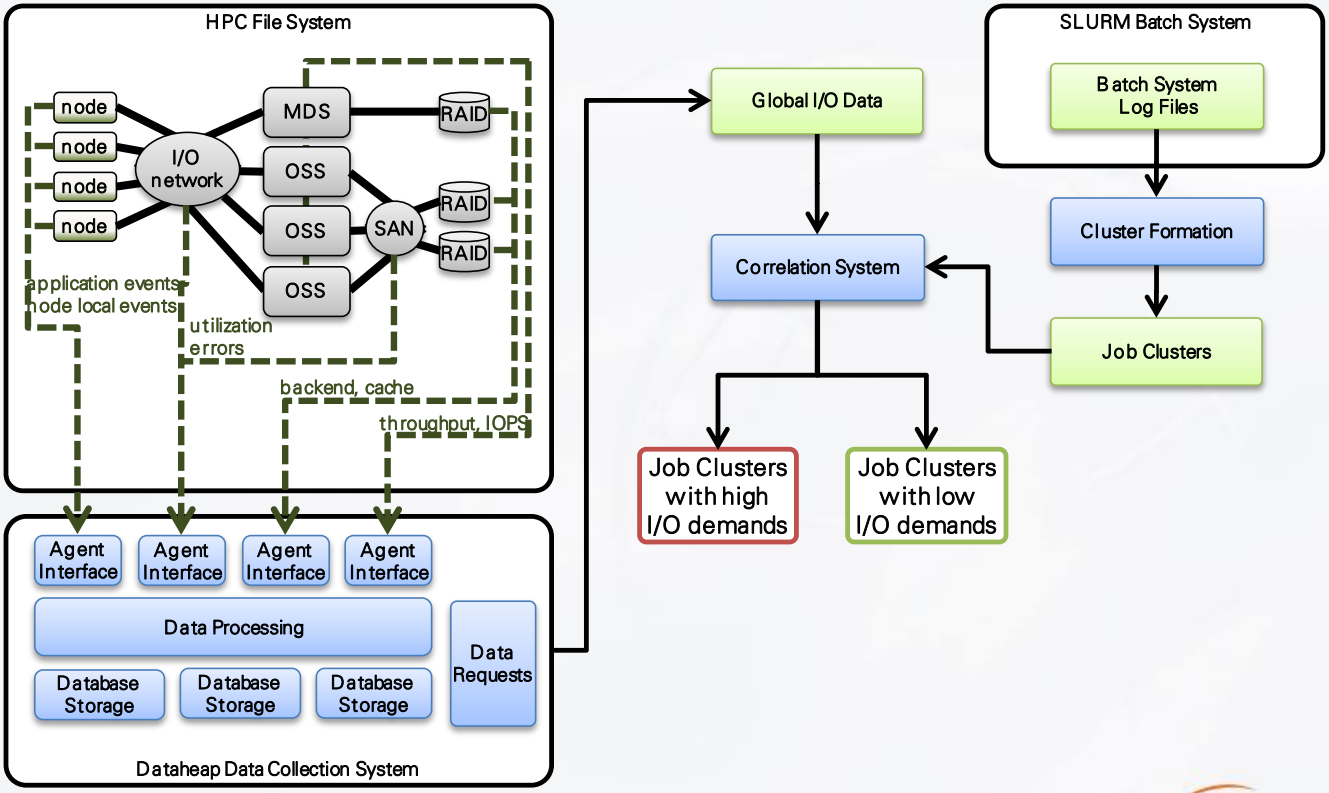
dCache @ ZIH

- 3 active usergroups (Grid-VOs, no single users)
  - ATLAS, SNO+ (physics), biomed (life sciences)
- almost only analysis jobs
- ~7 TB usage (mostly ATLAS)
- ~1 TB read/day for site-local computation
- every few months new data is bulk-transfered from upper-tier data centers
- almost no writes/day (< 1 GB)
- all data stored on disk (Lustre), no HSM (as usually utilized by dCache)

Performance Analysis for I/O Activities (1)

