HEPiX Storage Working Group
- progress notes 10/2012 -

Andrei Maslennikov
CASPUR/CINECA

October, 2012 – Beijing

Summary

- Raison d'être
- Activities July-October 2012
- Current results
- Discussion

Raison d'êtr e

- The group was created in the end of 2006 to make an assessment of the most popular HEP storage solutions and to compare them.
- In the period of 2007-2011 we ran two major storage questionnaires and performed 11 series of comparative performance measurements with realistic use cases.
- The group is trying to track the storage technologies and to perform evaluation of some of them, along with the periodic reassessment of the situation at the participating sites.

Activities July-October 2012

- At the end of July the group had received a new block of worker nodes at KIT to potentiate the load farm in the storage laboratory. The nodes were installed and configured by mid-September, and new test series began immediately.
- This time we were concentrating on the most diffused solutions rather than on the new ones. The goals were to upgrade all components to the latest levels and obtain an updated snapshot of the situation.
- In parallel, in the second half of September, we have prepared a new edition of Storage Questionnaire

Credits 2012

- The test laboratory at KIT was built on the top of hardware kindly provided by Karlsruhe Institute of Technology (rack and network infrastructure, load farm) and by CASPUR (disk server). CERN contributed with funds to cover a part of human hours.

- These people participated in provisioning, funding, discussions, laboratory building, preparation of test cases and test framework, tests, elaboration of results and collecting data for Storage Questionnaire:

BNL	C.Caramarcu, Y.McCarthy, O.Rind, T.Wong, D.Yu
CASPUR	A.Maslennikov(Chair), M.Calori (Webmaster)
CEA	J-C.Lafoucriere
CERN	M.Lamanna, A.Wiebalck
DESY	M.Gasthuber, P.van der Reest
INFN	G.Donvito, V.Sapunenko
IN2P3	P-E.Brinette, Y.Calas, J-Y.Nief,L.Tortay
JLAB	S.Philpott
GSI	H.Göringer, T.Roth
KIT	J.van Wezel, Ch-E. Pfeiler, M.Alef, B.Hoeft
RAL	M.Bly
SARA	J.Saathof, R. Starink
TRIUMF	Th.Lindner, S.McDonald

Storage Questionnaire

Summary 2012

- As before, the data for Questionnaire were collected in electronic format using a special web form. Participants were asked to describe their most representative storage areas.
- The current data may be consulted at the folowing URL:

  http://w3.hepix.org/storage/questionnaire1.php (u/pw: hepix/hepix)
- Some quick observations:

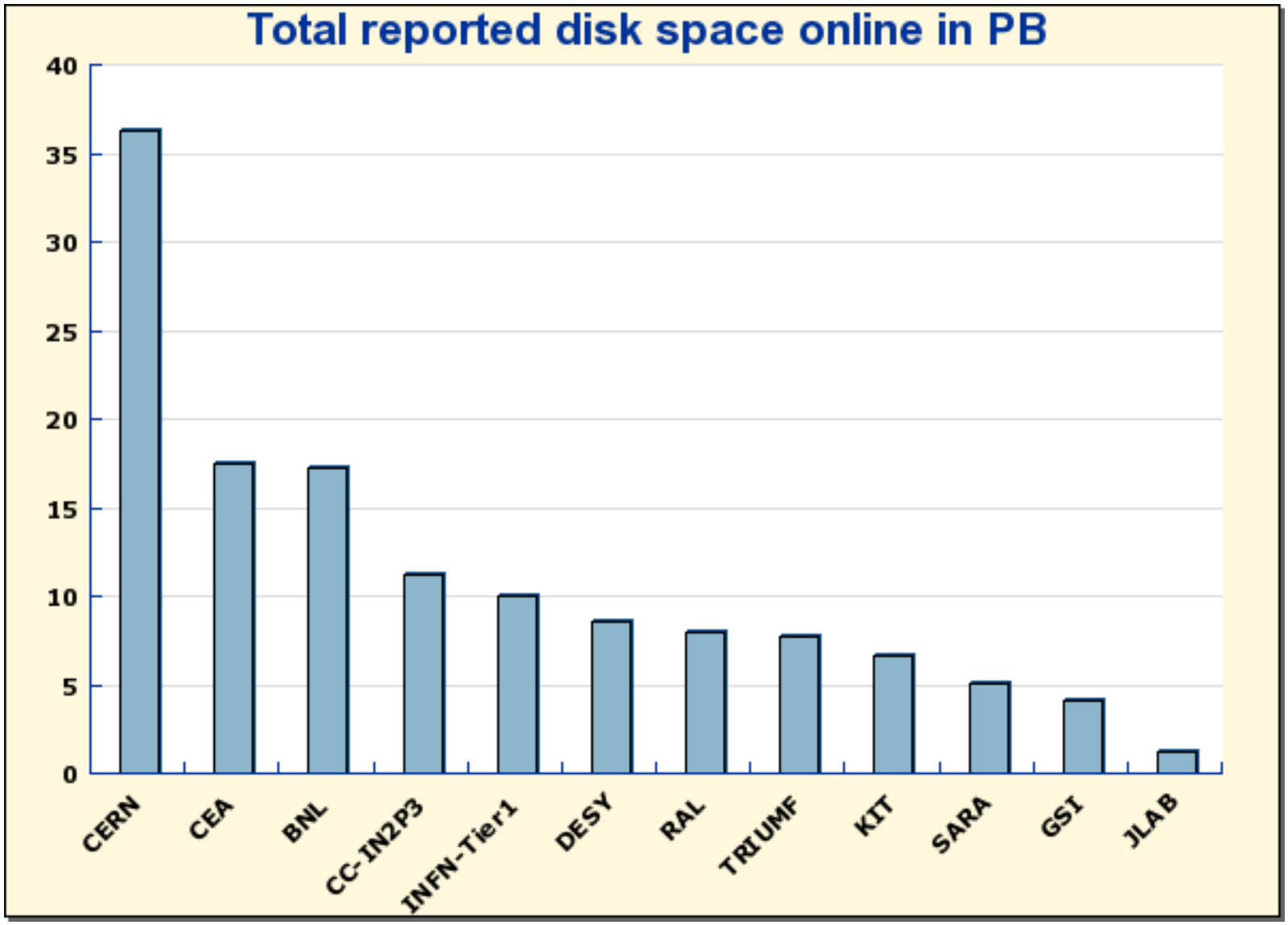
  - Three main data store/access technologies (dCache, Xrootd and Lustre)

  - New since 2010: EOS (CERN), SONAS (DESY), ZFS via NFS (JLAB)

