ALICE / LHCb Storage Needs for Run3

U. FUCHS / CERN

ALICE & LHCb storage systems for Run 2 and 3

Storage needs & systems

- For Run 2: In Production
- For Run 3: Under design / development

Future ? Numbers ?

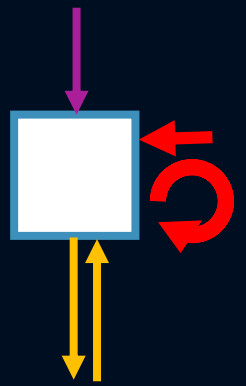
Cum Grano Salis !

Disclaimer:

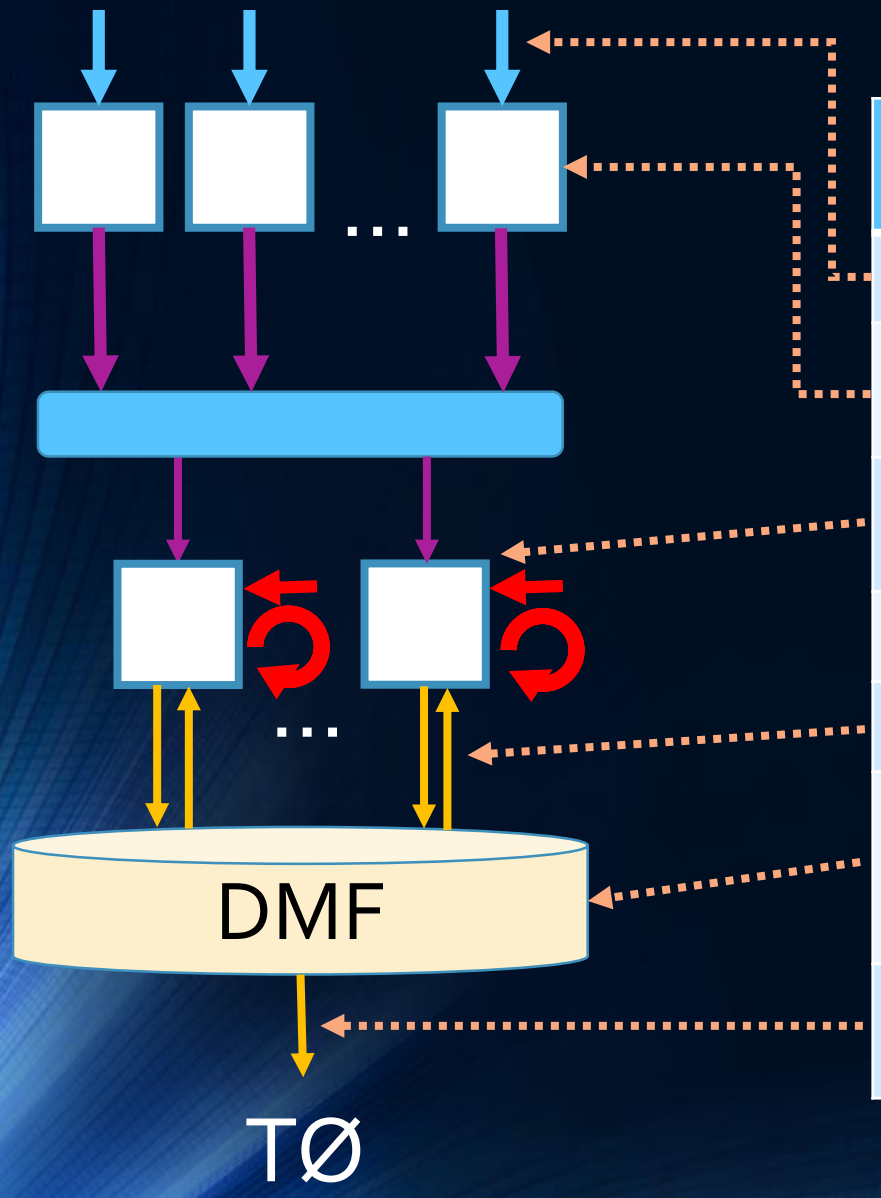
This talk is not suited for people on a low-sodium diet.

