Spectrum Scale Erasure Code Edition

New Storage Options for Spectrum Scale

Nikhil Khandelwal – nikhilk@us.ibm.com

Special thanks for contributing material to:

Stephen Edel, Client Technical Architect

Bill Owen, Senior Technical Staff Member
Spectrum Scale Development

Lin Feng Shen, Senior Engineer
Spectrum Scale Development

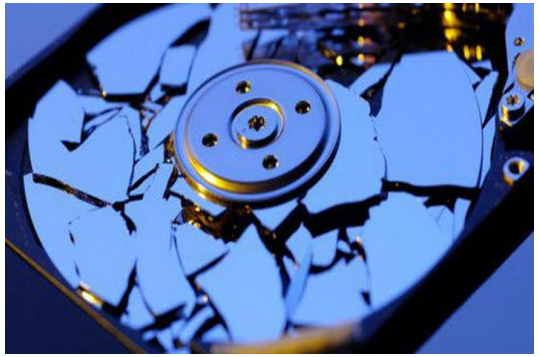
Table of Contents

* For latest hardware requirements see:
https://www.ibm.com/support/knowledgecenter/en/STXKQY_ECE_5.0.3/com.ibm.spectrum.scale.ece.v5r03.doc/b1lece_min_hwrequirements.htm

Storage-rich Servers are Growing Rapidly, Driven by…

- **Supplier mandates**
  - “We buy from Dell, HP, Lenovo, SuperMicro – whoever is cheapest at that moment”
  - “Our designated configuration is HPE Apollo”
  - “We assemble our own servers that are OCP compliant”
- **Technical and architectural mandates**
  - “This is for an analytical grid where the IT architecture team only allows x86”
  - “We need a strategic direction for scale-out storage”
  - “Only storage rich servers are acceptable, no appliances”
  - “We use storage arrays today and we are forced by upper management to go with storage rich servers”
- **Cost considerations**
  - “We want the economic benefits of commodity hardware”
  - “We don’t want to pay for high-end or even mid-range storage”

Challenges with Commodity Server Based Distributed Storage



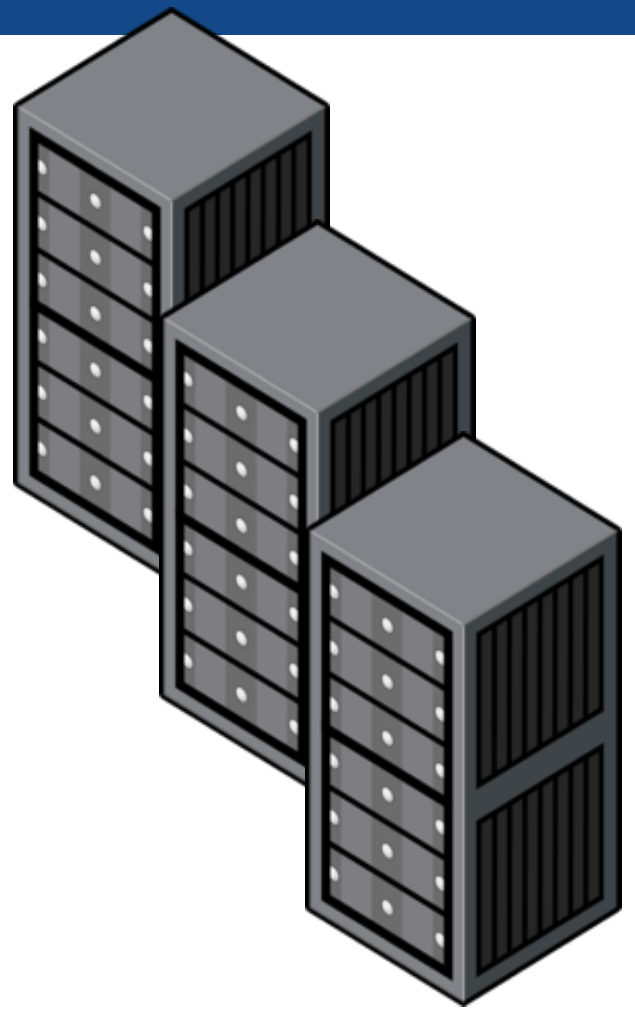
- **Poor storage utilization**: Hadoop/Spark and other applications often have 3 replicas to protect data from hardware or software failures, resulting in low storage efficiency (33%) and thus higher costs
- **High failure rates**: Higher failure rates of commodity hardware means poor reliability, less availability, longer disk rebuild times and more impact to performance w/o SW RAID.
- **Data integrity concern**: With large volumes of data in commodity drives in commodity servers, the possibility of silent data disk corruption becomes much higher than in traditional storage systems at a smaller scale
- **Scalability challenges and data silos**: Some distributed storage systems may not be able to scale or be managed easily when approaching Petabyte capacity server farms, resulting in inefficiencies and potentially unnecessary data movement
- **Missing enterprise storage features**, e.g. data life cycle management, snapshots, backup/restore, disaster recovery, disk management, encryption, etc.
- **NAS Only Support**: Lack of POSIX compliant capabilities, limited performance of traditional file systems

Spectrum Scale Erasure Code Edition

What is Spectrum Scale Erasure Code Edition?

A new Spectrum Scale offering that brings all of the benefits of “Data Management Edition” *plus* **Spectrum Scale RAID**

- Spectrum Scale running in storage rich servers connected to each other with a high speed network infrastructure
- Bring your own hardware – select any hardware that meets minimum requirements
  - Provides Storage devices can be HDD, SSD, NVMe or a mixture
- features of an Enterprise Storage Controller all in software
  - Enterprise ready storage software used in Spectrum Scale Elastic Storage Server (ESS)
- Restricted GA June 2019 *



* Required to verify supported hardware configuration

IBM Spectrum Scale Erasure Code Edition

Delivers all the capability of Spectrum Scale Data Management Edition

- Enormous scalability with Software-based declustered RAID protection
- Very high performance no additional RAID hardware
- Enterprise manageability

Plus: Durable, robust, and storage-efficient

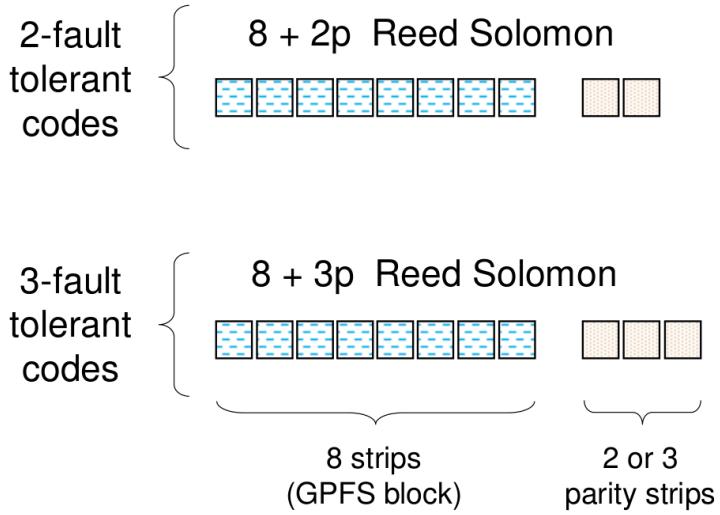
- Distributes data across nodes and drives for higher durability *without* the cost of replication
- End to end checksum identifies and corrects errors introduced by network or media
- Rapid recovery and rebuild after hardware failure

Plus: Delivered at hyperscale

- Hardware platform neutrality -Supports the user's choice of commodity servers and drives
- Disk Hospital manages drive issues before they become disasters
- Continuous background error correction supports deployment on very large numbers of drives

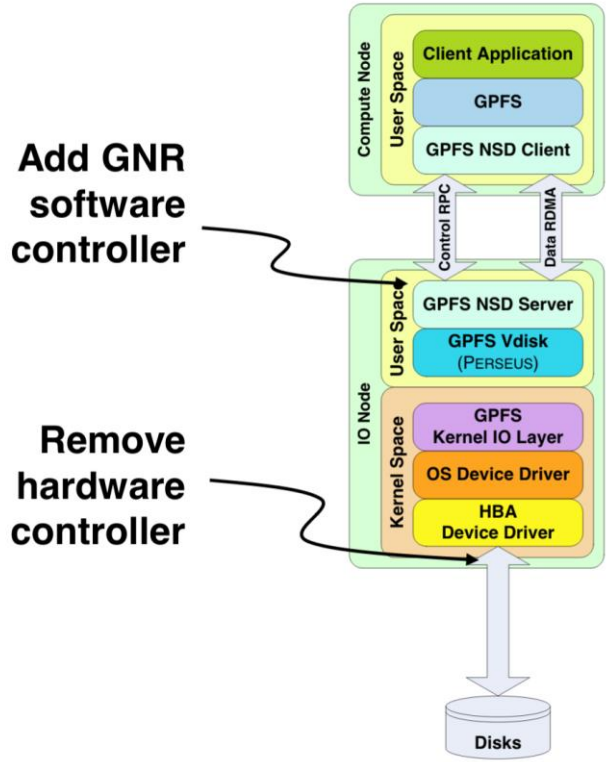
Erasure Coding Edition Reed Solomon Code Options

- ECE supports several erasure coding options and brings much better storage efficiency, with 8+3p and 8+2p Reed Solomon Code
- New erasure coding options include 4+2P and 4+3P
- Better storage efficiency means less hardware and SW costs, which can help customers to save on budgets without compromising system availability and data reliability.
- ECE erasure coding can better protect data compared with traditional RAID5/6
  - e.g. a configuration of 3 nodes of fault tolerance in an 8+3p mode, with 11 or more nodes, which can survive concurrent failure of multiple servers and storage devices.
- IBM's ECE high performance erasure coding can be used in a first tier of storage or stand-alone. High performance on commodity servers is a key differentiation compared with other erasure coding implementations in distributed storage systems.



