SSSD23
multi tenant file system consolidation

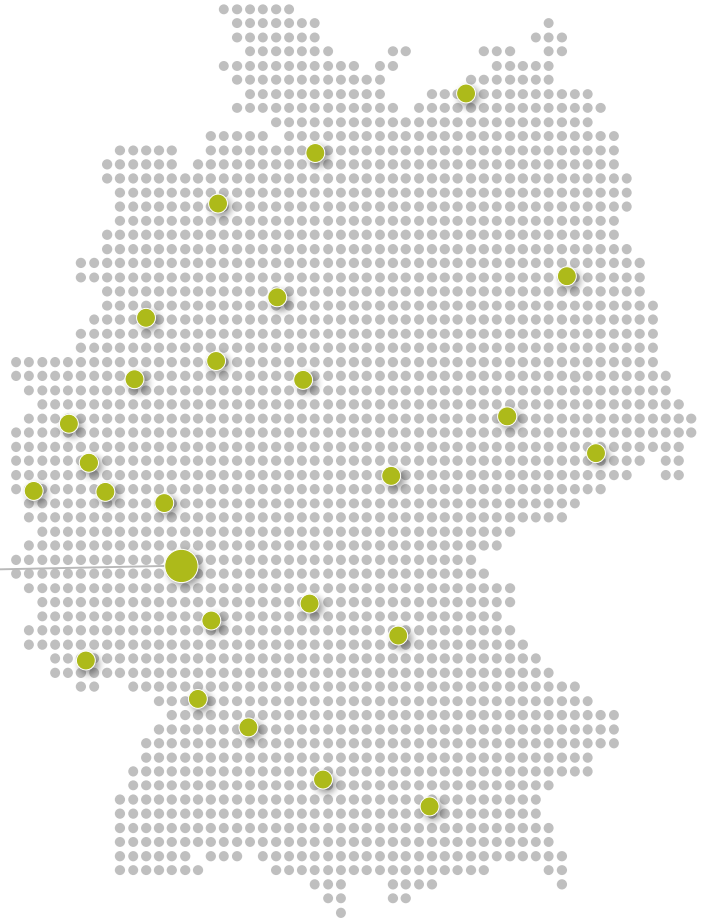
jochen.zeller@sva.de

SVA ISS Team

ABOUT US

/ COMPANY

26
Locations



7 TOP
Industries

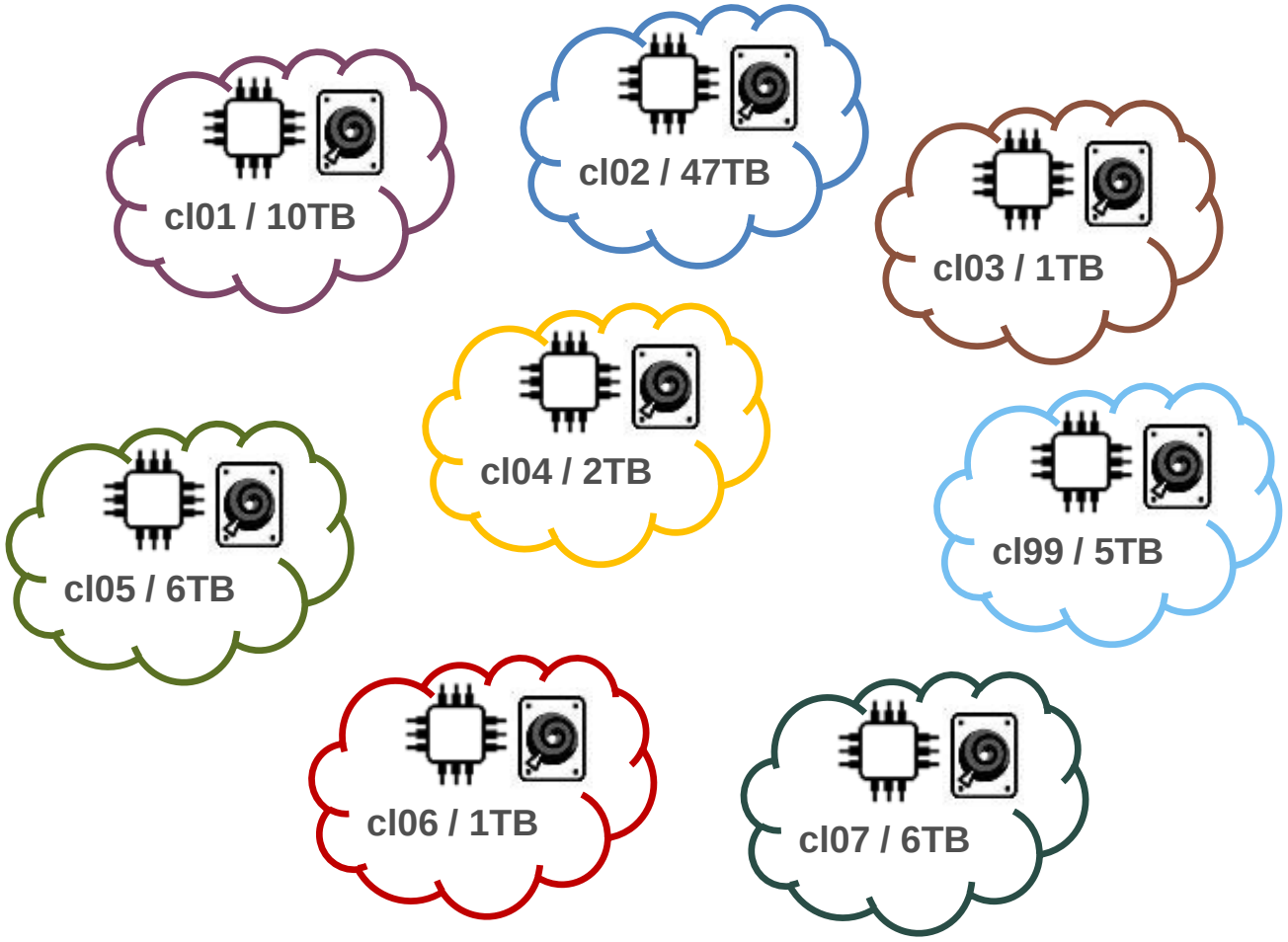
- Automotive
- Retail
- Public
- Machine & Plant Construction
- Finance & Insurance
- Telecommu-nication
- Healthcare

WHAT MAKES US UNIQUE?

