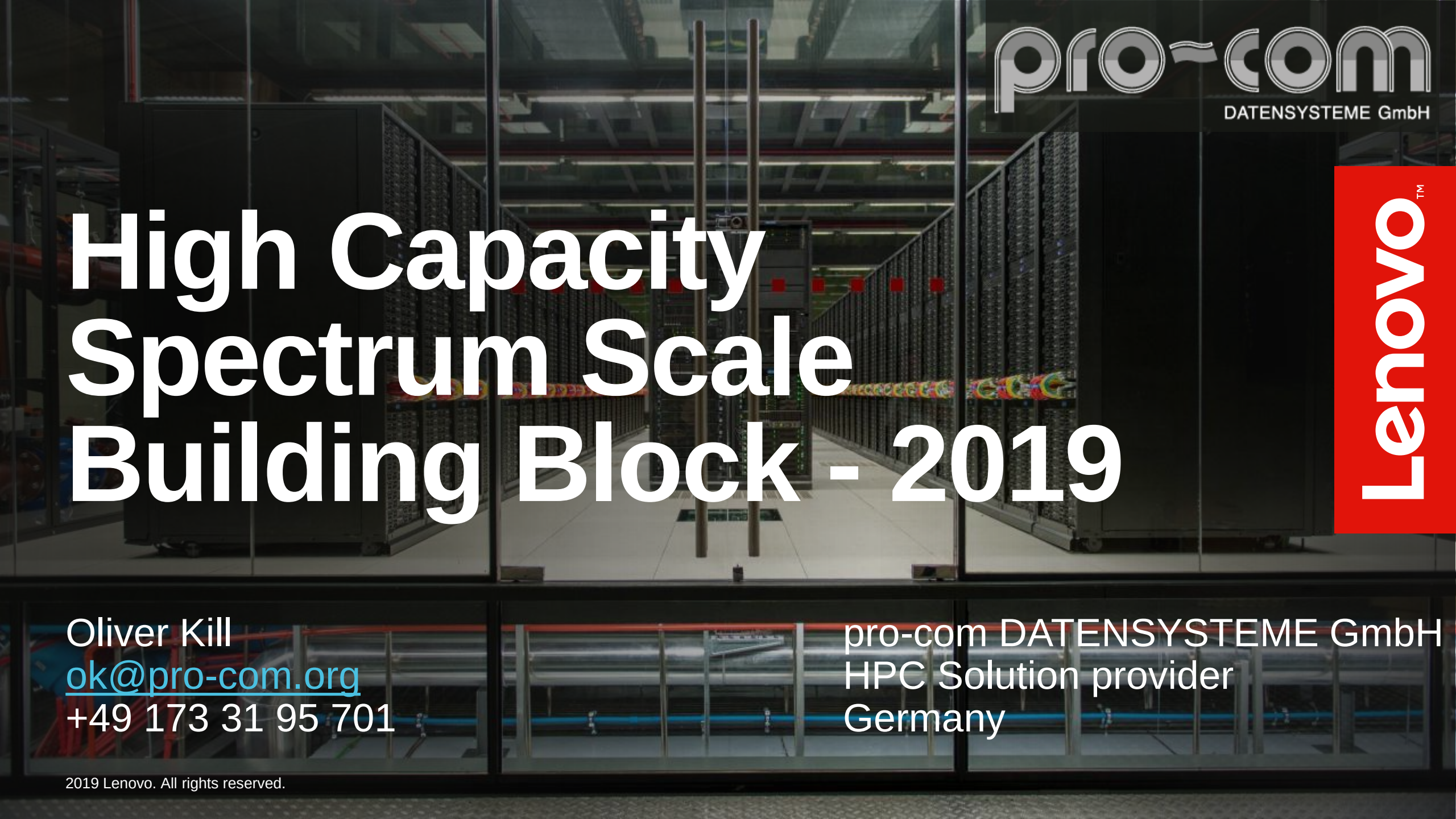


High Capacity Spectrum Scale Building Block - 2019

Oliver Kill
ok@pro-com.org
+49 173 31 95 701

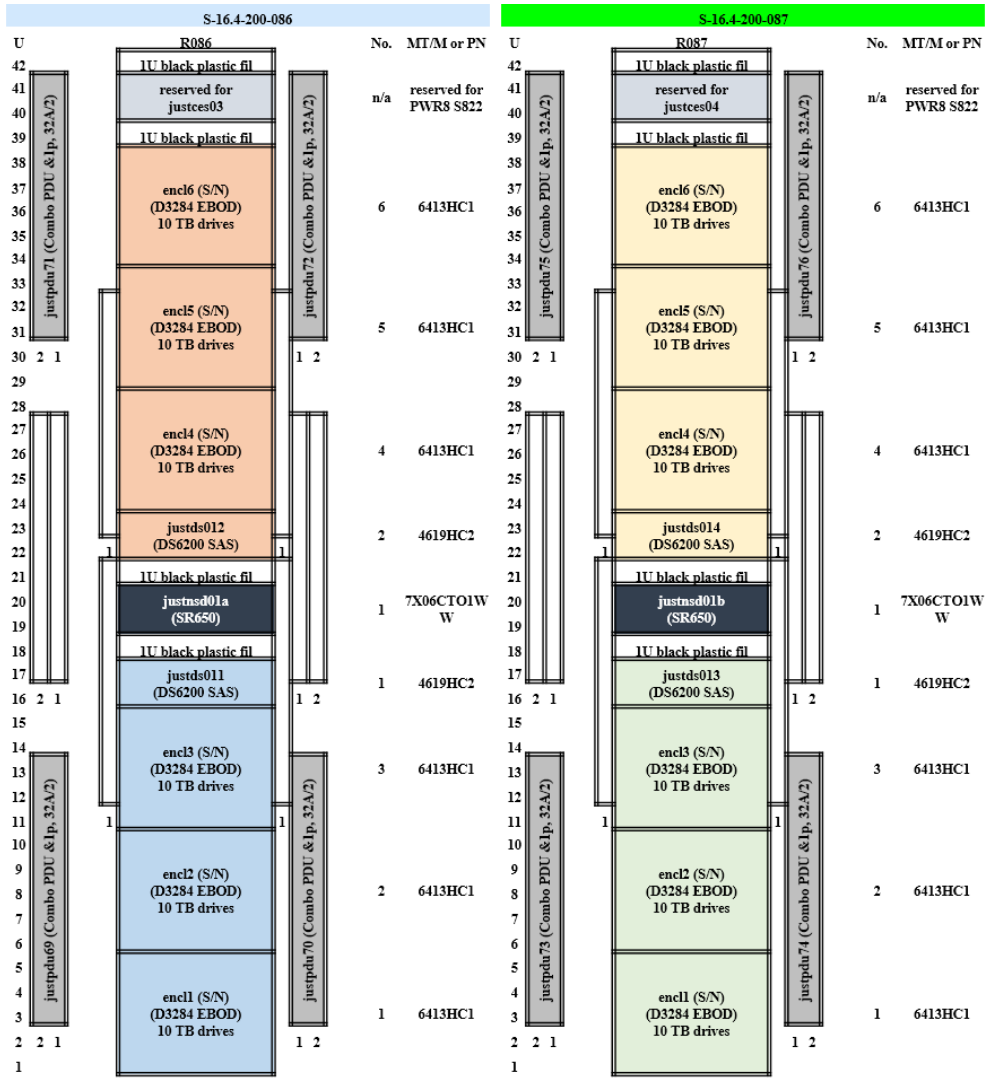
pro-com DATENSYSTEME GmbH
HPC Solution provider
Germany

