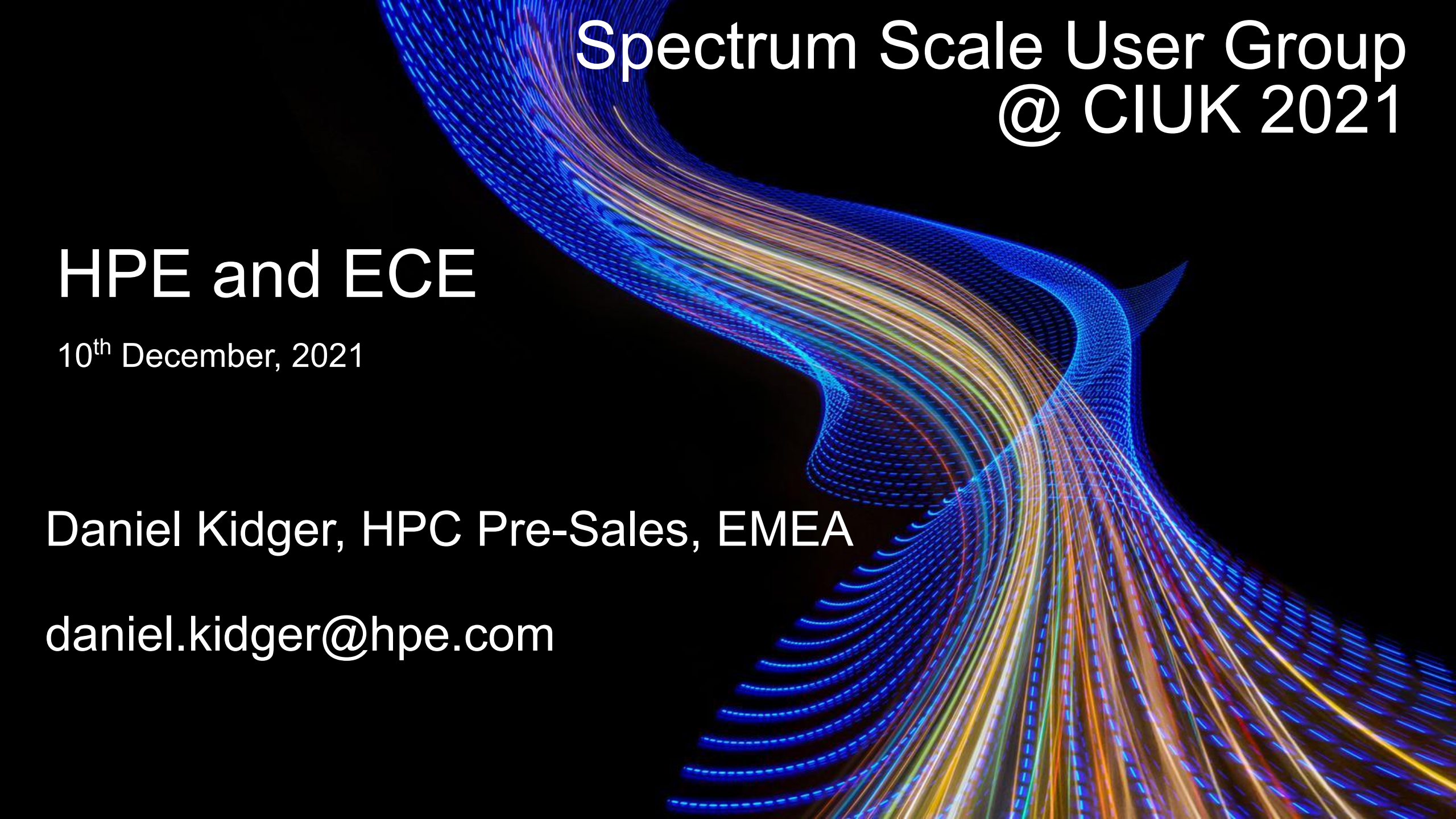
Spectrum Scale User Group @ CIUK 2021



HPE and ECE

10th December, 2021

Daniel Kidger, HPC Pre-Sales, EMEA

daniel.kidger@hpe.com

HISTORY OF SPECTRUM SCALE AT HPE / CRAY

- HPE introduced their Spectrum Scale ECE product in April this year (2021)
- But before then…