/ SVA ...

- Owner-operated
- High employee satisfaction and loyality
- Long-term technical support for solutions delivered by us
- Customer satisfaction and recommendation above average

/ SO MANY CLUSTER AND FILE SYSTEMS



- 2 x AIX VM as NSD server per cluster
- SVC as backend storage, SVC vdisk mirror
- Many file systems per cluster
- Not every client mounts all file systems
- Scale clients run applications on RHEL

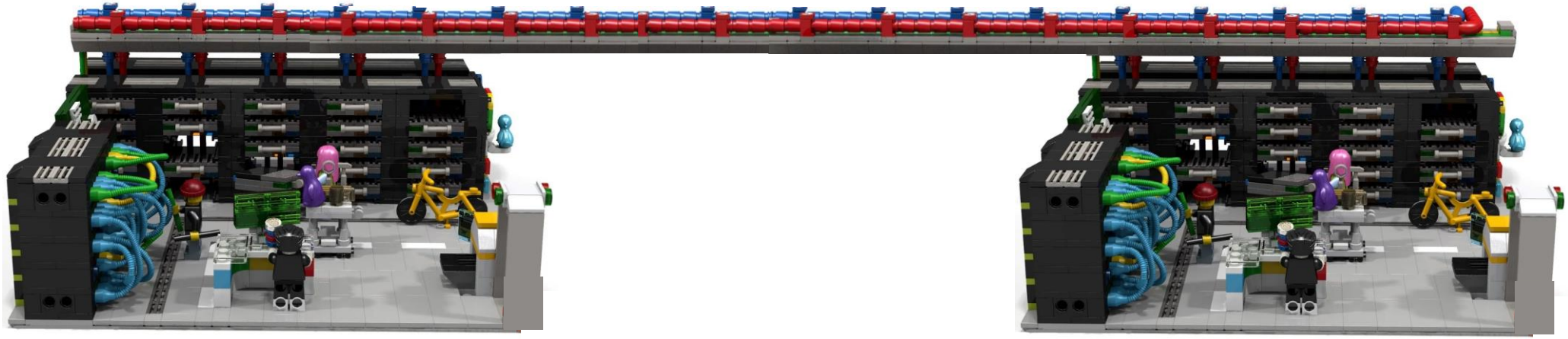
cluster	file systems	capacity (sum)
8	116	78TB

The task:

- Replace IBM POWER by x86
- Replace IBM AIX by RHEL
- Replace SVC
- The separation of data must remain

/ GENERAL CONDITIONS

- Pure Storage, FlashArray, is the customers successor for SVC
- Mirroring, high availability between two data centers, is necessary
- Performance requirements: 2GB/sec overall is fine
- Are virtual NSD servers an option, like the AIX VMs were before?
  - 16 NSD server VMs are cheaper than 16 rack servers
- Storage compression would be nice, current SVC compression is 1:2
- Spectrum Scale clients access the file system by Ethernet / tcpip
- NFS shares would be great