Data Flow: ALICE

- Strategy Run 2:
  - Raw data from the DAQ is moved to Tzero for reconstruction /analysis
- Strategy Run 3:
  - Compute nodes build time frames, run calibration, fast tracking, store
  - Data stored on-site
  - When cpu cycles are available (during data taking, fills, pauses ..):
    - Read-back data, re-calibrate, reconstruction.



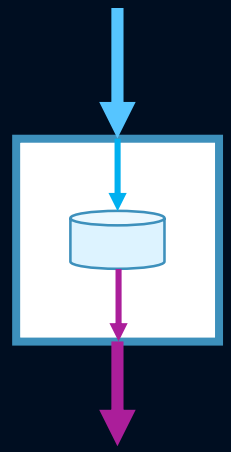
Schematic Data Flow: ALICE



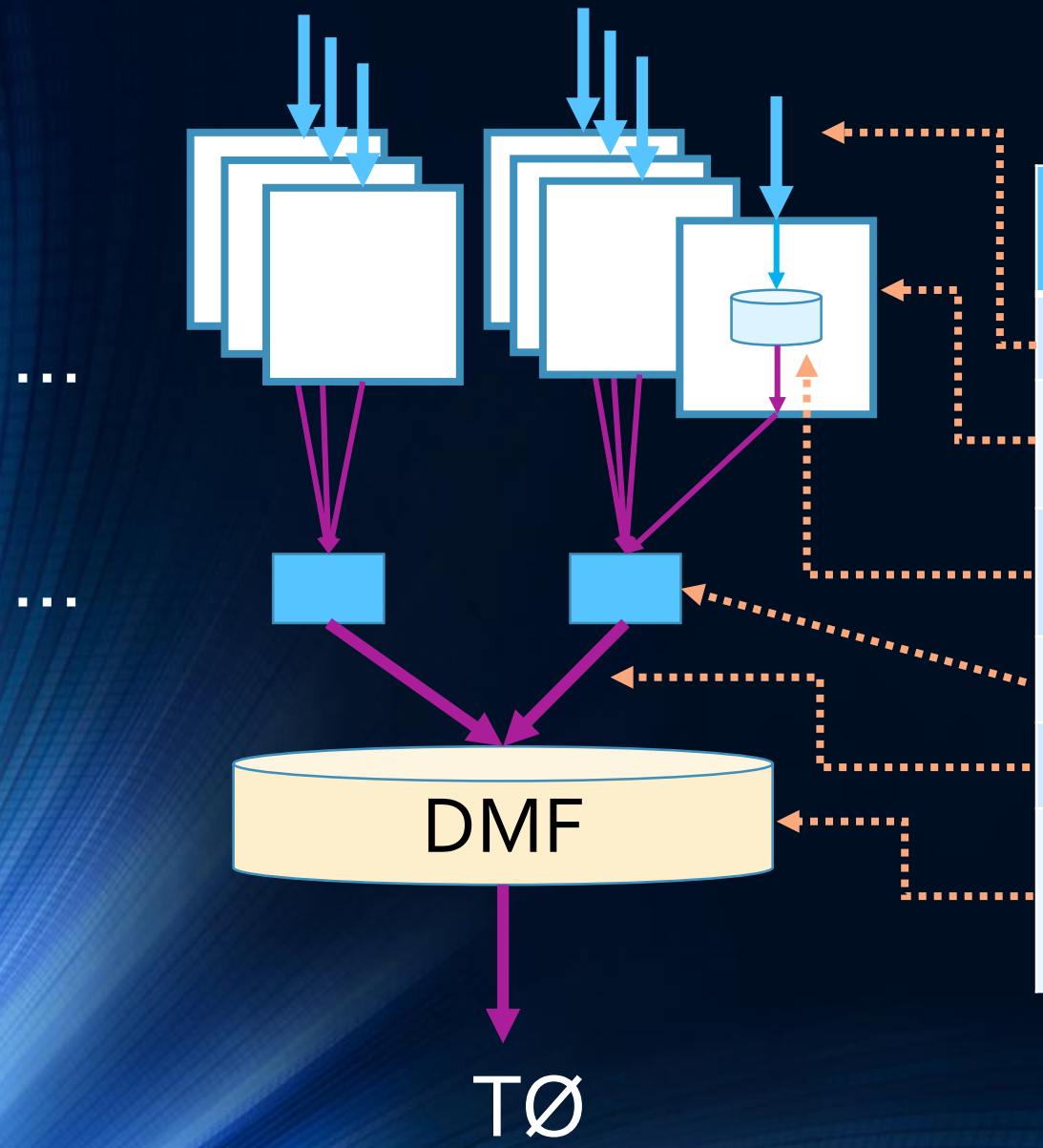
	Run 2	Run 3
Total Data Input	60 Gbps	> 5 Tbps
Readout Servers	200 0.1-10 Gbps/server	250 24 Gbps/server
Event Builder	20 4-5 Gbps/server	1500 3-5 Gbps/server
	L2: External source	EBuild./FTrack./Calib.
Event Builder Out	Max 700-1000 MBps	> 80 GBps
Data Mgmt Facility	1 PB (1.5 days data) avg. (12w+10r) GBps 1200 disks	~100 PB (1y of event and analysis data) avg. (80w+60r) GBps
Data Mover	10 servers max. 1 GBps / server	1/3 of event data / all analysis copied to T0

