JÜLICH
FORSCHUNGSZENTRUM

Lustre at FZ Juelich -

Status and Goals

September 9, 2009 Otto Büchner

Agenda

- FZJ and JSC
- Jugene and Juropa
- Juropa configuration
- Lustre on Juropa
- First Lustre experience
- Needs for parallel file systems
- Cooperation SUN FZJ

Research Centre Jülich

- Founded 1956
- 4300 employees
- 1200 scientist
- Budget € 436 million

main research:

- Health
- Energy & Environment
- Information
- Key Technologies

Jülich Supercomputing Centre (JSC)

Part of the
Institute for Advanced Simulation

- Supercomputers
- Computational Science
- Grid computing

JSC runs the most powerful Supercomputer in Europe

JUGENE - Juelicher Blue Gene/P

72 Racks
3728 compute nodes
294912 processors
144 TB main memory
1 Peta flop peak performance

Filesystem GPFS

JuRoPA - Jülich Research on Petaflop Architectures

Two Parts:

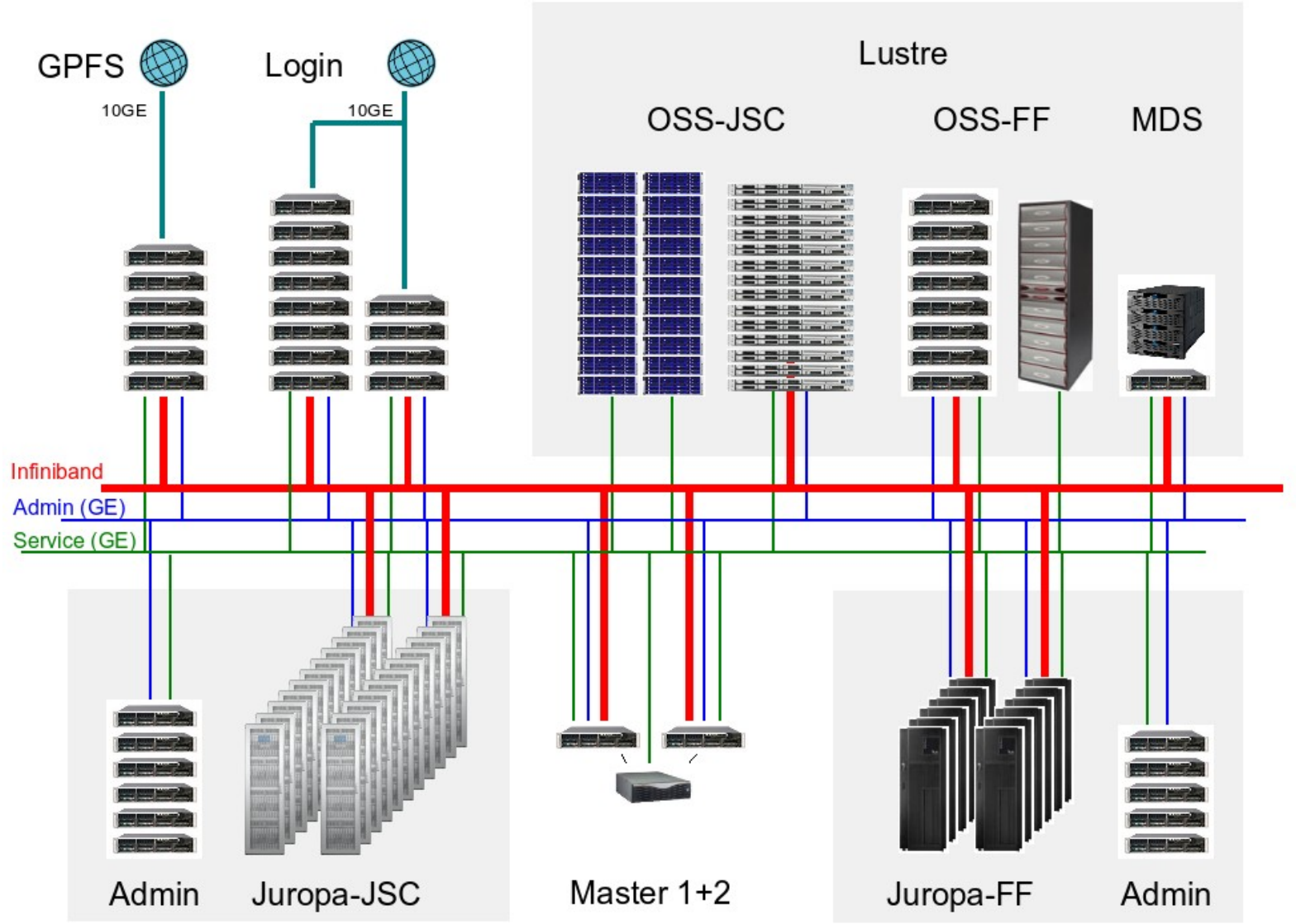
- Juropa JSC
- HPC – FF

- 3288 nodes
- 26304 cores
- 79 TB memory
- 308 TF peak performance

Cooperation:

Bull, Sun, Partec, Intel, Mellanox, Novell, FZJ

JuRoPA - Architecture



JuRoPA – Lustre Storage Pool MDS

2 MDS Server
- Bull NovaScale R423-E2
- (2 x Nehalem-EP quad-core)

- 14 home file systems
- 1 work file system
- mutual fail-over
- 40 TB for mirrored meta data
  on one EMC CX4-240

JuRoPA – Lustre Storage Pool OSS SUN

7 building blocks for home fs
one consists of

- 4 Sun Storage j4400
- 2 Sun Fire X4170 Server
- 8 OST mutual fail over
- serves 2 home fs

in total 500 TB user data

JuRoPA – Lustre Storage Pool OSS Bull

1 building block for work
- 8 Bull NovaScale R423-E2
- each 7OSTs
- 2 DDN SAA9900
- serves 1 file system

360 TB work data
aggregated data rate ~20 GB/s

JuRoPA – Lustre problems

start production 8/6 with not GA version

- unstable, system hangs
- high load on OSS and MDS
- lost data because of OST crash