- Spectrum Scale was a choice on ClusterStor (when it was at Seagate)
  - Hence the team of people that moved with ClusterStor to Cray had much experience

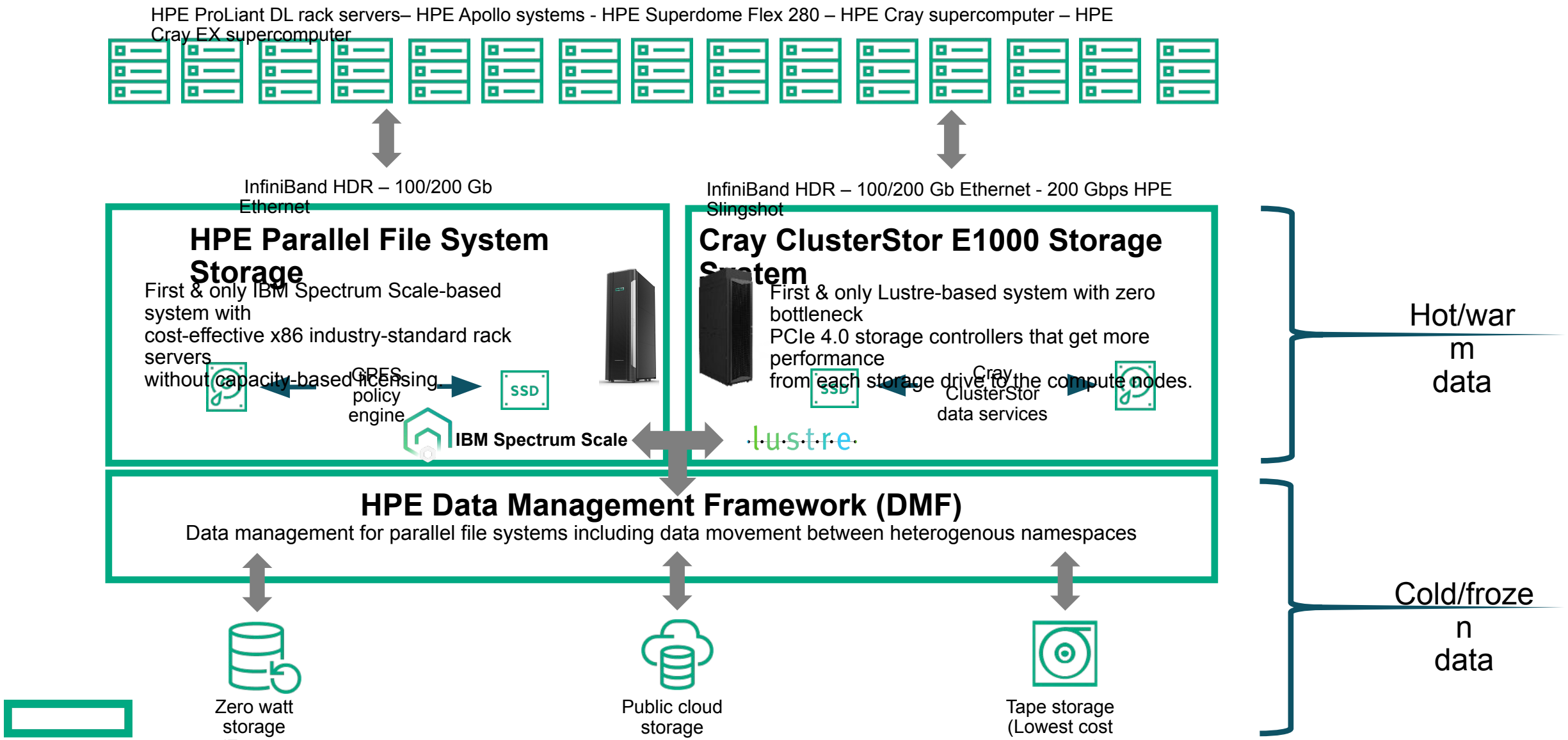
- Cray then offered IBM's ESS as part of an OEM offering

- Independently HPE have often delivered HPC clusters with DDN GridScaler.