Data Flow: LHCb

- Strategy :
  - Read-out nodes record locally
  - L2 processing done on r/o node
    - Between fills: at 100%
    - During fills: if CPU cycles available: at reduced speed
  - Advantage
    - No shared file system needed up to R/O node level
  - Issues
    - Nodes process at its own pace
      - Further processing of specific runs delayed until all nodes finish
    - Node down / drive failure
      - Further processing of concerned runs delayed until node is repaired

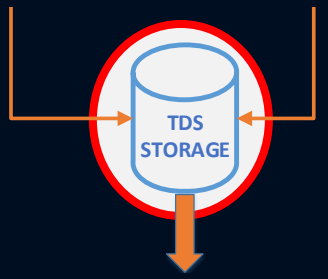


Schematic Data Flow: LHCb



	Run 2	Run 3
Total Data Input	60 GBps	40 TBps
Readout Servers	2000 30 MBps/server	8000 5 GBps/server
Internal Storage	2-4 TB 7 MBps x2 (rw)	14-32 TB 27 MBps x2 (rw)
I/O aggregation	4 Proxy Servers	Proxy Servers
Total I/O	Max 13 GBps	> 30 GBps
Data Mgmt Facility	340 TB (1 wk data) avg. (1w+2r) GBps 140 disks	4 PB (1 wk data) avg. (10w+20r) GBps >3000 disks

Transient Data Storage, “The Can”



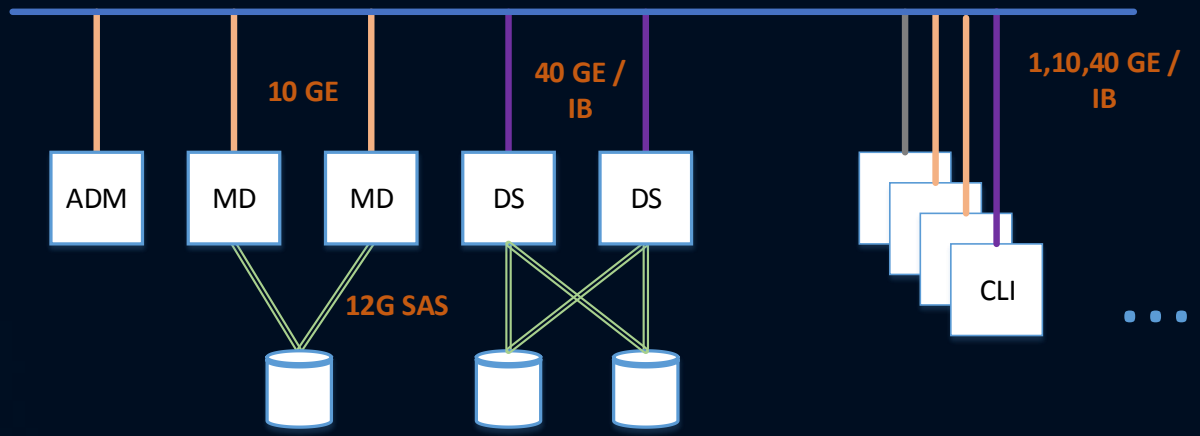
- High-Capacity, High-Throughput file system
  - ~100PB, ~200GBps, 10^9 files
- High number of clients: ~2000
- Few candidates on the market, retained:
  - Lustre (v2.6.32)
  - GPFS (v4.1.1)
  - CEPH/RADOS (Hammer)
  - EOS/CERN (tbd)