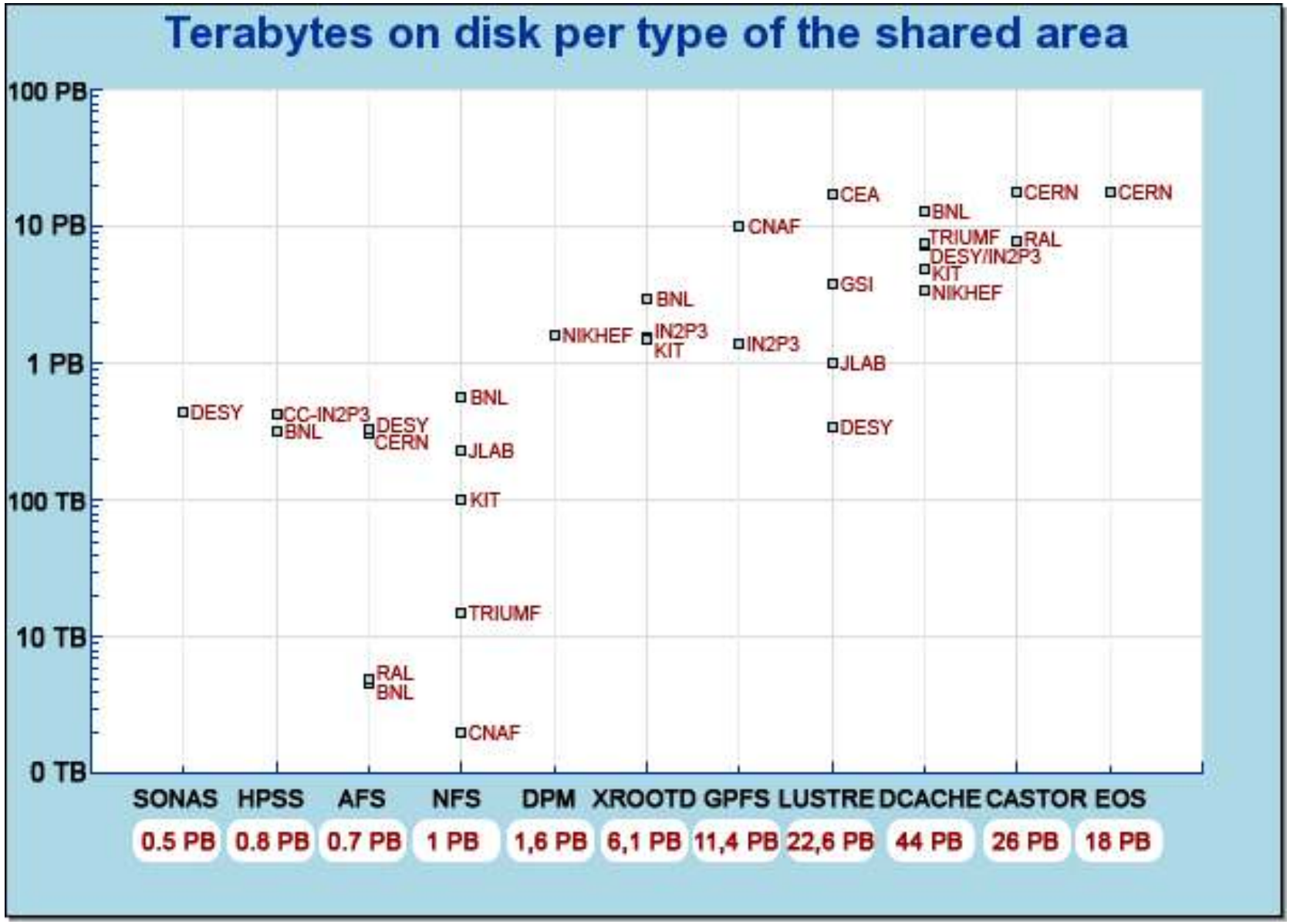
  - Sites differ a lot in terms of ratio Nclients/Nservers for the same type of filestore . There's hardly a way to locate any reliable pattern. While the use cases are similar across sites, the hardware is highly not uniform..

- Next slides: Total Volume, Breakdown Per Solution, Anatomy Of Datastores

Data volume 2012



Storage technology breakdown 2012



Anatomy of reported datastores (p.1)

ORG	SA	Tot PB	Ncli	Nsrv	Ncli / Nsrv
RAL	AFS	0,005	750	3	**250**
CERN	AFS	0,31	15000	55	**272,72**
BNL	AFS	0,005		4	
DESY	AFS	0,33	4000	58	**68,96**
RAL	CASTOR	8	750	500	**1,5**
CERN	CASTOR	18	6000	1300	**4,61**
CC-IN2P3	DCACHE	7,3	8000	100	**80**
KIT	DCACHE	5	1350	82	**16,46**
DESY	DCACHE	7,3		70	
NIKHEF-SARA	DCACHE	3,5	450	48	**9,38**
BNL	DCACHE/ATLAS	10	820	80	**10,3**
BNL	DCACHE/PHENIX	3,262	1200	608	**1,97**

Anatomy of reported datastores (p.2)

ORG	SA	Tot PB	Ncli	Nsrv	Ncli / Nsrv
INFN-Tier1	GPFS	10	2000	130	**15,38**
CC-IN2P3	GPFS	1,4	1100	43	**25,58**
CC-IN2P3	HPSS	0,432	1200	12	**100**
BNL	HPSS	0,325	1300	25	**52**
CEA	Lustre	2,5	5000	100	**50**
JLAB	Lustre	1	1200	24	**50**
DESY	Lustre	0,351	110	9	**12,22**
GSI	Lustre	2,3	1000	147	**6,80**
CEA	Lustre shared FS	15	4000	71	**56,33**
GSI	lustre-Hera	1,5	500	50	**10**
BNL	NFS	0,58	3000	8	**375**
KIT	NFS, GPFS	0,1	1350	5	**270**
TRIUMF	NFS, iSCSI, CIFS	0,015	10	1	**10**

