Spectrum Scale Strategy Days 2022

USING SPECTRUM SCALE FOR MEDIA AND BROADCAST WORKFLOWS

Felix Bartelt
Senior Developer for Storage Solutions
Rohde & Schwarz

ROHDE&SCHWARZ
Make ideas real

Rohde & Schwarz

THE COMPANY

Founded 1933 in Munich

16 % of net revenue goes into R&D

Presence in over 70 countries
storage development team located in Hanover, Germany

Independent family business

Over 10,000 employees all over the world

Over 2 billion USD revenue in fiscal year 2020/2021

Rohde & Schwarz

BUSINESS FIELDS

Test & Measurement

Aerospace | Defense | Security

Cybersecurity

Broadcast & Media

Delivery & Distribution

Studio Production

Post Production

MEDIA PRODUCTS

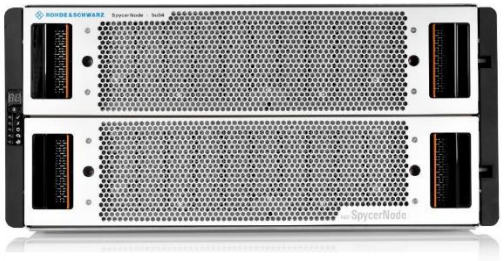
Mastering
edit and finalize uncompressed video
for digital cinemas and streaming platforms

Editing
create and edit media files for
TV and film productions

Production
ingest and playout media for Live and
Studio productions

R&S Clipster

Storage



R&S SpycerNode

R&S VENICE

OUR STORAGE HISTORY

1997 — ProntoVision

(world's first) uncompressed HDTV Disk Recorder
Real-Time HDTV I/O
at 1.3 Gbit/s

2004 — SAN

Storage Area Network
Shared Filesystem
StorNext

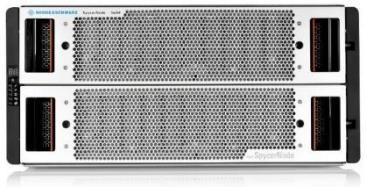
2007 — SpycerBox

SAN-in-a-Box
build-in:
metadata manager,
Fileserver,
FibreChannel
disk export

2013 — SpycerBoxCell

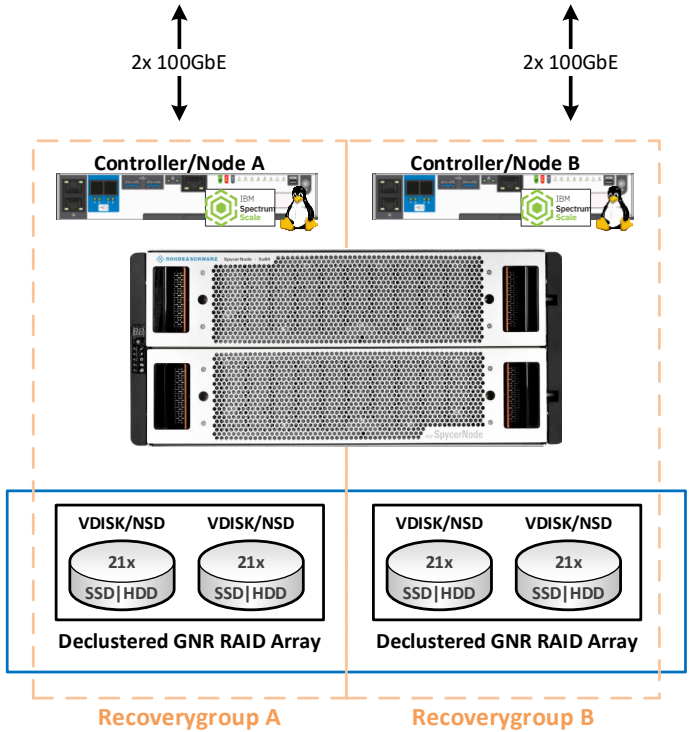
High-Density Storage
30x 2.5" disks @ 1U
bandwidth 3GB/s per unit
Quantum StorNext
IBM GPFS