Spectrum Scale ECE – Scaling Out Architecture

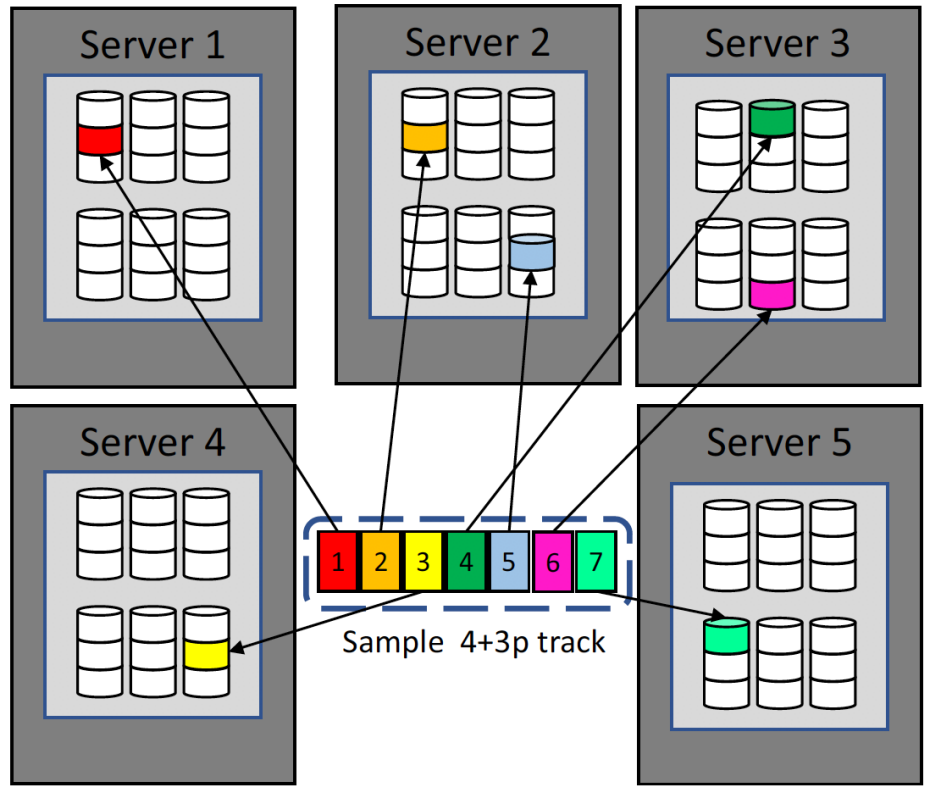
- Spectrum Scale ECE software replaces any server based hardware controller or software RAID
  - Use any hardware that meets minimum requirements*
- Scale capacity by:
  - Adding drives to each server
  - Adding servers to a recovery group server set
- When adding servers, you are also scaling other critical resources
  - CPU, Network and PCI bandwidth, memory



* See following slides

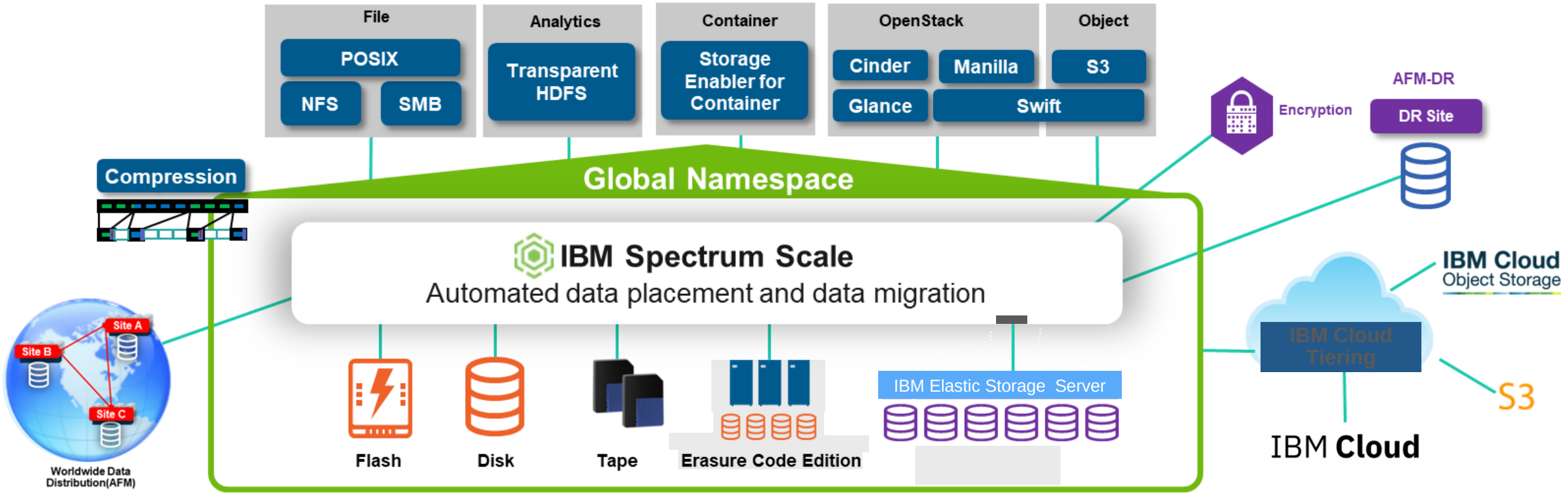
Data Distribution Across Nodes

- Tracks are distributed across the nodes in a recovery group
- Ability to tolerate multiple node and device failures while preserving access to the data
  - Number of concurrent failures is determined by the stripe size and the number of nodes



ECE Storage Pool Coexistence with IBM Spectrum Scale

Unleash new storage economics on a global scale



Highest Performance Storage with Diverse Access Protocols *using hardware that you select*

Consolidate all your unstructured data storage on spectrum scale with unlimited and painless scaling of capacity and performance

ECE is based on Proven IBM Spectrum Scale software

The software in ECE has been field-proven in over 1000 deployed ESS systems

ESS is the storage power behind the fastest supercomputers on the planet

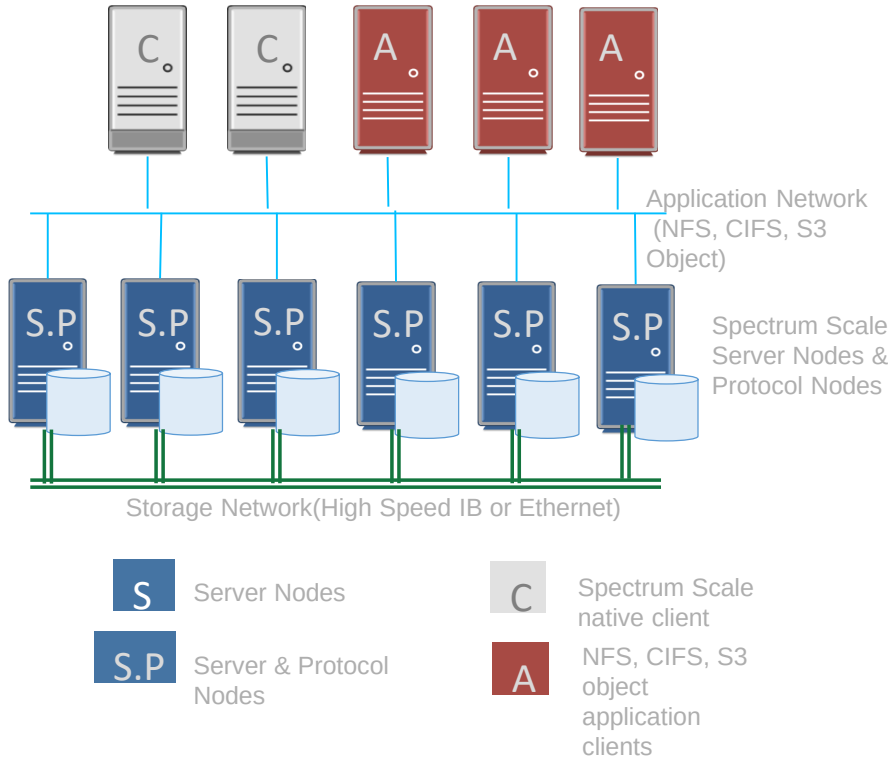
- Summit and Sierra supercomputers at Oak Ridge National Laboratory and Lawrence Livermore National Laboratory are ranked the #1 and #2 fastest computers in the world
- They are helping to model supernovas, pioneer new materials, and explore cancer, genetics and the environment, using technologies available to all customers