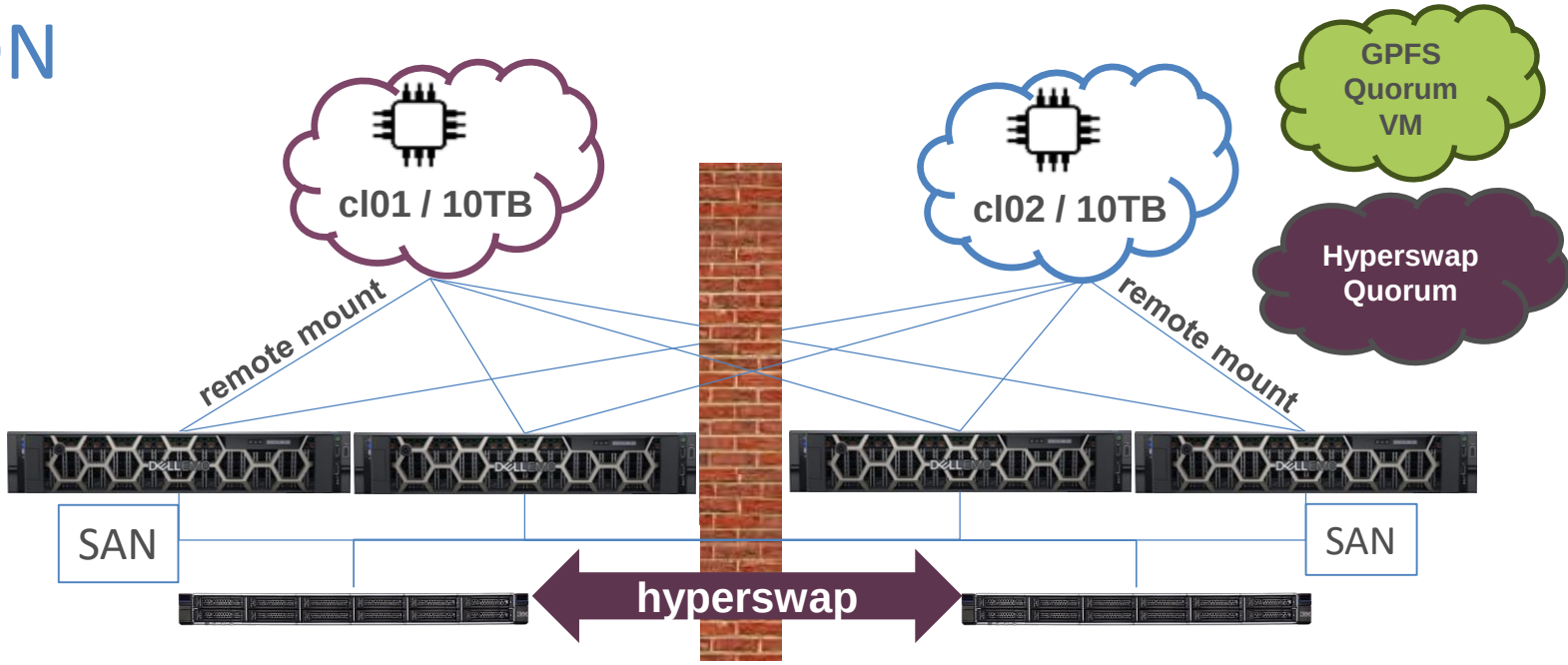
/ PLANNING WORKSHOP

cluster	file systems
8	116

- Let’s start with a vision: We will have only one file system
- With a single file system, we just need two or four NSD servers
  - File system consolidation with separation means a remote cluster setup
- What about the storage?
  - Alternatives to Pure Storage? IBM?
  - ESS is to expensive for this small configuration with low performance requirements
  - IBM FS5200 could be a great (and cheap?) solution
    - Mirroring with HyperSwap (no additional license cost)
      - Reduces the Spectrum Scale licenses by half compared to Spectrum Scale mirroring
    - FCMs with compression

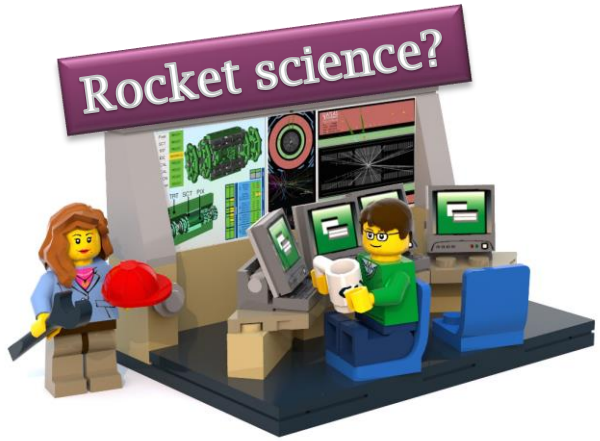
/ THE HARDWARE SOLUTION

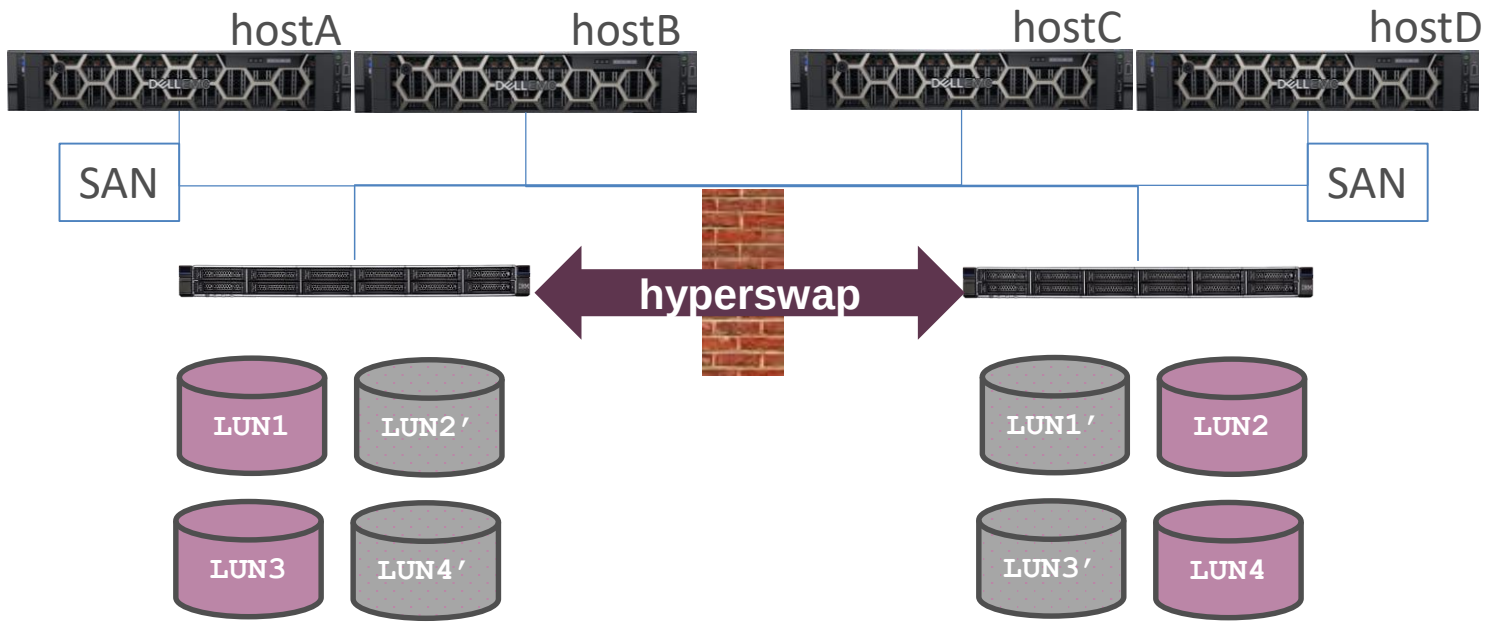
- 4 NSD server + 1 quorum VM
  - 2 x 32Gbit FC
  - 2 x 25Gbit Ethernet
- 2 x IBM FS5200 in Hyper Swap cluster
  - 70TB @ 12 x 9,6TB FMC
  - 140TB with 1:2 compression, 120TB assigned
- SAN for storage access and hyperswap (already present)



