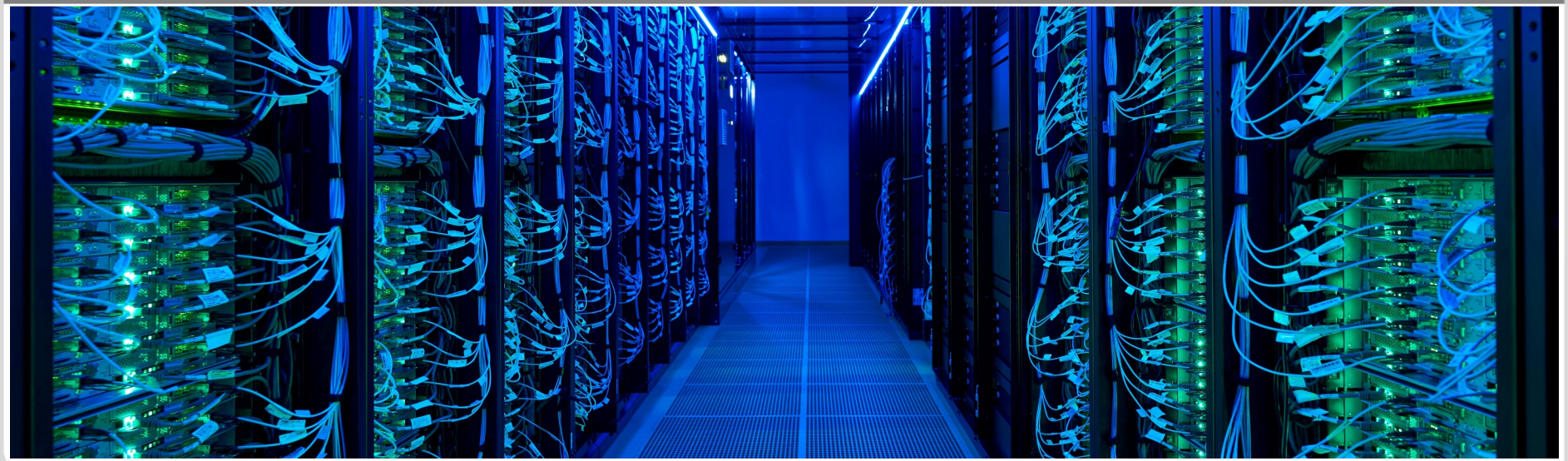
Karlsruhe Institute of Technology

Mixing HDD and Flash Storage on Parallel File Systems

Roland Laifer

SCIENTIFIC COMPUTING CENTRE - SCC



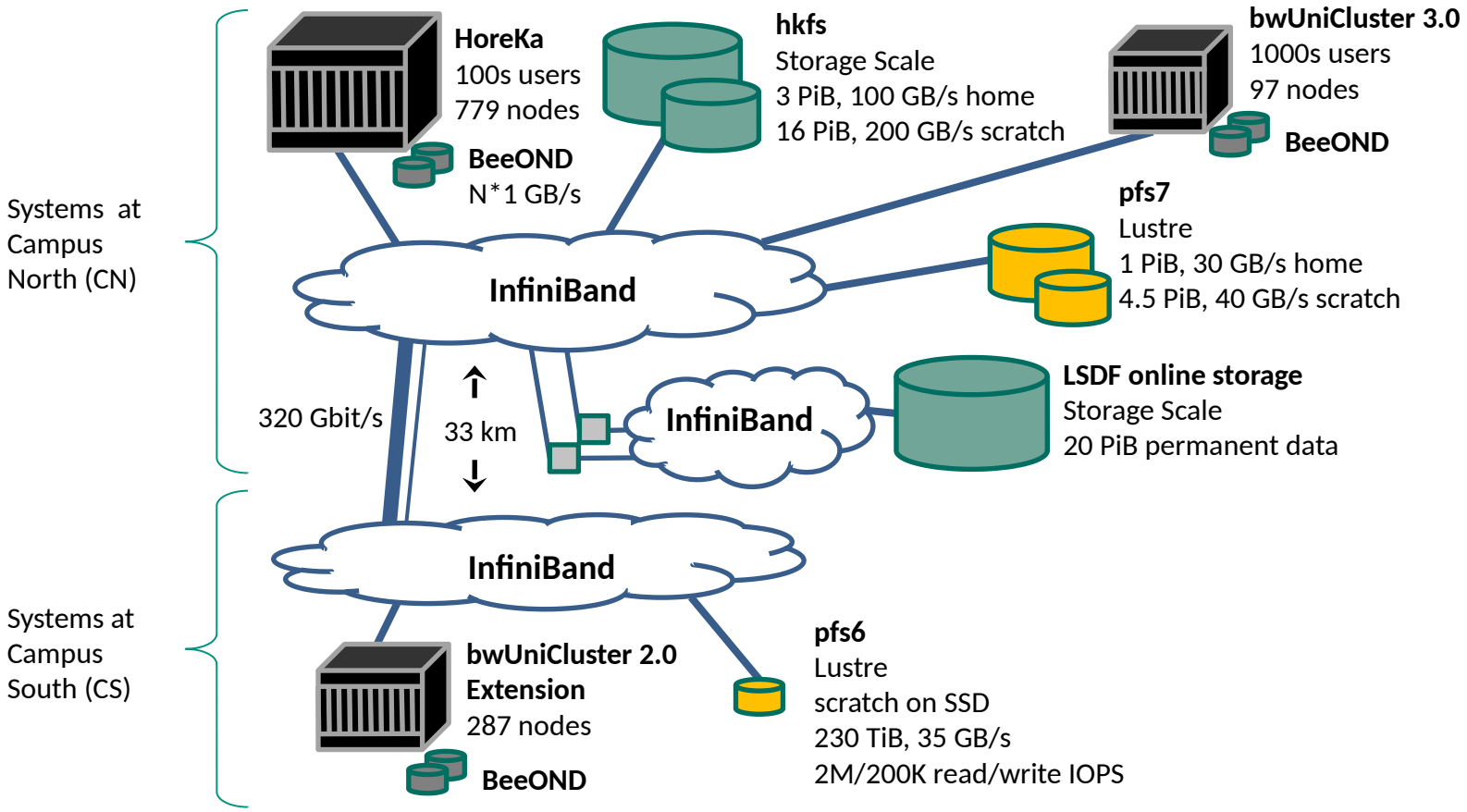
Background

- Karlsruhe Institute of Technology (KIT)
  - Merger of University and Research Center Karlsruhe
  - More than 10000 employees and 22000 students
- Scientific Computing Center (SCC)
  - Computing center of KIT
- Scientific Computing und Simulation (SCS)
  - Department of SCC, operates HPC systems
    - Tier 3 system bwUniCluster 3.0, part of bwHPC
    - Tier 2 system HoreKa, part of NHR
- Roland Laifer
  - HPC file system administrator since 30+ years

About this talk

- Why not using different parallel file systems (PFS) for HDD and flash?
  - Actually, different PFS make sense in some cases
  - With different PFS users have to select the right file system
    - Typically this is not done
- Why mixing HDDs and flash (SSDs) makes sense
  - If high capacity is needed flash is still much more expensive than HDDs
  - SSDs are much faster for small files and random I/O
    - Even with the overhead of a PFS
    - Note: For AI workloads if possible use local file system on SSDs of nodes
- ☆ This talk will present options for mixing HDDs and SSDs

HPC and parallel file systems at KIT



Details of PFS hardware

- pfs6 – Lustre full flash work
  - Used with HPC workspace tools on bwUniCluster 3.0 and HoreKa
  - DDN SFA400NV (4 servers) with 24 SSDs
- hkfs – IBM Storage Scale home and work on SSDs and HDDs
  - Used as home/project and workspace on HoreKa
  - 10 servers with 20 Lenovo DE6000 RAID systems and 2260 HDDs
  - 4 IBM ESS3200 with 96 SSDs
- pfs7 – Lustre home and work on SSDs and HDDs
  - Used as home and workspace on bwUniCluster 3.0
  - DDN SFA400NVX2 (4 servers) with 48 SSDs (QLC) for home
  - DDN SFA400NVX2 (4 servers) with 24 SSDs and 328 HDDs for work

Option 1: Placement rules

- Placement rules to determine if files are stored on SSDs or HDDs
  - Possible with policy engine of IBM Storage Scale
  - Examples
    - Place files with extension .txt or .sh on SSD pool
    - Place files with extension .iso or .jpg on HDD pool
    - Place all files below software directory on SSD pool
  - Note: Placement is decided at creation time
    - Placement depending on file size is not possible
- Problem: General rule for good placement of all files not possible

Option 2: Migrate data

- Migrate data between SSD and HDD pool
  - Used for our hkfs IBM Storage Scale file systems
    - No experience but would be possible for Lustre, too