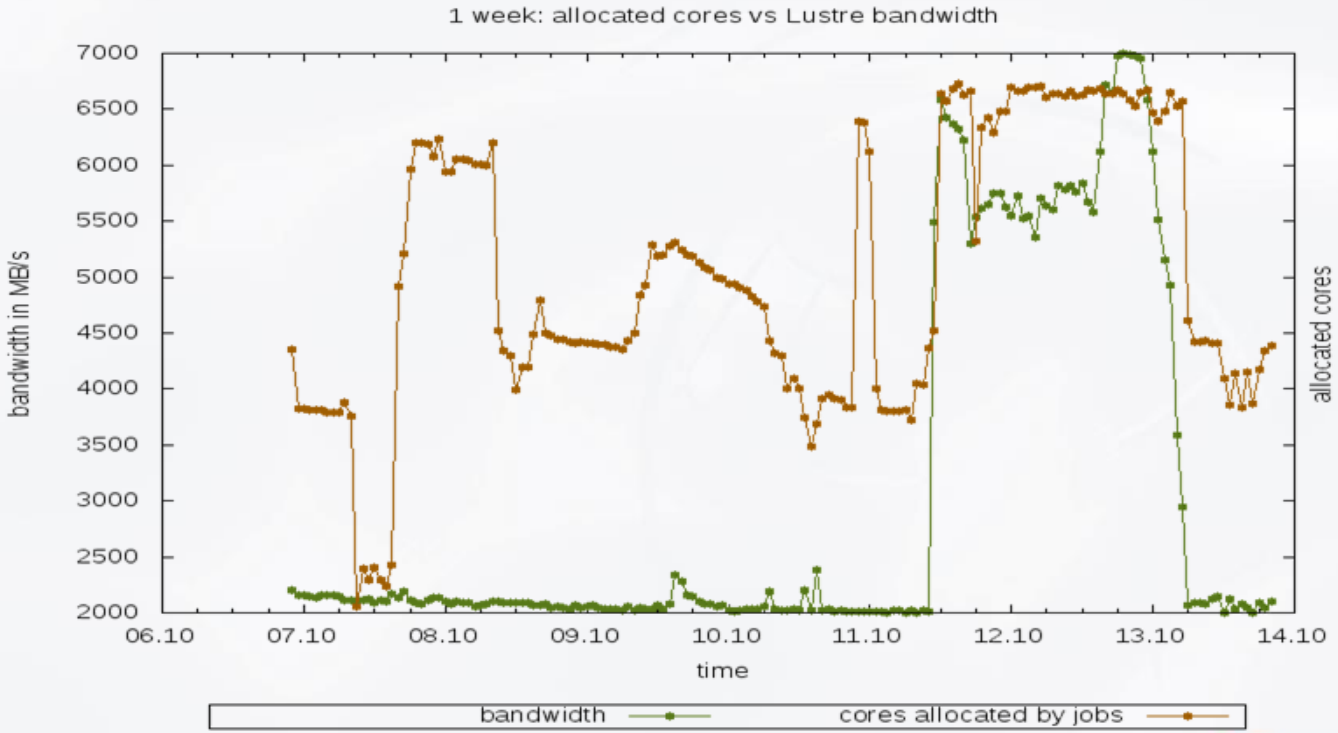
- What is of interest
  - Application requests, Interface types
  - Access sizes/patterns
  - Network/Server/Storage utilization
- Level of detail:
  - Record everything
  - Record data every 5 to 10 seconds
- Challenge:
  - How to analyze this?
  - How to deal with anonymous data?