Startpoint

- Recap last GUG
- New Block Design
- Systems
- Performance Numbers
- Best Practices (so far)

Last Time on GUG@ISC-2018

- Introduction Lenovo High Capacity Spectrum Scale Storage Block…
  - 2 Server
  - 4 DS Storage Units
  - 1008 Hardrives
  - 2 Racks
- Based on a Costumer RFP
- Solid Performance
  - 15 GB/s Write / 20 GB/s Read per Block
  - 3,75 GB/s Write / 5 GB/s Read per Controller



pro-com DATENSYSTEME GmbH

New Systems 2019

Update 2019

- Lenovo and Netapp announced Global Strategic Alliance

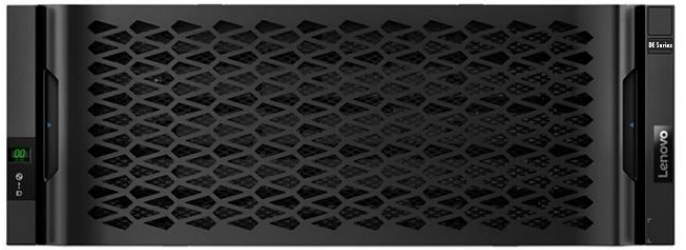
- Update of our Configuration

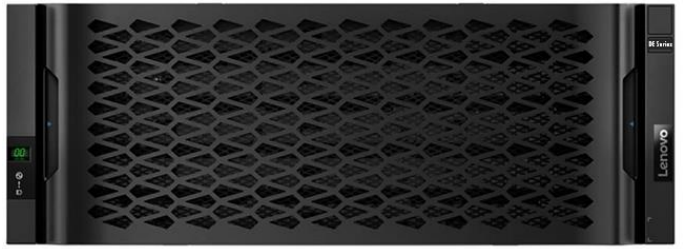
- Targets
  - Increase Density
  - Increase Performance
  - Same Price per TB Range

DE6000-H / DE600s

DE6000-H

- Dual Controller
- RAID 0,1,5,6 / Dynamic Disc Pool
- FSWA Support
  - Full Stripe Write Acceleration
- 16-64 GB Cache
- 12GB SAS / 32 GB FC / 25 GbE iSCSI
- Up to 480 Drives
- 2U and 4U Versions
  - 24x 2.5“ or 60x 3.5“