2019 — SpycerNode

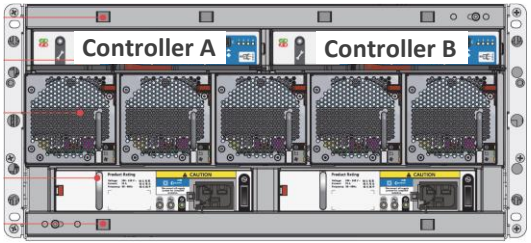


High Availability Media Storage
IBM SpectrumScale
Filesystem + RAID

R&S SPYCERNODE



- 5U chassis
- 2 drawers with 84 drives total
- two redundant embedded storage controllers
- embedded Spectrum Scale services:
  cluster manager, RAID, file system, CES
- can be expanded with JBODs to add more capacity



WORKFLOWS FOR MEDIA PRODUCTIONS
(SOME EXAMPLES)

- Post Production
  creation of encrypted Digital Cinema Packages (DCP)
  color grading and editing, quality control
  film scanner/digitizer

- Studio Ingest and Playout
  recording of live productions (e.g. sport events, talk shows,...)
  live playout/broadcasting for TV channels

- Broadcast Editing
  creation and editing of editorial content (video clips, graphics, audio, text,...)

- VFX Rendering
  creation of visual effects (fire, smoke, water,...)
  rendering of 3D animations

POST PRODUCTION WORKFLOWS

► Post Production

creation of encrypted Digital Cinema Packages (DCP)
color grading and editing, quality control
film scanner/digitizer

Editing & Mastering
Clipster | Rohde & Schwarz

Color grading
DaVinci Resolve | Blackmagic

Film Scanner (Telecine)
4K Scanity | DFT

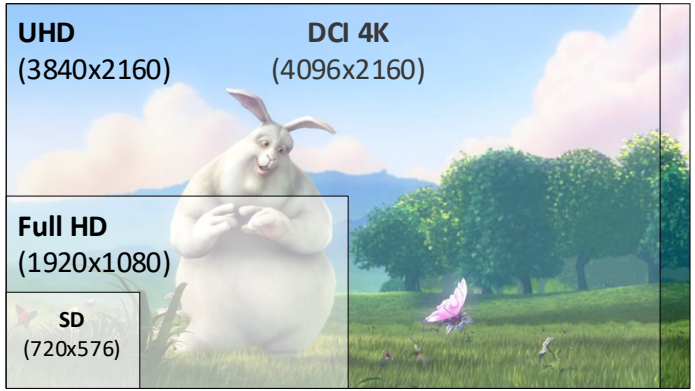
POST PRODUCTION STORAGE

Requirements:

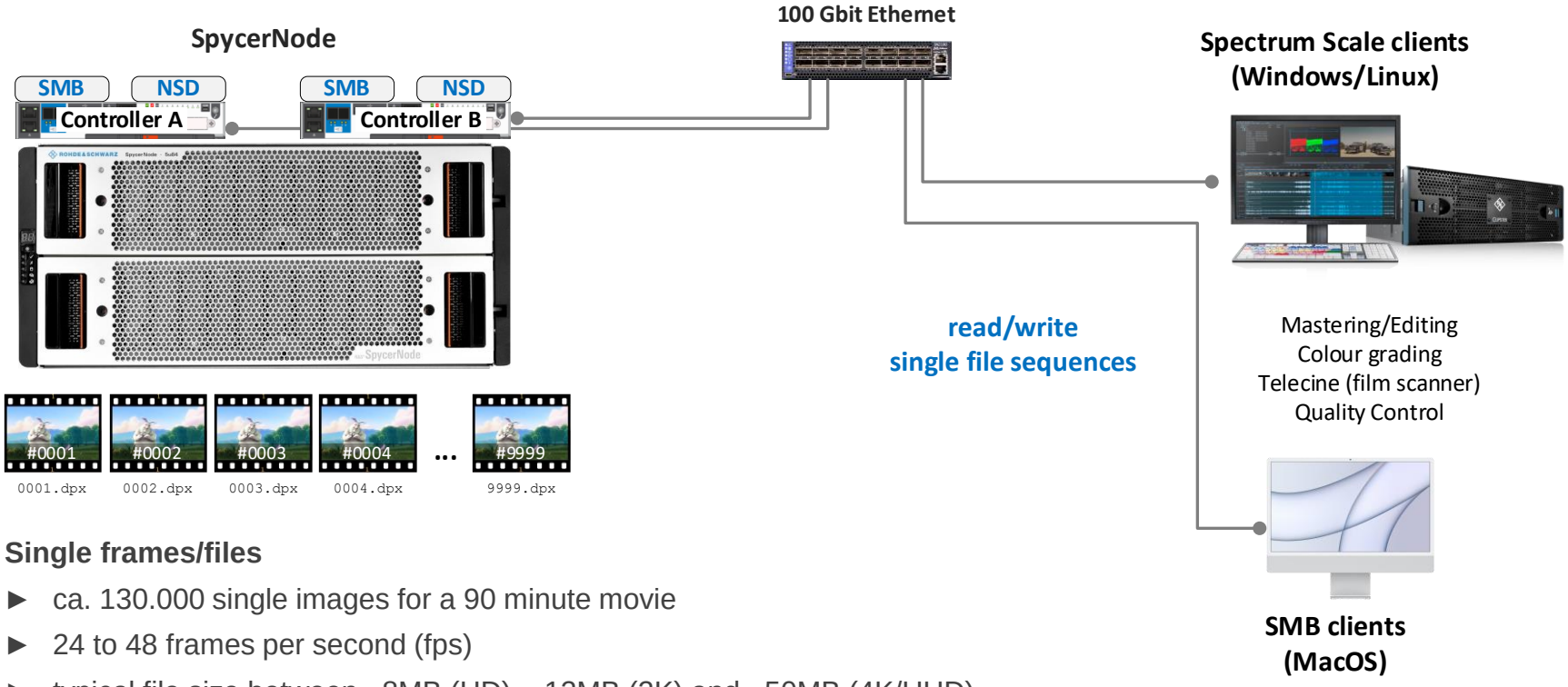
- fast reliable storage for video workstations for editing, color grading or mastering, telecines (film scanners), Windows and MacOS, rarely Linux
- demand for high single client storage transfers (up to 5GB/s)
- working with uncompressed image sequences (lots of single files)
  file sizes: between ~8MB (HD) and ~50MB (4K/UHD)
  file formats: e.g. dpx, exr, tiff
  frame rates per second: 24/48 fps (cinema), 25/30 fps (TV)

Challenges:

- many customers originally used FibreChannel infrastructures struggle with transition to ethernet networks
- price sensitive (fast affordable storage)



POST PRODUCTION STORAGE



Single frames/files

- ca. 130.000 single images for a 90 minute movie
- 24 to 48 frames per second (fps)
- typical file size between ~8MB (HD), ~12MB (2K) and ~50MB (4K/UHD)

POST PRODUCTION STORAGE

Tuning and Configuration:

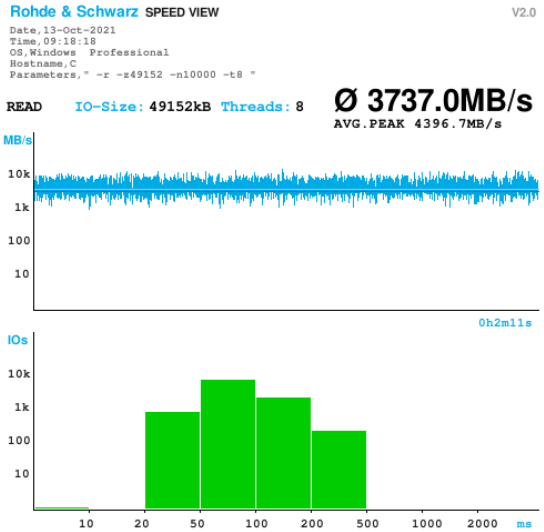
- Jumbo frames (MTU 9000) will give you 10-15% more throughput
  needs to be enabled on client and switch side
  helpful check: `ping -M do -s 8972`
- Increase RX/TX buffers for the 100GbE NICs to prevent package drops
- Set '`maxTcpConnsPerNodeConn=4`'
  (this nearly doubled the throughput to a single Linux client compared to a single TCP connection)
- Enable multithreaded IO in the client applications (transfer multiple IOS/files in parallel)
- With RDMA/RoCE we have seen up to 8GB/s for a single Linux client
  (largely reduced CPU load and latency, requires lossless ethernet configuration for all the switches and clients)
- MacOS connected through CES SMB, limited to ~2.5GB/s
- configure ACL inheritance for directories to set global file permissions without AD integration
  without global ID mapping: each native windows client writes with its individual user and group id
  set ACLs using `mmputacl/mmgetacl`

