Karlsruhe Institute of Technology

Experiences with Lustre, Spectrum Scale and BeeOND at KIT

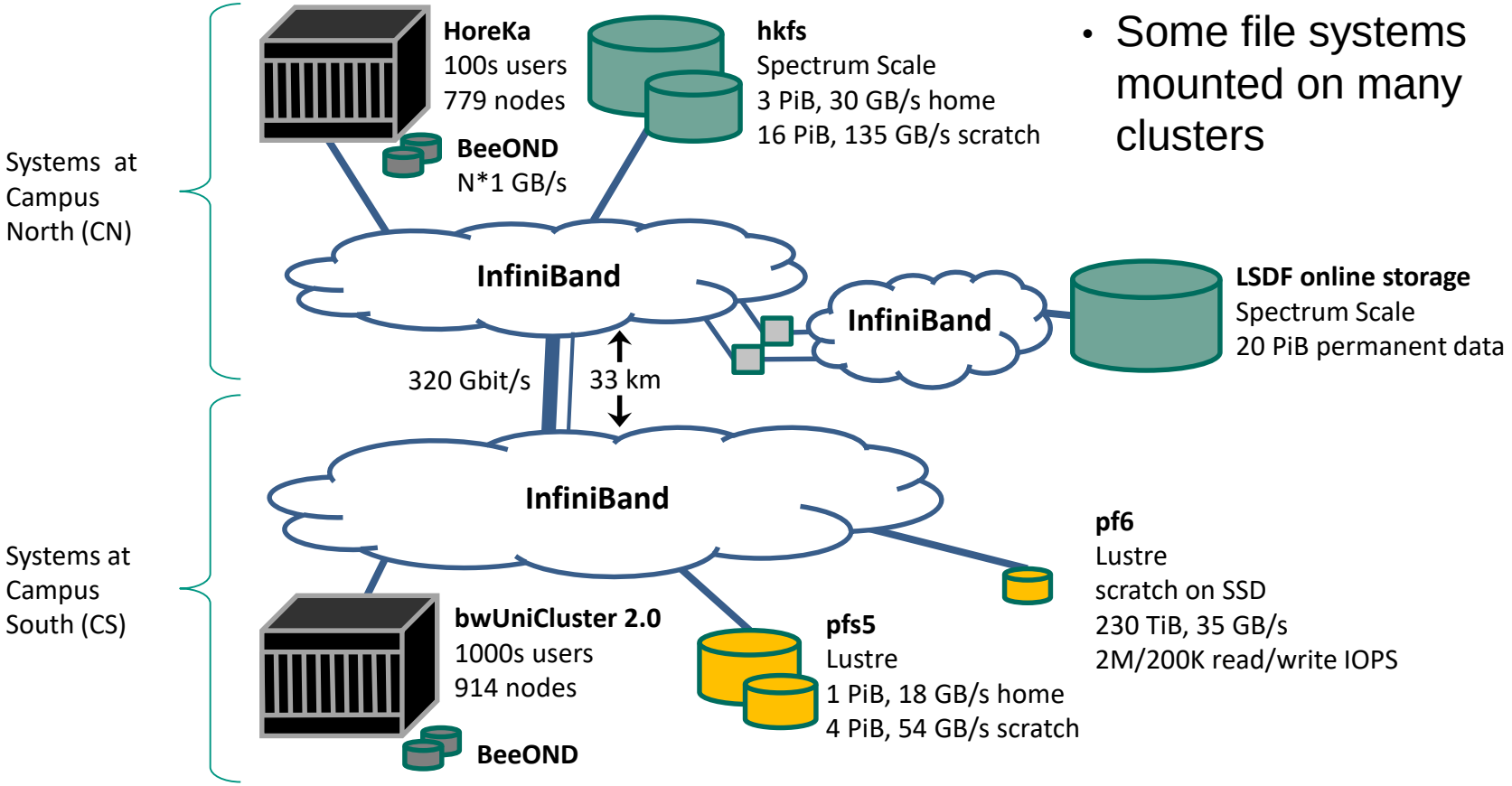
Roland Laifer

STEINBUCH CENTRE FOR COMPUTING - SCC

Overview

- HPC and parallel file systems at KIT
- Details of KIT's Lustre, Spectrum Scale and BeeOND installations
- Pros and cons of Lustre, Spectrum Scale and BeeGFS