ECE delivers the same capabilities on commodity compute, storage, and network components

ECE USE CASES

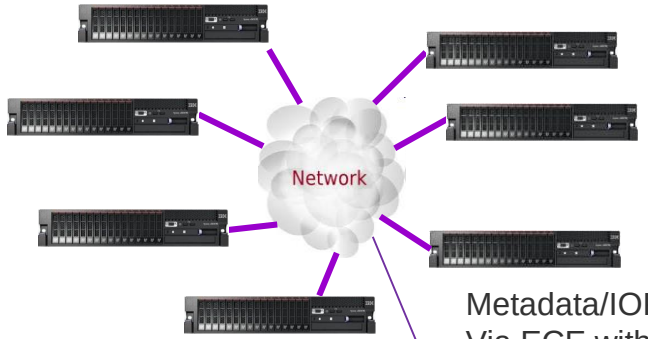
ECE Use Case: Dedicated High Performance file serving

- Spectrum Scale Server high performance file services deployed on storage rich nodes communicating to native Spectrum Scale clients
- Deploy IBM Spectrum Scale Protocol services to allow customers to access ECE with NFS, SMB and Object.
- Dedicate High speed IB or Ethernet for NFS/SMB/storage communication
- Accelerate data processing by leveraging enterprise NVMe drives to deliver high throughput and low latency
- Each ECE storage server is typically configured with several NVMe drives to store and accelerate Spectrum Scale metadata and small data I/O, combined with a number of HDD drives to store user data.
- With the high performance design of ECE, it can deliver high performance file serving to the customer workloads.



ECE Use Case: High Performance Compute tier

- ECE's high performance erasure coding provides the capability of being a tier 1 storage device that can then tier to different storage medias (e.g. flash drives, spinning disks, tape, cloud storage, etc.) with different performance and cost characteristics.
- The policy based Information Life Cycle management feature makes it very convenient to manage data movement among different storage tiers.
- In this example, the ECE high performance compute tier is composed of NVMe drives to store and accelerate Spectrum Scale metadata and the set of hot data for high performance computing and analytics. The second tier can consist of NL SAS drives for lower $/TB and fast sequential performance.



Metadata/IOP Tier via
Via ECE with NVMe

High Sequential
Performance
Tier via ESS
or build your own

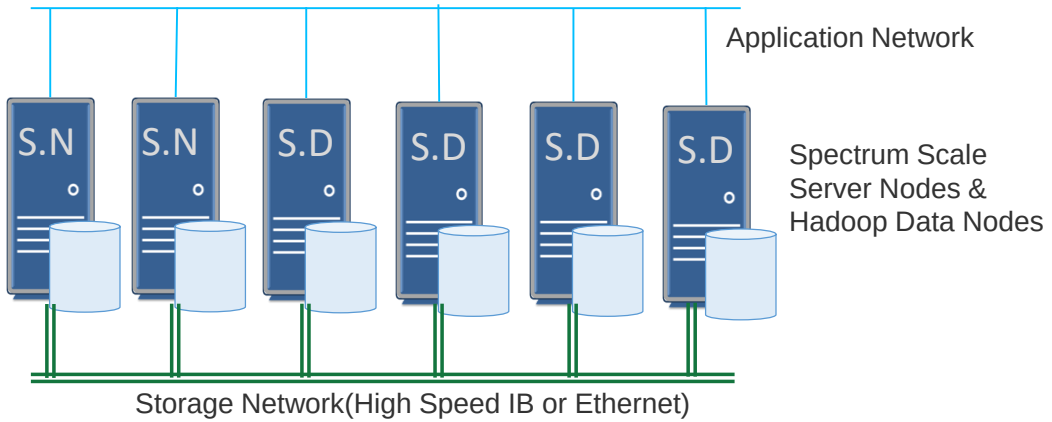
ECE Use Case: Analytics

Deployment Model:

- Spectrum Scale Server and Transparency nodes (Name Node and Data Node) are deployed in storage rich server
- Dedicate High speed IB or Ethernet for storage communication (optional but highly recommended)

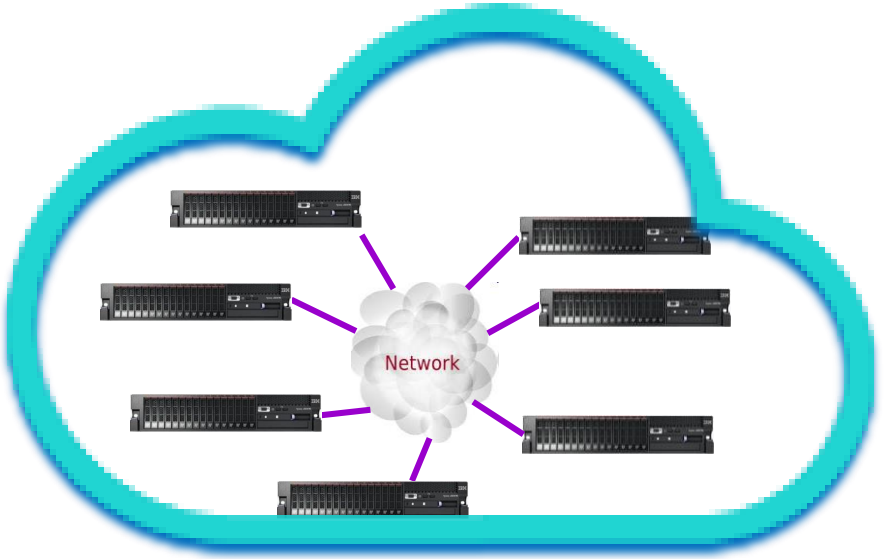
Use Case:

- Analytics workload based on HDP (or even Cloudera)
- Enterprise storage of HDFS alternative



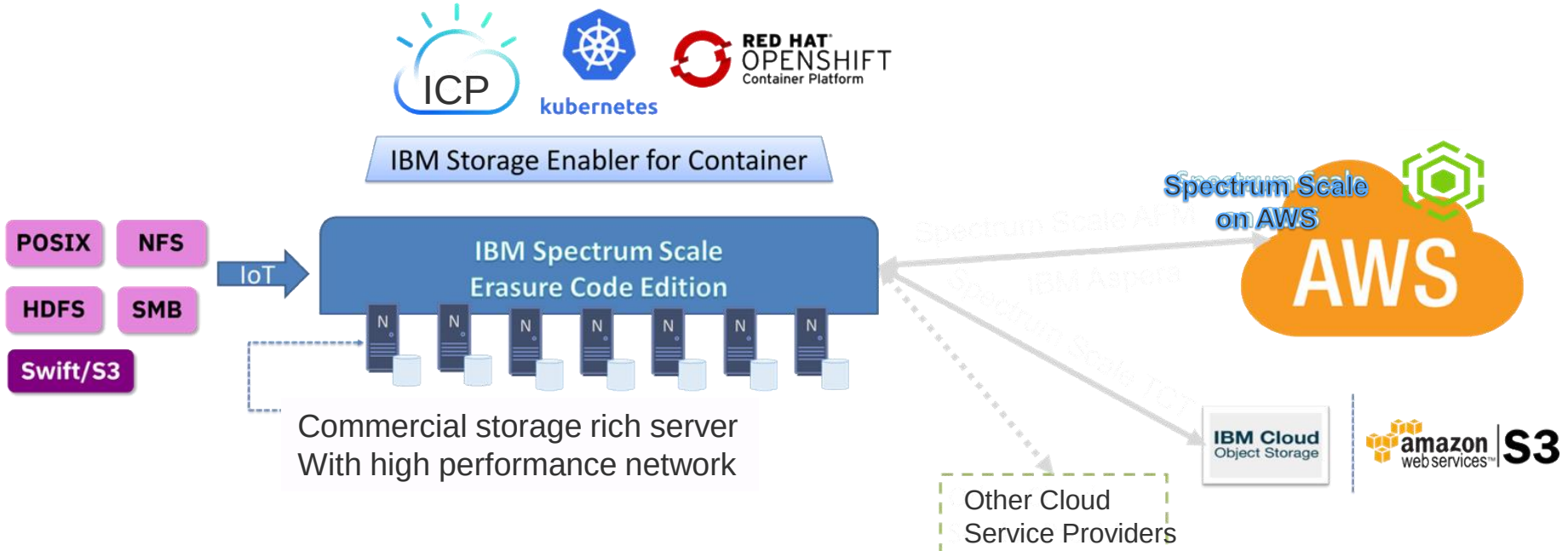
ECE Use Cases: High Capacity Cloud Storage