I/O and Job Correlation: Architecture/Software

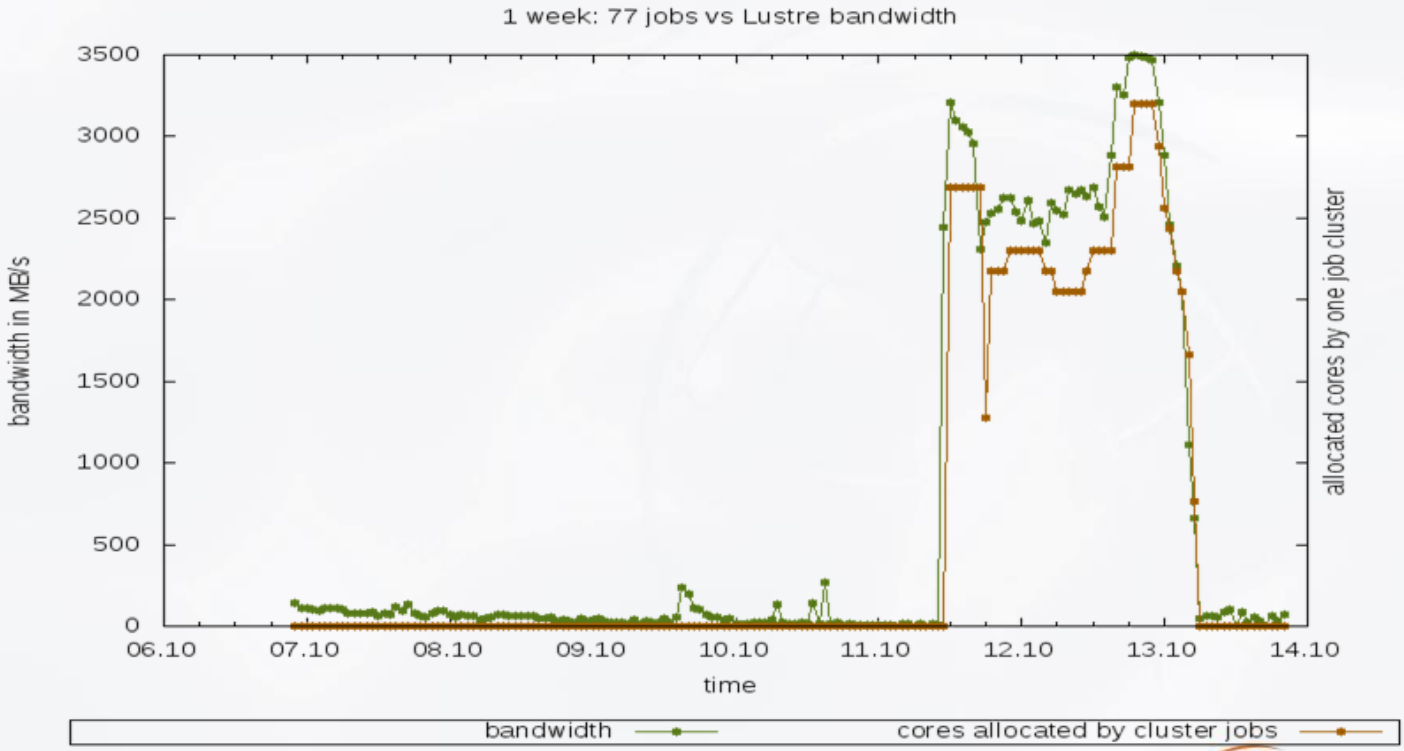


https://fusionforge.zih.tu-dresden.de/projects/dataheap/

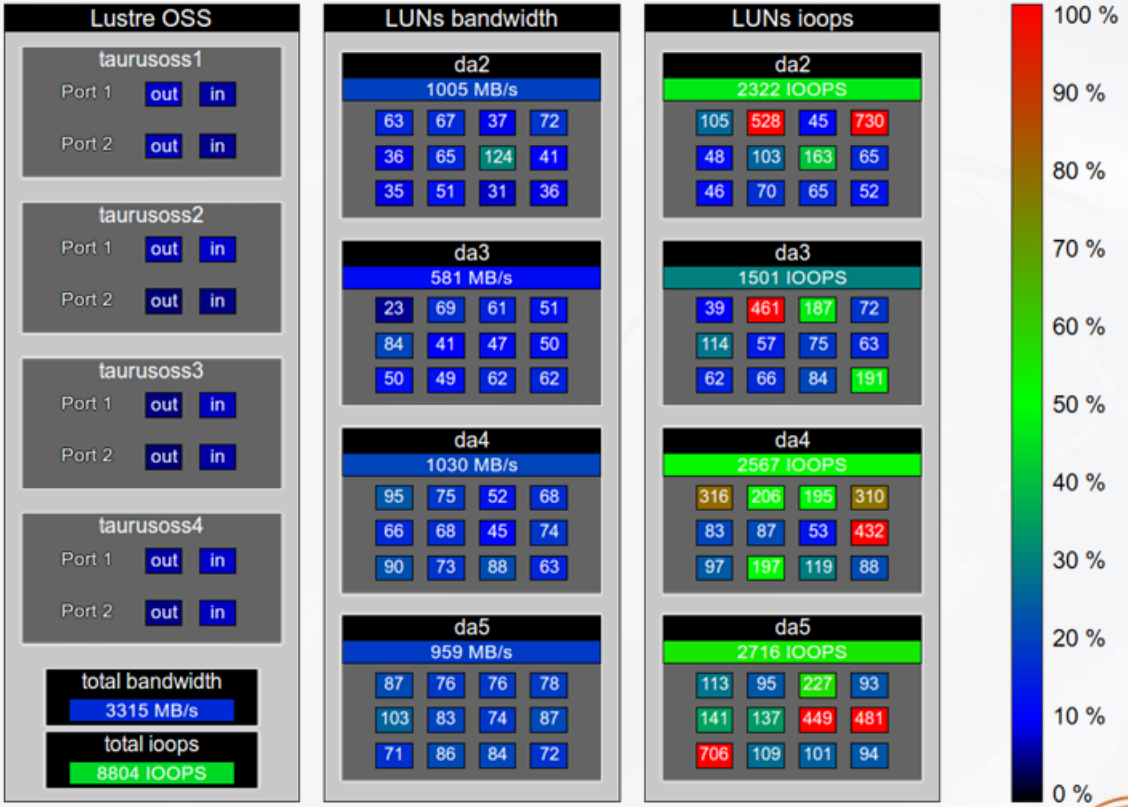
I/O and Job Correlation: Starting Point (Bandwidth)