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

Storage Test, Setup

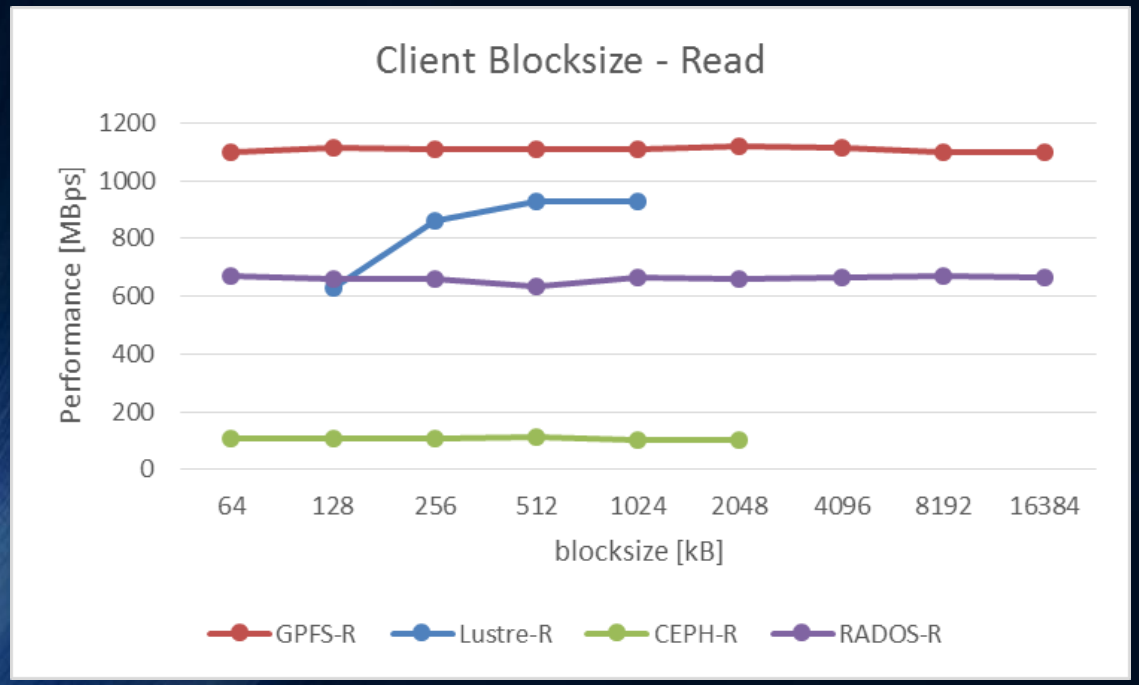
- Test environment

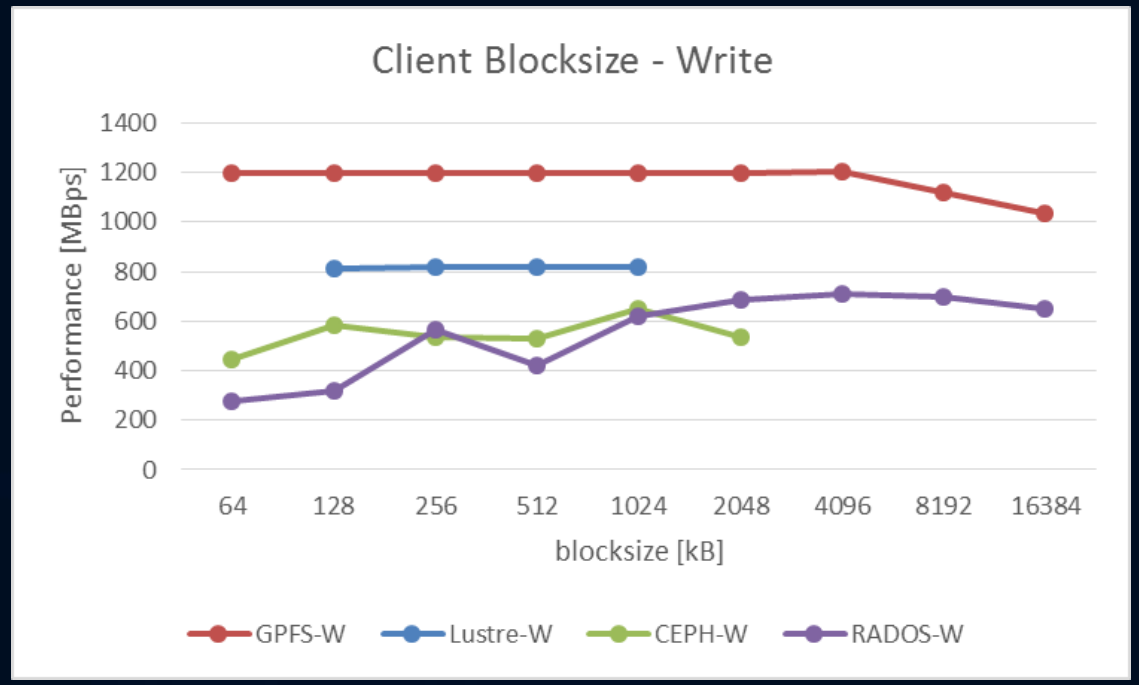


- 6 LUNs, 500MBps ea, per storage chassis
