Diamond Light Source site report

Tina Friedrich

Diamond Light Source Ltd

24 April 2012

Diamond Light Source

Science Computing group

new Lustre file system

old Lustre file system MDT upgrade

HPC facilities

monitoring

Current work / Future plans / Outlook

Many thanks to my colleagues – Greg Matthews, Frederik Ferner, Mark Godwin, David Simpson, and Nick Rees – for their contributions to this talk.

Diamond Light Source

Diamond Light Source is the UK's national Synchrotron facility.

- third generation light source (561.6 m storage ring; 3GeV)
- located at the Harwell Science and Innovation Campus, south Oxfordshire
- opened in 2007
- 32 beamlines (28 currently)

We have $\sim$ 5000 users per year ($\sim$ 1500 visits).

Science Computing group

The Diamond Science Computing group supports the following services:

- storage and file sharing (home areas, data areas,. . .)
- software areas, package repositories
- high performance computing
- systems monitoring and reporting
- installation and configuration management
- remote access (using NoMachine NX)
- various services (LDAP, printing, version control, status displays. . .)

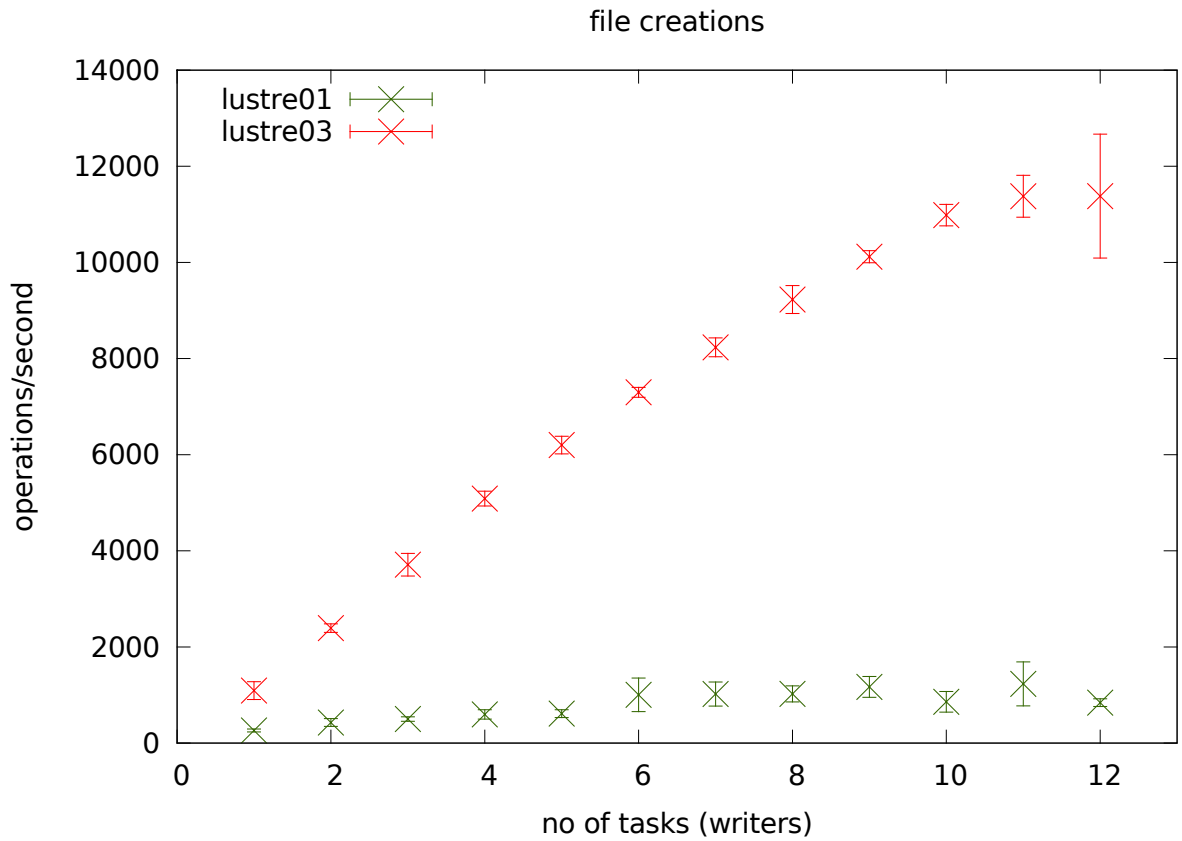
We currently look after ~500 servers and ~500 workstations. We rely heavily on central provisioning and management (kickstart, cfengine). OS is Red Hat Enterprise.

lustre03

We commissioned a second production Lustre file system (”lustre03”).

- 600TB raw (~400TB usable)
- DDN SFA 10K for OSTs, EFI 3015 for MDT
- PE R610 for OSS and MDS
- 4 OSSs in active-active fail-over pairs
- servers connected to core networks via 2x 10Gbit Ethernet bonded links
- throughput ~5.5GB/s

Much better meta data performance than our ”old” Lustre file system (lustre01):