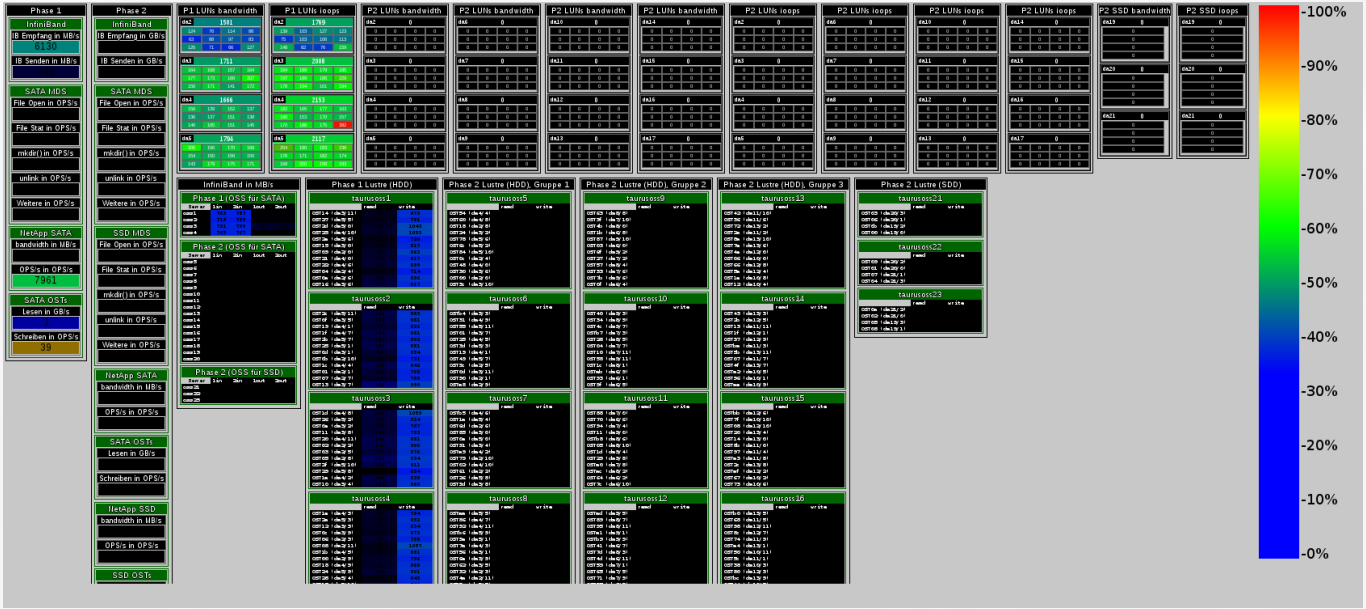
I/O and Job Correlation: Result



Visualization of Live Data for Phase 1



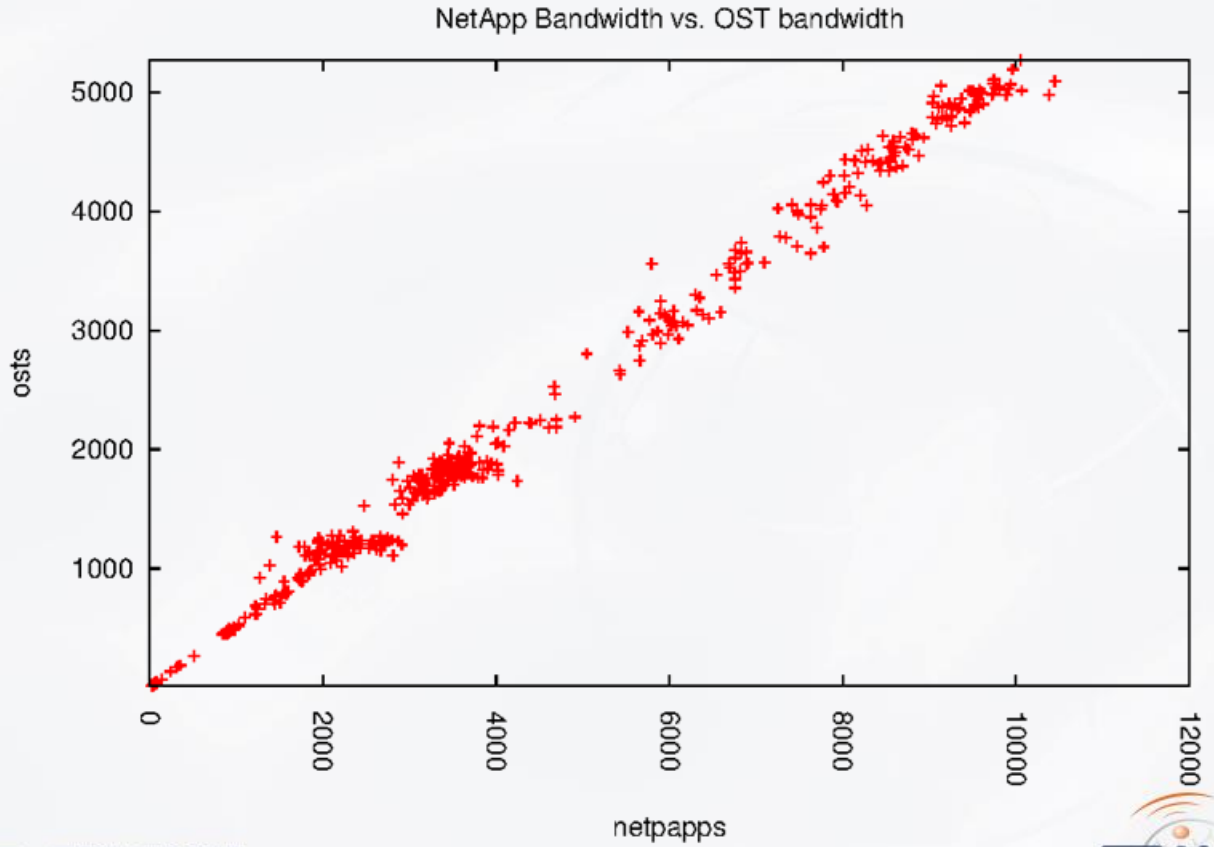
Visualization of Live Data for Phase 2



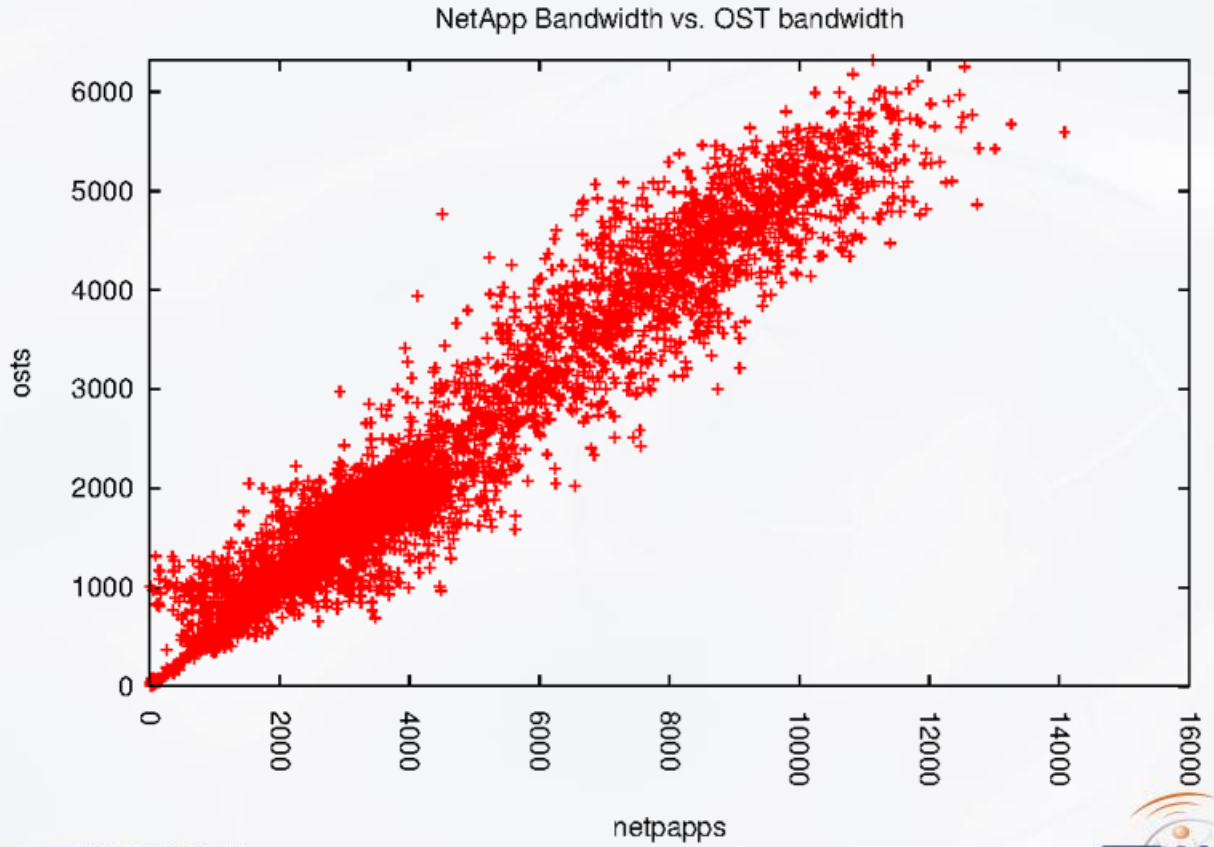
Performance Analysis for I/O Activities (2)

- Amount of collected data:
  - 240 OSTs, 20 OSS servers, …
  - Looking at stats and brw_stats (how to store a histogram in database?)
  - → ~75.000 tables
  - about 1 GB of data per hour
- Analysis is supposed to cover 6 month
  - 4 TB data
  - No way to analyze this with any serial approach