- With space efficient erasure coding and extreme end-to-end data protection design and implementation, ECE can deliver the essential cost effective and data reliability value-adds to large scale cloud storage systems.
- The ECE storage system for high capacity cloud storage may be composed of an NVMe storage pool to store and accelerate GPFS metadata and small data I/O's, or all high capacity drives for lowest $/TB
- An ECE storage system can also be low cost cloud storage connected to an on-site Spectrum Scale cluster with AFM



ECE Use Cases: Hybrid Multi-Cloud Storage including Containers

- ECE can provide a high performance on-prem Scale Out Filesystem and leverage containers and Kubernetes to support IBM Cloud Private, Red Hat OpenShift as well as leveraging AFM and TCT to a multitude of private/hybrid/public clouds
- Data comes from both data center and public cloud which need to be stored in a single name space to provide storage service for container
- IBM Spectrum Scale runs in both on-perm data center and AWS public cloud are connected by Spectrum Scale AFM to provide a single name space.
- Spectrum Scale with IBM Storage Enabler for Container providers storage service for container

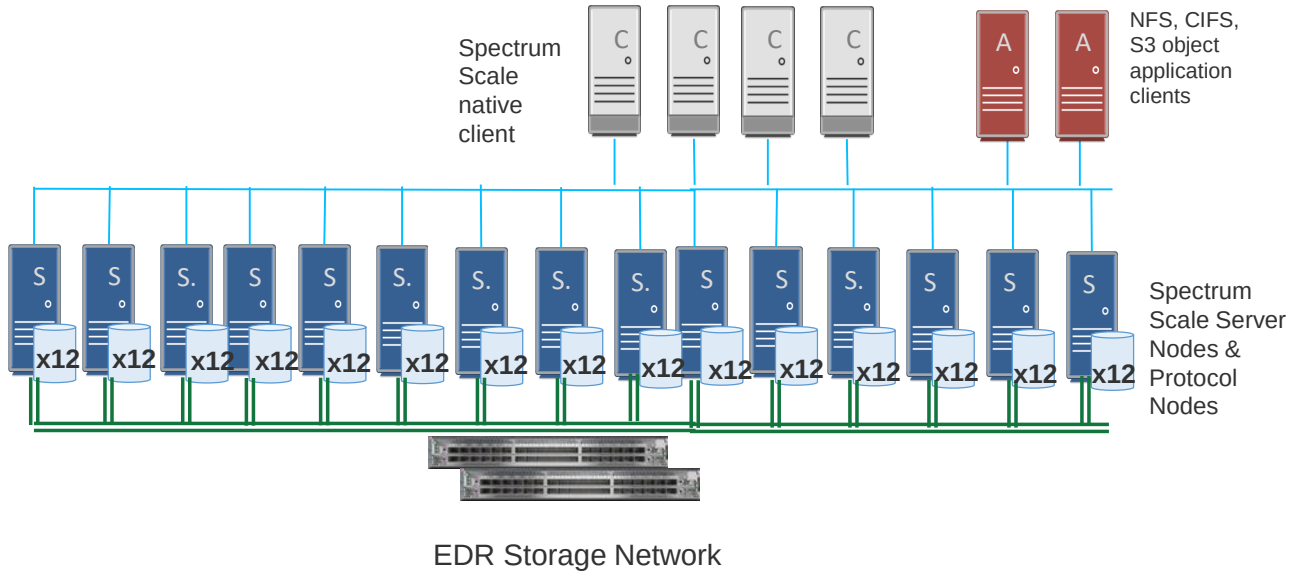


Sample High Performance Design w/NFS/SMB Export

REQUIREMENT:
- 200TB Usable
- HPC -100GB/sec throughput
- SMB and NFS Access

DESIGN
- # Nodes needed – 15
- 12 drives per node
- 1.5TB drive size (e.g. Intel 4800)
- 2 x 100Gb/sec network cards per node
- Usable capacity per node – 35TB
- Usable Network BW per node – 10GB/sec

•Overall est. Read BW – 160GB/sec
•Overall est. Write BW – 135GB/sec



Recap: Key Features of Erasure Coding

Erasure Coding Overview

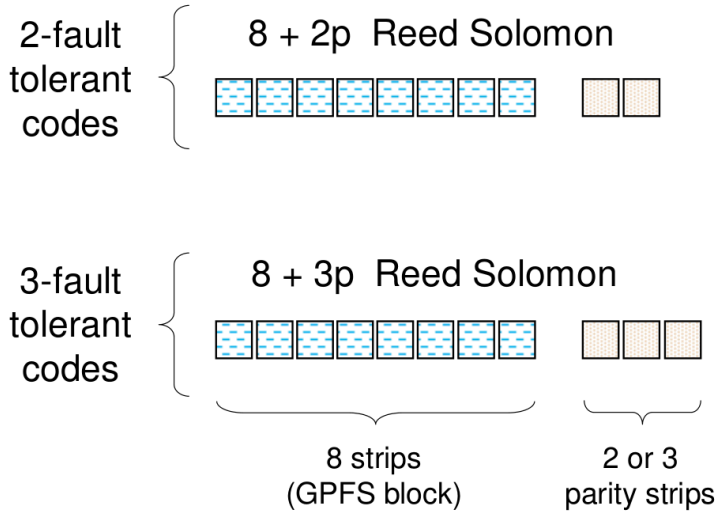
Spectrum Scale's Reed-Solomon implementation

- For every block of data, we slice the block into K strips of equal size
- We then calculate N parity strips using RS encoding functions

- When writing data blocks, after calculating the N parity strips we store K + N total strips
  - We then distribute the data and parity strips as widely as possible across racks, servers and devices in order to minimize impact of any failure

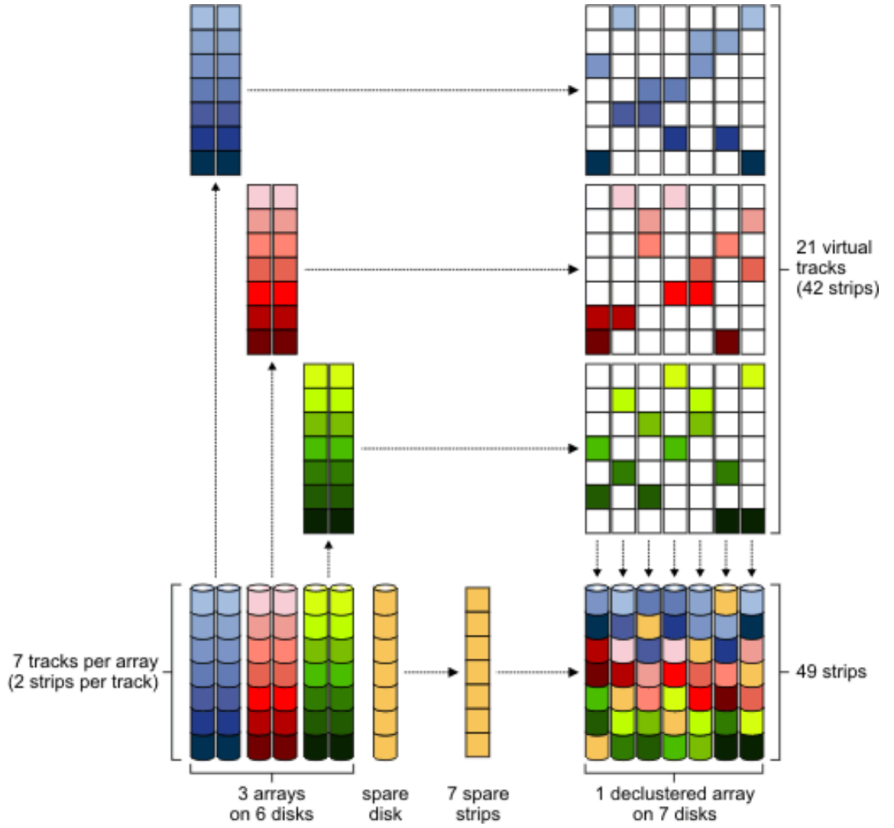
- When reading data blocks:
  - Normal case is to read and aggregate the K data strips adding no extra overhead
  - Only rebuild data from parity strips when a lost or corrupt strip is detected
  - When possible detect and rebuild lost or corrupt data in the background