  - MD3660 chassis with 30 disks 4TB
- Centos 7
- Infiniband FDR only tested for Lustre

Test Results, Linear Access

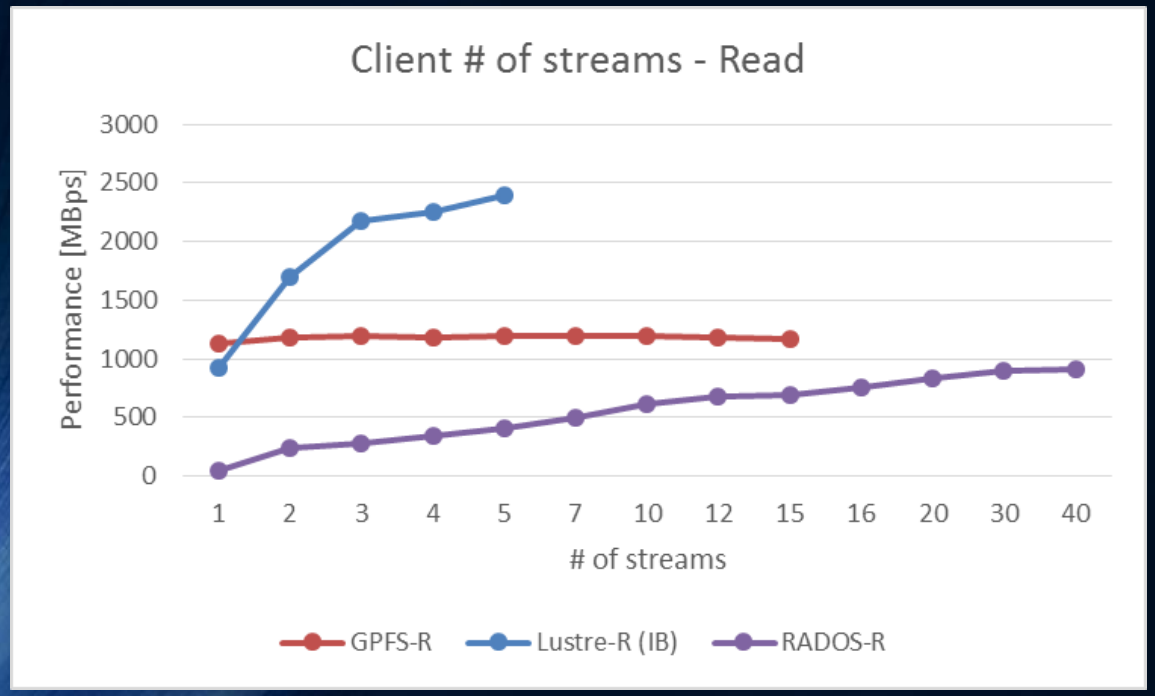
- Performance vs (Application) Block Size
  - 1 client on 10GE/IB
  - 1 stream

