High Performance Storage for Science Workloads

THE DATA STORAGE CHALLENGE OF PHYSICS IN THE 21st CENTURY

U. FUCHS / CERN
CERN / ALICE

WHO WE ARE AND WHAT WE DO
CERN

- CERN is the world's largest particle physics laboratory funded by 21 European member states
- ~2000 staff, ~9000 visiting physicists
- Physics goals are to study elementary particles and fundamental forces
- Particles physics requires:
  - special tools to create new particles: particle accelerators: LHC
    "Steer a beam of 85 kg TNT through a 3mm hole 10000 times per second"
  - special instruments to study new particles: the experiments: ALICE, ATLAS, CMS, LHCb
The ALICE Experiment

Detector characteristics
Width: 25m
Diameter: 15m
Weight: 10'000t
ALICE Physics

• The Readout Challenge
  • 40 million collisions per second
  • 100 million readout channels
  • Terabytes of data per sec to be read and dealt with

What we see

Reconstructed tracks with pt > 26 GeV

What we are looking for
CERN Computing History

- Due the analysis and storage needs, CERN has a long lasting history in computing
  - First version of machine X
  - First network, first Ethernet n/w
  - First “internet” (TCP/IP)
  - Development of the WEB (http)
  - ...

![CERN Computing History Image]
A typical data acquisition system, 2018

• 8500 links read by 250 servers

• >2000 port network, ~2.5 TBps

• 1500 servers for real-time data formatting (+4 GPUs)

• ~100 PB, $10^9$ files, ~450 GBps

• Data Management facilities, Tier-0 storage
The ALICE storage system – first test results

PUSHING THE LIMITS
Transient Data Storage, “The Can”

- High-Capacity, High-Throughput file system
  - ~100PB, ~450GBps, $10^9$ files
- High number of clients: ~2000
- Mixed access (r and w) per client
- There are also good news: Operations strictly linear
- Few candidates on the market, three retained:
  - Lustre (v2.6.32)
  - GPFS (v4.1.1)
  - CEPH/RADOS (Hammer)
File Systems Considerations

• Lustre
  • Clustered File system: data servers, meta-data server
  • Beware of MDS bottlenecks, whole meta data should fit in memory to avoid disk i/o

• GPFS
  • All i/o striped over all servers/LUNs
  • Distributed meta data or separate MDS server possible

• RADOS
  • Object storage, “get”-“put”-“list” interface
  • Underlying storage pools made for redundancy and zero data loss

• CEPH
  • POSIX file system interface on top of Rados data stores
Transient Data Storage, Tests

• Tests are still ongoing
  • Out-of-the-box tests finished
  • Now working with vendors to tune the system

• Test: linear workload (r/w), big files, big block sizes
  • It’s all about throughput
  • We’re not (yet?) testing iops performance

• Mixed workloads, IOPS
  • Not our primary concern
  • Tests will be done in collaboration with other institutions
Transient Data Storage, Test Setup

- Test environment

- 6 LUNs, 500MBps ea, per storage chassis
  - MD3660 chassis with 30 disks 4TB

- Centos 7

- Infiniband FDR only tested for Lustre
Test Results

- Performance vs (Application) Block Size
- 1 client on 10GE/IB
- 1 stream
Test Results

- Performance vs # of streams
  - 1 client on 10GE/IB
  - x streams
Test Results

• Performance vs # of clients
  • x clients on 10GE/IB
  • 1 stream
Wrap-up
WHERE WE ARE TODAY
Observations

- GPFS, LUSTRE: hitting network limitations
- CEPH, RADOS: made for safety, not speed
  - Future versions will improve performance, so they say.
  - Linear writes cause 30-40% read i/o on disk level (journal)
- Data Integrity
  - Lustre: based on underlying LUNs (e.g. RAID)
  - GPFS: good protection through “declustered RAID”
- Rebuild Times:
  - Lustre: based on underlying LUNs
  - GPFS: Minutes
- All-SSDs and 100G networking will change the picture
Conclusions

• There’s a big storage challenge ahead of us
• If you want to play the BIG game, there are only few solutions, choose wisely.
Thank you.