Karlsruher Institut für Technologie

Steinbuch Centre for Computing

KIT Site Report (HEPiX Spring 2011)

Manfred Alef, Jos van Wezel, Stephanie Böhringer

STEINBUCH CENTRE FOR COMPUTING

Contents

- **GridKa:**
  German LCG Tier-1

- **LSDF:**
  Large Scale Data Facility

GridKa

GridKa Hardware Status

- **CPU expansion:**
  106,000 HS06 available
  - 2010:
    - 55 HP ProLiant DL h1000 G6 chassis (220 DL170h G6 nodes)
      - 2x Intel Xeon E5520 (quad-core, 2.26 GHz Nehalem)
        - SMP (Hyperthreading) enabled
      - 24 GB RAM
        - 12 job slots per node with 2 GB each
      - 2x 500 GB disk
      - 120 HS06 per node
    - 1 DELL PowerEdge R815 **(→ study of WNs with many cores)**
      - 4x AMD Opteron 6174 (12-core, 2.2 GHz)
      - 96 GB RAM (2 GB per job slot)
      - 6x 500 GB disk
      - 400 HS06 per node

GridKa Hardware Status

- **CPU expansion:**
  106,000 HS06 available
  - 2011:
    - 29 Supermicro Twin^2 (116 X7DCL nodes)
      - 2x AMD Opteron 6168 (12-core, 1.9 GHz)
      - 72 GB RAM (3 GB per job slot)
      - 3x 500 GB disk
      - 183 HS06 per node

Performance investigations → separate talk in Computing track

GridKa Hardware Status

- **Experiences with cluster hardware and on-site service (by hardware vendor):**
  - ...

GridKa Hardware Status

- **Storage update 2010/2011:**
  - 9 DDN 9900 systems available
    - 150 enclosures x 60 disks
  - dCache (+ ~2 PB)
    - Atlas and CMS instance:
      Chimera namespace provider
    - LHCb and others:
      PNFS namespace provider
    - Stable operation since summer 2009
  - Xrootd (+ ~1,5 PB)
    - Alice only
    - BestMan SRM, tape access

GridKa Hardware Status

- **Storage update 2010/2011:**
  - Tape:
    - All Libraries (STK, IBM, GRAU) and drives now managed via eRMM for library and drive virtualisation
    - Single TSM server
    - Tape DB is at 5 GB
    - Interface to dCache and Xrootd via TSS

Details → separate talk by Dorin Lobontu in Storage track

GridKa Hardware Status

- **Storage totals:**
  - 90 dCache pool nodes in 5 dCache instances
  - 40 Administrative servers
  - 16 GridFTP doors
  - 11 xrootd servers
  - 15 NFS servers (for shared areas)
  - 10 PB usable disk space
  - 6 PB Tape storage
  - 52 LTO drives
  - 3 Libraries (IBM, GRAU, STK)

GridKa Operating Issues

- **Kernel issues on WNs:**
  - kswapd: ~100% system CPU utilization
  - /proc filesystem partially unreadable
  - Processes changed to D state when reading from /proc (pbs-mom, ssh, monitoring tools, ...)
  - Automated in-node system management tasks failed because of freezing shutdown scripts
    - Problem resolving (manual reset) from outside the box required
  - PBS server stalled when connecting to frozen mom
    - Server reboot required to resume

  - Correlated to fileserver (e.g. shared software area) issues

GridKa Operating Issues

- **Kernel issues on WNs:**
  - Trouble-shooting:
    - Manual or IPMI-based reset of affected nodes (because of hanging in-node procedures)
    - Detection of affected nodes using Ganglia monitoring recordings (check for high system load)
  - Final problem solution:
    - Improved monitoring of critical fileservers (shared software areas)
    - On-call service solves fileserver issues at first, then immediately checks for affected nodes to prevent PBS trouble

GridKa Operating Issues

- **PBS issues:**
  - First batch system at GridKa:
    OpenPBS
  - Migration to PBS-Pro in 2002 to escape from frequent failures
  - "Hidden limit" of PBS-Pro has been crossed in 2009
    - Very slow response, timeouts at CE side
    - Trouble with WNs which run into strange state (kswapd issue)
    - "Black hole" WN issues
    - Large quantities of log files (up to 10 GB per day), nevertheless insufficient amount of messages regarding batch operation
    - ...

GridKa Operating Issues

- **PBS issues:**
  - Trouble-shooting:
    - Cluster-split in 2010 as a work-around
    - Update to new PBS-Pro release 11 in 2011
      - 
        - Faster scheduling engine (job DB instead of flat files)
      - 
        - New license manager (third one since release 9) was buggy
        - Old {bugs|bad features} are still existing

LSDF

LSDF Status in a Nutshell

- Provide support and infrastructures for data intensive science
  - Open for all scientific communities
  - Currently: biology (microscopy, gene sequencing), climate research, geology, materials research (synchrotron radiation)
  - High Energy Physics is covered by GridKa
  - Software developments and community specific solutions
  - Leveraging experiences from WLCG computing
- Hardware resources:
  - 10 Gb Campus Data Acquisition Network
  - Hadoop cluster of 58 HP nodes,116 TB internal storage
  - DDN SFA10K (1 PB) with DELL R610 servers
  - IBM DS5300 (7 PB) with IBM x3650
- Software:
  - Hadoop, SoFS, GPFS, TSM

LSDF Key Factors

- Integration of facility and software support
- 10 Gb dedicated network and expansion to at least 7 PB in 2012
- Archival of measurement data for at least 10 years
- Integrated computing
- Ubiquitous access
- Storage for universities in the state of Baden-Württemberg
- Source of many publications in cooperation with users of the facility