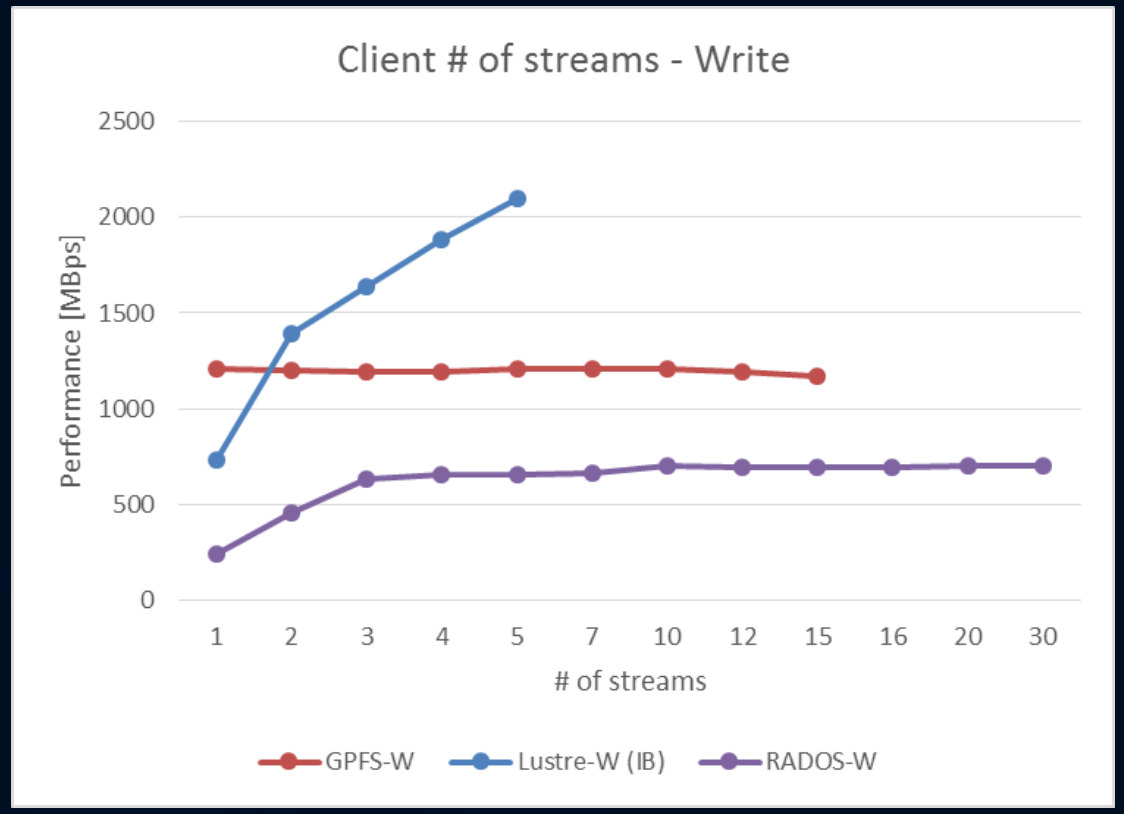




Test Results, Linear Access

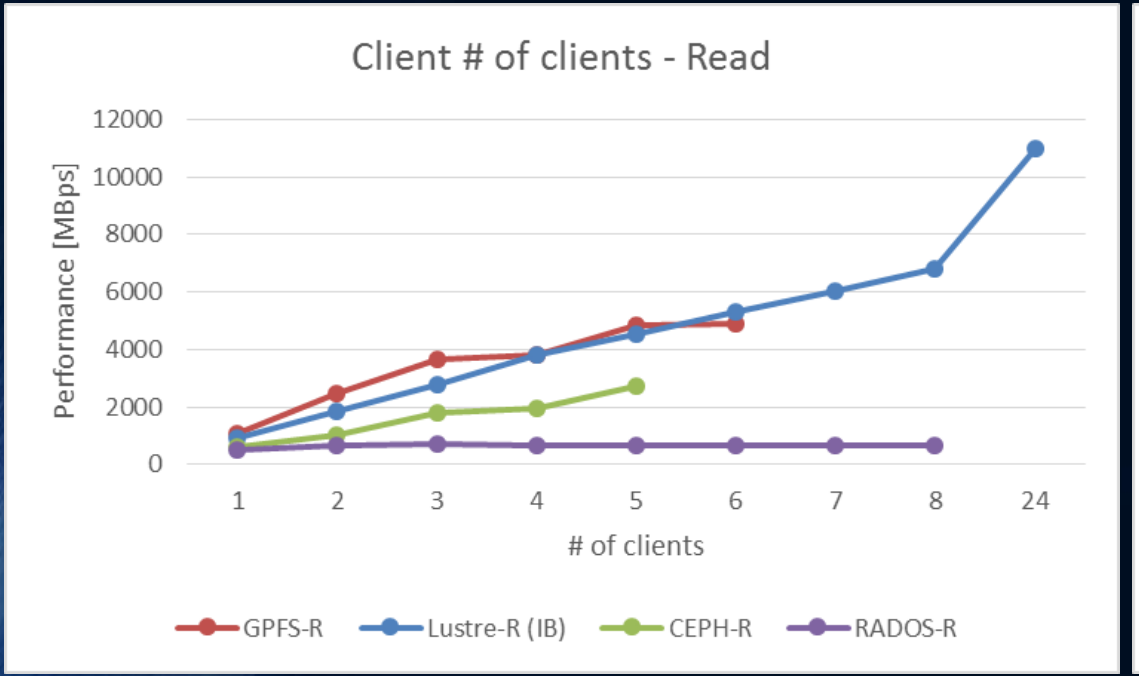
- Performance vs # of streams
  - 1 client on 10GE/IB
  - x streams

