UNIVERSITÄTS-
RECHENZENTRUM

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Challenges and Migrations with Spectrum Scale at Heidelberg University

Oliver Mattes, Sven Siebler – IBM Spectrum Scale Strategy Days 2020

Agenda

- Heidelberg University and their researchers
- Scientific Data Storage Service
- Storage Hardware, funding and procurement
- Using Spectrum Scale Encryption
- Disaster Recovery with AFM-DR
- Planned Data Migration with AFM
- Lessons Learned

Heidelberg University

Scientific Environment

URZ UNIVERSITÄTS-RECHENZENTRUM

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

- Founded in 1386, oldest university in today's Germany
- 12 Faculties, ~28 700 students, ~14 000 employees and 527 professorships
- Successful in the German Excellence Initiatives

- 2 Medical Faculties (Heidelberg, Mannheim)
- 2 University Medical Centres

- Cooperations with other Universities, Industry and Business
- High international Cooperations and Exchange

- Other research centres: Deutsche Krebsforschungszentrum (DKFZ), European Molecular Biology Laboratory (EMBL), MPIs (4x), HITS,. . . (cross-cutting projects and workgroups)

Changing Workflows

Growth of Scientific Data

- Our users: Researchers of Medicine, Biology, Physics,. . .
- Generating, storing, processing/analyzing and archiving
- Previously on paper, usb drives, external harddisks, small NAS
- Explosion of the amount of data, e.g. because of increasingly powerful systems
  - Microscopy: 800 MB, 800 GB or up to 84 TB data per sample
  - Processing time: 2h/sample instead of 60h, automated sample handling
  - Data capacity of TB to PB per year and research group

Challenges:

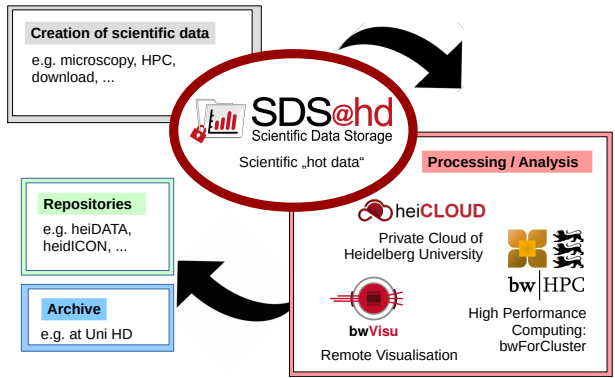
- How to store these amounts of data?
- How to process further this data and where?
- What to do with the data afterwards?

SDS@hd Scientific Data Storage

Central service of the Scientific Data Life Cycle

- ”Landesdienst” for scientists of Baden-Württemberg universities
- ”hot data” (current research)
- Central data storage
- Avoid local data-silo effects
- Tight combination with other services
- Support of multiple access protocols
  - NFSv4 wih Kerberos
  - SMB 2.x, 3.x
  - SFTP
- Cross-site usage

Data Life Cycle:



Large Scale Data Faculity LSDF2

Funding, Procurement of Hardware

UNIVERSITÄTS-
RECHENZENTRUM

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

- Previous System: IBM SONAS, out of service, Capacity reached
- HW funded by DFG and Baden-Württemberg (Art. 91b GG), 2 Proposals and Funding Rounds
- Open bidding, official procurement
- Resulting in heterogenous system architecture

2016-03:	HPE Server & Seagate ClusterStor G200 ($\rightarrow$ Cray $\rightarrow$ HPE) IBM Spectrum Scale Advanced Edition 4.2 ($\rightarrow$ Encryption)	7,8 PB brutto
2017-07:	Dell EMC MD3460 + MD3060e (HW only)	3,6 PB brutto
2019-03:	Dell EMC ME4084 + ME484 (HW only)	2,5 PB brutto
2019-10:	NEC with Dell EMC ME4084 (HW only)	11,2 PB brutto

  - Multiple additional Protocol, Admin and ISKLM Server (HPE and Dell, only HW)
  - **since 2017:** IULA - IBM Spectrum Scale Licences (with Helmholtz-Gemeinschaft) currently IBM Spectrum Scale Advanced Edition 5.0
  - **2017-12:** Quantum Tape Library (~8 PB)

- **Current total size**: 25,1 PB brutto + 8 PB Tape

Encryption

Motivation/Setup

UNIVERSITÄTS-
RECHENZENTRUM

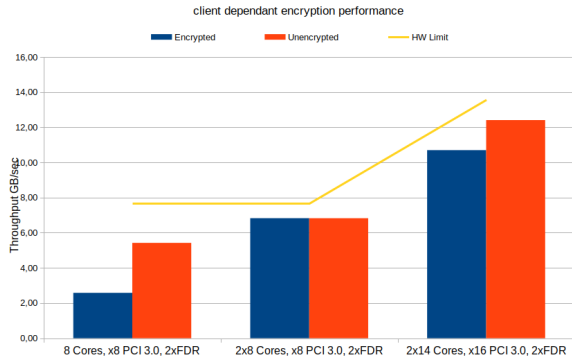
UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

- Needed for some genomic projects and medicine data
- Using 3 SKLM Instances
- Encryption only “data at rest“
  - Not sufficient for all use cases