- HPE wanted to offer HPC clusters with an in-house choice fully-supported choice of both Lustre and Spectrum Scale

- Hence PFSS was born…
  - ECE solution based on 1U disk rich servers: NVMe and HDD

HPC STORAGE AT HPE



WHY ECE ?

- HPE is not trying to compete with ESS from IBM

- Where does ECE fits in the Spectrum Scale Landscape?

- Customer says:
  - “We only buy servers buy from Dell, HP, Lenovo, SuperMicro …”
  - “This is for an analytical grid where the IT architecture team only allows x86”
  - “Only storage rich servers are acceptable, no appliances”
  - “We don’t use SAN or Fibre Channel, only Ethernet.
  - We want equipment that could be re-purposed if our needs change,
  - We need to grow storage with a small granularity of maybe 100 GB

- So why not HDFS, Ceph, Netapp, tc.?
  - Customer needs high performance / Scalability / 2 site Replication / Snapshots / tiering / ‘AFM’ etc.

ECE INTRODUCES SOME NEW TERMS

Storage Pool

NSD

Failure Group

8+2P

Recovery Group
(RG)

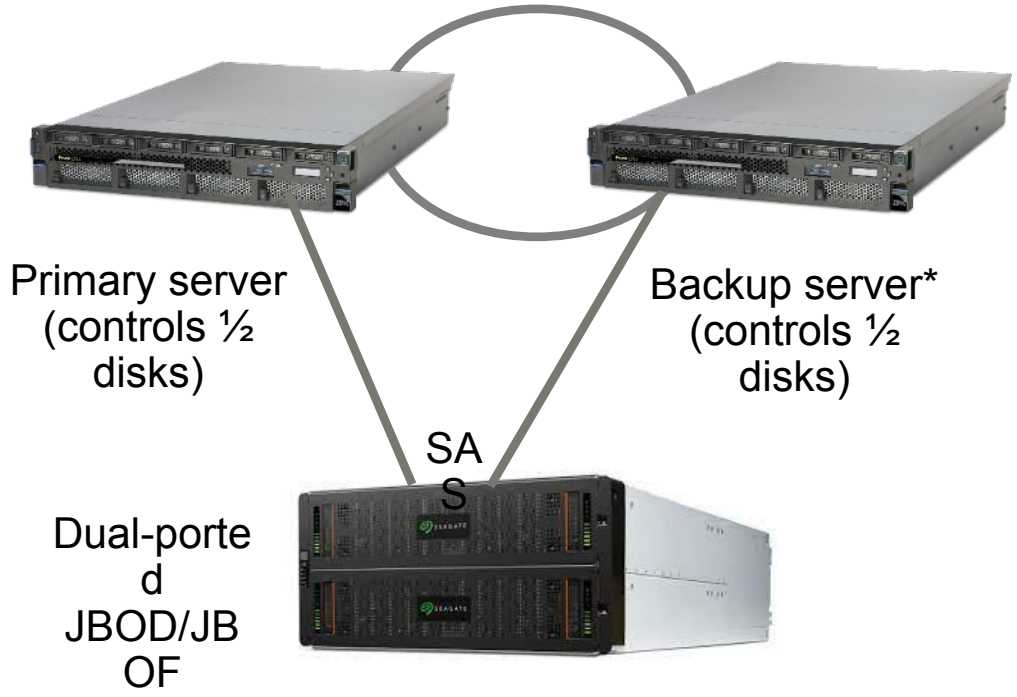
Pdisks and Vdisks

Declustered Array (DA)

Vdiskset (VS)