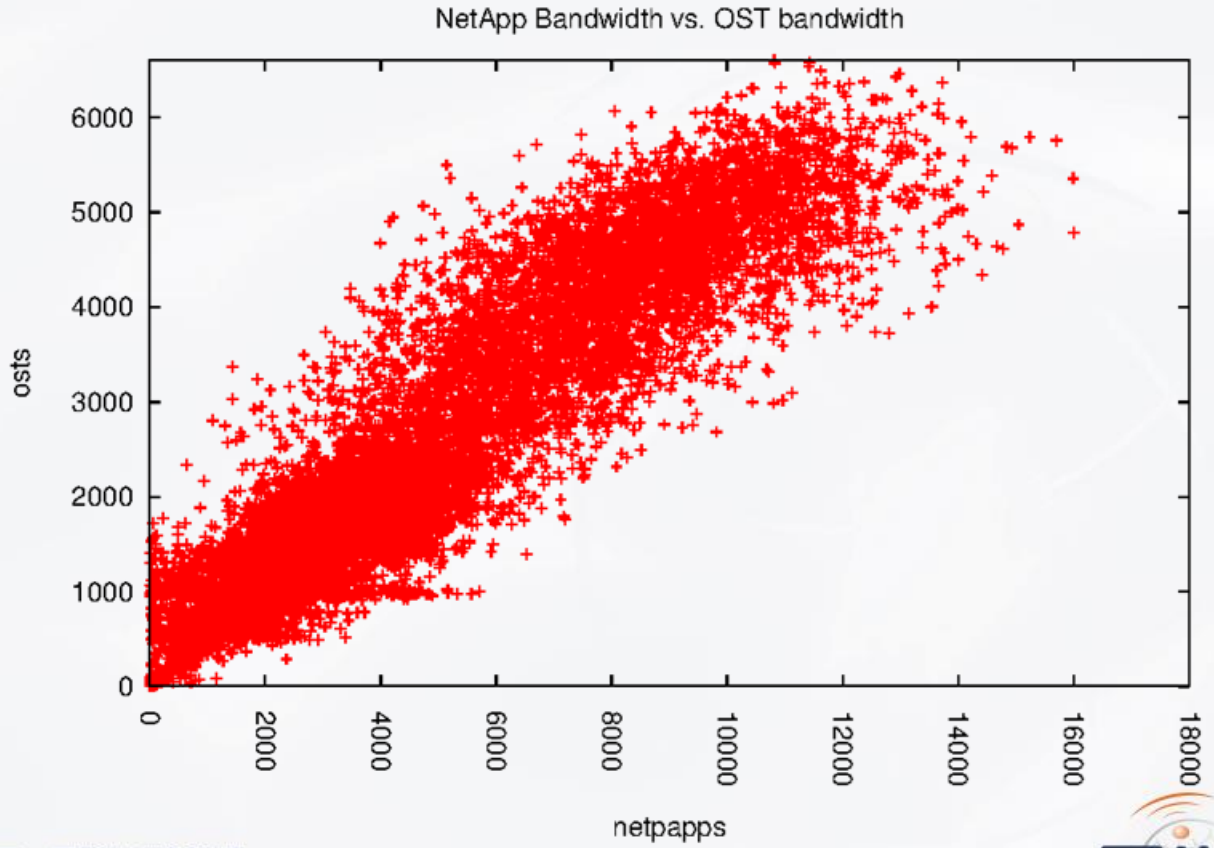
Metric Example: NetApp vs. OST Bandwidth (10 minutes)