  - Placement rules to initially store all files on SSDs
  - Policy rules for migration to HDDs
    - Use weight to select files based on access time and size
    - Start if SSD pool usage is above 70%, stop if below 50%
  - Currently no rule to migrate data back
- Problem: Data creation might be faster than migration
  - Mainly happened after small files were selected and migration was slow
  - Solution: Adapt placement rule, store on HDDs if usage is above 80%
- Disadvantage: Additional I/O for data migration

Experiences with data migration

- Experiences with policy runs
  - Scanning the whole file system is fast
    - 700 million files and directories in 15 minutes
    - Servers with appropriate role participate in scanning and migration
  - Policy runs might fail if another administrative command is running
- General observations
  - Much more data is actively used than expected
    - Policy might find no data on SSDs which was not accessed during last day
    - Still a lot of read activity from HDD pool
  - No way for users to check if data is located on HDD or SSD pool
    - Would be useful for performance validation

Option 3: Place small part of files on SSDs

- Place first KBs of all files on SSD pool and rest on HDD pool
  - Used with Progressive File Layout (PFL) on Lustre pfs7 work file system
    - lfs setstripe -E 128K -c 1 -p <fs>.ddn_ssd -E 4G -c 1 -p <fs>.ddn_hdd -E 16G -c 4 -p <fs>.ddn_hdd -E -1 -c -1 -p <fs>.ddn_hdd </path/to/fs>
    - Files below 128 KB are completely located on SSD pool
    - User quota limits on SSD pool to prevent heavy SSD usage
  - Currently IBM Storage Scale can only store files on one pool
    - Request to extend capabilities (IBM RFE) has just been created
  - Note: Initially make sure to create enough inodes on SSD pool

Experiences with using PFL on SSD/HDD pool

- Why this configuration makes sense
  - Large files mostly located on HDD
    - HDD pool provides good streaming performance
  - All small files located on SSDs
    - Small file access creates random I/O which benefits from SSDs
  - Benchmarks confirmed our performance expectations
- General observations
  - Up to now no issues detected
    - Only one month of production experience on bwUniCluster 3.0

Summary

- Currently mixing SSDs and HDDs on PFS usually makes sense
  - For price and performance reasons
- Talk presented possible solutions
  - Using policy engine of IBM storage Scale
  - Using Progressive File Layout of Lustre
- My talks about parallel file systems
  - http://www.scc.kit.edu/produkte/lustre.php
  - roland.laifer@kit.edu