A NEW WAY FOR IBM SPECTRUM SCALE ON X86 RACK SERVERS

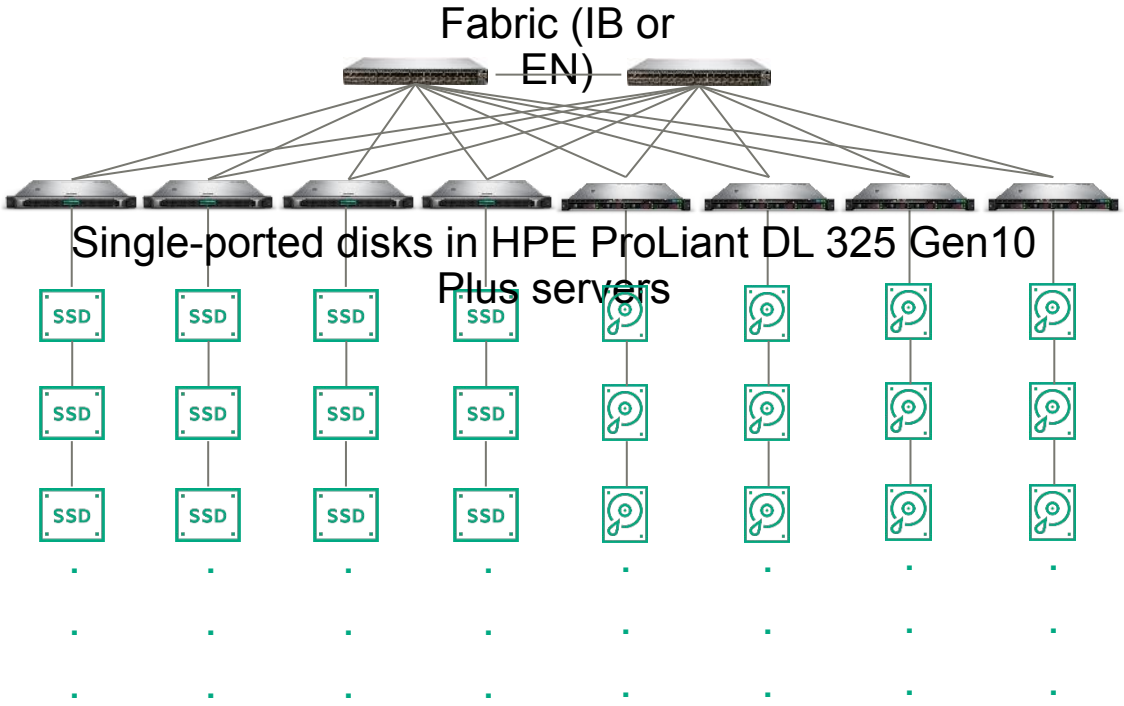
Traditional Way



- Twin-tailed disks, dual HA servers
- Parity de-clustered RAID (GNR) within the JBOD/JBOF
- Very scalable, but expensive (expensive HA HW plus SW licenses)
- Data unavailable if both servers fail