Anatomy of reported datastores (p.3)

ORG	SA	Tot PB	Ncli	Nsrv	Ncli / Nsrv
TRIUMF	T2k-dcache	0,45	0	1	
TRIUMF	Tier1-dCache	7,2	1000	54	**18,51**
CC-IN2P3	Xrootd	1,58	3000	31	**96,77**
KIT	xrootd	1,5	1350	10	**135**
BNL	XRootD/STAR	3,02	1200	545	**2,201**
FNAL	BlueArc	0	2000	6	**333,33**
DESY	CIFS	0,11	4000	2	**2000**
INFN-Tier1	CNFS	0,002	2000	4	**500**
NIKHEF-SARA	DPM	1,6	430	18	**23,89**
CERN	EOS	18	6000	600	**10**
GSI	gStore	0,24	350	17	**20,58**
CC-IN2P3	iRODS	0,5		9	**0**
DESY	SONAS	0,442	300	20	**15**
JLAB	ZFS via NFS	0,23	1200	3	**400**

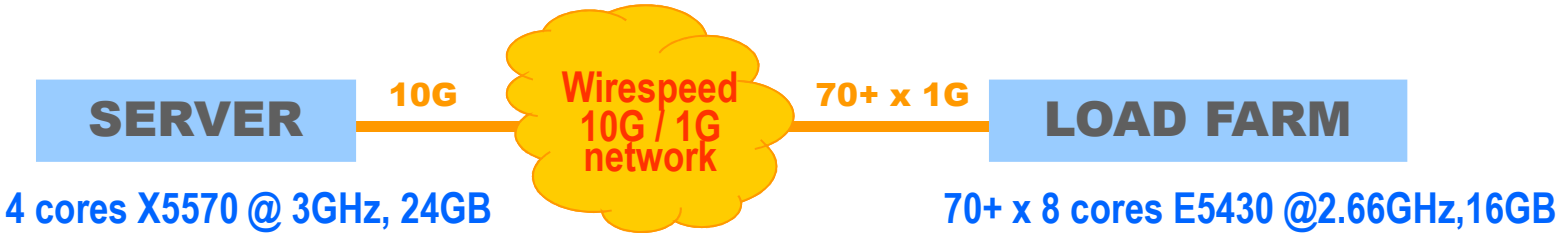
Storage Laboratory

Disclaimer

- We are constantly dealing with the “moving target”: data formats and use cases are evolving, hardware base is changing, new versions of storage access and archival software replace the old ones. This implies that results obtained in the storage laboratory are and will always remain a subject to change.

- Whatever we report should hence aways be seen as “work in progress”. We are not trying to provide any final recommendations but are rather sharing with you our findings and are ready to accept any advice and feedback.

Hardware setup 2012 at KIT



4 cores X5570 @ 3GHz, 24GB
3 Adaptec 5805 8p RAID controllers
24 Hitachi drives of 1 TB
1 Intel 82598EB 10G NIC

70+ x 8 cores E5430 @2.66GHz,16GB

This setup reperesents well an elementary fraction of a typical large hardware installation and has basically no bottlenecks:

- **Each of the three Adaptec controllers may deliver 600+ MB/sec (R6)**
- **Ttcp memory-memory network test (1 server – 10 clients) shows full 10G speed**
- **70+ worker nodes allow to employ use cases with non-pronounced I/O boundness**

Details of the current test environment

- RHEL 6.3+/64bit on all nodes (kernels 2.6.32-279.9.1 on clients and 2.6.32-220.4.2.el6_lustre on server)
- Lustre 2.2
- GPFS 3.4.0-17
- NFS v4 (native RH6.3)
- Xrootd 3.2.5
- Glusterfs 3.3.1