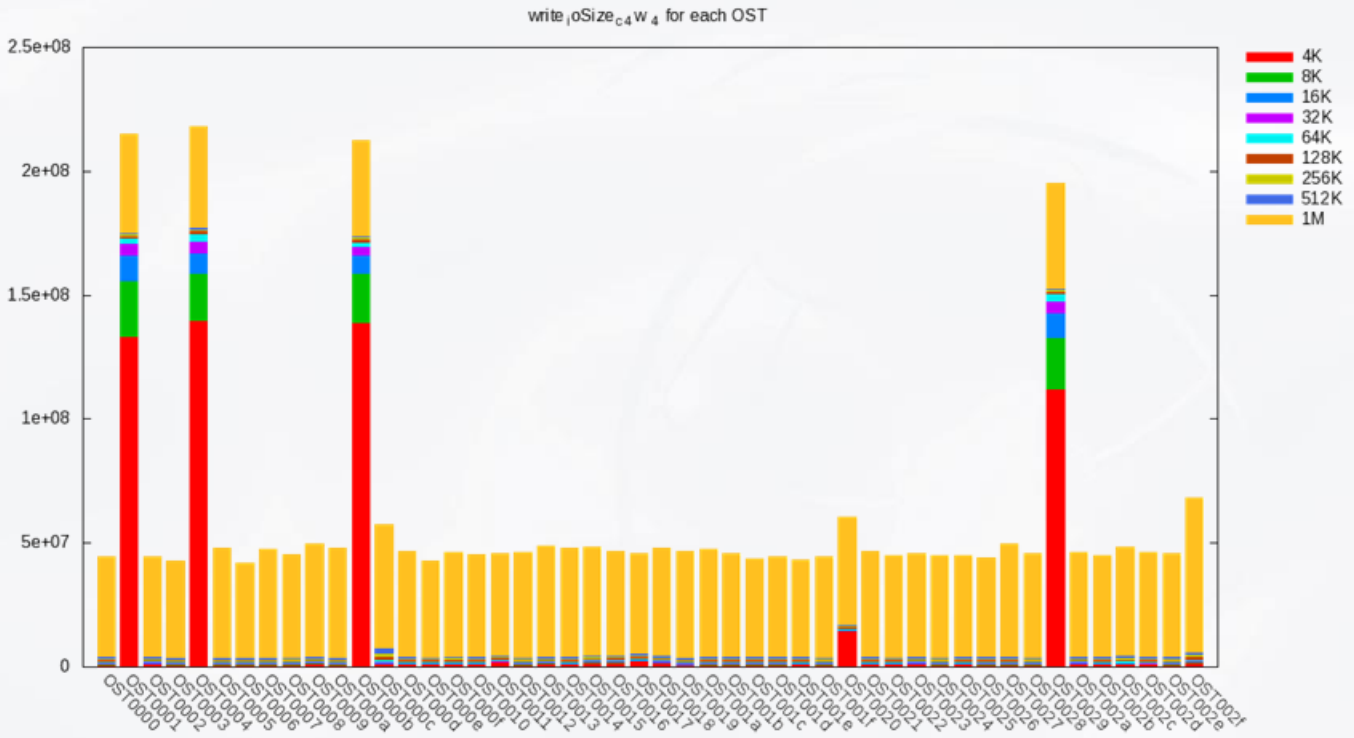
Metric Example: NetApp vs. OST Bandwidth (60s)