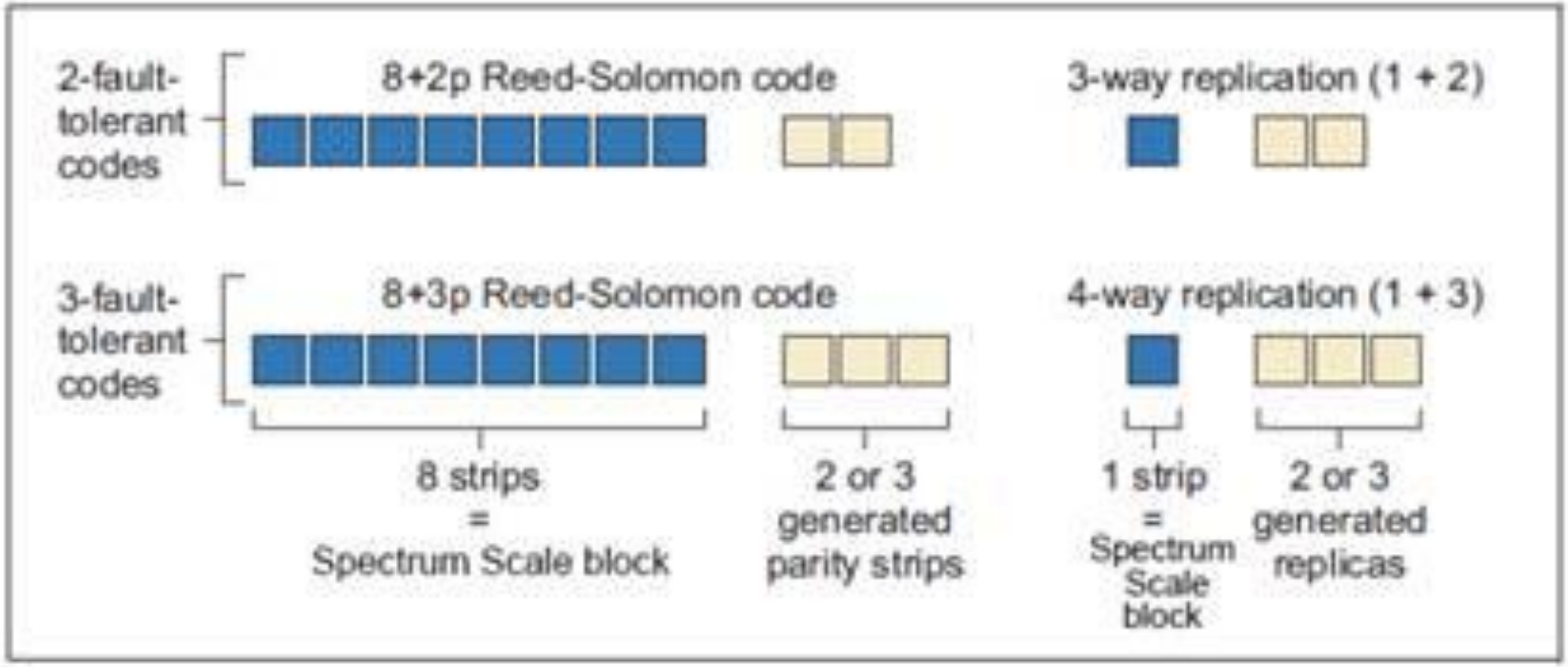
*Active-active

NEW WAY



- IBM Spectrum Scale Erasure Code Edition provides network-dispersed erasure coding, distributing data and metadata across the internal disks of a cluster of HPE ProLiant DL325 Gen10 Plus servers.
- Some scalability restrictions for the high-end, but very cost-effective
- Tolerates concurrent failure of an arbitrary pair of servers and disks

ERASURE CODING CHOICES



Erasure Codes

Protection Type	Usable capacity (approximate)
4-Way Replication	25%
3-Way Replication	33%
4+3P	57%
4+2P	67%
8+3P	73%
8+2P	80%

Fault tolerance for 'small' systems

Number of Nodes in Group	4+2P	4+3P	8+2P	8+3P
4-5	Not Recommended 1 Node	**1 Node + 1 Device**	Not Recommended 0 Nodes	Not Recommended 1 Node
6-8	**2 Nodes**	**2 Nodes*** **(Limited by RG descriptors)**	Not Recommended 1 Node	Not Recommended 1 Node + 1 Device
9	**2 Nodes**	**3 Nodes**	Not Recommended 1 Node	Not Recommended 1 Node + 1 Device
10	**2 Nodes**	**3 Nodes**	**2 Nodes**	**2 Nodes**
11+	**2 Nodes**	**3 Nodes**	**2 Nodes**	**3 Nodes**

What do you mean by “1 Node and 1 disk” ?

