We make IT happen

FROM INFRASTRUCTURE TO SOLUTION

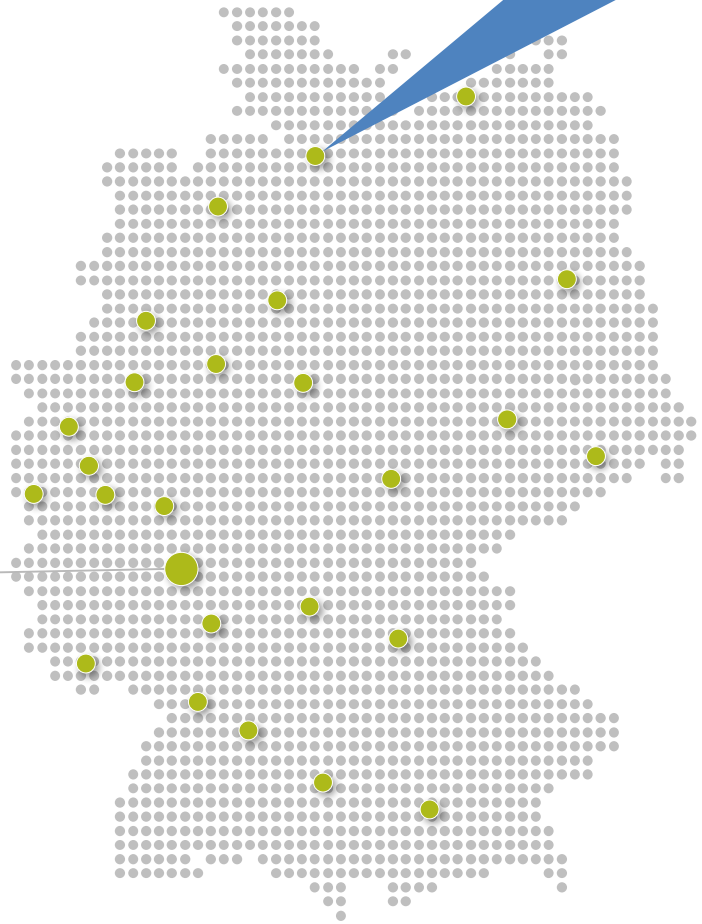
Not enough money for an all flash HPC storage - A brief cheating guide

jochen.zeller@sva.de

/ COMPANY

During ISC:
Meet SVA @ G703

26
Locations



7 TOP
Industries

- Automotive
- Retail
- Public
- Machine & Plant Construction
- Finance & Insurance
- Telecommu-nication
- Healthcare