POST PRODUCTION STORAGE

Benchmarking tools

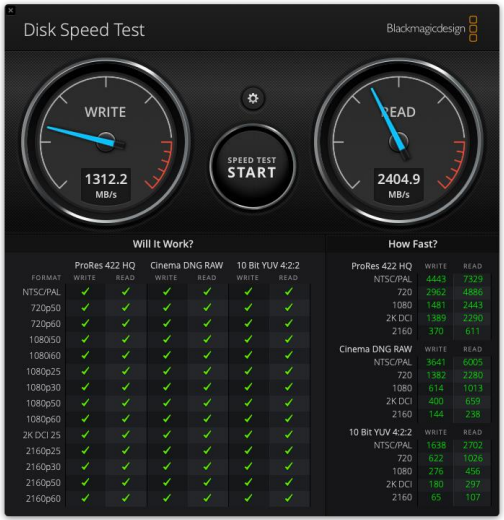
frametest (Linux/Win/MacOS)

developed originally by SGI to simulate single file video sequences



Disk Speed Test (MacOS)

Blackmagic



AJA System Test (MacOS/Win)

AJA

STUDIO INGEST AND PLAYOUT WORKFLOWS

► Studio Ingest and Playout

recording of live productions (e.g. sport events, talk shows,...)
live playout/broadcasting for TV channels

Broadcast Playout
Control room „TeleZüri“ | CH Media

Studio Ingest
Talkshow „Anne Will“ | Studio Berlin

STUDIO INGEST AND PLAYOUT STORAGE

Requirements:

- **Live video** ingest and playout for Broadcast and Studio productions
- **High Availability and full redundancy** (storage + network + video servers)
- **Seamless Failover**: No interruption of video IO transfers at any time
  → do not drop a single video frame even in case of a failure!
  → keep IO latencies below a guaranteed threshold!

Challenges:

- central storage for all media
- **Why we cannot use the built-in replication of Spectrum Scale?**
  Node failures in a Spectrum Scale cluster cause too long IO interruptions
  (timeouts for node failover and recovery ~20s-90s)
- **Why we cannot use caching?**
  Short term changes to next played video clips + a video clip can be shorter than the failover time

STUDIO INGEST AND PLAYOUT STORAGE

No seamless failover → Blocking of IOs → Missing video frames → Black frames on Air!

Why are Black Frames on Air a Problem? → It is very expensive!