Consider 4+3P on 5 storage nodes

4+3P = Take 4 chunks and turn them into 7 things

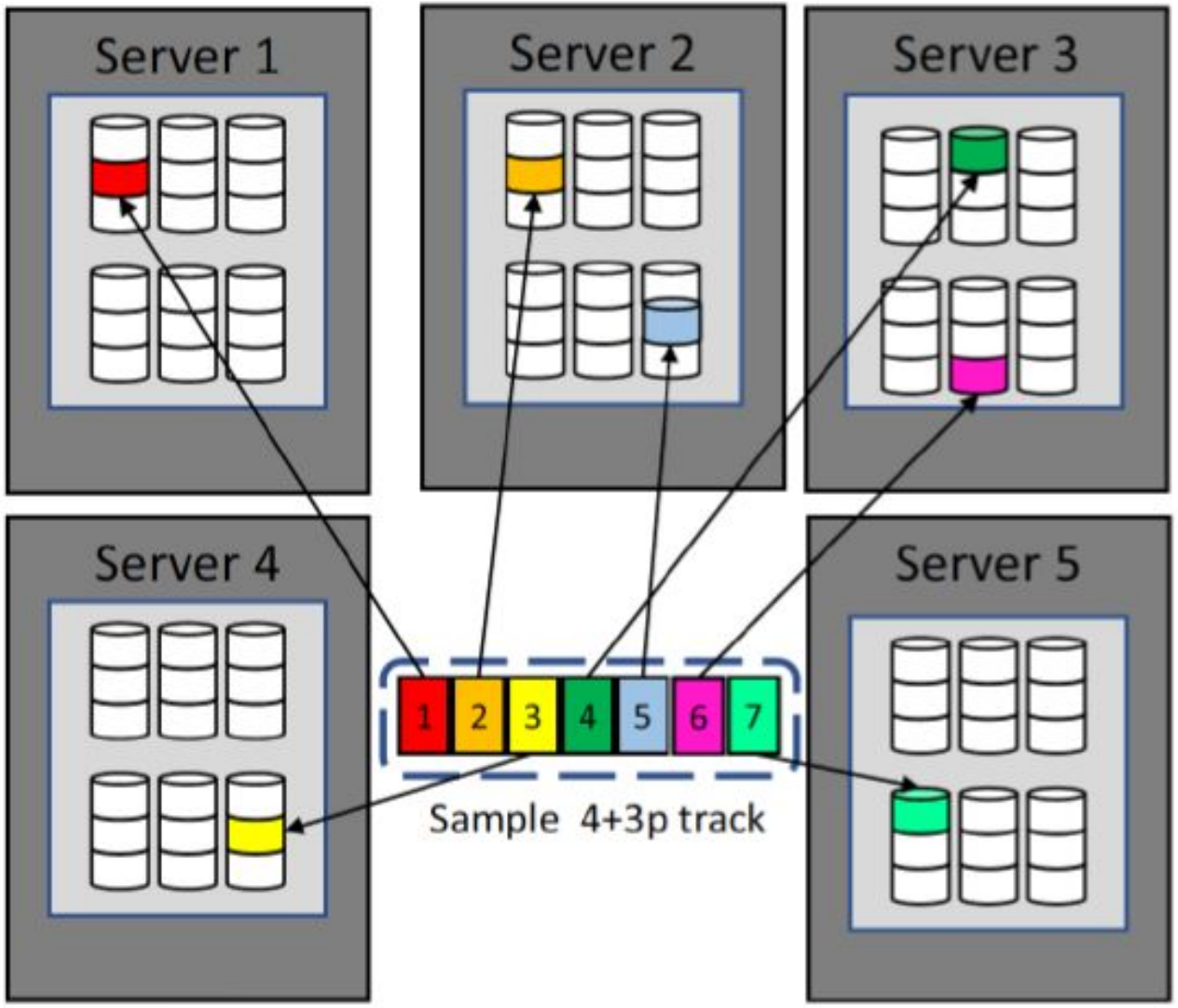
Write these 7 things to the 5 servers as ‘thinly’ as possible

Only need to read 4 to reconstruct the original

But any of the 4 will do – they are symmetric.

We can lose any of the 5 servers and any of the disks in the other 4 ie 1N +1D

(but can’t lose both server 2 and server 3)



HPE’S IMPLEMENTATION OF ECE

Based on standard HPE 1U servers

TWO TYPES OF STORAGE SERVERS

Combination of both in the same file system is supported

Flash Storage Server

HPE ProLiant DL325 Gen10 Plus with 16 x SFF slots

- # of NVMe SSD per server: 3, 4, 6, 8, 10, 12, 14 or 16
- Capacity points of NVMe SSD in TB: 3.84, 7.68 or 15.36
- (2) InfiniBand HDR/Ethernet 200 Gb 1p adapters

HDD Storage Server

HPE ProLiant DL325 Gen10 Plus with 8 x LFF slots

- # of SAS 7.2K RPM HDD per server: 3, 4, 6, or 8
- Capacity points of HDD in TB: 4, 8, 12 or 16
- (1) InfiniBand HDR100/ Ethernet 100 Gb 2p adapter
- Factory installed 1.6 TB NVMe capacity to serve file system metadata and small files from fast NVMe Flash

Same configuration rules for both:

At least 4 and up to 32 storage servers in identical configuration in a RAID cluster.