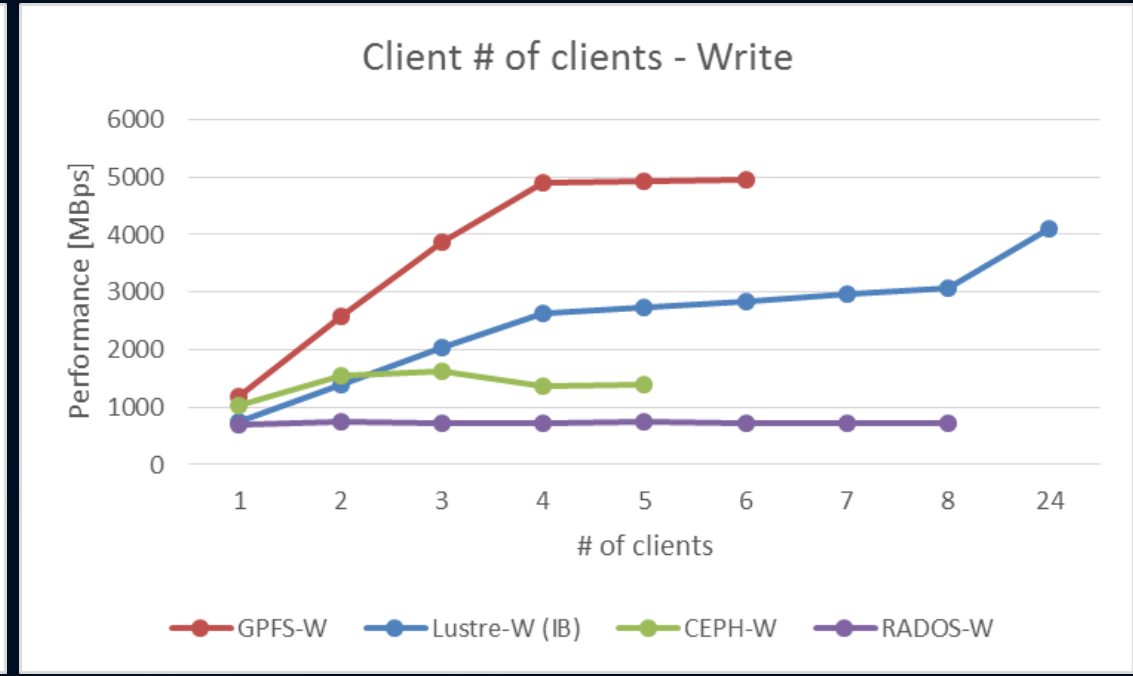




Test Results, Linear Access

- Performance vs # of clients
  - x clients on 10GE/IB
  - 1 stream





Storage Solutions

- There are challenges
  - Capacity
  - Performance
  - Latency, access, metadata, ..
- There are solutions
  - Depending on today's biggest mystery: the data access pattern
  - .. And feature needs: Access policies, migration, remote replication, tiering, ..
- There are surprises (nice on the outside, ..)

Thank you.