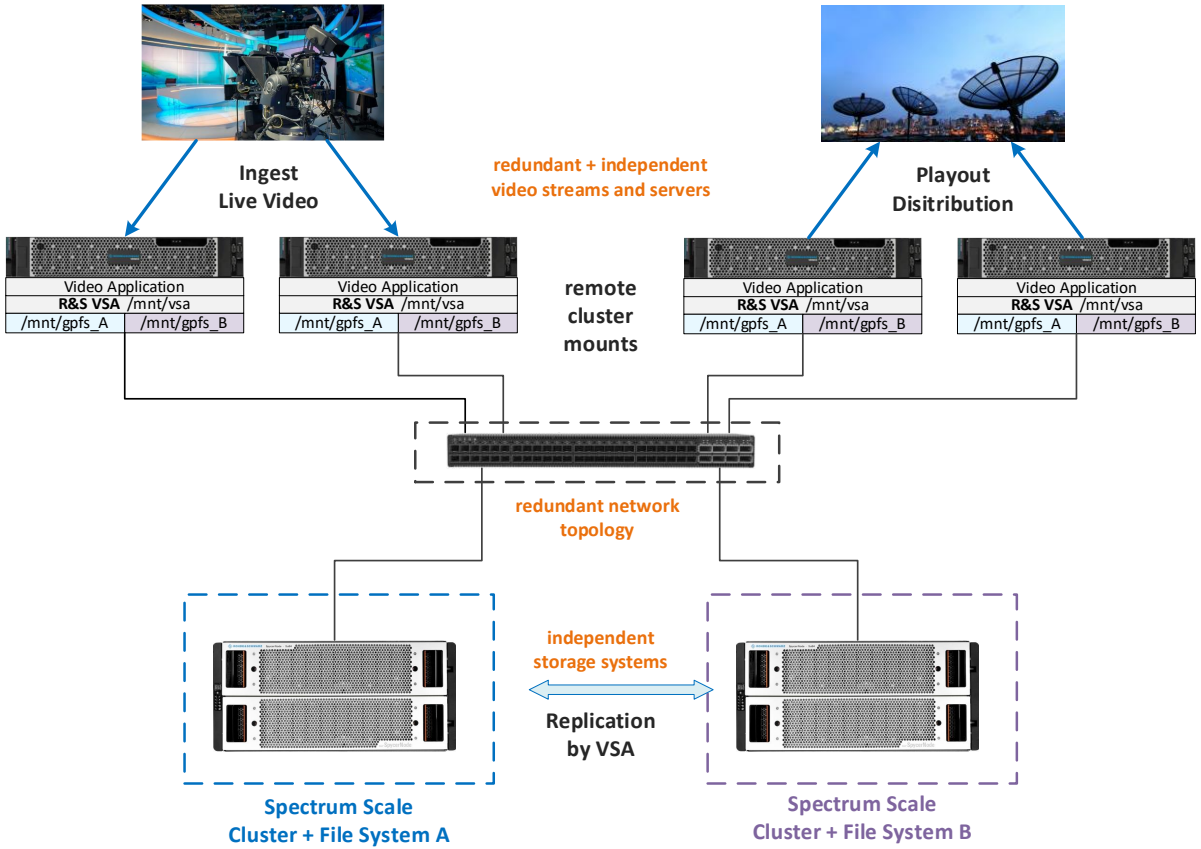
Some numbers (from 2018):

- Advertising Private Broadcast (Germany) after 8pm → 30 seconds about 60.000€ (2000€/s)
- Advertising Sunday Afternoon Formula1 Race → 30 seconds about 150.000€ (5000€/s)
- Advertising during Super-Bowl Final (USA) → 30 seconds 5.000.000 US$ (166667$/s)
- Costs German Crime Movie (Tatort) → 17.000€/min

R&S Solution:

- Separated storage clusters
  → no interference between storage sides
  → external mirroring with **R&S VSA** software