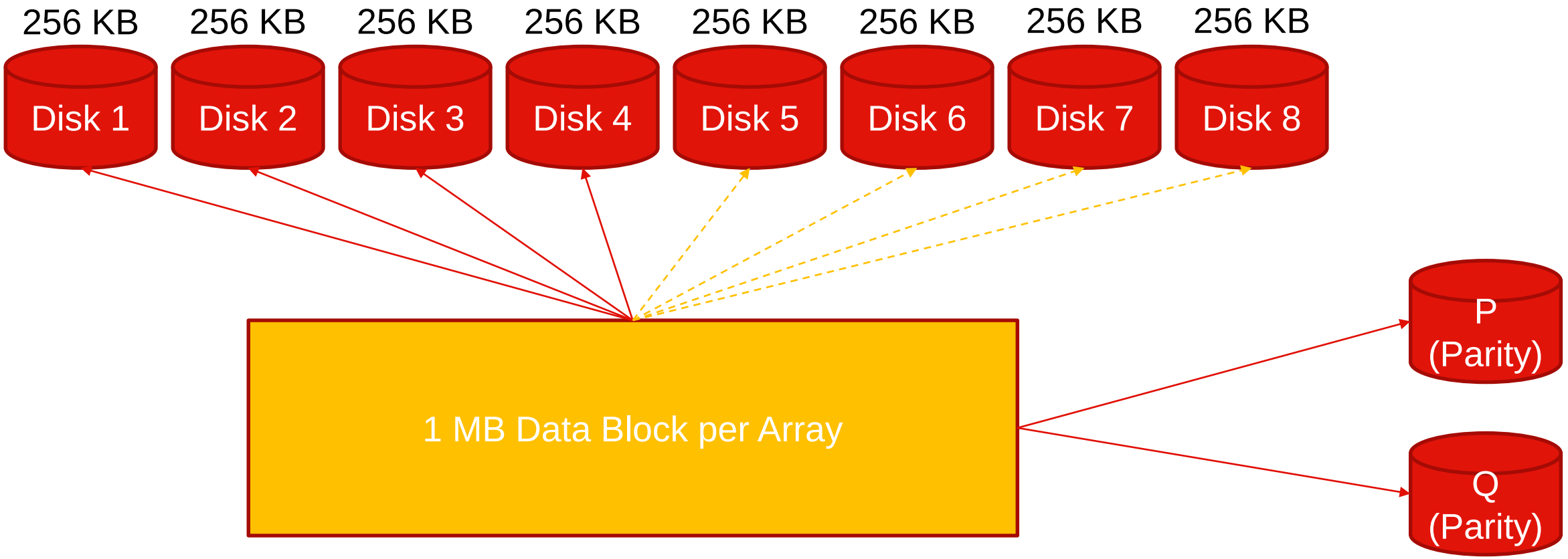
DE600s Expansion Chassis

- Dual ESM
- 60x 3.5“

DE6000-H – FSWA

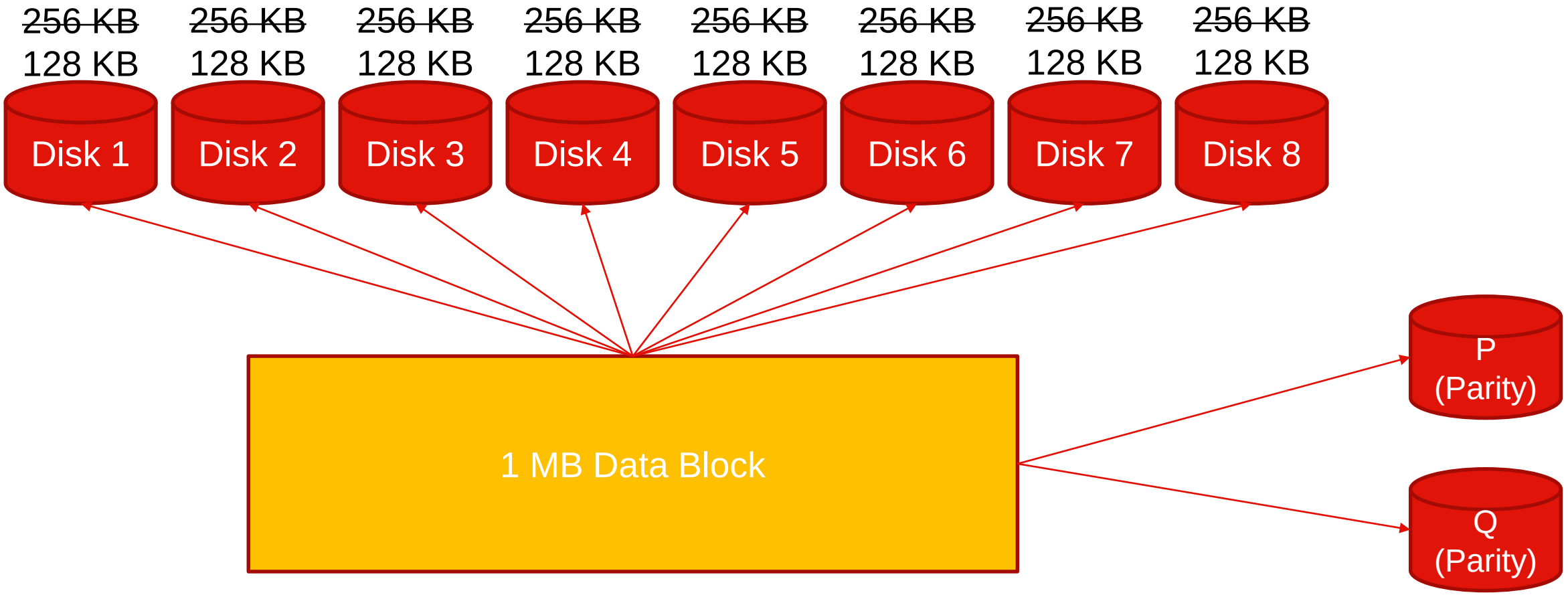
- Full Stripe Write Acceleration (FSWA)
- Only with classical RAID6
- Massive performance increase
- Depending on Workload