HPE PFSS

- HPE’s Spectrum Scale ‘Appliance’
- Built with standard 1U servers
- Minimum of 4 plus a switch

- No client licenses
- Customer has a perpetual right to use
- L1/ L2 support from us
- L3 comes from IBM

- GA was April 2021

PFSS BUILDING BLOCKS : BOTH PROLIANT DL326 GEN10

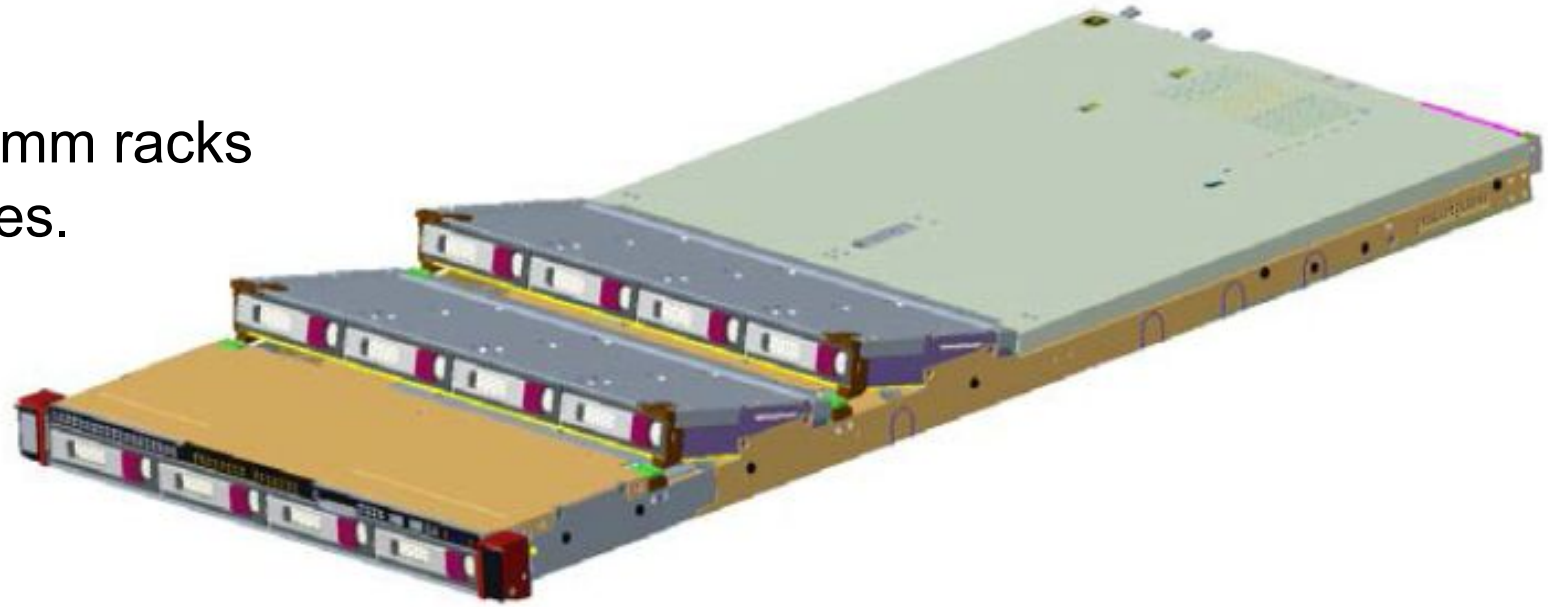
- 1U
- 3 to 16 **NVMe** 2.5” drives
- Dual 200 Gbit network
- 16 * 15.36TB = 253 TB raw (**1 PB in 4U**)

- 1U
- 3 to 8/12 **HDD** 3.5” drives
- Dual 100+ Gbit network
- 12 * 18TB = 216 TB raw (**1 PB in 5U**)