/ COMPANY

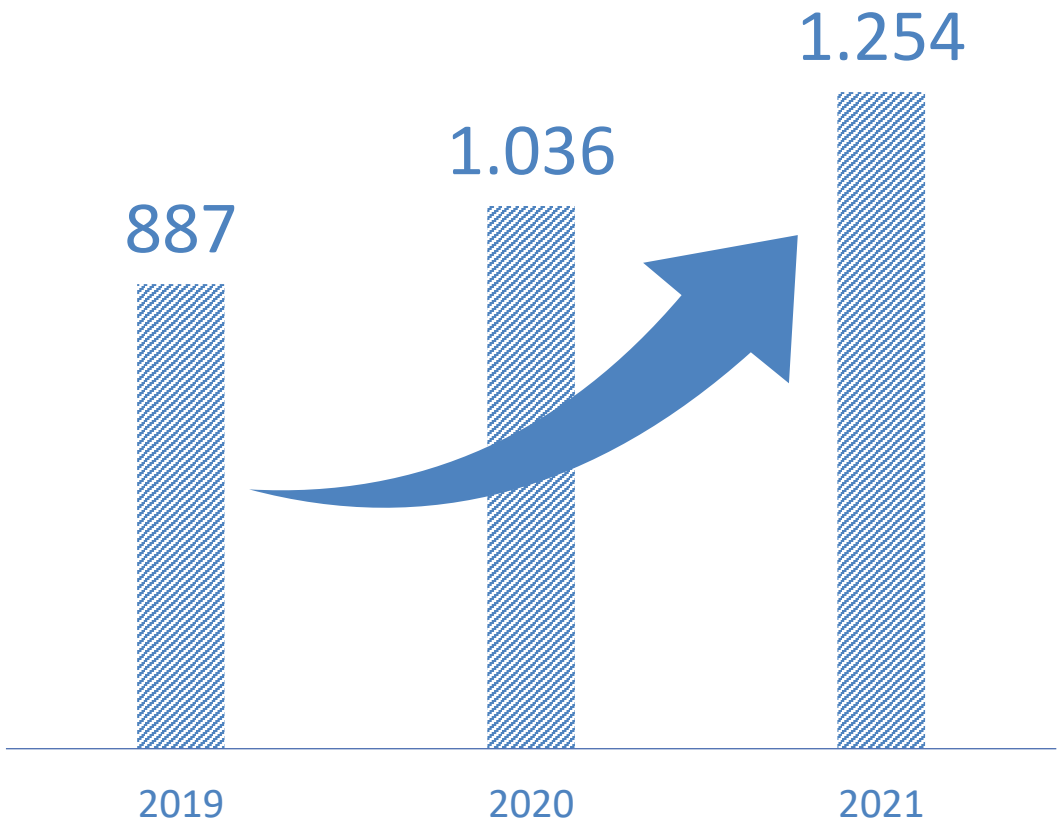
Employees

2.300
2021

1.500
2020

1.000
2019

Sales volume in Mio €



/ WHY ALL FLASH SCRATCH AND DATA AND HOME AND …?

- Flash = low latency IOPS = more IOPS per second = (probably) higher bandwidth = higher CPU efficiency
- But higher cost per TB compared to high-capacity NL-SAS
- By the way, do not mix up IBM's new ESS3500 ultra speed NVMe system with ESS3500 welding tool
- How we combined flash and SAS/NL-SAS so far:
  - copy data to flash and start processing
  - Use pools with flash and NL-SAS and the policy engine to place your data appropriate

/ WHY ALL FLASH SCRATCH AND DATA AND HOME AND …?

- Data placement and data flow with pools and policy engine, example:

system pool: metadataOnly

nvme pool: default for all files

ESS3200. NVMe.

But: automatic „make hot again“ not available

Applications / projects which mostly generates files >10GB should (must) point to data pool (by fileset rule)

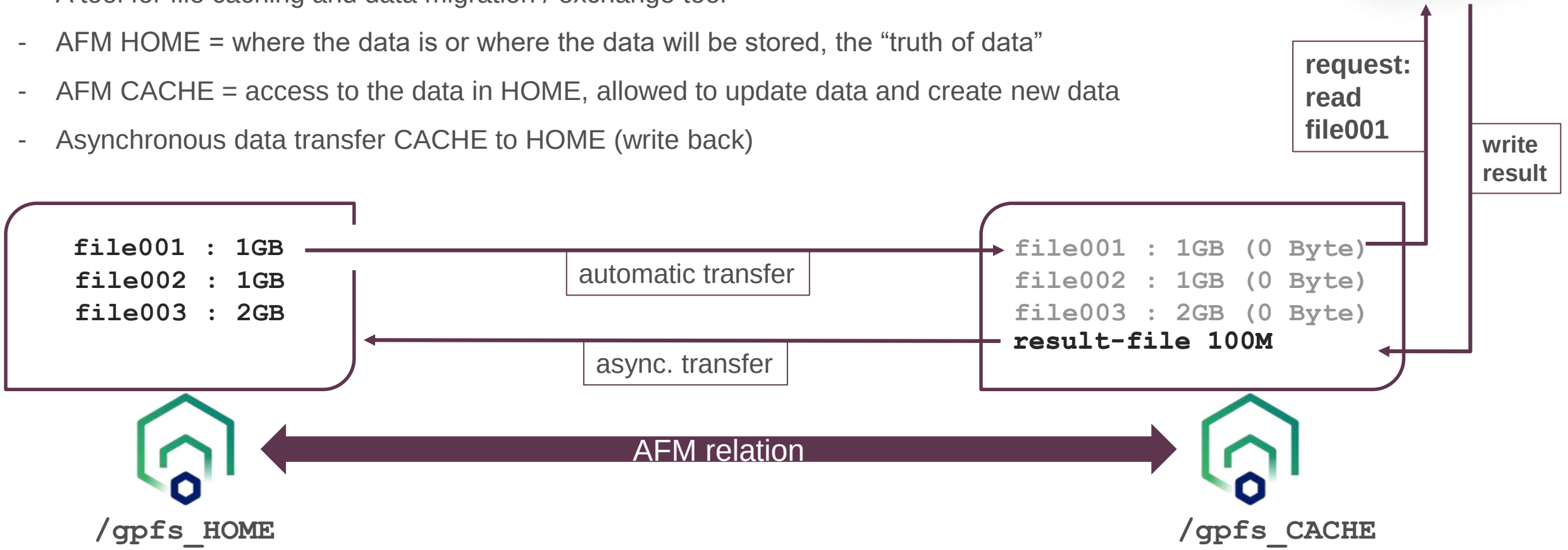
ESS5000. NL-SAS.