Current use cases

- New CMS use case: CMS Hammercloud CMSSW_5_3_1, mostly CPU bound (Giacinto Donvito /INFN)
- Nova use case (NOVA-1): Nova/ANA standalone analysis job with condensed output stream – bidirectional I/O (Andrew Norman /FNAL)
- In preparation: New ATLAS Hammercloud (Wahid Bhimji /U Edinburgh)

How the tests are performed

- Configure the server and client parts of a solution under test;
  Load the data files into the data area under test;
- Run increasing number of jobs per server; each of the jobs is processing a dedicated non-shared set of event files;
- In each of the measurements start all the jobs simultaneously and then kill them simultaneously, after some predefined period of smooth running;
- Calculate the processing speed in terms of events/second (these speed numbers may then be compared directly for all solutions under test;
- While the jobs are running, measure the average data traffic on the server;
- Try to tune each of the solutions under test to get the largest possible processing speeds;

Tunables

We report here, for reference, some of the relevant settings that were used so far.

Diskware: One large software RAID-0 MD device (chunk size 512K) configured over 12 RAID-1 hardware LUNs (all three controllers were engaged)

Lustre: No checksumming, No caching on server
OST threads: “options ost oss_num_threads=512”
Read-aheads on clients: standard(40MB)

GPFS: 12 NSDs, one large file system
-B 512K –j cluster - maxMBpS 1250 - maxReceiverThreads 128
nsdMaxWorkerThreads 128 - nsdThreadsPerDisk 8 - pagepool 2G

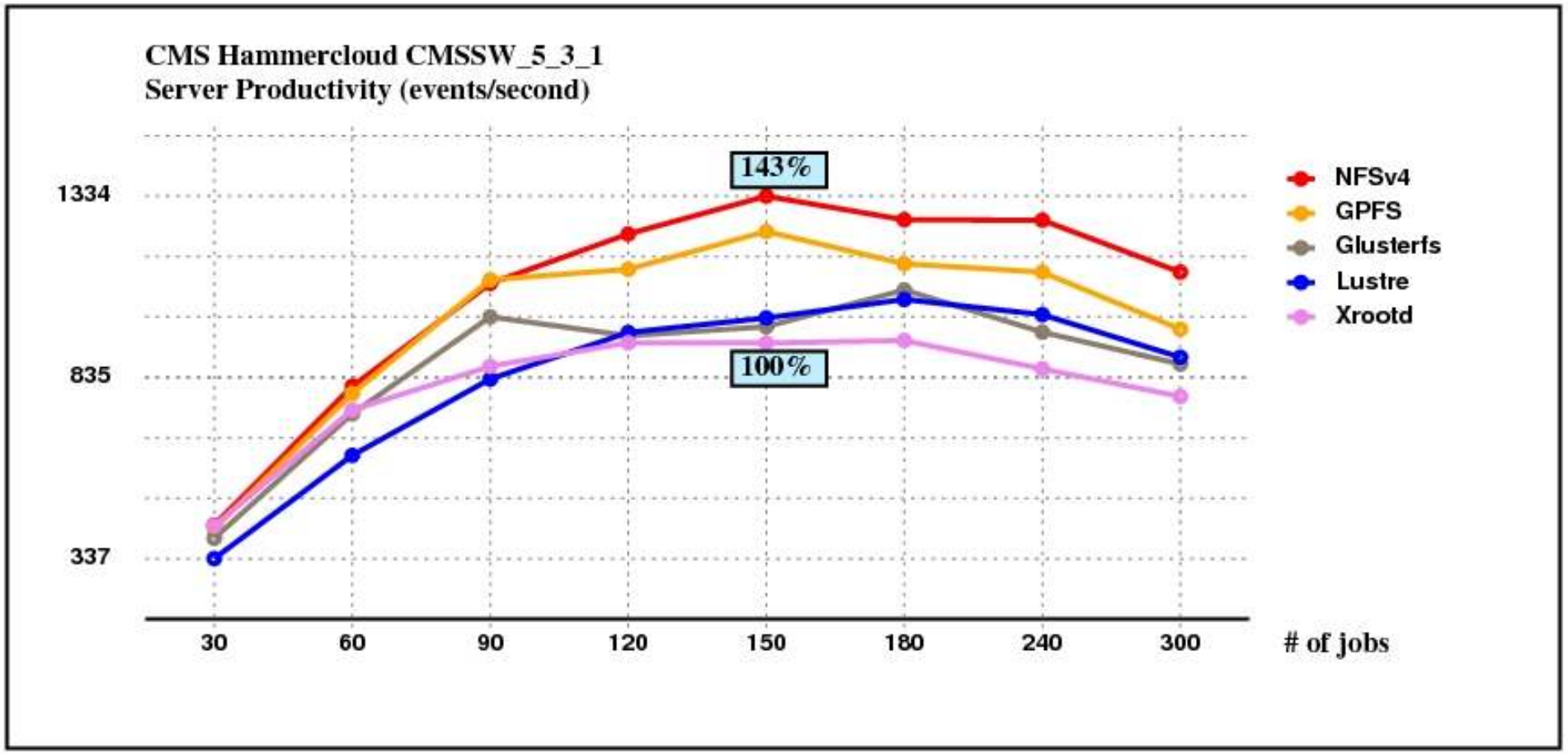
Xrootd: 1 large XFS filesystem
xrd.sched mint 800 maxt 800 avlt 800 idle 0