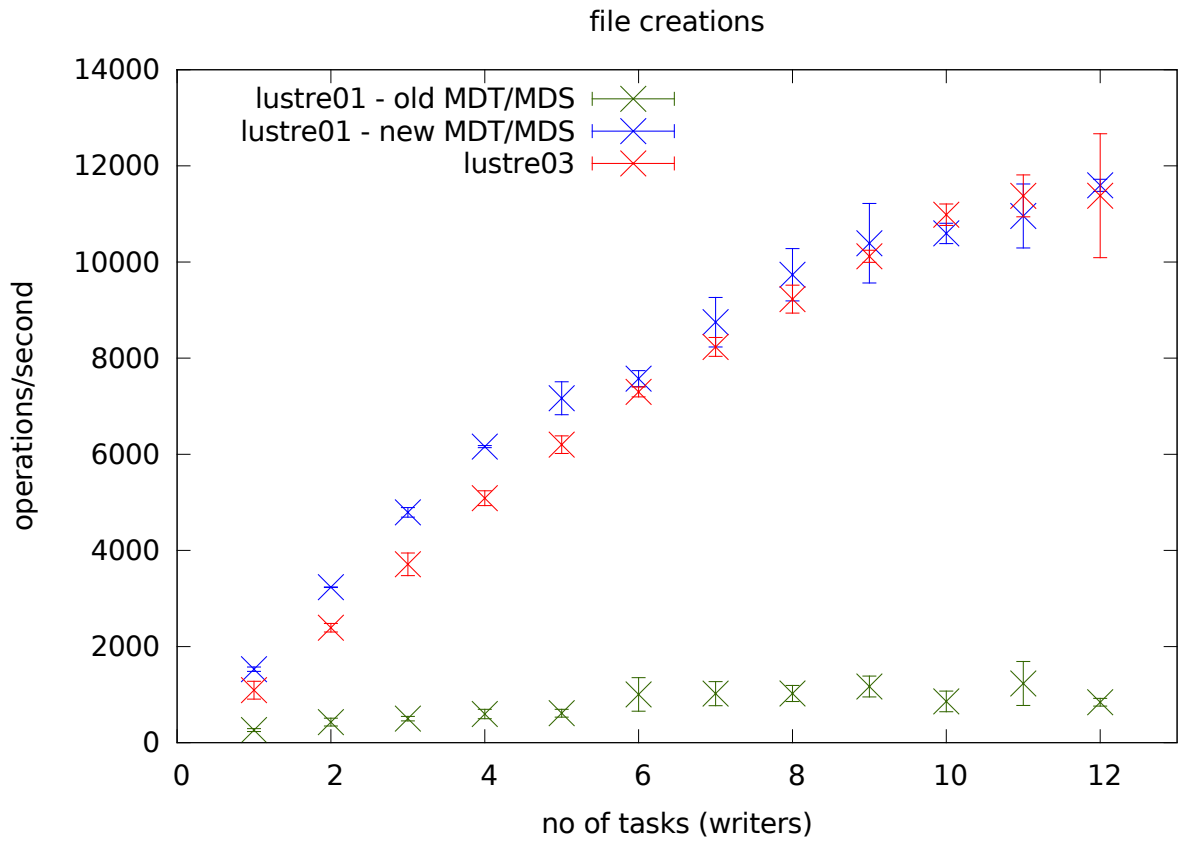
(measured using mdtest, creating 39060 objects per client)

Recent Upgrade: MDT for lustre01

Earlier this year we upgraded our MDS/MDT for our ”old” Lustre file system (lustre01, commissioned end of 2008).

- replaced MD3000 with MD3200 as MDT
- upgraded the MDS servers from PE2970s to R610s.
- MDT transfer using dd

This significantly improved meta data performance:



(again measured using mdtest)

HPC facilities

No major new developments regarding our clusters. Scheduler (still) is Sun Grid Engine 6.2u4.

- purchase another 40 compute nodes — Viglen HX425T^{2}i Quad HPC nodes (Supermicro X8DTT-F boards), Intel Xeon X5650 CPUs
- recently purchased 12 new GPU nodes (HP SL390, Nvidia Telsa M2090)
- purchased another 10 Nvidia M1070 1U boxes
- retired our oldest set of cluster nodes (29 dual core IBM x3455 nodes)

Now got a total of 1120 cores in production (will be 1264 with the HP nodes), and (apparently) 29568 GPU cores

Users – or so I am told – are happy with the system.

monitoring

Started to look into replacing our current monitoring setup – Nagios – in earnest.

Looking for a product that can do both monitoring/alerting and data gathering/reporting (replace both Nagios & Ganglia). Should also be somewhat more convenient in terms of adding new machines (discovery functionality).

After some research, decided to trial Zenoss — quite feature rich, relatively well established, quite widely used, plus has the advantage of being able to utilise Nagios plugins.

Works quite well; had some issues with performance (mainly due to the fact that we collect & keep data in high granularity).

SCAAP review

We had a computing strategy review (”Scientific Computing Audit and Advisory Panel”).

Up to now, investments in Science computing were usually tied in with e.g. construction phases.

Early in 2011, we submitted a long term plan for computing expenditure, to allow better forward planning in developing the infrastructure. We proposed a steady-state expenditure of around £500k – 600k per year, along with bringing the group strength up to 6 people.

The directors agreed in principle; they suggested to call an external panel to review our proposal.

SCAAP review — panel recommendations

The basic view of the Panel is that IT is the enabler of science – no IT, no science. The quality of the scientific output from Diamond is determined as much by the quality of its IT as it is by the quality of its accelerator, storage ring and detectors.

Without an appropriate level of IT support the scientific goals of any institute cannot be delivered. Experience from comparable scientific computing sites [. . .] is that about 10% of the scientific budget needs to be devoted to the IT infrastructure.

To allow for strategic planning and growth part of the scientific budget must be devoted to IT, the suggested level of £600k per annum is realistic.

As a consequence:

- we now have a ”proper” budget
- we hired more people (group grew to 6 staff)

Outlook

- started migration from CFEngine 2 to CFEngine 3
- upgrade to Red Hat 6
- commission GPU cluster upgrade
- finalise commissioning of Zenoss
- commission new storage facilities
- implement a data management procedure
- (maybe) upgrade of core network
  - new core switches, with some 40/100 Gbit beamline uplinks
  - commodity, low data rate beamlines, . . .
- new detectors coming our way — 100Hz Pilatus, Excalibur. . .

Thank You!