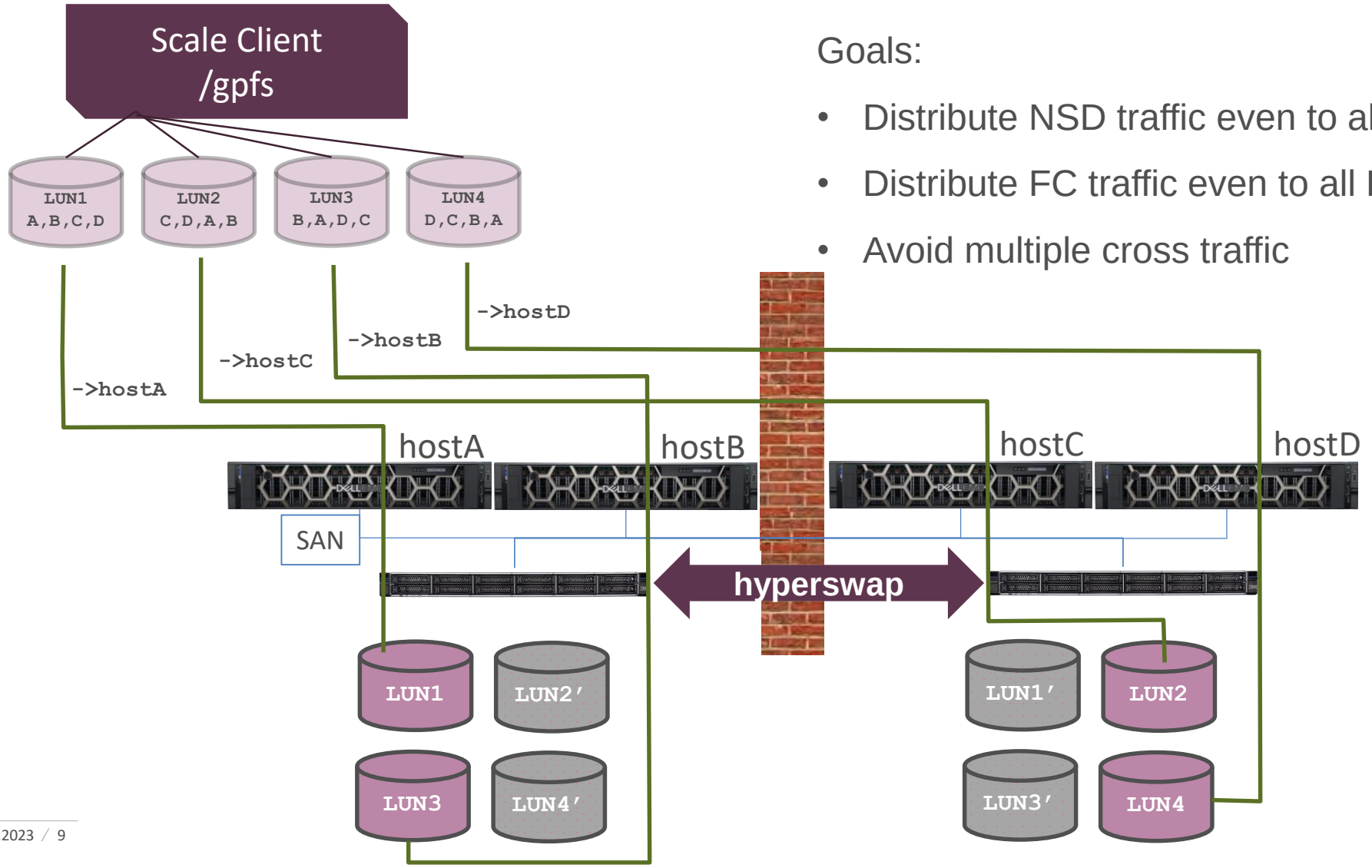
/ HYPERSWAP AND MULTIPATHING

- Both FS5200 build a cluster and both provide the same LUNs
- The paths of a LUN are on one site “active” and on the other site “enabled”
- Host multipathing uses ALUA - Asymmetric Logical Unit Access
- Hyperswap can change the active site for a LUN automatically