NFS: 256 server instances
mount -o vers=4,rsize=32768,wsize=32768

Glusterfs: 1 brick with 1 large ext4 filesystem, tuning is still being investigated

Current results

CMS Use Case



NFS V4 on RHEL6 looks suprisingly good for CMS use case.
The gain may be as high as 40% compared to Xrootd.

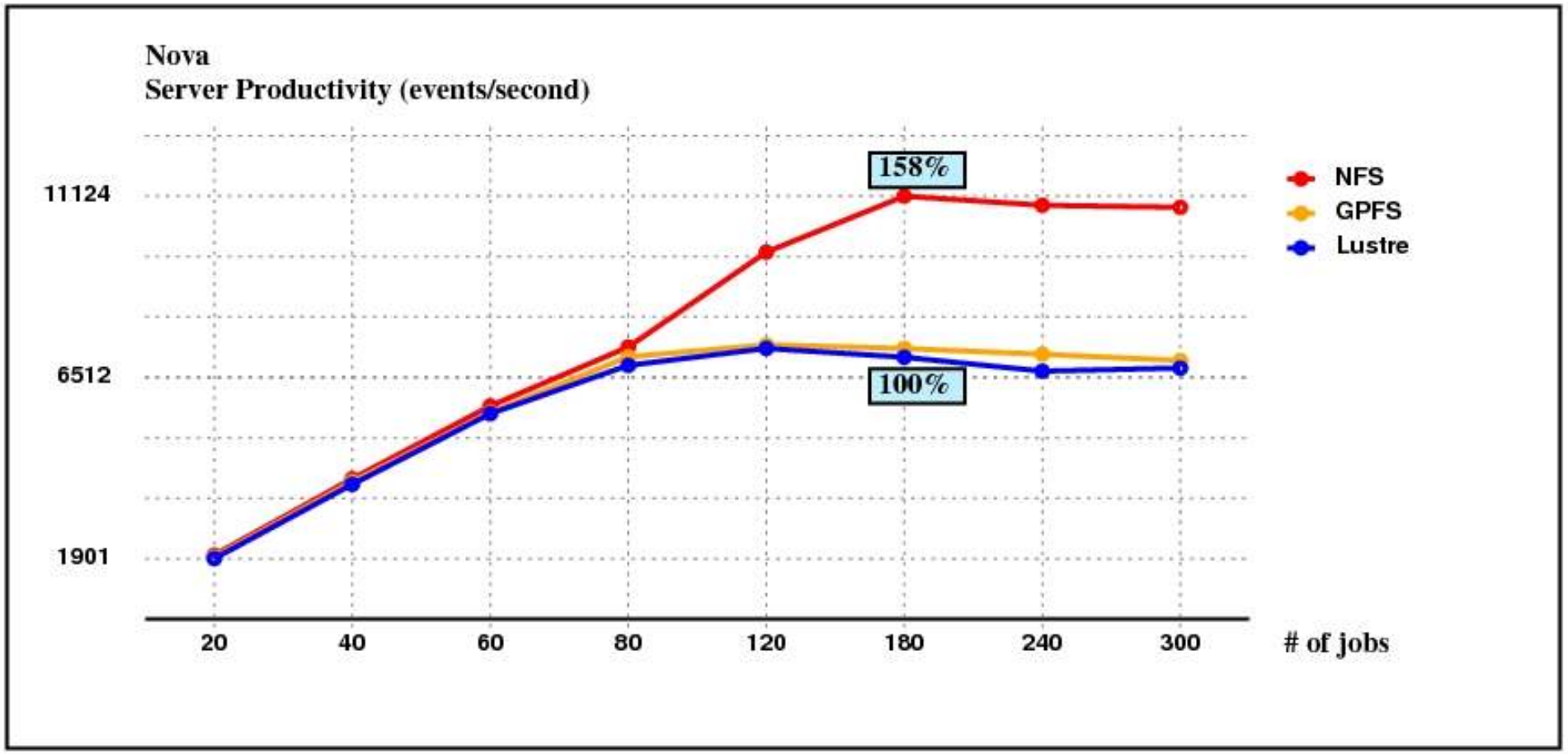
Glusterfs does really well, yet without any special tuning.
We might be able to further improve its performance.