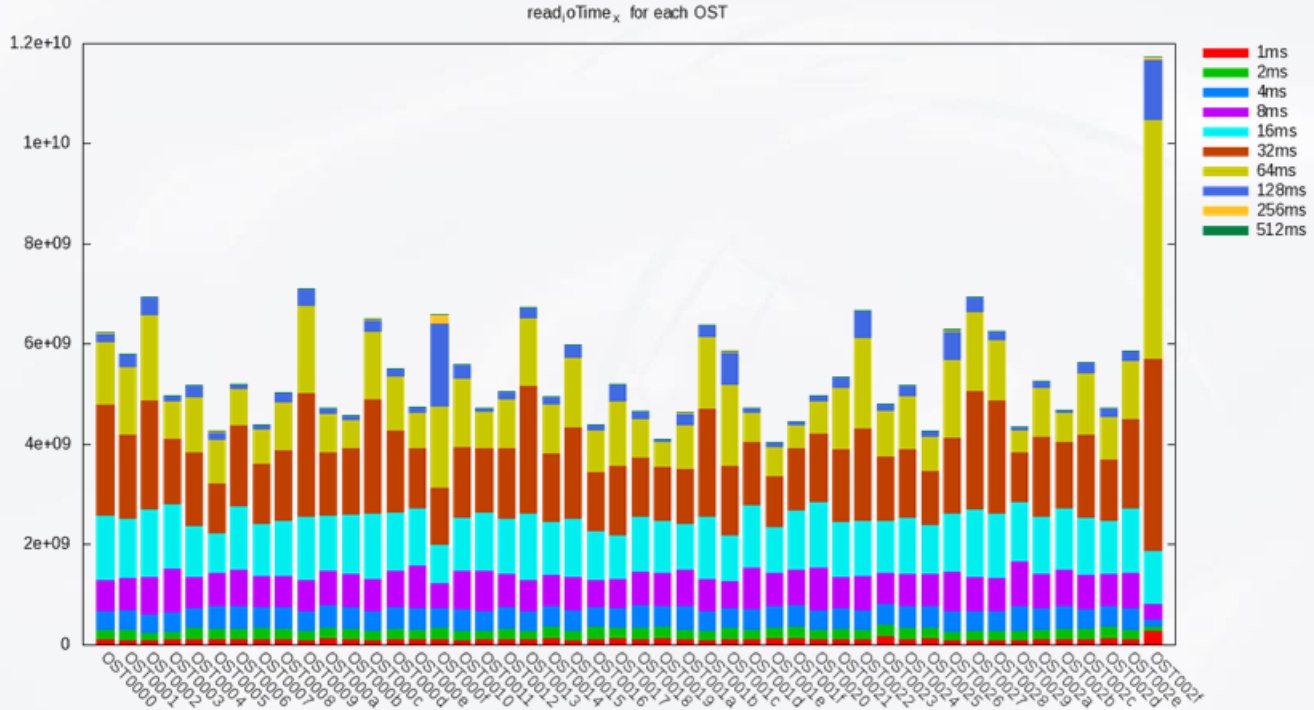
Metric Example: NetApp vs. OST Bandwidth (20s)



Imbalanced use of storage targets (1)



Imbalanced use of storage targets(2)



Conclusion / Questions

- I/O performance analysis and system design is our expertise
- we run many services …