- Reed-Solomon error correction and erasure coding allows data to survive loss (erasure) or data error (correction) in up to N strips



Spectrum Scale Erasure Code Advantages

- Declustered erasure coding provides for data and parity to be distributed over all the disks and nodes in the declustered array for fastest performance out of the chosen media
  - Faster and more intelligent rebuild operations, using more drives in parallel
  - Prioritize normal vs critical conditions to better use node resources
  - Spare capacity is also distributed across all drives and nodes, so no dedicated spare disks are needed
- Improved storage efficiency and performance
  - 8+2P and 8+3P utilize less overhead vs 100% - 200% for 2X-3X replication
  - Patented algorithms optimize I/O data paths, read and multi-layer write caching



Spectrum Scale Erasure Code - Disk Hospital

- Identify device problems before hard drive failure:
  - Dead or misbehaving disks
  - Connectivity issues
  - Media errors
  - Slow drives
- Attempt corrective action to revive sick or failing devices:
  - Power cycle non-responsive drives
  - Recompute and rewrite corrupted data
  - Rediscover disk connectivity
- Maintain “health record” for each device
  - If device is accumulating too many errors, remove from service
  - If device is persistently slow, remove from service

Spectrum Scale Erasure Code - Integrity Management

- *Every IO* has a checksum added to data trailer
- For writes, verify data integrity when data passes from
  - Client (compute node) to storage node
  - Storage node to storage media
  - Writes also include a sequence number in the metadata to detect dropped/skipped writes
- For reads, verify data integrity when data passes from
  - Storage media to storage node
  - Storage node to client
- A background scrub task periodically detects and fixes silent data corruption on the storage devices
- Automatic data rebuild on failure, automatic rebalance on recovery or when new storage is added
- Rebuild has minimal impact on system performance
  - Rebuild is distributed across disks and nodes
  - Rebuild can be deferred with sufficient protection
- Failure domain for high hardware failure tolerance

ECE Hardware Requirements

ECE Hardware/Architecture Requirements

ECE software is hardware platform neutral, but there are hardware requirements.

- An ECE storage system must have at least 4 servers, and up to 128 servers (128 is a test limitation in the first release.)

- Customers can create multiple ECE recovery groups. Each recovery group limits the number of servers to between 4 and 32.

- Customers may scale out their ECE storage system with one server, multiple servers or a whole building block.

- Every server in a recovery group must have the same configuration in terms of CPU, memory, network, storage, OS, etc.

- For SSD and NVMe drives, it is recommended to use a file system block size of 4M or less with 8+2P or 8+3P erasure codes, and 2M file system block size or less for 4+2P or 4+3P erasure codes.

- Minimum Declustered Array (DA) size DA is : At least one DA must contain 12 or more drives [1]
  - A DA is a subset of the physical disks within a recovery group that have matching size and speed.
  - A recovery group may contain multiple declustered arrays, which are unique (that is, a pdisk must belong to exactly one declustered array
  - The minimum DA size is met by each node contributing a uniform number of disks. That means a 4 node RG must have one DA with 3 or more drives per node. A twelve node RG could have one drive per node, but that drive must be a “fast device”, either SSD or NVMe.

- Each node must have at least one fast device (NVMe or SAS SSD)

- All nodes/HBA's/drives in a Recovery Group must meet minimum firmware level requirements specified in Hardware Selection and Sizing Guide. (Tested number of drives per Recovery Group: 512)

- To deliver the best performance, stability and functionality the next chart lists the minimal hardware requirements for each storage server. (This list will be expanded over time).

ECE Hardware Requirements for each Storage Server
as of August 2019*

CPU architecture	x86 64 bit processor with 8 or more processor cores per socket. Server should be dual socket with both sockets populated
Memory	•64 GB or more for configurations with up to 24 drives per node. For NVMe configurations, it is recommended to utilize all available memory DIMM sockets to get optimal performance. •For server configurations with more than 24 drives per node, contact IBM® for memory requirements.
Server packaging	Single server per enclosure. Multi-node server packaging with common hardware components that provide a single point of failure across servers is not supported at this time.
System drive	A physical drive is required for each server's system disk. Recommend RAID1 protected and have a capacity of 100 GB or more.
SAS Host Bus Adapter	LSI SAS HBA, models SAS3108, SAS3216 or SAS3516.
SAS Data Drives	SAS or NL-SAS HDD or SSDs in JBOD mode. SATA drives are not supported at this time.
NVMe Data Drives	Enterprise class NVMe drives with U.2 form factor.
Fast Media Req.	At least one SSD or NVMe drive is required in each server for IBM Spectrum Scale Erasure Code Edition logging.

* For latest hardware requirements see:
https://www.ibm.com/support/knowledgecenter/en/STXKQY_ECE_5.0.3/com.ibm.spectrum.scale.ece.v5r03.doc/b1lece_min_hwrequirements.htm

ECE OS & Network Requirements for each Storage Server

as of August 2019*

Operating system	RHEL 7.5 or 7.6. See IBM Spectrum™ Scale FAQ for details of supported versions.
Network Adapter	Mellanox ConnectX-4 or ConnectX-5, (Ethernet or InfiniBand)
Network Bandwidth	25 Gbps or more between storage nodes. Higher bandwidth may be required depending on the workload requirements.
Network Latency	Average latency must be less than 1 msec between any storage nodes.
Network Topology	To achieve the maximum performance for a workload, a dedicated storage network is recommended. For other workloads, a separate network is recommended but not required.

* For latest hardware requirements see:
https://www.ibm.com/support/knowledgecenter/en/STXKQY_ECE_5.0.3/com.ibm.spectrum.scale.ece.v5r03.doc/b1lece_min_hwrequirements.htm

TECHNICAL SPECIFICATIONS / IMPLEMENTATION

https://www.ibm.com/support/knowledgecenter/STXKQY_ECE_5.0.3/ibmspectrumscaleece503_welcome.html

High Level ECE Configuration – 6 Steps

1.Create a node class that contains a set of identical storage servers that belong to a single recovery group. There should be a minimum of 4 nodes and maximum of 32 nodes in a recovery group:
mmvdisk nc create --nc <nodeclass-name> -N <node-list>

2.To maintain quorum availability in the IBM Spectrum Scale cluster, exercise caution when you recycle nodes. The example below uses "--recycle one" so that nodes are recycled one at a time.
mmvdisk server configure --nc <nodeclass-name> --recycle one

3.Create a recovery group:
mmvdisk rg create --rg <rg-name> --nc <nodeclass-name>

4.Define one or more vdisk sets:
mmvdisk vs define --vdisk-set <vs-name> --rg <rg-name> --code <erasure-code> --block-size <bsize> --set-size <set-size>

5.Create the vdisk sets that you defined:
mmvdisk vs create --vs <vs-name>

6.Create and mount the file system:
mmvdisk filesystem create --file-system <fs-name> --vs <vs-name> mmmount <fs-name> -N <nodes-to-mount-on>

For details on each command and the supported arguments. see the *mmvdisk* topic in the *IBM Spectrum Scale: Concepts, Planning, and Installation Guide*.