Numbers seen for CMS Use Case

	30 jobs	60 jobs	90 jobs	120 jobs
Xrootd	62 MB/sec 427 EV/sec	109 MB/sec 746 EV/sec	134 MB/sec 867 EV/sec	137 MB/sec 932 EV/sec
Lustre	96 MB/sec 337 EV/sec	183 MB/sec 622 EV/sec	249 MB/sec 831 EV/sec	275 MB/sec 959 EV/sec
Gluster	337 MB/sec 393 EV/sec	661 MB/sec 735 EV/sec	896 MB/sec 1002 EV/sec	865 MB/sec 950 EV/sec
GPFS	250 MB/sec 421 EV/sec	495 MB/sec 791 EV/sec	681 MB/sec 1104 EV/sec	736 MB/sec 1134 EV/sec
NFS V4	82 MB/sec 439 EV/sec	156 MB/sec 812 EV/sec	211 MB/sec 1094 EV/sec	255 MB/sec 1230 EV/sec

	150 jobs	180 jobs	240 jobs	300 jobs
Xrootd	141 MB/sec 930 EV/sec	144 MB/sec 938 EV/sec	138 MB/sec 860 EV/sec	130 MB/sec 784 EV/sec
Lustre	302 MB/sec 999 EV/sec	320 MB/sec 1050 EV/sec	310 MB/sec 1008 EV/sec	305 MB/sec 891 EV/sec
Gluster	954 MB/sec 976 EV/sec	972 MB/sec 1076 EV/sec	943 MB/sec 961 EV/sec	993 MB/sec 873 EV/sec
GPFS	768 MB/sec 1237 EV/sec	781 MB/sec 1149 EV/sec	741 MB/sec 1126 EV/sec	715 MB/sec 969 EV/sec
NFS V4	259 MB/sec 1335 EV/sec	262 MB/sec 1269 EV/sec	266 MB/sec 1269 EV/sec	260 MB/sec 1126 EV/sec

Nova Use Case (R/W)



NFS V4 in this case is saturating visibly later in respect to others.

Glusterfs tests for this use case are not yet complete.

Immediate plans

- Will finish with these test series in October - November.
  Final results will be published on the HEPiX web site.
- Next plans include evaluation of Gluster with ZFS backend, and of S3.
- Will run the recent ATLAS use case for all solutions to get a full picture.
- Storage web site needs updating, will clean it up before the Spring 2013 HEPiX meeting. Volunteers are very welcome.
- Will try to widen the Questionnaire and include more sites.

Discussion