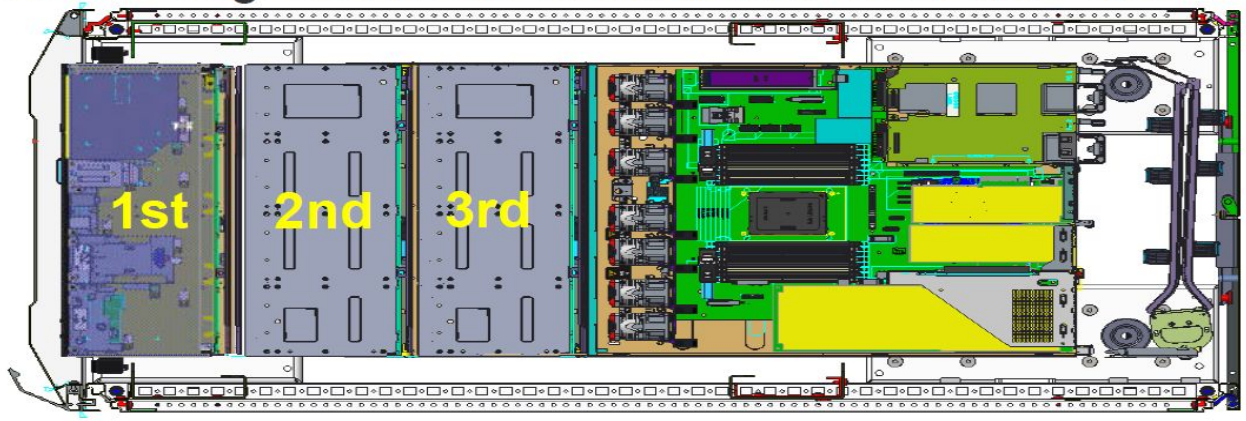
Filesystem Sizing:
SSD: from 11 TB Usable (max c. 34 PB)
HDD: from 27 TB Usable (max c. 21 PB)

PFSS HDD BUILDING BLOCK

- 8 drive or 12 drive
- 12 drive is deep – needs 1200mm racks
- 1U but still hot swappable drives.
- Single socket AMD CPU



3 HDD cages chassis in industrial 1200 mm rack



3RD BUILDING BLOCK : PROTOCOL SERVER

- 1U
- “Standard” Linux server
- AMD Rome CPU (16,32 or 64 core)
- 128, 256 or 512 GB RAM
- Dual 100+ Gbit network

- Not needed for traditional HPC or AI clusters.
- Only required if the customer needs NFS /SMB / S3 access to the filesystems
- (For very small NFS/SMB requirements just uses generic NFS/SMB under Linux instead)
- Recommend 2 minimum for HA failover, maybe 4 or 6 for large systems
- Uses floating IP address for seamless failover (CES = Cluster Export Services)
- Built on top of Ganesha NFS / SAMBA with HA additions by IBM

PFSS LICENSING

- Per storage server !

- not per-TiB,
- not per disk

Hewlett Packard
Enterprise

THANK YOU

INTRODUCING HPE PARALLEL FILE SYSTEM STORAGE

Fusion of the leading enterprise parallel file system with the leading industry-standard servers in the HPE factory

- **HPE storage with embedded IBM Spectrum Scale ECE file system running on HPE ProLiant DL325 Gen10 Plus servers** for
  - Clusters of HPE Apollo 2000/HPE ProLiant DL rack servers and HPE Apollo 6500 AI solutions
  - Verticals that often reject Lustre as they need enterprise grade storage features (e.g. Financial Service, Life Sciences, etc.)
  - Home directory storage for large HPC solutions where /scratch directories are on Lustre-based Cray ClusterStor E1000
- **Provides high performance parallel data access for compute nodes concurrently**
  - Native IBM Spectrum Scale client installed on compute nodes
  - NFS/SMB via Cluster Export Services (CES)
- **High speed connectivity to compute nodes** via
  - InfiniBand HDR100/Ethernet 100 Gb
  - InfiniBand HDR/Ethernet 200 Gb
- **Available in All Flash, All HDD or mixed configurations** based on workload profile
- **Provides broad set of enterprise storage functionality** like
  - Enterprise-grade system availability (“5 Nines”) incl. non-disruptive hardware & software upgrades, online expansion/contraction of the file system, etc.
  - Snapshots, compression, data replication, end-to-end data encryption, end-to-end data integrity (from disk to client), audit features for compliance
  - Protocol support beyond POSIX for NFS, SMB, HDFS, Object (S3, SWIFT) and (soon) Nvidia GPUDirect Storage
  - Data life cycle management - policy based data movement and curating and auto-tuning
- **Single price for the HPE storage system** (no file system license per terabyte or per storage drive)
- **Base warranty 3 years** for hardware & 1 year for software – HPE Pointnext Tech Care and HPE Datacenter Care are available

(in 4 node starter configuration)