Encryption

Motivation/Setup

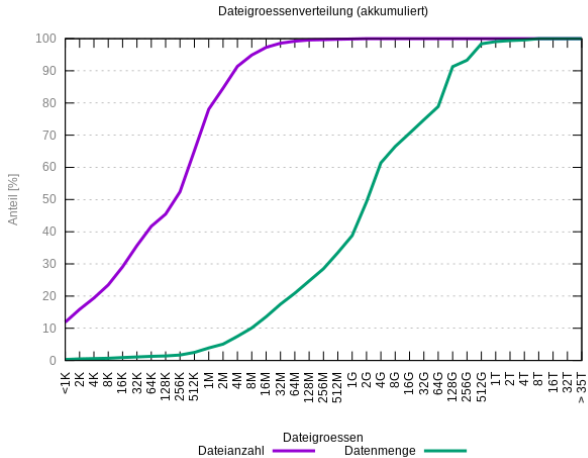
- Needed for some genomic projects and medicine data
- Using 3 SKLM Instances
- Encryption only “data at rest“
  - Not sufficient for all use cases
- Limitations
  - Performance strongly dependend on CPU ressources



Encryption

Motivation/Setup

- Needed for some genomic projects and medicine data
- Using 3 SKLM Instances
- Encryption only “data at rest“
  - Not sufficient for all use cases
- Limitations
  - Performance strongly dependend on CPU ressources
  - No ”data in inode“ available!



Disaster Recovery (AFM-DR)

Motivation

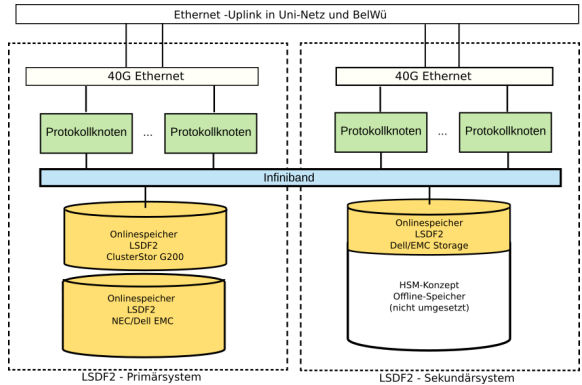
Goals

- Increase service availability
- Prevent restore of data
- Using secondary system during regular maintainance windows
- Minimal user impact
- Scale out solution (100s Filesets, $>$100Mio Inodes)
- Simple and reliable, cost efficient

Disaster Recovery (AFM-DR)

Infrastructure

- Spectrum Scale Advanced Edition
- 4x AFM Gatewaynodes, 4x NFS Nodes (16/28 Cores, 384GB RAM)
- 40GE Ethernet Network and FDR Infiniband
- Using NFSv3 export over 40GE
- TSM-HSM on secondary site to tape
- r/w primary system (Cache)
- r/o secondary system (Home)



Disaster Recovery (AFM-DR)

Issues/Limitations

CES Stack makes it complicated

- ”mmcesdr“ for protocol failover
  - Code level has to be the same on primary/secondary
  - Not working reliable (in our tests)
  - IBM: deprecated feature
- Own implementation of protocol failover/failback needed (REST API)
  - Creating afm-relationships
  - correct en-/disable user access (prevent data corruption!)
  - Maintain/synchronize CES exports
  - Creating psnaps
  - ...

Disaster Recovery (AFM-DR)

Use-Case - G200 Upgrade

Failover (mmafmctl failoverToSecondary)

- Planned downtime of 5 days
- Failover needs 3min per fileset (~ 5-6h in total)
- User access over secondary system without problems

Failback (mmafmctl failbackToPrimary)

- works out of the box in 80% of the filesets
- Balancing Gateway/NFS nodes is difficult (5.x: afmHashVersion)
- Large filesets (>50Mio Inodes) could not failback → RoleReversal needed

⇒ Failback finished in total after 3 months, but no user disruption

Planned Data Migration

LSDF2 → LSDF2.2

Goals/Aims/Tasks

- New filesystem - update version (v4 → v5)
- Unencrypt most data
- Minimal user disruption

Planned Data Migration

using AFM (Local Update Mode)

Goals/Aims/Tasks

- New filesystem - update version (v4 $\rightarrow$ v5)
- Unencrypt most data
- Minimal user disruption

AFM Advantages

- Expecting good performance via RCM
- Prefetching of ”hottest” data possible
- Remaining data will be migrated after switching to new system (transparent)
- Minimal downtime for users, no additional changes needed
- Local-Update reduces data synchronisation to G200
- Snapshots on G200 still available

Lessons Learned

Hardware

- “Closed” appliances could be problematic
- If possible use only hardware, to ensure homogenous software stack over time

Encryption

- using only for specific use cases! mostly client-site encryption is the better solution
- no security benefit, if disks stay on site

AFM-DR

- use fast SSDs for meta data, even on secondary site!
- large filesets needs special treatment
- experience is important
- it does work, but expect the unexpected

Thank you for your attention!

Questions?

- SDS@hd: `https://sds-hd.urz.uni-heidelberg.de`

URZ UNIVERSITÄTS-RECHENZENTRUM

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Challenges and Migrations with Spectrum Scale at Heidelberg University

Oliver Mattes, Sven Siebler – IBM Spectrum Scale Strategy Days 2020