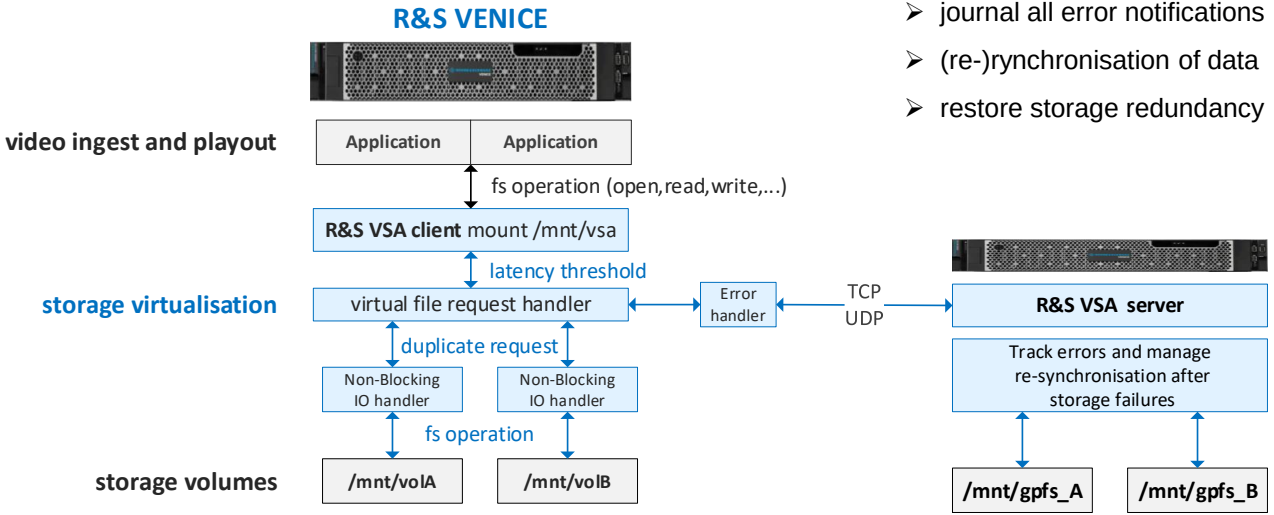
STUDIO INGEST AND PLAYOUT STORAGE



R&S VSA (VIRTUAL STORAGE ACCESS) FOR SEAMLESS FAILOVER

R&S VSA client

- duplicate all file system operations to both storage volumes
- **Non-Blocking architecture** Guarantee specified max. IO latencies (<1s) even in case of a storage failure or performance degradation
- return read data from fastest storage side
- acknowledge written data after a specific threshold time if successfully written to at least one storage volume
- asynchronous error handling send notifications about errors to external server instance, storage inconsistencies are forwarded to all clients



R&S VSA server

- redundant (active/passive)
- journal all error notifications
- (re-)rynchronisation of data
- restore storage redundancy

BROADCAST EDITING WORKFLOWS

► Broadcast Editing

creation and editing of editorial content (video clips, graphics, audio, text,...)

Editing
Media Composer | AVID

Graphics
CREATE | Rohde & Schwarz PixelPower

BROADCAST EDITING STORAGE

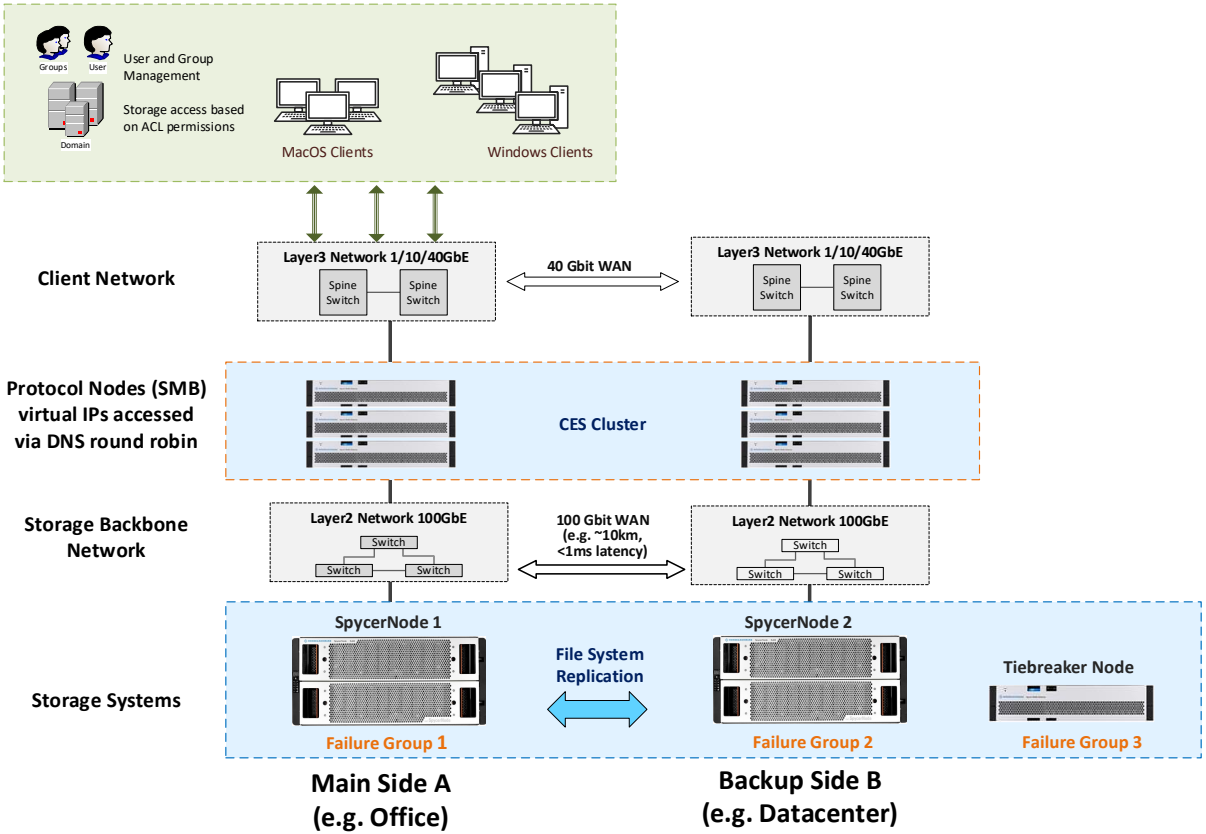
Requirements:

- 24/7 operation, no downtime, full redundancy, only small maintenance windows
- parallel editing of compressed media files for e.g. news productions
- most clients connect through SMB protocol nodes

Challenges:

- Stretched cluster with replication over two locations
- integration into large AD domain and user management
- complex network topologies (multiple sides, cascaded switches)
- maintenance and software updates during live operation

BROADCAST EDITING STORAGE



BROADCAST EDITING STORAGE

Tuning and Configuration:

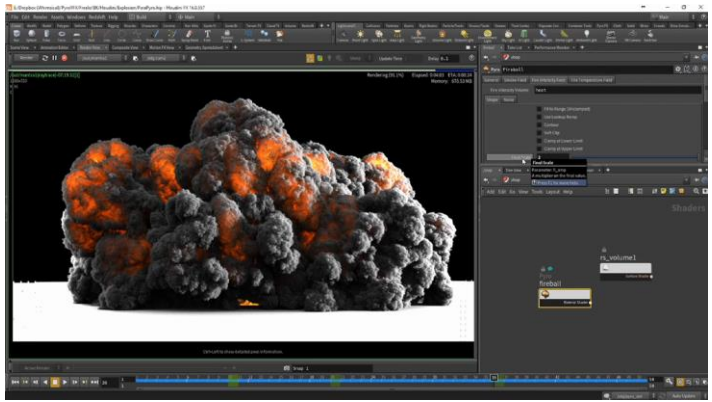
- Round Robin DNS for virtual CES IPs
  one single global name to access all SMB shares, also provides load balancing
  needs to be configured in the local AD/DNS server
- software rollout planning for minimal interruptions of production workflows
  Recovery group failover takes ~80 seconds during which time al IOs to the file system will be blocked
  cttdb database update requires to stop all CES services
- set `readReplicaPolicy=local`
  to prefer reading data from the local side to the clients
- enable ‘fruit’ modul for MacOS SMB clients
  support for alternate data streams (ADS) used by MacOS, also improves the browsing speed in the Finder