Erasure Code Options and Failure Tolerance

Number of Nodes in RG	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

Installation and Hardware Pre-check

The IBM Spectrum Scale Erasure Code Edition precheck, integrated in the installation toolkit installation, deployment or upgrade precheck. The ECE check: standalone, publicly available, open source

For IBM Spectrum Scale Erasure Code Edition, the pre-check includes the following on all scale-out nodes:

- Check CPU requirements (Server cpu type/number of sockets/number of cores
- Check memory requirements (Server memory & DIMM utilization)
- Confirm consistent, allowable disk topology
- Check OS and firmware levels
- Check whether the networking requirements including the required NIC and SAS adapters are met
- Check whether the required syscall parameters are set correctly

Installation toolkit-related prerequisites

- Ensure that networking is set up in one of the following ways.
- DNS is configured such that all host names, either short or long, are resolvable.
- All host names are resolvable in the /etc/hosts file. The host entries in the /etc/hosts file must be in the following order:<IP address> <Fully qualified domain name> <Short name>
- Passwordless SSH must be set up using the FQDN and the short name of the node

Hardware precheck - verify minimum levels and consistency across Recovery Groups via toolkit

Test results saved with install log for installation record

https://www.ibm.com/support/knowledgecenter/en/STXKQY_ECE_5.0.3/com.ibm.spectrum.scale.ece.v5r03.doc/b1lece_install_intro.htm

High Level Networking Installation Steps

- Network precheck between every ECE storage node
  - Average latency < 1 msec
  - Maximum latency < 2 msec
  - Standard Deviation < 0.33 msec

- Network KPI check for network assessment
  - Standalone (based on nsdperf)
  - Publicly available
  - Open source (nsdperf becomes opensource software)

Again, for ECE, it’s all about the network

- Spectrum Scale ECE is highly network dependent
- NSD servers receive a request (ex. Write), and will need to send the write data and parity data to pdisks on other nodes
- Latency on the network plays a large role in performance
  - A High speed, low latency storage network is essential
- Keep CES, AFM, TCT and other services on separate networks
- Ensure storage network (backend) is as fast as or faster than client network (frontend)
- Use the mmnetverify connectivity all option in the mmnetverify command in the IBM Spectrum Scale: Command and Programming Reference to ensure that your network is configured for use by IBM Spectrum Scale

ECE Installation Steps

Phase 1: Cluster definition

- Installer node is defined by the user. Setup type is specified as ece by the user.
- Scale-out nodes and other node designations are done by the user. Other types of nodes that can be designated include protocol, GUI, call home, and file audit logging. But you must add a client node for each of these functions.
- Recovery group is defined by the user.
- Vdisk set and filesystem are defined by the user. [Vdisk set & filesystem definition can be done after the installation phase

Phase 2: Installation - This phase starts upon issuing the ./spectrumscale install command.IBM Spectrum Scale Erasure Code Edition packages including the IBM Spectrum Scale Erasure Code Edition license package are installed.

- IBM Spectrum Scale Erasure Code Edition cluster is created.
- Quorum and manager nodes are configured.
- Server and client licenses are applied.
- Node class is created.
- Recovery group is created.

Phase 3: Deployment - This phase starts upon issuing the ./spectrumscale deploy command.Vdisk sets are created.

- File systems are created.
- Protocols are deployed, if applicable
- For more information reference the IBM Knowledge Center here:

https://www.ibm.com/support/knowledgecenter/en/STXKQY_ECE_5.0.3/com.ibm.spectrum.scale.ece.v5r03.doc/bl1ece_scaleecewithtoolkitoverview.htm

Quorum/Manager node rules

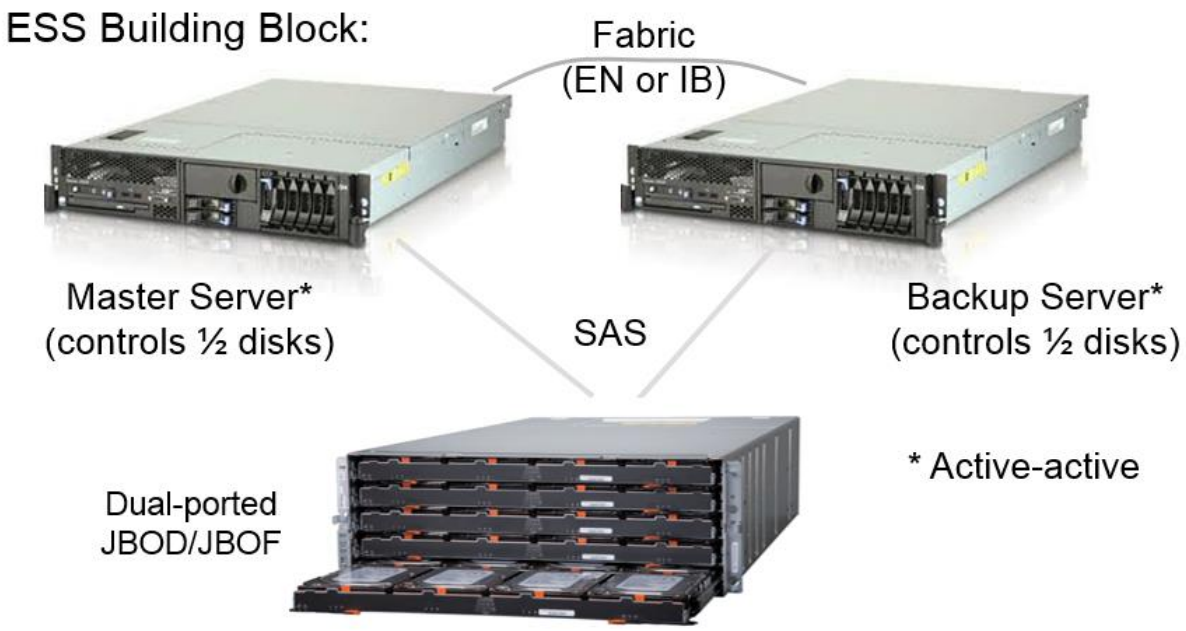
- In case of a single recovery group, the following quorum node rules apply.
  - When the number of scale-out nodes is 4, the number of quorum nodes is set to 3.
  - When the number of scale-out nodes is 5 or 6, the number of quorum nodes is set to 5.
  - When the number of scale-out nodes is 7 or more, the number of quorum nodes is set to 7.
- If the number of recovery groups is more than 1 and less than or equal to 7, 7 quorum nodes are distributed across recovery groups in a round robin manner.
- If the number of recovery groups is more than 7, 7 recovery groups are selected as quorum holders.
- If there is no recovery group or quorum node defined in the cluster configuration, the installation toolkit displays the following message.
  - "You have not defined any recovery group in the cluster configuration. Installer will automatically define the quorum configuration. Do you want to continue"
- If you specify yes then quorum nodes are distributed according to the single recovery group rule.
- If you are adding a new recovery group in an existing cluster or if you want to add a new node into the existing node class, the existing quorum configuration is not modified by the installation toolkit.
- For an existing cluster, if you want to have quorum on a different node or a different recovery group then you must use an IBM Spectrum Scale command such as **mmchnode** to change this configuration.
- Every scale-out node has the manager mode designation. Scale-out nodes in a recovery group are equivalent so any of them can pick up the cluster manager or the file system manager role.

Architectural Comparison with ESS

Hardware Architecture Comparison

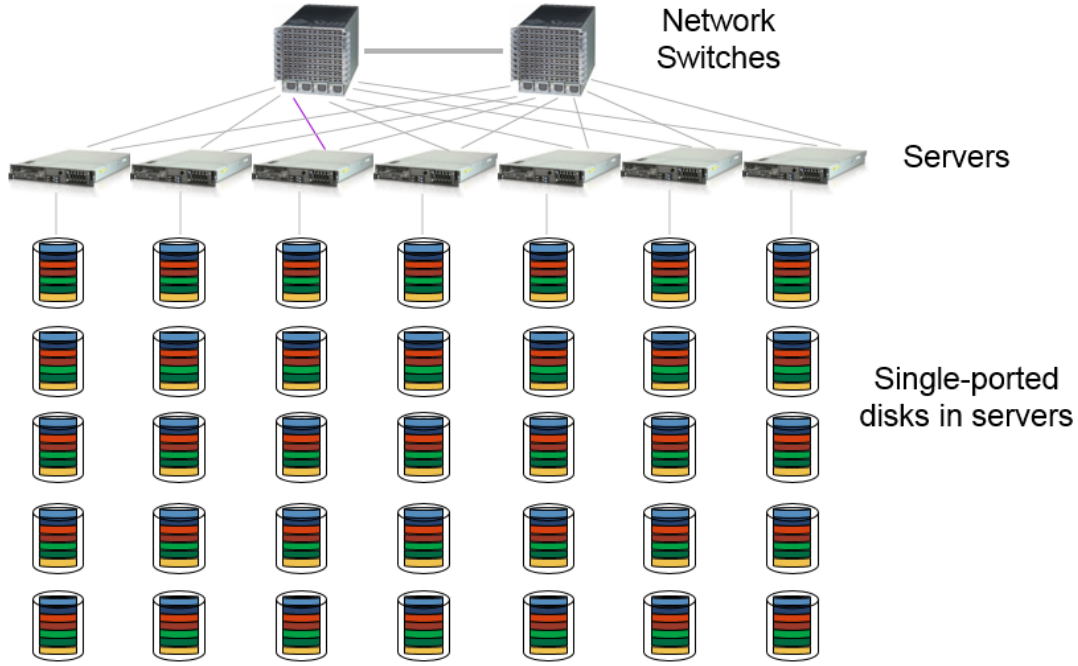
ESS

- Twin-tailed disks, dual servers – provide very high availability
- However, in case when a failure of both the master and backup servers happens it results in data unavailability