daily: migrate files >128MB from pool nvme to pool data (with QoS!)

threshold: if pool nvme > 60% migrate files (to 50%) to pool data, weight by filesize (without QoS!)

/ WHAT IS AFM? COULD THIS HELP?

- AFM – Active File Management
- A tool for file caching and data migration / exchange tool
- AFM HOME = where the data is or where the data will be stored, the “truth of data”
- AFM CACHE = access to the data in HOME, allowed to update data and create new data
- Asynchronous data transfer CACHE to HOME (write back)



/ AFM - EASY TO SET UP

- Let's use AFM to run as many IOPS as possible on flash
- One Spectrum Scale Cluster with two file systems **`/hpc`** and **`/hpc_slurm`**
- The plan: AFM relation between **`/hpc/scratch`** (HOME) and **`/hpc_slurm/scratch`** (CACHE)
- Define AFM gateway nodes with mmchnode command, e.g. **`# mmchnode --gateway -N <node1,node2,node3,…>`**
- /hpc/scratch is an already existing independent fileset, export this for AFM access:

```
# mmafmconfig enable /hpc/scratch
```

`/hpc`
`/hpc/scratch`

AFM enabled

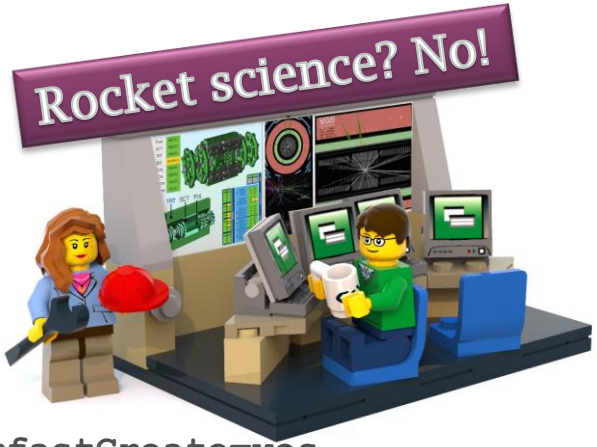
`ESS5000`
`NL-SAS`
`1.3 Pbyte @ 15GB/sec`

`gateway nodes`

`/hpc_slurm`

`ESS3200`
`NVMe`
`105 Tbyte @ 80GB/sec`

/ AFM - EASY TO SET UP



- Create AFM fileset **/hpc_slurm/scratch:**

```
# mmcrfileset hpc_slurm scratch -p afmMode=IW,afmTarget=gpfs:///hpc/scratch,afmfastCreate=yes
--inode-space new --inode-limit 500M:100M --allow-permission-change chmodAndUpdateAcl
# mmlinkfileset hpc_slurm scratch -J /hpc_slurm/scratch
```

- Prefetch inodes to CACHE (not necessary, but helpful):

```
# mmafmctl hpc_slurm prefetch -j scratch --directory /hpc_slurm/scratch --metadata-only
```