/ SPECTRUM SCALE AND HYPERSWAP – DATA ACCESS PATH

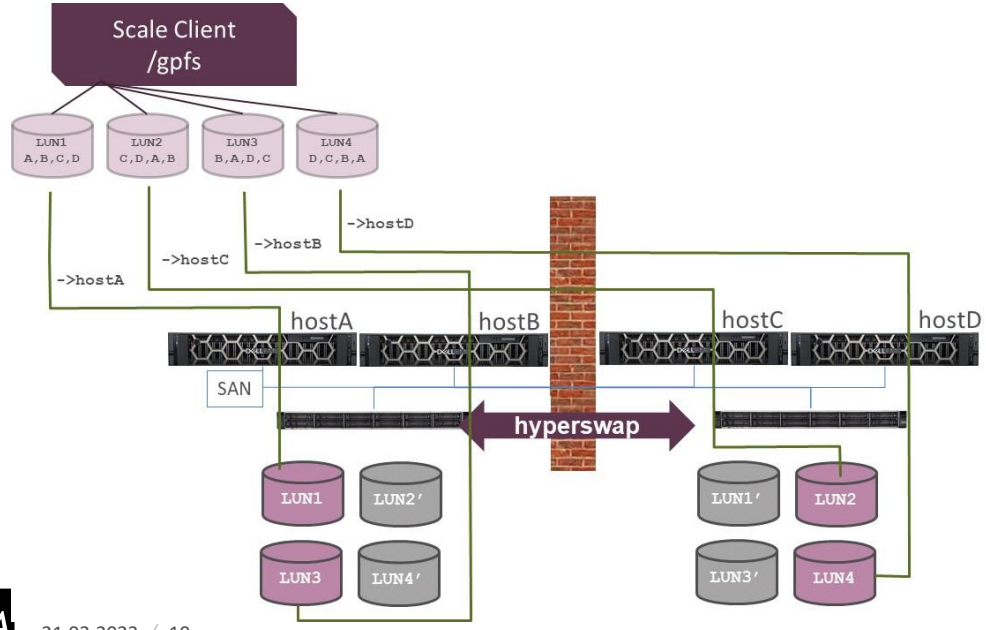


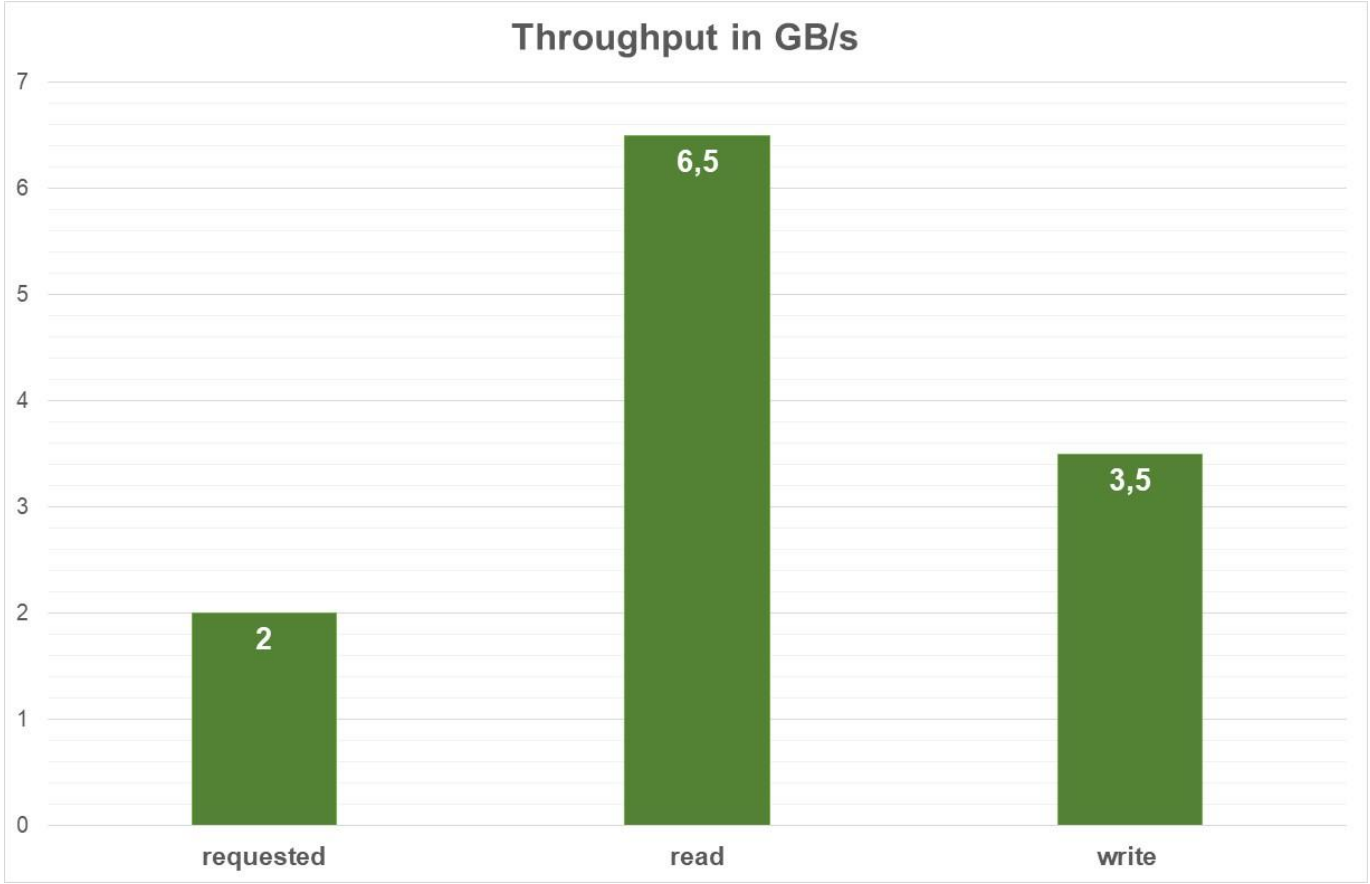
Goals:

- Distribute NSD traffic even to all 4 NSD servers
- Distribute FC traffic even to all FS5200 and controllers
- Avoid multiple cross traffic