FSWA



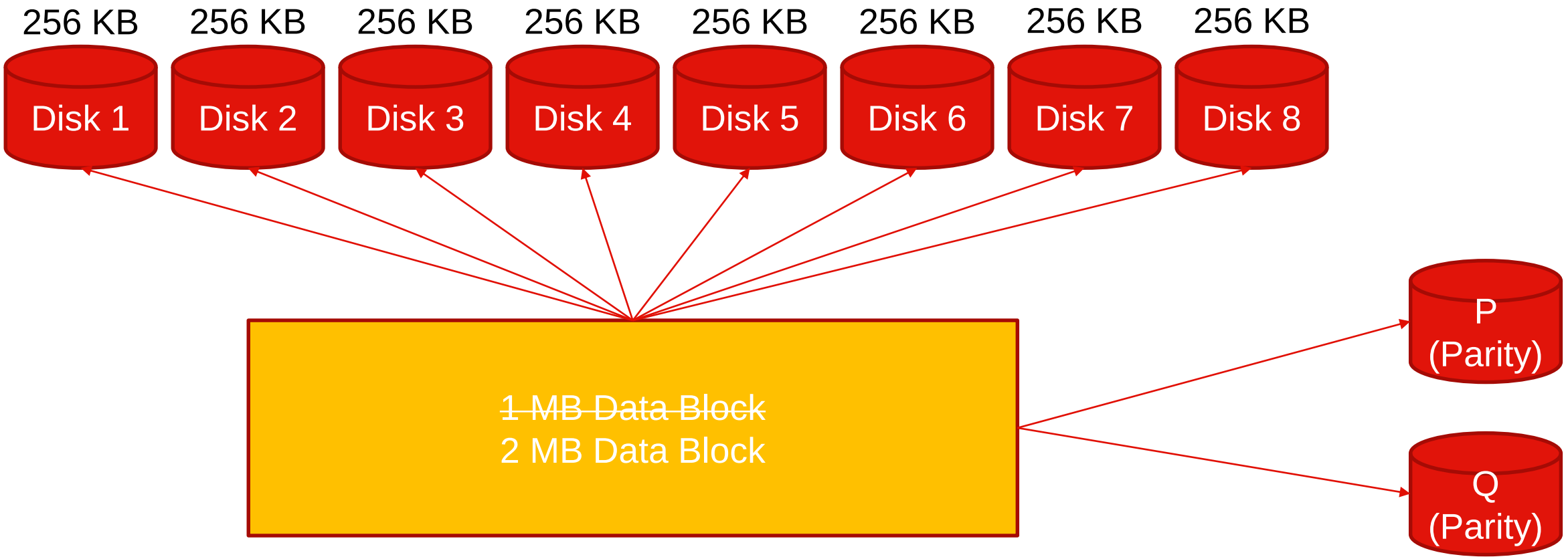
Valid values for the segment size are: 8, 16, 32, 64, 128, 256 and 512 KB

FSWA



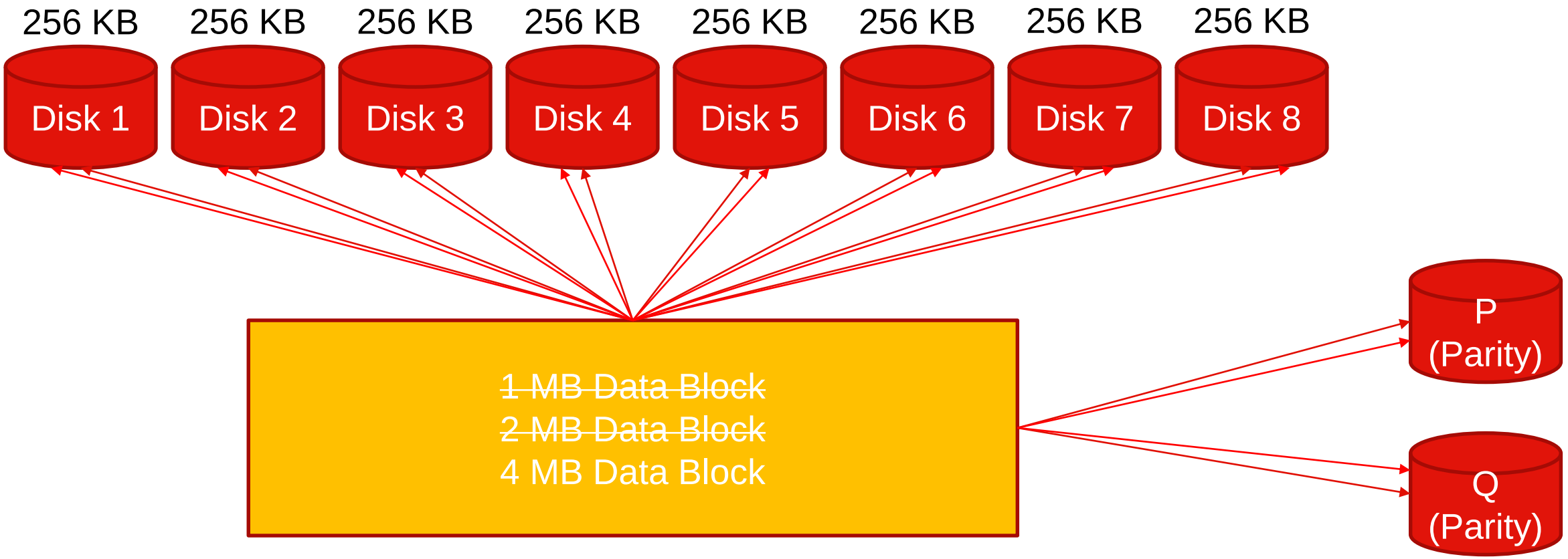
Valid values for the segment size are: 8, 16, 32, 64, 128, 256 and 512 KB