/hpc
/hpc/scratch

ESS5000
NL-SAS
1.3 Pbyte @ 15GB/sec

gateway nodes

/hpc_slurm
/hpc_slurm/scratch

ESS3200
NVMe
105 Tbyte @ 80GB/sec

/ PROJEKT SETUP – DATA ACCESS AND DATA FLOW

ESS3200 105TB
24 x 7,68TB NVMe
4x HDR200 IB
4x 100GBE

/hpc_slurm

access

HPC compute

AFM

/hpc

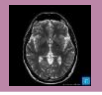
CES
access

write

all data

ESS5000 SL2 1.3PB
182 x 10TB
4x HDR100 IB
4x 100GBE

prefetch
files



run compute job

--dependency=after:prefetch+1

/ FILE PREFETCHING

- Why prefetching?
  - Prefetched files are already transferred to /hpc_slurm. No further wait times for data movement when compute resources are allocated. Files that are not already prefetched will be prefetched during job runtime.
- How to:

```
# sudo -i mmafmctl hpc_slurm prefetch -j scratch --list-file /tmp/file-list-job4711
--gateway gwnode1
```

The files to prefetch will be send to the queue, mmafmctl returns as soon as the files are queued!

Four and a half options:

1. Just wait n seconds and start job (doubtful, but easy)
2. Wait for callback event “afmPrepopEnd” and start job (more complicated, more reliable)
3. Use mmafmctl or mmpmon and verify the queue (not easy with parallel prefetching)
4. Do not prefetch and keep some filesets in CACHE, let AFM work for you (probably slower)
5. Find your own solution – with Scale there is more than one way to skin a cat

/ AFM - TUNING, TIPS, TRICKS

- Set fileset quotas on each AFM cache fileset. The quota is used as eviction watermarks. The hard limit = high water mark, the soft limit = low water mark. Example: soft=70T hard=80T
  → AFM start eviction at 80T and stops at 70T. Algorithm is LastRecentUse, LRU

- Don’t forget to increase the inode limit! Default is 100,000 inodes

- One fileset is handled by one gateway node. If this node is not available, AFM initiates a takeover to another gateway node from the list. Manual distribution is allowed, you can specify `mmcrfileset … -p … afmGateway=`

- You can adjust prefetching on a fileset level by afmPrefetchThreshold.
  → “0” - means full file prefetch on access (DEFAULT)
  → “1-99” - in percent, when this amount of the file was read, than full file prefetch starts
  → “100” – fetch only what should be read

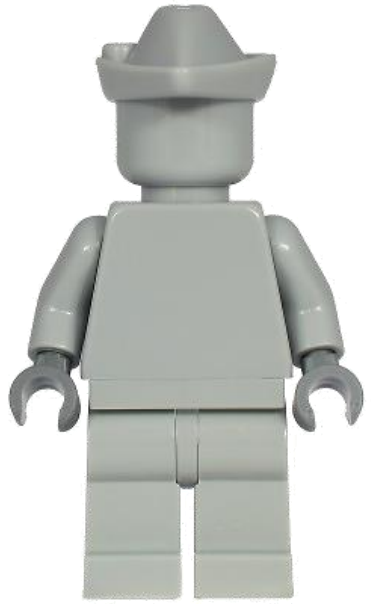
/ AFM - TUNING, TIPS, TRICKS

- Take a look to some other options and adopt them to your needs (the defaults are best practices, and as you now, best practices must fit for all installations → best practices makes you average)

  `mmchfileset` …

  Refresh: **`afmAsyncDelay, afmDirLookupRefreshInterval, afmDirOpenRefreshInterval, afmFileLookupRefreshInterval, afmFileOpenRefreshInterval`**

  Threads: **`afmNumFlushThreads, afmNumReadThreads, afmNumWriteThreads`**



/ SUMMARY

- Meanwhile, AFM is a reliable tool
- Easy to set up, easy to monitor (mmhealth, mmafmctl)
- An almost perfect solution to integrate flash into a huge NL-SAS based file system
- Best with scheduler integration and great without any additional automation

The good thing about Spectrum Scale is that you can design and adjust a lot.

The bad thing about Spectrum Scale is that you can design and adjust a lot.

/ CONTACT

JOCHEN ZELLER

IT Architect

Technical Leader IBM Spectrum Scale

Phone.: +49 151 180 256 77

Mail: jochen.zeller@sva.de