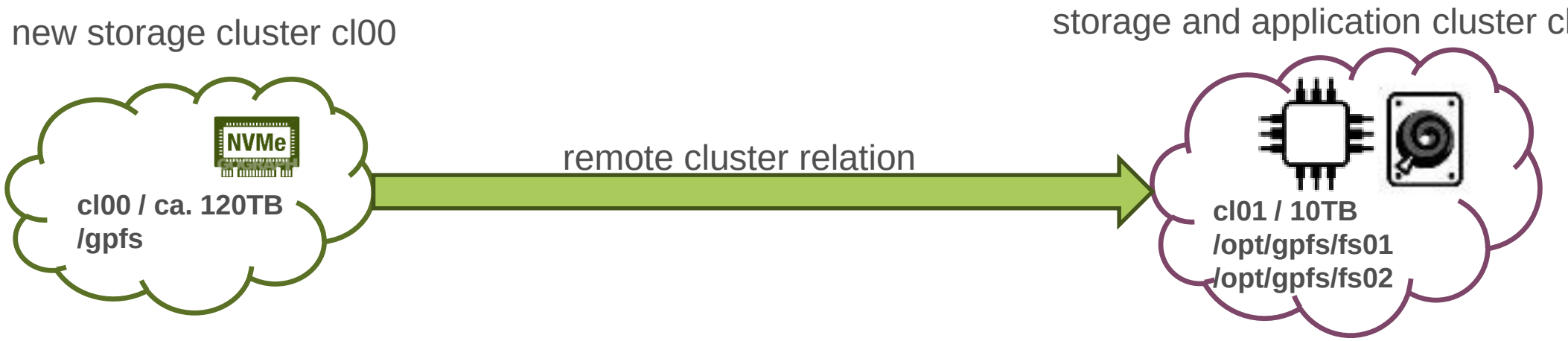
/ PERFORMANCE

- FS5200 HyperSwap has a reduced write performance compared to a single FS5200
  - Host writes to one FS5200, this must copy the data to the second FS5200. When the second FS5200 reports "sync", the IO is reported by the first FS5200 as sync to the host.
- 50% of the client IO’s go via Ethernet to the other data center (ISL bound)





/ TO THE ACTUAL TOPIC: MULTI TENANCY / REMOTE MOUNT




```
# mmcrfileset gpfs cl01_fs01
# mmlinkfileset gpfs cl01_fs01 -T /gpfs/cl01/fs01
# mmsetquota gpfs:cl01_fs01 --block 100G:80G
# mmauth grant cl01 -f gpfs --fileset cl01_fs01
```

```
# mmremotefs add gpfs -f gpfs -C cl00 -T /gpfs -A
# mmmount gpfs
# <stop application>
# rsync -a /opt/gpfs/fs01 /gpfs/cl01/fs01    ***
# mmumount /opt/gpfs/fs01
# mmchfs fs01 -T /opt/gpfs/fs01_old
# mount --bind /gpfs/cl01/fs01 /opt/gpfs/fs01
# <start application>
```

*** *or AFM for larger file systems*

```
mmlinkfileset gpfs cl01_fs01 -T /gpfs/cl01/fs01
```

/ REMOTE FILESET LOOK AND FEEL

- Mount one (or more) remote file system
- Share access of one or more filesets
- Content of root fileset is always visible / accessible
- In our case, a "**ls**" to the file system root **/gpfs** shows all directories **cl01,cl02,cl03, … cl99**

```
# ls
cl01    cl02    cl03    cl04    cl05
```

- Because all subdirectories in the clXY directories are filesets (fs01, fs02, …), they are only visible if they are shared by the source cluster.

```
# cd cl02
# ls
fs01    fs02    fs09
```