FSWA



Valid values for the segment size are: 8, 16, 32, 64, 128, 256 and 512 KB

FSWA

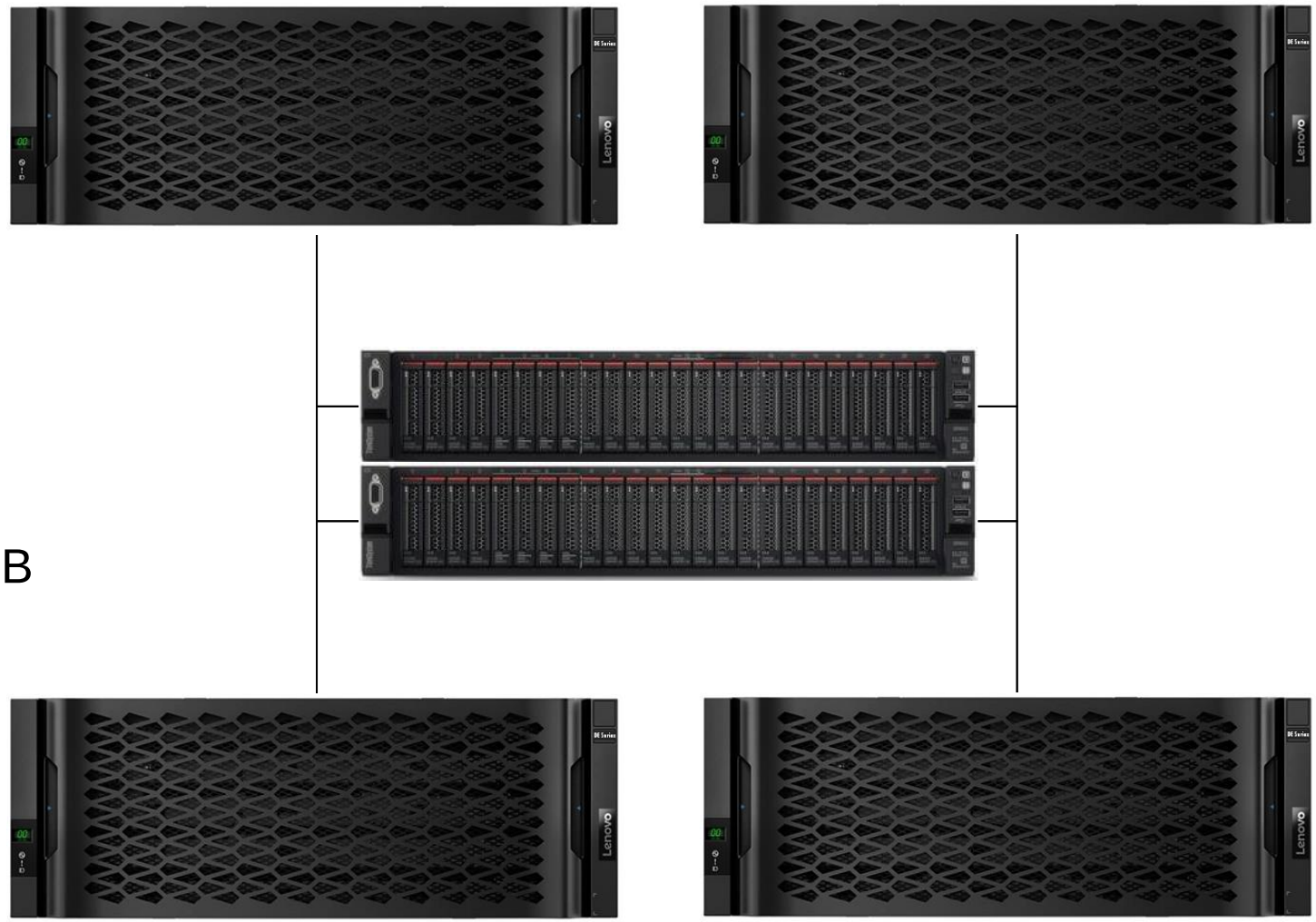


Valid values for the segment size are: 8, 16, 32, 64, 128, 256 and 512 KB

New Block 2019

Setup

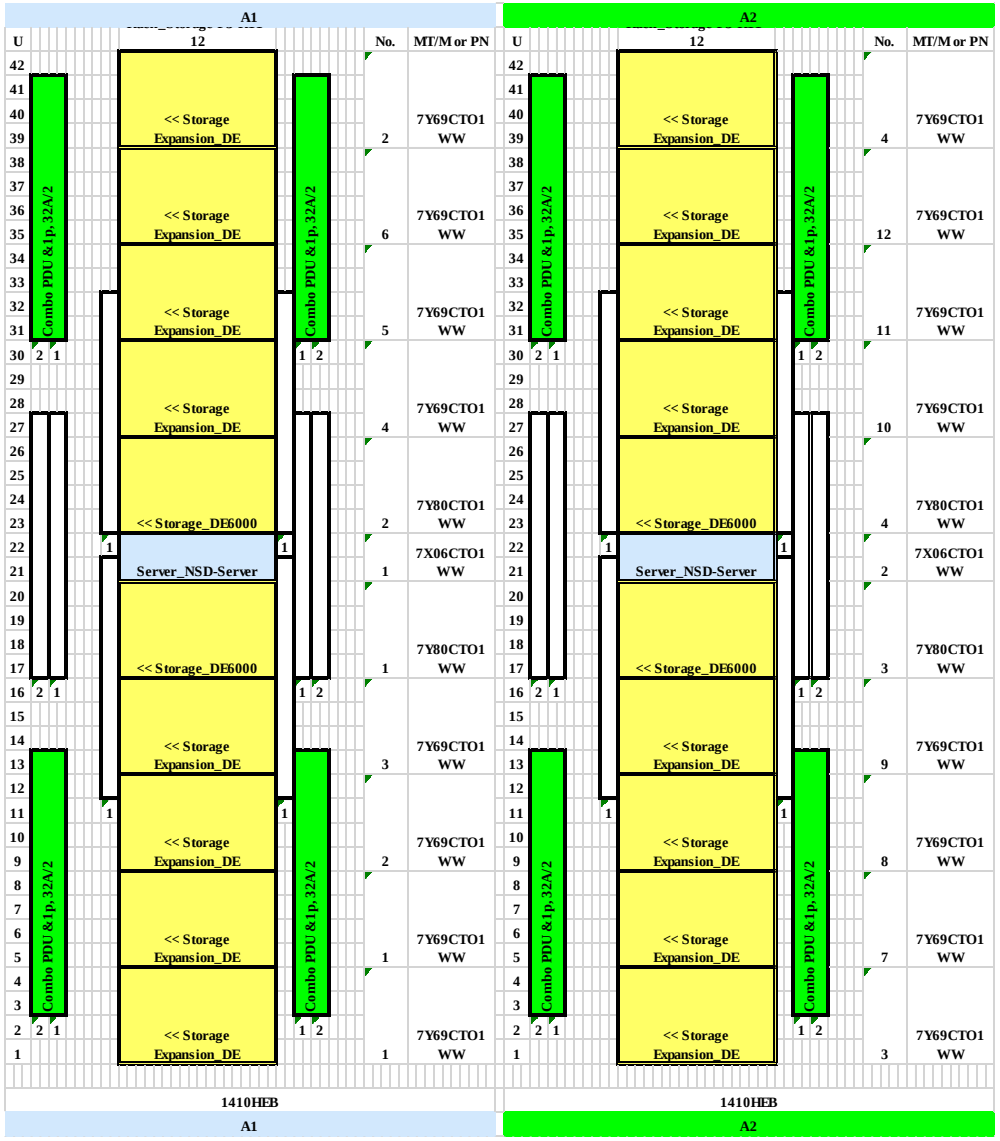
- How to Connected?
  - SAS direct attached
- How to Scale
  - Up to 300 Drives per Disk System
  - 1200 Drives per Block
- How to Balance
  - 16 SAS Ports 12GB / 4x 100 GbE / IB Ports
  - 32 PCI-E Lanes vs. 32 PCI-E Lanes



1200 drive configuration in two racks

- 2x NSD – Server
  - Lenovo ThinkSystem SR650
  - 2x Intel Xeon 6242 – 16 Core 2.8 GHz
  - 12x 32 GByte Memory
  - 2x High Speed Adapter (IB / OPA / 100 GbE)
  - 4x 12 GBIT Quad Port SAS Adapter

- 4x Back End Storage
  - Lenovo ThinkSystem DE6000-4U
  - Dual Controller – 8x 12 GB SAS Ports
  - RAID 0,1, 5, 6, 10 – DDP Technology
  - Each 300 Drives via 4x DE600s Expansion Unit



pro-com DATENSYSTEME GmbH

Benchmarks

New vs. Old

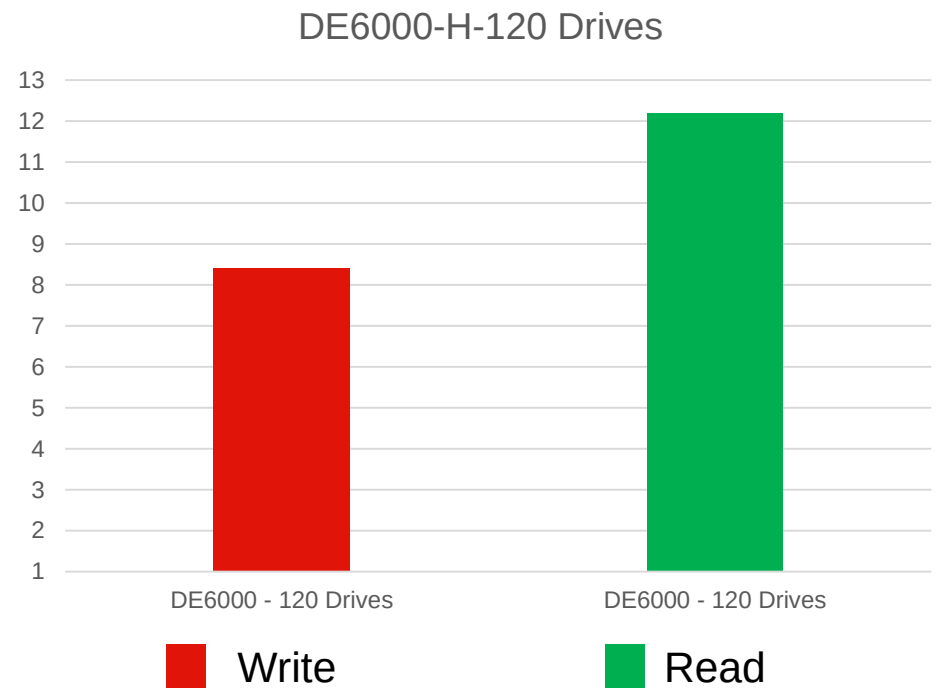
Benchmarks

- Started begin of June 2019 ☺

- Focus
  - DE6000-H Scaling
    - 60,120,180… Drives
  - DE4000-H Scaling
  - SSD Performance
  - Classical Raid FSWA vs. DDP
  - Directly mapped LUN to SAS link vs. DMM round-robin
  - …
  - …
  - …