HPC and parallel file systems at KIT



- Some file systems mounted on many clusters

Lustre details (pfs5 / pfs6)

- Hardware from DDN
  - pfs5: 4 ES7990 (8 OSS), 4 MDS with 2 SFA200NV, 1 MGS
  - pfs6: 1 ES400NVX (4 MDS/OSS)
  - 2 DDN Insight servers for monitoring
- Administrative details
  - Firmware and Lustre software upgrades done by DDN
  - Project quotas used according to financial share of organizations
- Stability
  - Runs very stable since years with extreme user base
  - Lustre jobstats steadily used to find and educate power users
  - Good DDN support in critical cases

Spectrum Scale details (hkfs)

- Hardware from Lenovo
  - 6 NSD metadata servers, each with 8+1 internal NVMe
    - 3-way metadata replication by Scale across servers
  - 5 NSD building blocks
    - each with 2 servers, 4 DE6000 + 4 expansion enclosures
  - 5 nodes with Scale GUI for monitoring, NFS export, backup
- Administrative details
  - Server side software and firmware upgrades done by pro-com
  - Multicluster setup, root ssh from compute nodes not allowed
- Stability
  - Good support by pro-com and IBM
  - Half-dead clients might hang up the complete file system
    - Better monitoring in newest version, real fix later

BeeOND details

- Hardware
  - Uses internal SSDs on compute nodes
- Administrative details
  - Support contract for BeeOND with ThinkparQ
    - Tuning during installation greatly improved performance
  - Configuration and software build is done by KIT
    - Own workaround since root ssh from compute nodes is not allowed
    - Use sparse file and loopback device for targets to allow quick deletion
  - Adapted SLURM to create/destroy in prolog/epilog if FS is needed
- Stability
  - Rarely used, no issues, no support needed
  - For huge jobs timeout might appear during BeeGFS mount

Lustre pros and cons

+ Standard features very stable due to huge HPC user base
+ Lustre jobstats allow easy performance monitoring
  + on job, user and host basis
+ LNET routers provide powerful networking options
- Check if RHEL and MOFED version is supported
  - currently needs to be done before every upgrade
- Currently no easily usable snapshots
- Possibly missing features for enterprise usage
  - Windows client
  - Commands to rebalance or fix replication

Spectrum Scale pros and cons

+ Most features stable due to huge HPC / industry user base
  + Snapshots, CIFS/NFS export and HSM available since many years
+ Supports metadata replication
  + Useful for online upgrades/extensions, additional hardware options
+ QoS on commands (e.g. rebalancing) very helpful
+ Many outstanding features
  + Windows client, AFM, multi cluster
- High license costs, frequently changing license policies
- Spectrum Scale client needs fixed amount of memory
- Normal configuration requires root ssh between all nodes
- Apparently performance monitoring needs special solution
  - Usually helpful on job and user basis

BeeGFS pros and cons

+ BeeOND very useful to provide on-demand FS for jobs
+ Administration is fairly easy
+ Relatively cheap support (for BeeOND)
- Commits after data is stored in server memory
  - Reason for some good performance rates
  - SPoF solutions require buddy mirroring, i.e. high hardware costs
- BeeOND normally requires root ssh between all nodes
- Relatively small feature list
  - Some features (e.g. quotas) only available with support contract

Summary

- Lustre and Spectrum Scale are a good choice for HPC systems
  - Can be used as parallel home and scratch file system
  - Check the details to find which solution fits best for you
- BeeOND provides a good on-demand file system for jobs
- Supported products from system vendors are recommended
  - Finding good hardware, driver and software levels is challenging
  - Hard job to find and fix critical issues
- Using multiple PFS on the same system caused no issues
- /tmp on local SSDs allows to reduce load on PFS
- My talks about Lustre
  - http://www.scc.kit.edu/produkte/lustre.php
  - roland.laifer@kit.edu