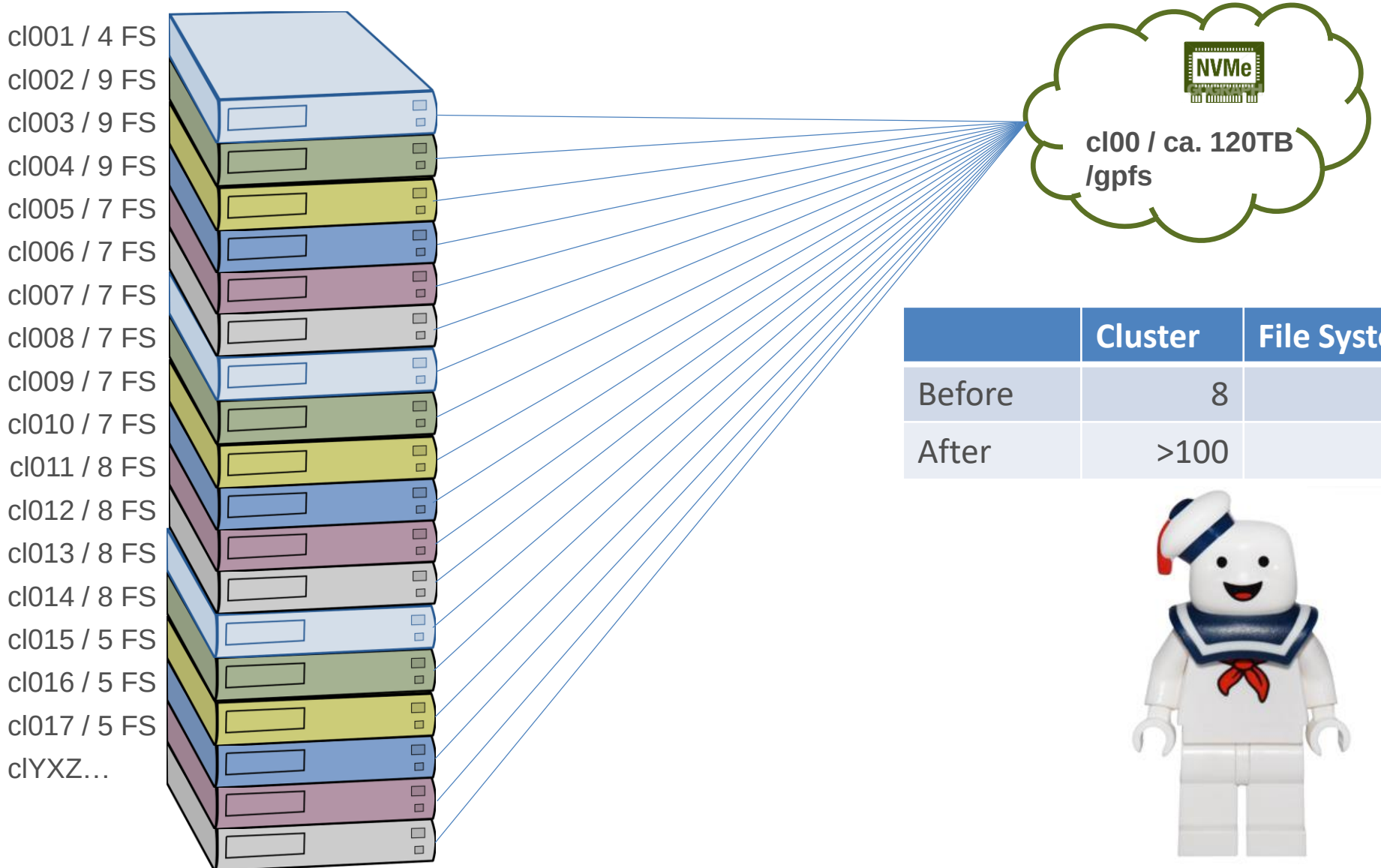
/ HOW TO SETUP THE REMOTE CLUSTER

- **THE** question: many multi node clusters or even more single node clusters?
- The problem with multi node cluster:
  - You have to take care of the cluster quorum
  - This means that at least three nodes are not equal to the others
- Decision: single node cluster
- Why: The client cluster setup and remote relationship will be automated with Ansible anyway. Whether execute 8 times or more than 100 times does not make much difference. By the way:

Q5.6:
What is the limit of remote clusters that can join a local cluster?
A5.6:
There is not really a limit. The smallest cluster possible is a single node cluster, which means that 16,383 clusters can join a local cluster (16384 - 1).

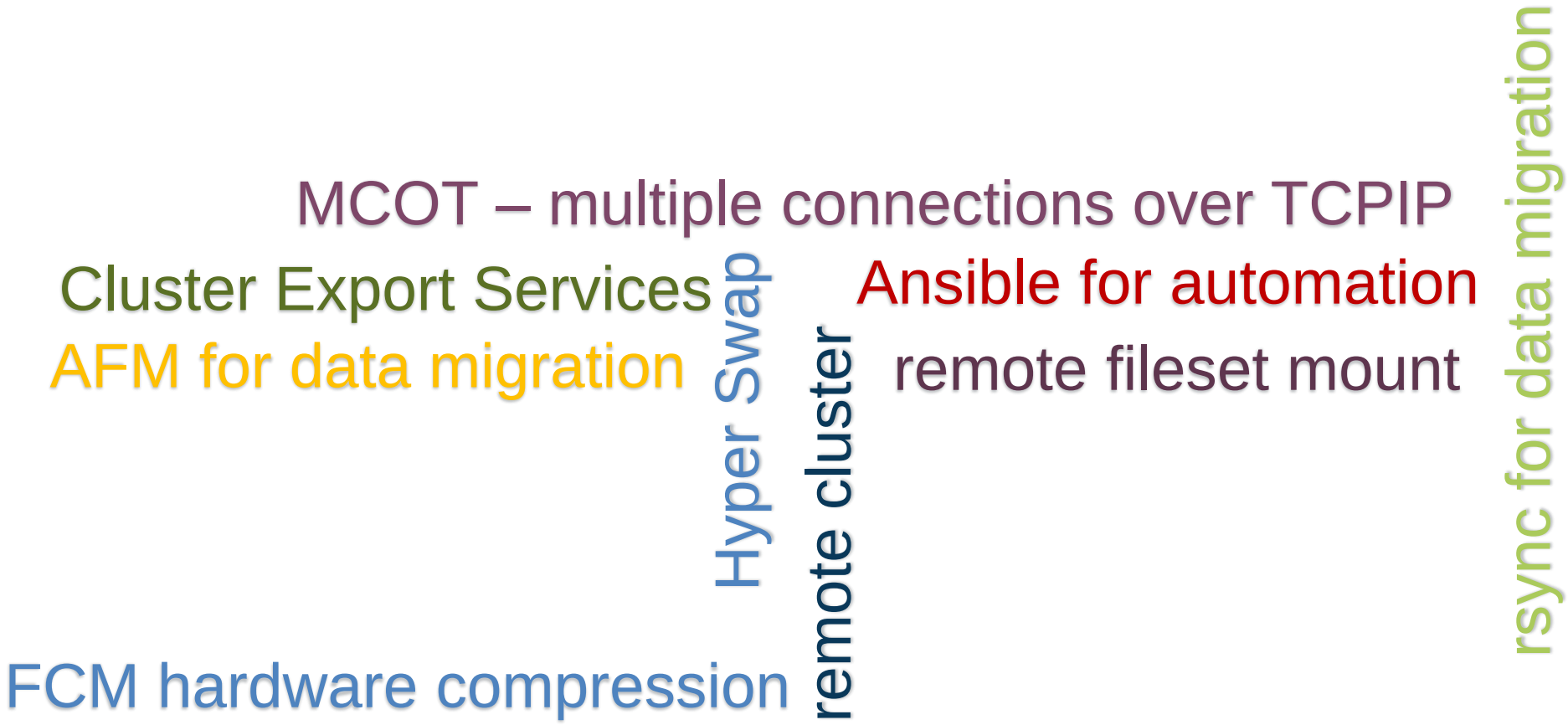
https://www.ibm.com/docs/en/STXKQY/gpfsclustersfaq.html#inclusters

/ MULTI TENANT FILE SYSTEM CONSOLIDATION. DONE.