- OST inconsistency
- planned upgrade to 1.8.1 next week

What do we expect from a parallel file system

- stability
- high performance
- extendable
- hsm manged storage
- data security
- open solution

JUST Juelicher Storage Cluster

- Upgrade in progress

- 4 meta data server
- 14 data server
- 4x10GB E for each server
- 5 PB capacity
- 6144 disks
- 6 PB Sun Tapes
- file system GPFS
- exports to Jugene, Jump Juropa, Deisa

what will be the successor ?

Cooperation FZJ and SUN

Copied from contract

Lustre

I. *Real-Time support*
   See OS-jitter sub-project
II. End-to-end data integrity
   1. Replace ldiskfs by ZFS as the base file system under Linux
      - Who: Sun
      - When: Might be started immediately
      - Resources: 1 year / 4 Lustre developers + 2 ZFS developers
   2. Extend ZFS checksum/end-to-end data integrity to the Lustre file system level
      - Who: Sun
      - When: included in general ZFS effort
      - Resources: included in general ZFS effort
   3. Optimize Lustre with end-to-end data integrity by checksum computation optimization
      - Who: Sun, FZJ
      - When: included in general ZFS effort
      - Resources: included in general ZFS effort + 3 month FZJ test engineer
   4. Benchmark and tune the Lustre/ZFS implementation
      - Prepare I/O benchmarks for ldiskfs/ZFS performance comparison
      - Tune Lustre according to FZJ needs
      - Who: FZJ
      - When: After availability of Lustre/ZFS
      - Resources: 2 month / 1 test-engineer

Hardware for evaluation

Small cluster with same HW as Juropa

- 2 MDS nodes
- 2 OSS nodes
- 2 J4400 JBODs
- 4 compute clients
- 1 Mellanox switch

First installation of Lustre 2.0 Alpha

- starting 9/2
- no time to test because of lustre problems

JÜLICH
FORSCHUNGSZENTRUM

Thank you for your attention

Questions ?