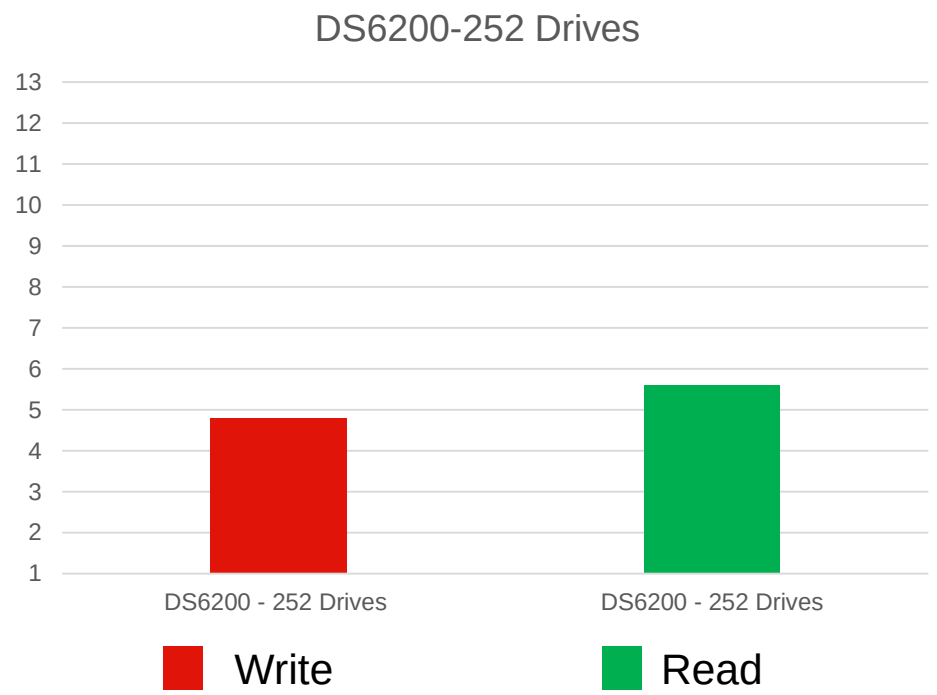
First Numbers DE6000

- DE6000
  - **120 Drives**
  - RAID 6 – (8+P+Q)
  - 12 LUNS
- NSD
  - 2x SR650
    - 2x 6134
    - EDR IB
- Clients
  - 24x Clients EDR
- Spectrum Scale
  - 5.0.3.0
  - 16M Blocksize
  - 16M IOR Transfer Size



Predecessor – System DS6200

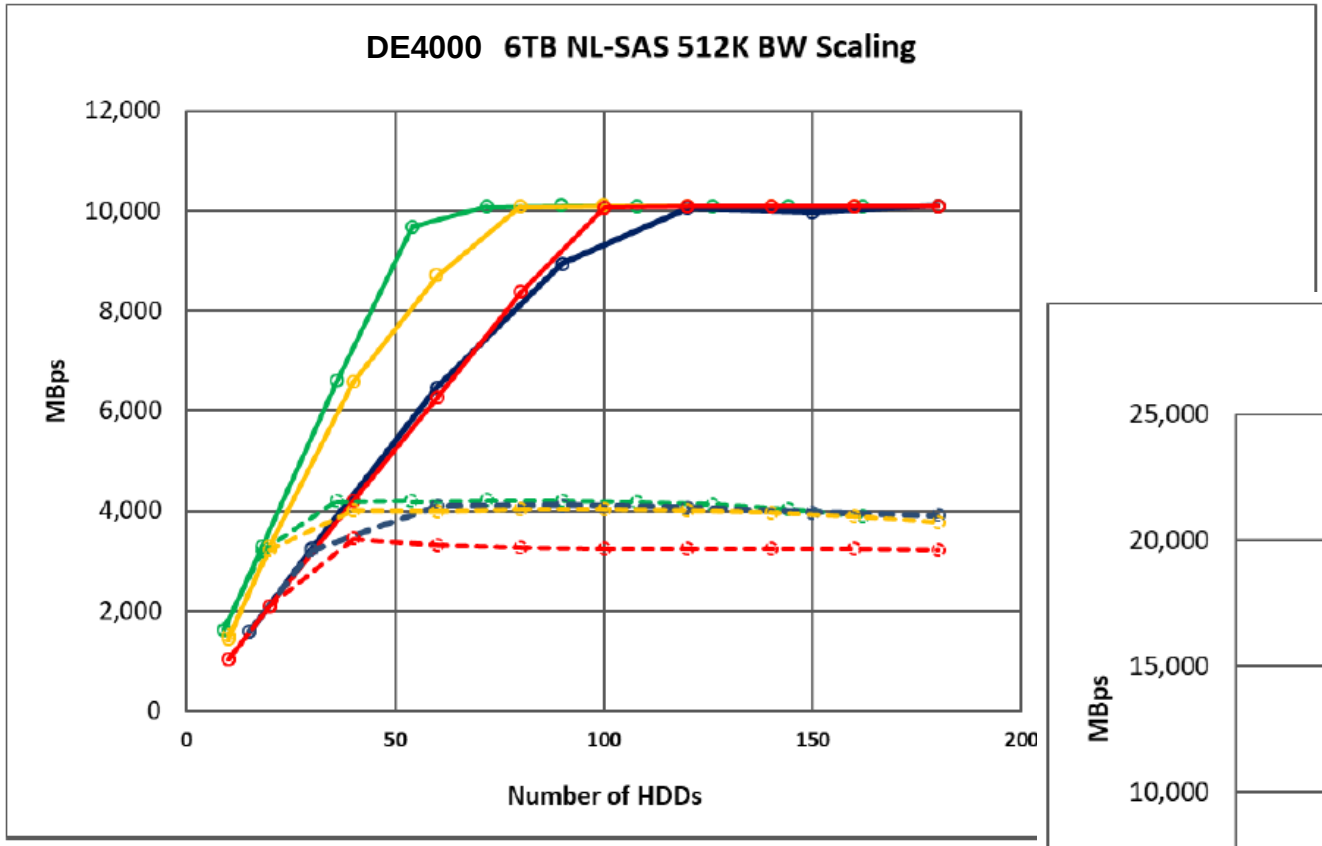
- DS6200
  - **252 Drives**
  - ADAPT RAID
  - 16 LUNS
- NSD
  - 2x SR650
    - 2x 6142
    - 100 GbE
- Clients
  - 24x Clients 100 GbE
- Spectrum Scale
  - 5.0.1.1
  - 16M Blocksize
  - 16M IOR Transfer Size