	Cluster	File Systems
Before	8	116
After	>100	1

/ A TECHNICAL AWESOME PROJECT



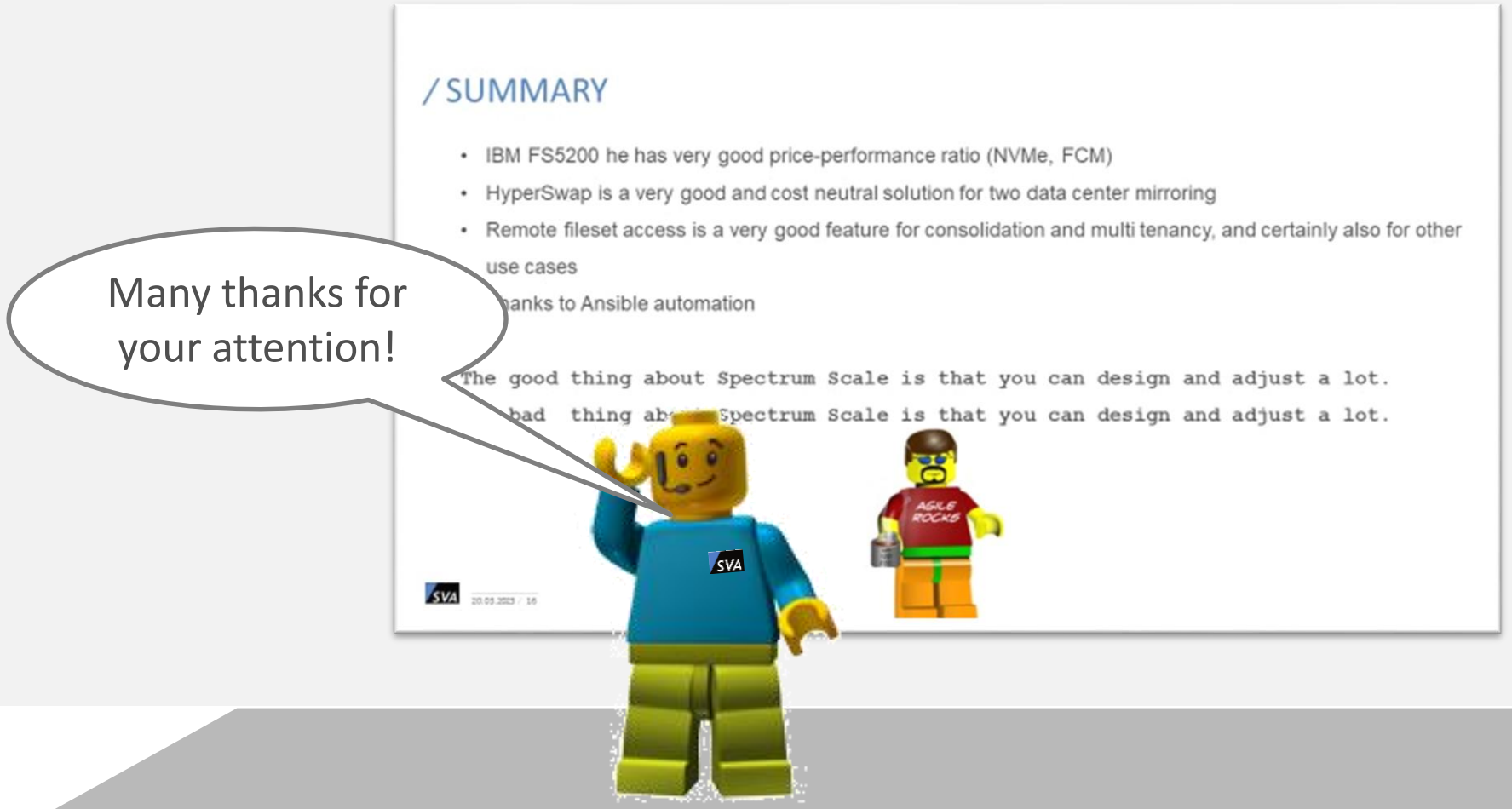
/ SUMMARY

- IBM FS5200 he has very good price-performance ratio (NVMe, FCM) for small configurations
- HyperSwap is a very good and cost neutral solution for two data center mirroring
- Remote fileset access is a very good feature for consolidation and multi tenancy, and certainly also for other use cases
- Thanks to Ansible 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.

/ THE END!


/ SUMMARY
• IBM FS5200 he has very good price-performance ratio (NVMe, FCM)
• HyperSwap is a very good and cost neutral solution for two data center mirroring
• Remote fileset access is a very good feature for consolidation and multi tenancy, and certainly also for other use cases
anks to Ansible automation
The good thing about Spectrum Scale is that you can design and adjust a lot.
bad thing ab Spectrum Scale is that you can design and adjust a lot.
Many thanks for your attention!
SVA
AGILE ROCKS

/ CONTACT

JOCHEN ZELLER

IT Architect

Technical Leader IBM Spectrum Scale

Phone.: +49 151 180 256 77

Mail: jochen.zeller@sva.de