CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

# Cray XC30 – A System Level Overview

CUG 2013

Nicola Bianchi
Colin McMurtrie
Sadaf Alam
Swiss National Supercomputing Centre (CSCS)

# Agenda

- **CSCS XC 30 Platform**
- **Configuration details**
- **System design and installation**
- **XC30 vs. XE6**
- **Early functionality and performance results**
- **Conclusions**

**What system did we receive?**

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

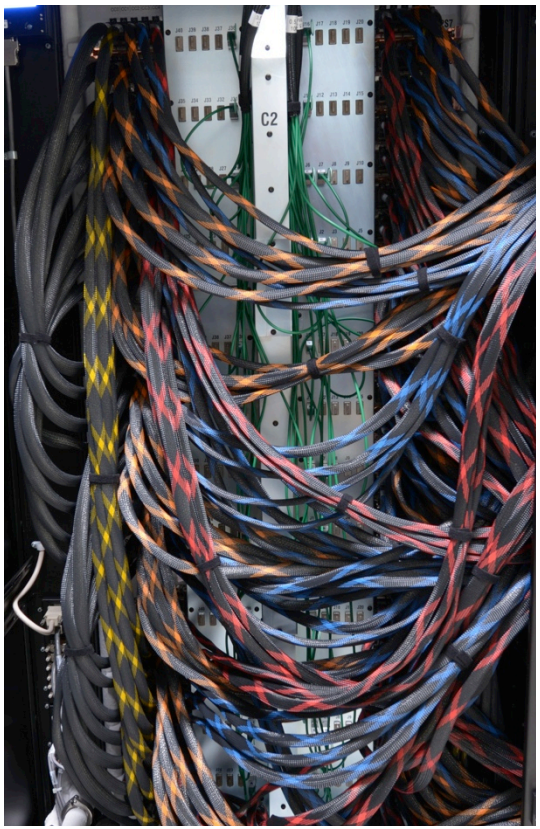
# XC30 Platform at CSCS

- **Currently largest XC30 worldwide**
- **12 cabinets**
  - 2256 compute nodes
  - 24 service nodes
- **5 esLogin servers**
- **Sonexion1600 Lustre Appliance**
- **SLURM workload manager**


**Configuration details?**

# Service Nodes

- **1 service blade per cabinet**
- **24 service nodes**
  - Boot + SDB
  - 4 login/Slurm frontend
  - 4 DVS
    - /users and /apps NFS projection
  - 14 LNET router
- **2x PCI Gen3 slots per service node**




**How is the Lustre environment configured?**

# Lustre

- **10 SSU Sonexion 1600**
  - 1.1 PB
  - 50 GB/s write
  - 20 OSSs
  - 80 OSTs
- **Dedicated FDR Fabric**
  - 2X 108 ports switch
  - 4x 36 ports top-of-rack switch
- **14 router nodes**
  - 12 OSS router nodes
  - 2 MDS router nodes


**HW installation**

# HW Installation

- **New rack Design**
  - Bigger than XE
  - Flat bottom, no pedestal, no high point loads
  - Easy and quick placement

- **Directly water cooled**
  - No XDP, less installation time
  - Easy pips attachment

- **Horizontal air flow**

- **Interconnect cabling simple snap-in connectors**

- **Planned time to power up the system: 4 days > real time 2 days**

**Comparison: XC30 vs. XE6**

# XC30 vs. XE6

- **Most of the administrative commands are the same**
  - xtbounce, xtcli, xtalive …
- **Smooth transition to the new architecture for sysadmins**
- **New HW, new names**
  - Cabinet controller, CC
  - Blade controller, BC
- **Main difference due to the new CLE5 + SMW7 environment**
  - we were used to CLE4.0 on XE6

**… and moreover**

# XC30 vs. XE6

- **CLES vs. CLE4 on XE6**
  - SMW logs location, new log aggregator
  - Controller logs (/var/opt/cray/log/controller)

- **Commands**
  - **xtzap** instead of xtflash
  - **cdump** instead of ldump
  - **hssclone, hssbootlink, hsspackage**
  - **xtccreboot**

- **CDT (Cray Developper Toolkit)**
  - craype-installer: easy to use
  - Automatically keep is sync main system & EsLogins

**Early impressions**

# Impressions & issues

- **Minor post installation problems**
  - Faulty IB cable
  - PCI Gen3 bus at 8x (not 16x)
  - HSN cable not correctly seated

- **SONEXION instability**
  - MDS crash (Adios I/O library), fixed
  - OSS failover problem, fixed


- **Lustre performace**
  - IOR write performance, now near 49 GB/sec
  - IOR read performance -40% less than write
  - Read performance still a problem, client related

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

# Impressions & issues

- **GPFS implementation**
  - Weird problem with aprun (#794091)

```
[nbianchi @ santis01]-[03:12:04]-[~]:-) salloc -N1
salloc: Granted job allocation 1250
[nbianchi @ santis01]-[03:12:10]-[~]:-) aprun -n1 date
Fri Feb  8 15:12:15 CET 2013
Application 19021 resources: utime ~0s, stime ~0s
[nbianchi @ santis01]-[03:12:15]-[~]:-) aprun -n1 date
aprun: getcwd: No such file or directory
aprun: Exiting due to errors. Application aborted
[nbianchi @ santis01]-[03:12:16]-[~]:-( exit
```

  - Unexpectedly disappeared, probably after a not directly related CLE patch

**Performance**

# Performance

- **IO performance**
  - User application with HDF5 library up to 28GiB/s write
  - Metadata benchmark (mdtest) better than any other Lustre FS at CSCS

- **Job placement / MPI / network**
  - I/O and MPI performance not affected by job placement
  - System default placement algorithm work well
  - Real jobs seem not to suffer any ill effects from the nominal difference in performance of the Dragonfly HSN
  - Unlike the XE/XK line there is no degradation in certain dimensions (i.e. y links half as slow as x and z)

**Conclusions**

# Conclusion

- **Piz Daint entered production on 1st of April 2013**
- **Less than 4 month to achieve this target**
- **The system, despite the youth, is stable and reliable**


**Questions?**

# Q&A