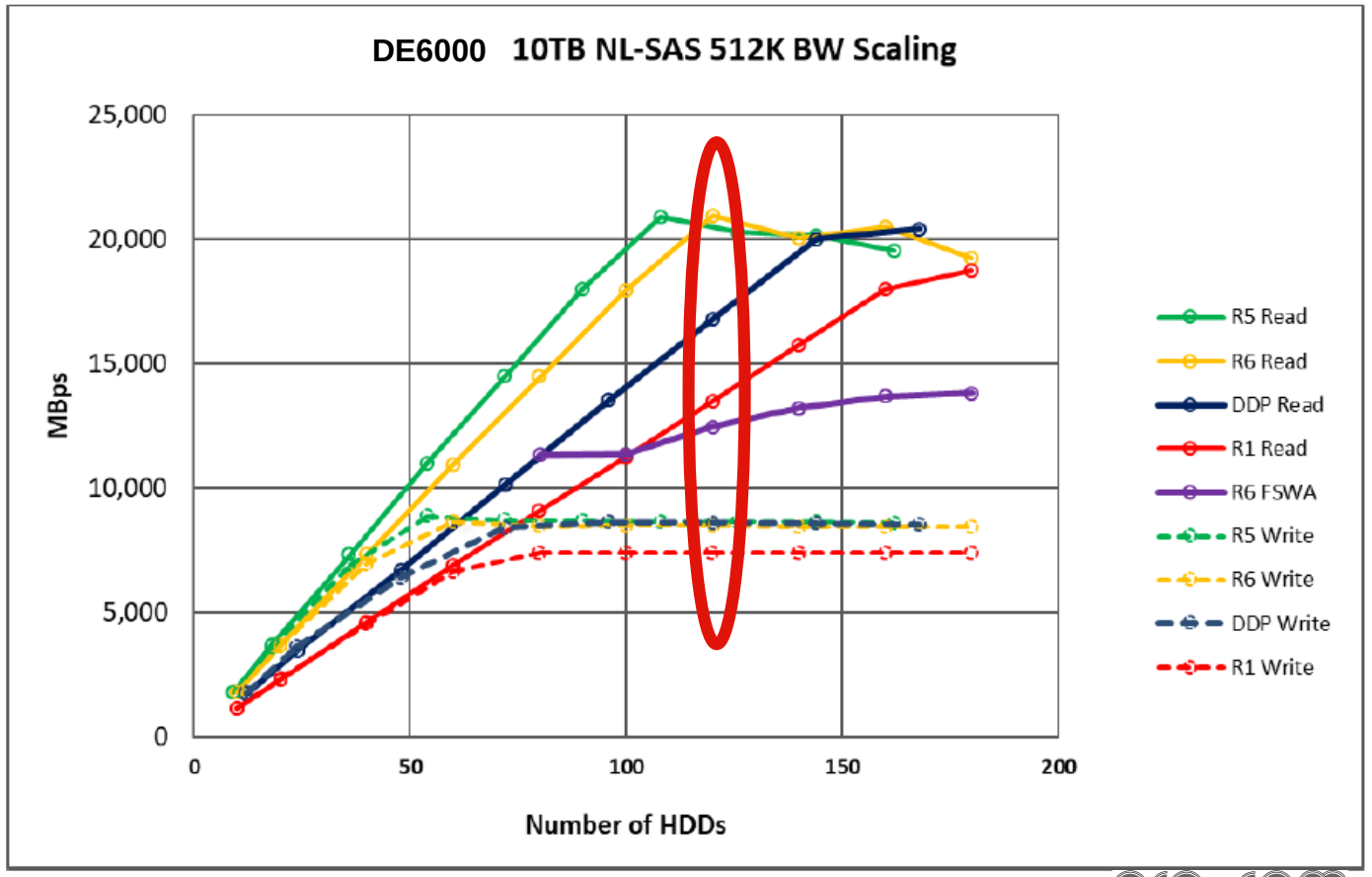


DS6200 vs. New DE6000-h

1,75 x higher Write Speed / 2x higher Read Speed

Some syntenic numbers

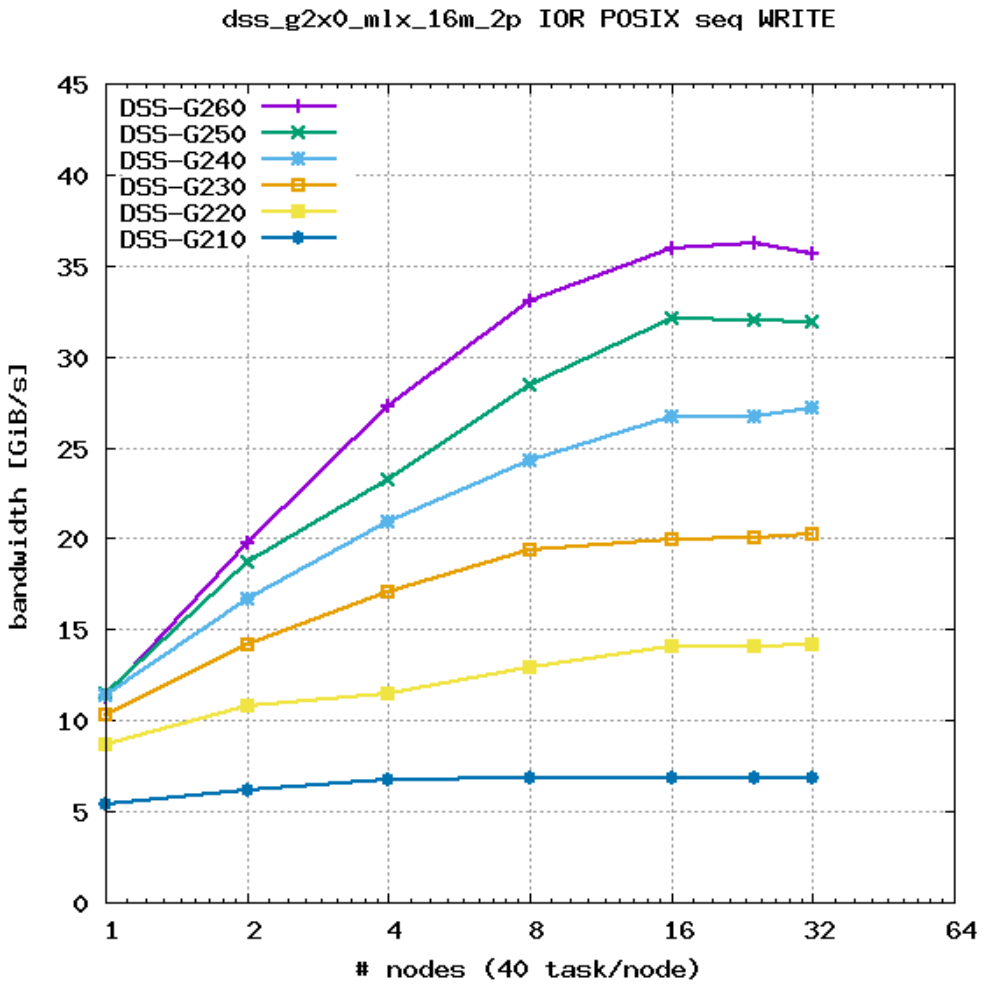


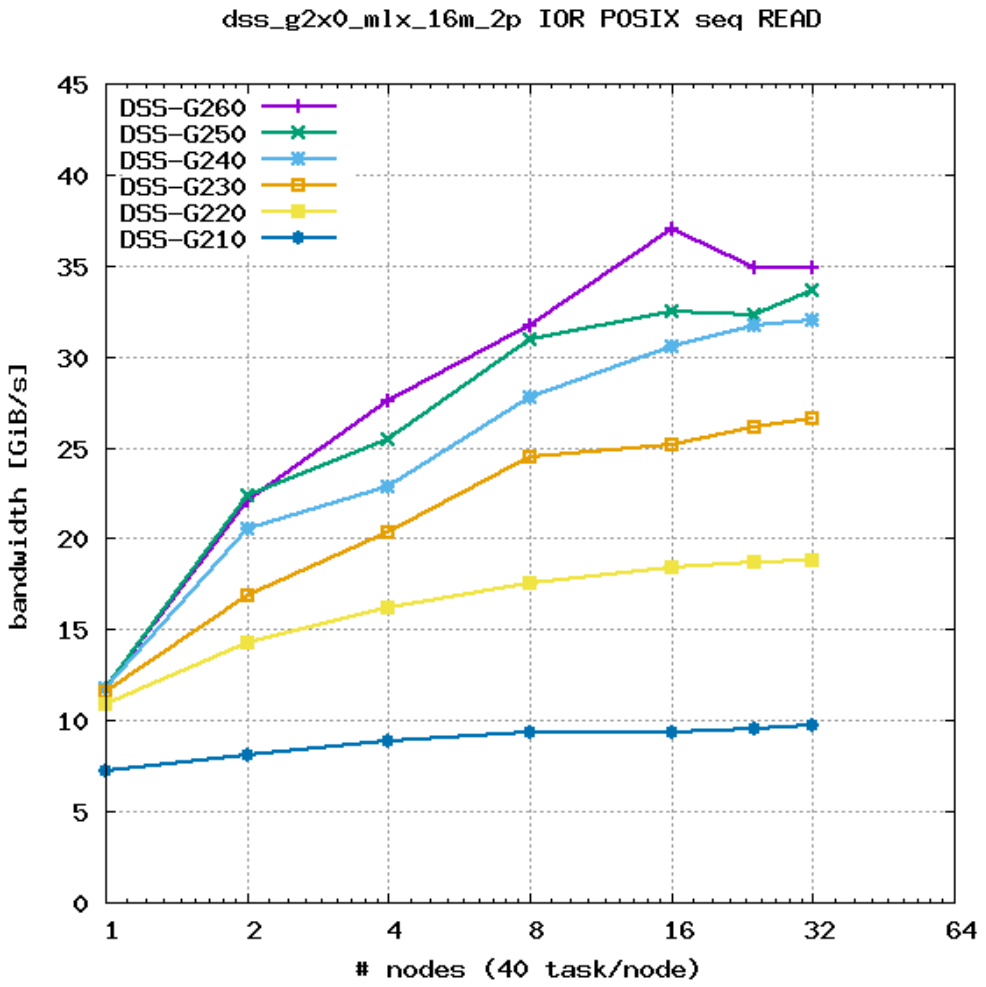


Bigger Picture

DE6000 vs. Nativ Raid

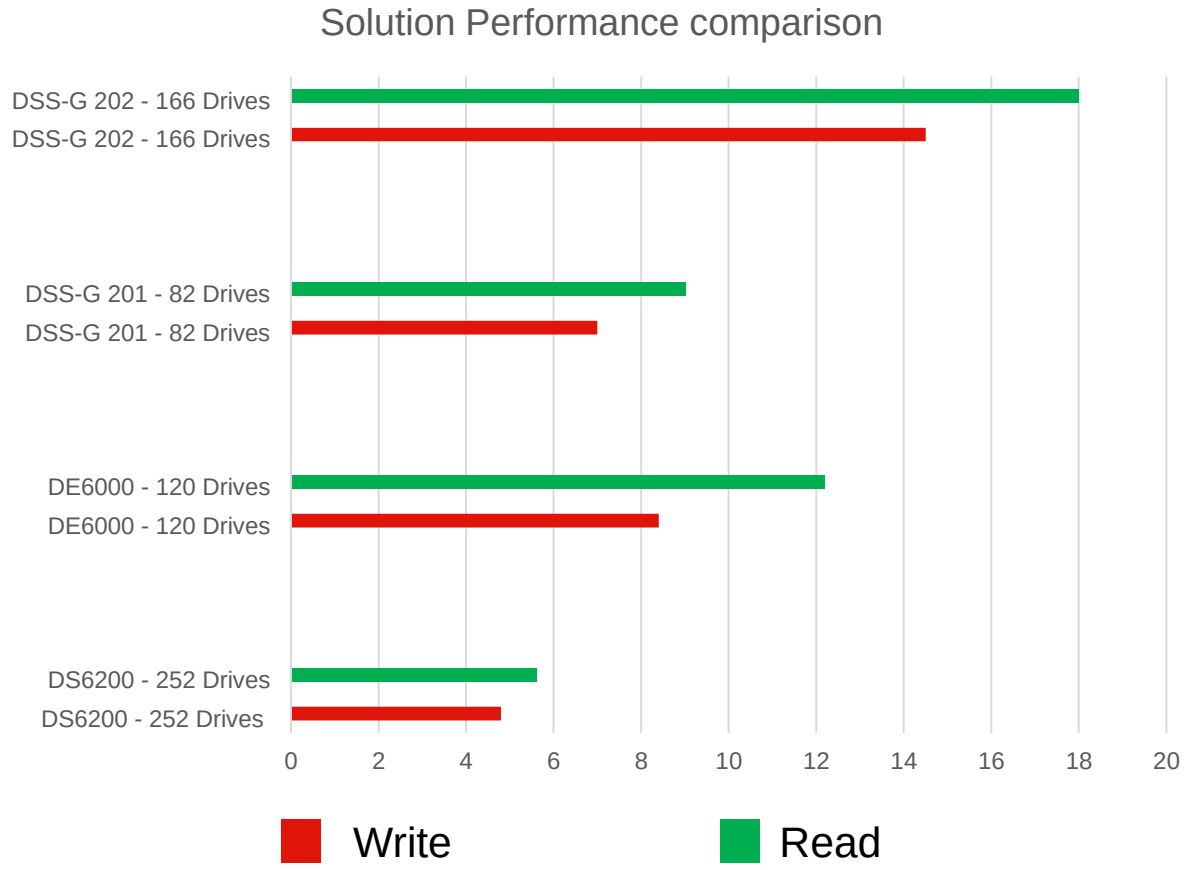
Some Nativ Raid Numbers





Hardware Controller vs. Nativ Raid

- Nativ Raid still King
  - Performance & Data Integrity
- DE6000
  - ~70 MB/s per Disk in Write
  - ~101 MB/s per Disk in Read
- Nativ Raid
  - ~85 MB/s per Disk in Write
  - ~108 MB/s per Disk in Read



16 MB Blocksize and 16 MB Transfersize

Best Practices – so far

- Planning and Implementation Guide
  - Installation, Configure and Tuning
  - Lenovo, Netapp and pro-com
  - Available late 2019
  - Lenovo Press

- Performance Level
  - RAID 6 with FSWA vs. Dynamic Disk Pools
  - Performance vs. Data Integrity

- Full Stripe Write Acceleration
  - Optimize Segment Size vs. LUNS vs. GPFS Blocksize

- LUNs QTY
  - multiple of 8 (to equally spread the LUNs across the Host Ports) - if not possible DMM round-robin

- Linux Disk I/O Settings:
  - Scheduler
    - Noop
  - reah_ahead_kb
    - 128 (Linux Default)
  - nr_request
    - 128 (Linux Default)
  - max_sectors_kb
    - 2048
  - …
  - …

thanks.

Different is better