ECE

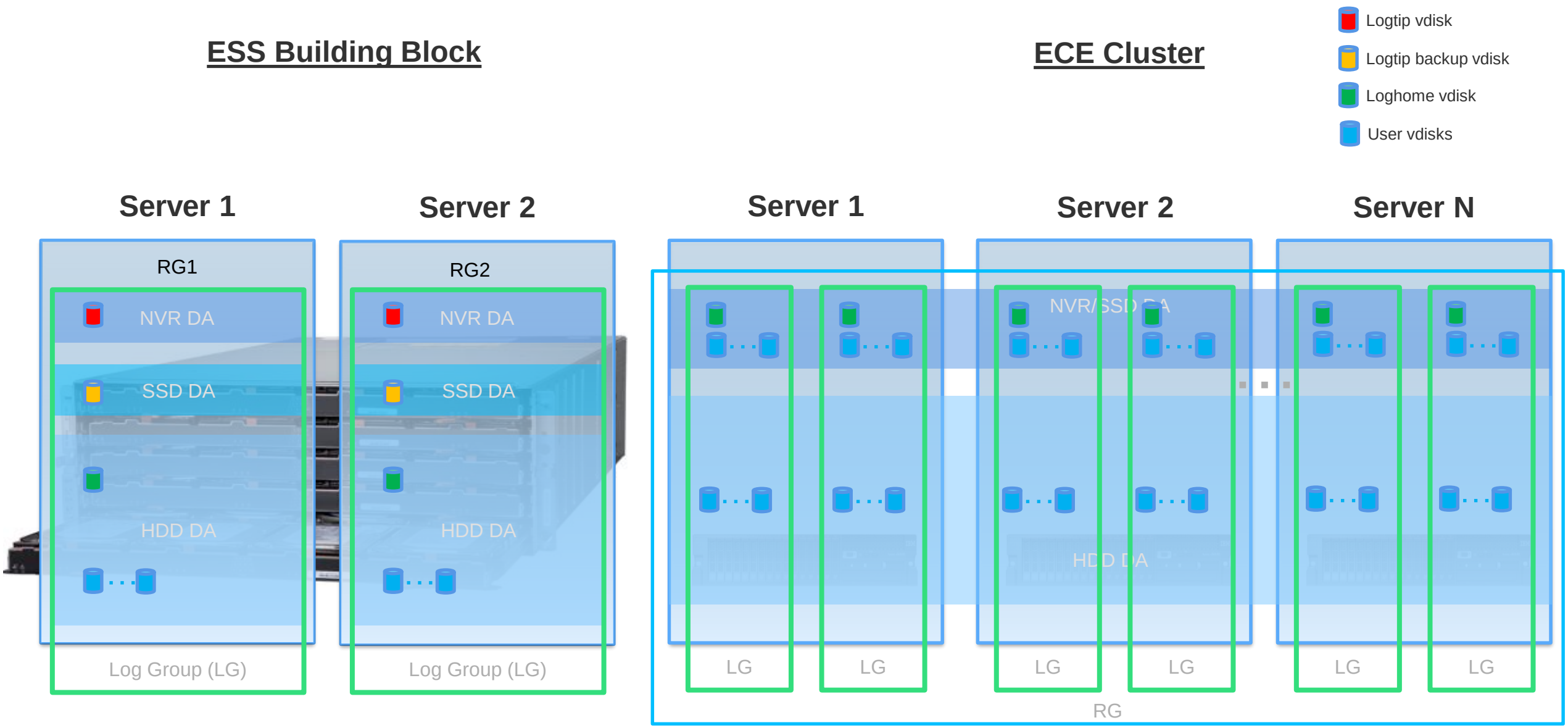
- Network RAID Internal disk rich commodity servers
- Tolerates concurrent failure of an arbitrary pair of servers (or 3 servers if 8+3p erasure code) and disks