VFX WORKFLOWS

► VFX Rendering

creation of visual effects (fire, smoke, water,...)
rendering of 3D animations

3D animation
movie „Sintel“ | Blender Foundation



Visual Effects
Houdini | SideFX

VFX STORAGE

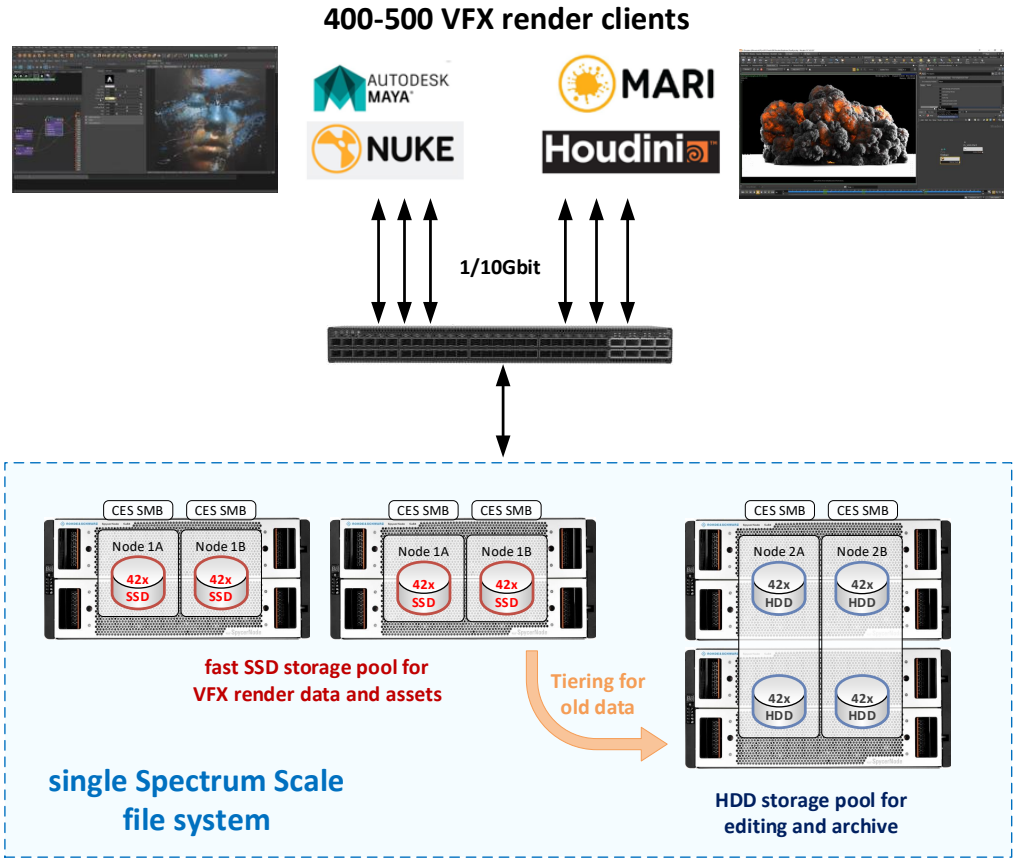
Requirements:

- Fast SSD storage for the creation (rendering) of visual effects (VFX)
- Move data to a slower (cheaper) HDD storage for archiving
- Parallel read/write access of multiple render clients to the same data

Challenges:

- 400-500 render clients connected through SMB
- Many IOPS and open files
- NSD servers also used as CES/SMB protocol nodes to reduce costs
- Automated tiering policies based on last modify date of files

VFX STORAGE


400-500 VFX render clients
AUTODESK MAYA
MARI
NUKE
Houdini
1/10Gbit
CES SMB
CES SMB
Node 1A
Node 1B
42x SSD
42x SSD
CES SMB
CES SMB
Node 1A
Node 1B
42x SSD
42x SSD
fast SSD storage pool for VFX render data and assets
Tiering for old data
CES SMB
CES SMB
Node 2A
Node 2B
42x HDD
42x HDD
42x HDD
42x HDD
HDD storage pool for editing and archive
single Spectrum Scale file system

VFX STORAGE

Tuning and Configuration:

- We saw a lot of transfer overhead due to discarded caches when multiple CES node access the same file
  → Disable read prefetching helps `(prefetchAggressivenessRead=0)`

- High load and a lot of updates in the cluster wide samba (ctdb) database resulting in slow SMB performance (over 26000 file locks and multiple locks per file)
  → Disable cluster wide file locking

Samba:

```
fileid:algorithm = hostname
gpfs:sharemodes = no
gpfs:leases = no
```

Spectrum Scale

```
locking = no
strict locking = no
```

THANK YOU

- Feel free to ask questions ☺