The Lustre® Community in Europe

Torben Kling Petersen, PhD
Principal Solutions Engineer
High performance Computing

xyratex

Lustre sites in EMEA

- Currently more than 120 Lustre file systems in production.
- 85% is academia
- Most is sub 200 TB
- Most is Lustre 1.8.x (still a number of 1.6.x around …)
- Most are home built and self supported …

- RFP requests for Lustre implementations are increasing
- Requirements are getting increasingly complex
- GPFS is still the biggest competitor

Requirements are changing

- Performance is still important ☺ but …
- Support for small files
- Enterprise reliability
  - Simplified management
- Back-up
- Desktop connectivity
- Energy efficient storage
  - Accurate measurements in real time
- Statistics and reporting
- Capacity and performance

Better support for Small File sizes

- Significant part of the data is sub 256k file sizes
- "Small file" repository should be part of the file system
- SSDs can help but …
  - Lustre is part of the problem
  - as is SAS/S-ATA/FC-AL and other protocols
- New ideas and methods are needed
  - But anything new needs tight integration

Enterprise Reliability

- Data integrity – Silent Data Corruption
  - T10-DIF
  - ZFS
- Hardware resiliency
  - HA on ALL components …
  - Self healing software …
- Manageability
  - Better tools
  - Better reporting
  - Lights out management
- Security (Kerberos, role based access control etc)
- No longer just scratch

Backing up and restoring data

- /home and /project increasingly common on Lustre
- Backing up not just data is imperative
  - Restore needs to be demonstrated (not just stated)
  - Restore must also restore striping etc, not just files ..
- HSM can (will) help but we need new tools
- Replication of entire file systems will be required
  - … and rsync won't cut it ..
- Disaster Recovery is also on the horizon
  - But 10s of PBs takes time even on fully saturated FDR

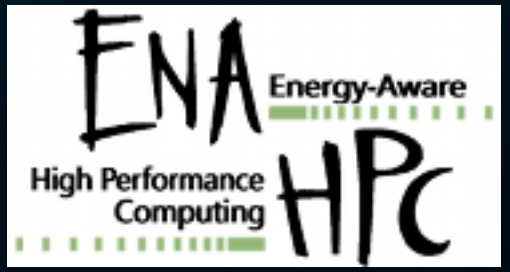
Desktop connectivity

- "I want to mount my Lustre file system on my Mac/WinPC"
  - Roll it yourself Samba server works, right ??
  - Native Win/Mac clients ???
- Complex workflows require non Linux platforms
  - DCC
  - O&G
  - Genomics
  - Financial
- CIFS/NFS connectivity no longer an add-on
  - Need a supported platform
  - Needs scalability

Energy efficiency

- Power cost BIG money
  - Storage is taking a bigger piece of the budgets
- New metrics are showing up
  - Watts per GB/s
  - Watts per TB
- EU legislation on power capping datacenters will make this worse as we get closer to ExaScale
- We now have European conferences dedicated to the subject :

International Conference on Energy-Aware High Performance Computing
September 2 - September 3, 2013
http://www.ena-hpc.org



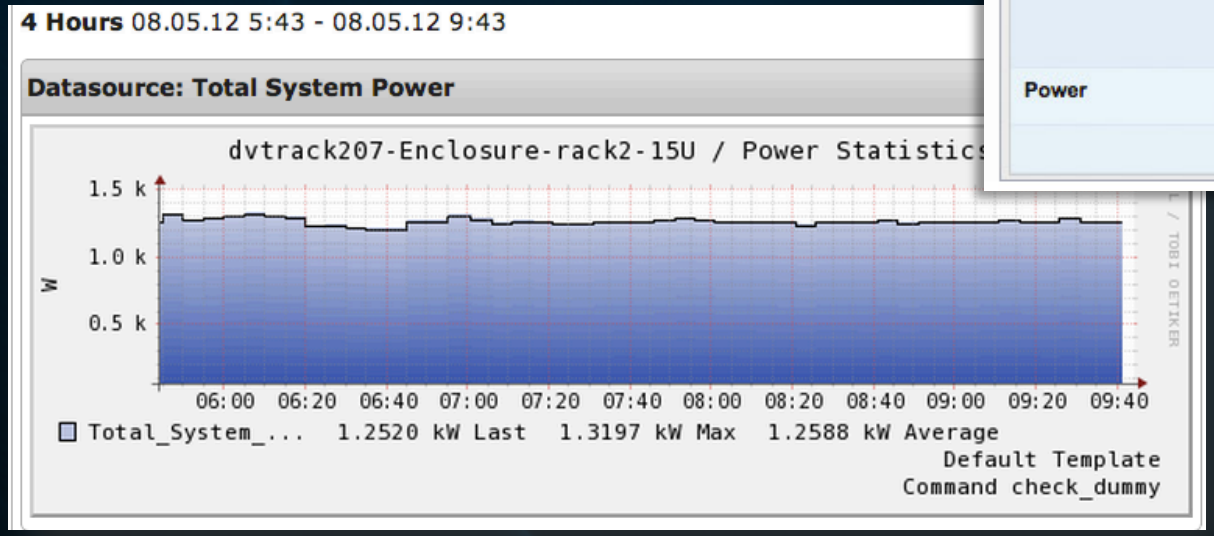
xyratex

Energy efficiency, cont'd

- Real time (and accurate) power consumption needs to be reported ...

Top System Statistics

	Metric		
Filesystem	Peak Read	9.87	GB/s
	Current Read	0	B/s
	Peak Write	9.55	GB/s
	Current Write	0	B/s
Metadata	Current Operations	0	Op/s
Storage	Number of OSTs in use	16	
	Number of Disks in use	192	
	Capacity in use	16.49	TB
	Capacity available	94.66	TB
Power	Current cluster usage	2.01	KW



xyratex

Statistics and Reporting

- Total usage per user/group
  - Bandwidth, volume, IOPS per project/time frame
- Inventory and changes over time (-> TCO)
- I/O behavior per job/user/group
- Tools to follow a job from submission to storage
  - Identify bad behavior (app, user etc)
  - Determine reasons for loss of performance (cf Vampir)
- Capacity planning

Summary and Conclusions

- European Lustre users place a high emphasis on enterprise features
  - But many feel they cannot influence the roadmap
- Performance is STILL important but not at any cost such as:
  - reliability
  - back-ups
  - replication
  - small file I/O
- Many sites are looking to replace GPFS with a different platform while retaining the same feature set
- Lustre is seen as a likely replacement but ....
  - the requirements of smaller sites must be taken seriously ..

Thank You

torben_kling_petersen@xyratex.com