Lustre at JSC

13 April 2011 | Frank Heckes
- Environment overview
- Installation History
- Monitoring
- Perspectives
Overview

- **FZJ, National Research Center**
  - Budget ~ 500 million €
  - ~ 4600 employees
  - Areas:
    - Live science, energy technology, neurobiology, solid state / nuclear physics, climate/meteorology, supercomputing
Overview
Overview

- **JSC in nutshell**
  - 100 employees
  - 2 Production Cluster
    - BlueGeneP - GPFS
      - First PRACE Tier-0 Center
    - JuRoPA - Lustre
      - PRACE Tier-1 Center
  - Two parallel FS in use: GPFS, Lustre
Overview
Overview

- **JuRoPA Cluster**
  (Jülich Research on Petaflop Architectures)
- **Involved Companies**
  *Bull, Sun (Oracle), Mellanox, Intel, ParTec, Novell*
- **Two parts**
  - European fusion community  (1/3)
  - JSC  (2/3)
  - Can act as 'one' cluster
- **Heterogenous user community**
- **High utilisation ( ~ 95 %)**
- **Life span till 2013/2014**
JuRoPA Architecture
Infiniband topology

JUROPA

x 92 building blocks
(24 nodes each)

648 port switch

648 port switch

648 port switch

648 port switch

648 port switch

648 port switch

HPC-FF

x 60 building blocks
(18 nodes each)
Overview

- **Operating System**
  - Compute nodes: SLES 11 SP1
  - Lustre server: SLES 11

- **Compute Component:**
  - 3288 compute nodes (26304 cores)
  - 2 Intel Xeon X5570 (Nehalem-EP) quad-core processors @ 2.93 GHz
  - 79 TB main memory
  - 308 Teraflops peak performance
  - 274.8 Teraflops Linpack performance
  - No. 10 in TOP500 list June 2009
  - No. 23 in TOP500 list Nov. 2010
Overview

- **FS component**
  - **HOME**
    - 7 x snowbird system
    - Building block: 2x Sun Fire X4170 + 4 x J4400 JBODs
    - 2 HOME FS, each ~ 29 TB, bandwidth 1GB/s
      Size adapt to backup/restore bandwidth
    - Total capacity: ~ 400 TB
  
- **Under construction**
  - 2 x DDN SFA 10000 + 8 OSS
    - Building block: 1 x SFA 10k + 4 Bull RS 423 nodes
    - Planned capacity / FS: ~ 24 TB
    - Total capacity: ~ 770 TB
Overview

- **FS component**
  - SCRATCH
    - 2 x SFA 10000 + 8 OSS nodes
    - Building block: 1 x SFA 10000 + 4 x Bull Novascale 423
    - ~ 800 TB, bandwidth 19 GB/s
  - MDS
    - 2 x Emc CX-240 + 4 MDS(MGS) nodes
    - Building block: 1 x Emc CX + 2 Bull Novascale 423
Installation History

- **Start with 1.8.0 GA**
  - Massive errors
    - 3 corrupted filesystem
    - Many OSS, MDS crashes
  - Very sensitive to IB errors

- **Lustre 1.8.1.1 + patches (SLES 11)**
  - version is stable, but very sensitive to IB and HW errors
  - OSS, MDS crashes
  - Large downtime (2 weeks) due defective MPT (SAS) driver
Installation History

- **Lustre 1.8.4**
  - stable version
  - Improved performance
  - More robust to IB errors
  - Fragmented I/O
    - Many iops not aligned to 1M blocks
  - Local flock feature enabled
Monitoring

**Functionality**
- Framework to execute bespoke scripts and programs
  - State of disk, FC- connection, mounts, ..., Temperature, ...NTP, DNS, ...
- Not scalable, but sufficient for current infrastructure

**Performance**
- Measurement with `collectl, sysstat, 'cat /proc...'`

currently on demand evaluation
- Latencies in RAID devices
Perspective

- **Unclear support situation**
  - Lustre support at Oracle??
  - New version after 1.8.5; bug fixes?

- **Improve backup procedure**
  - Use meta info for backup list

- **HSM support**
  - Integration of Lustre filesystems in Tivoli Storage Manager

- **End-to-End data integrity**
Perspective

- **Lustre 2.x Upgrade**
  - unclear
  - Cooperation originally planned with SUN; canceled by ORACLE
  - OSS/OST resources for data migration already allocated, due to delays

- **Improve knowledge**
  - Gap between SysAdmin – Developer
  - Lustre Internal Training needed
  - Plan is to contribute to Lustre 2.1++
Perspective

- FZJ / JSC is Initiator and Founding Member of EOFS (European Open File System)
Questions?