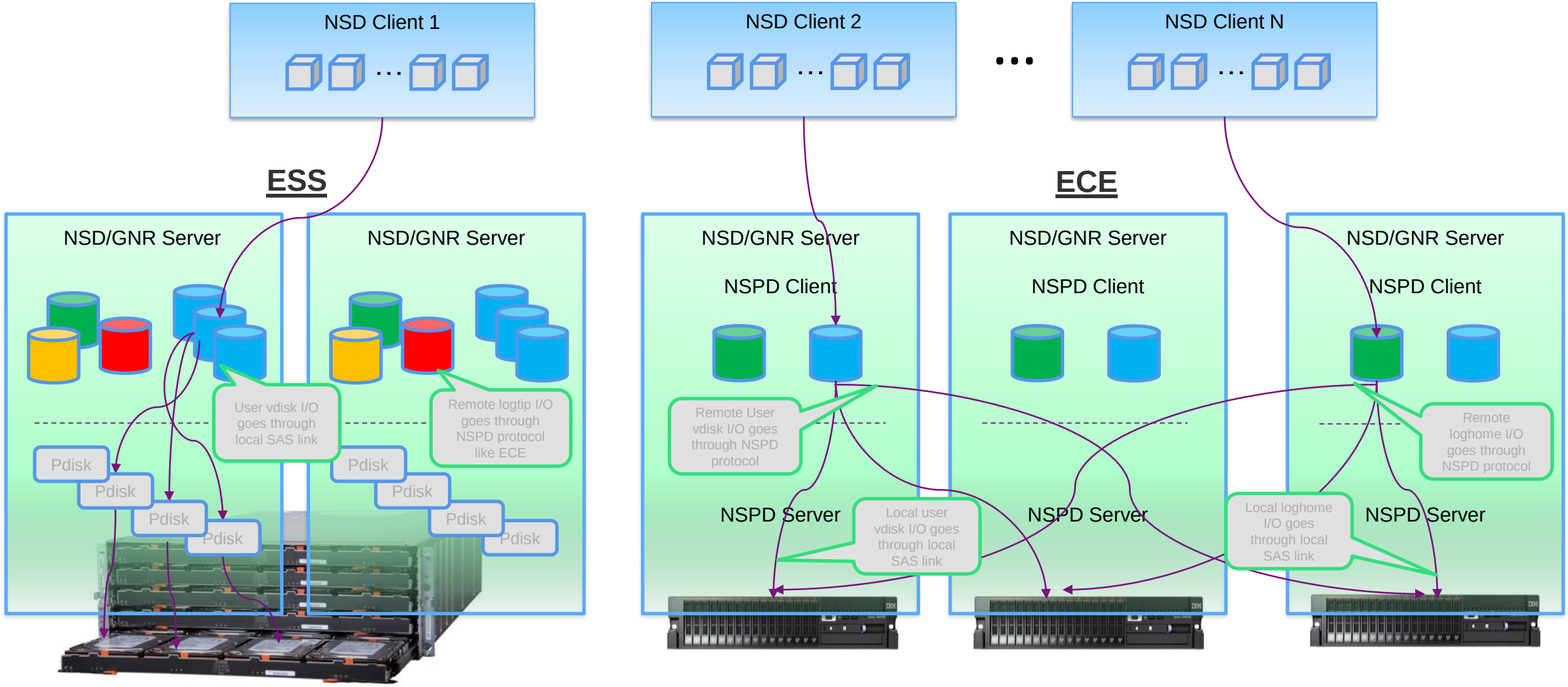
Hardware Resource Partitioning

ESS Building Block

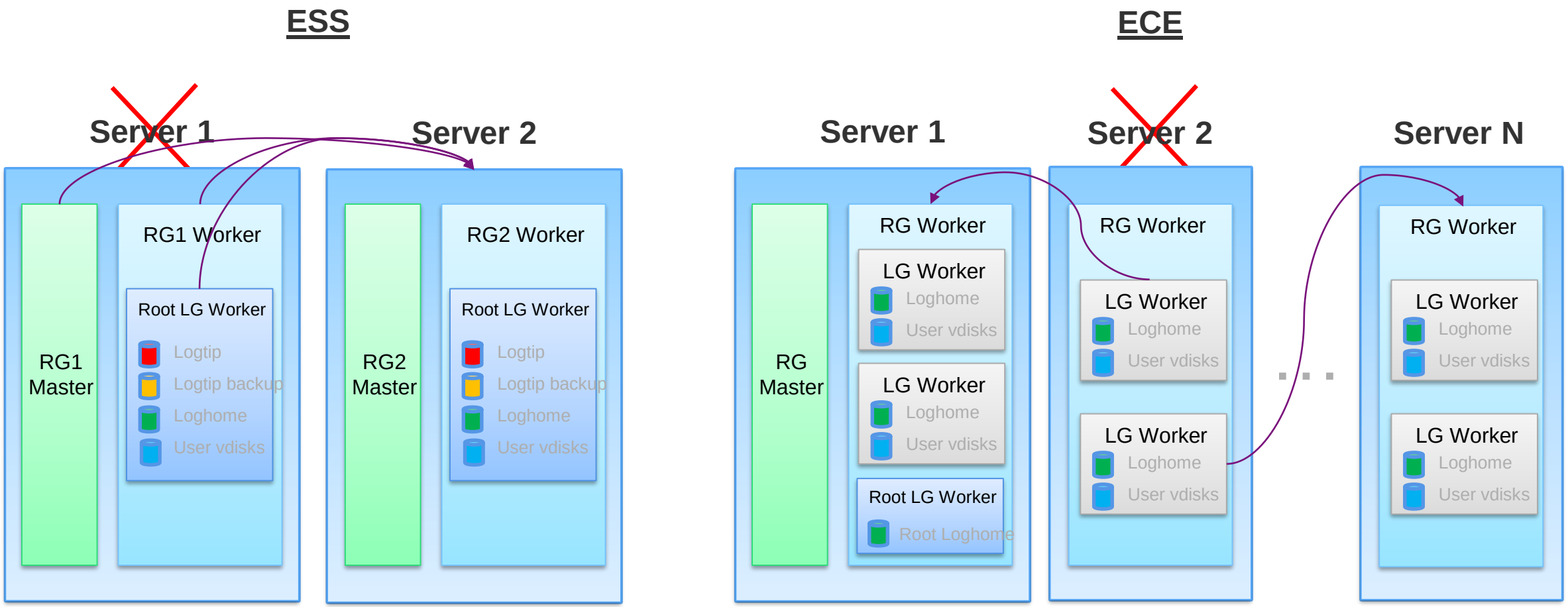
ECE Cluster



NSD I/O Path


NSD Client 1
NSD Client 2
NSD Client N
ESS
ECE
NSD/GNR Server
NSD/GNR Server
NSD/GNR Server
NSD/GNR Server
NSD/GNR Server
NSPD Client
NSPD Client
NSPD Client
User vdisk I/O goes through local SAS link
Remote logtip I/O goes through NSPD protocol like ECE
Remote User vdisk I/O goes through NSPD protocol
Remote loghome I/O goes through NSPD protocol
Local user vdisk I/O goes through local SAS link
Local loghome I/O goes through local SAS link
NSPD Server
NSPD Server
NSPD Server
Pdisk
Pdisk
Pdisk
Pdisk
Pdisk
Pdisk
Pdisk
Pdisk

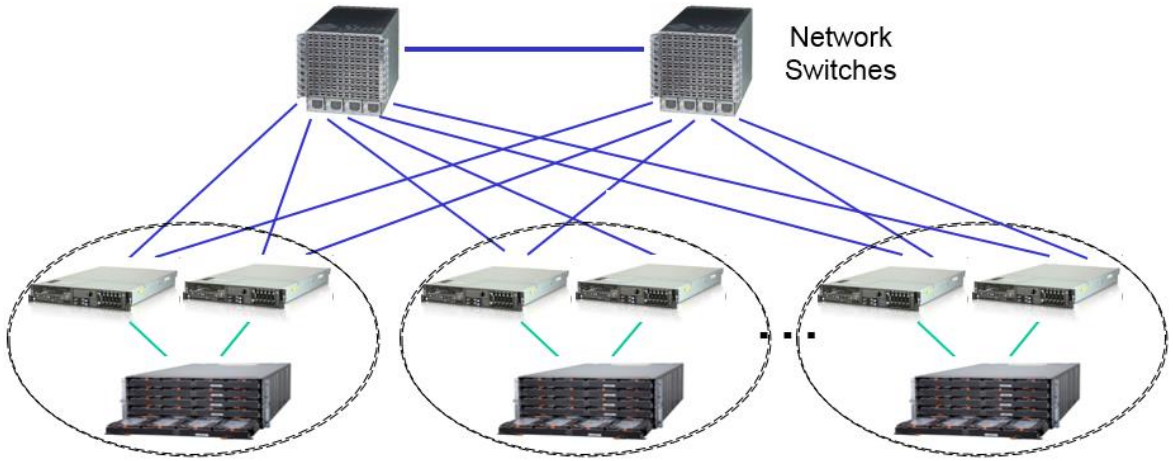
Software Failover Architecture



Scale-out Cluster with Multiple Building Blocks

ESS

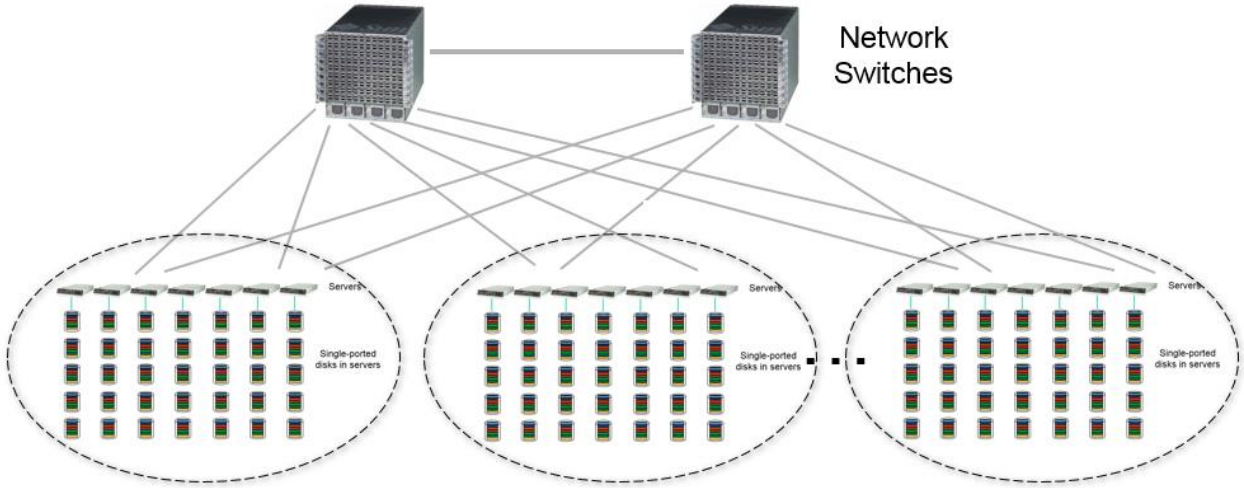
- Twin-tailed disks, dual servers building block
- Multiple ESS building blocks in the same Scale cluster



ESS Building Blocks

ECE

- Commodity servers based ECE cluster (building block)
- Multiple ECE clusters in the same Scale cluster



ECE Building Blocks

Licensing

ECE Licensing

- Spectrum Scale ECE is licensed by the usable TiB
  - usable capacity defined as the capacity presented to Linux, before applying erasure coding.
  - Thus the license pricing is independent of any choice of Error Correction width.
- Spectrum Scale ECE can also be licensed by the usable PiB with a discount
- ECE licenses ordered via Passport Advantage
  - The parts are Restricted initially pending review of a client’s requirements and design

Thank you

ECE FAQs

- **Can I buy ECE and use the licenses both with and without the erasure coding capability, i.e. both internal disk and SAN configurations?**
- Yes. An ECE license can be applied to ECE, DME or DAE. You will be compliant with licensing provided the total TBs deployed across all Editions does not exceed your ECE entitlement.

- **Will an ECE client be allowed (if the need arises in the future) to use its licenses within an IBM ESS appliance? If so, will the required capacity continue to be calculated as it is today with DAE/DME/Suite, i.e. as the net / RAIDed capacity reported by the ESS GUI?**
- Yes. ECE licenses will be treated the same as DME licenses.

- **Can ECE be offered within an ELA, provided the client meets OM's technical & business prerequisites?**
- Yes. Please contact Spectrum Scale offering management or Finance for approval of the Restricted part.

- **Can ECE be offered within an ESA, provided the client meets OM's technical & business prerequisites?**
- Yes. Note that inclusion of any edition of Scale within an ESA is subject to approval by Offering Management

- **Which options will IBM offer to trade-up existing Scale licensees to ECE from (a) DME/Suite, (b) DAE, (c) Advanced, (d) Standard?**
- Existing licenses for any Edition of Scale can be traded up. Trade-ups from DME or DAE will be "one TB for one TB" and based on the difference in price between the editions. Trade-ups from Advanced or Standard are similar to the existing process for trading up to DME

- **Are clients allowed to deploy ECE in the same Scale cluster as (a) DAE, (b) DME, (c) ESS DAE, (d) ESS DME? How are the current rules governing Multi-Clustering of different Scale Editions affected by the introduction of ECE?**
- The same rules apply to ECE as to DME. Different editions cannot exist in the same cluster. Multi-clustering is supported, with the same limitations as for DME. See the Knowledge Center.

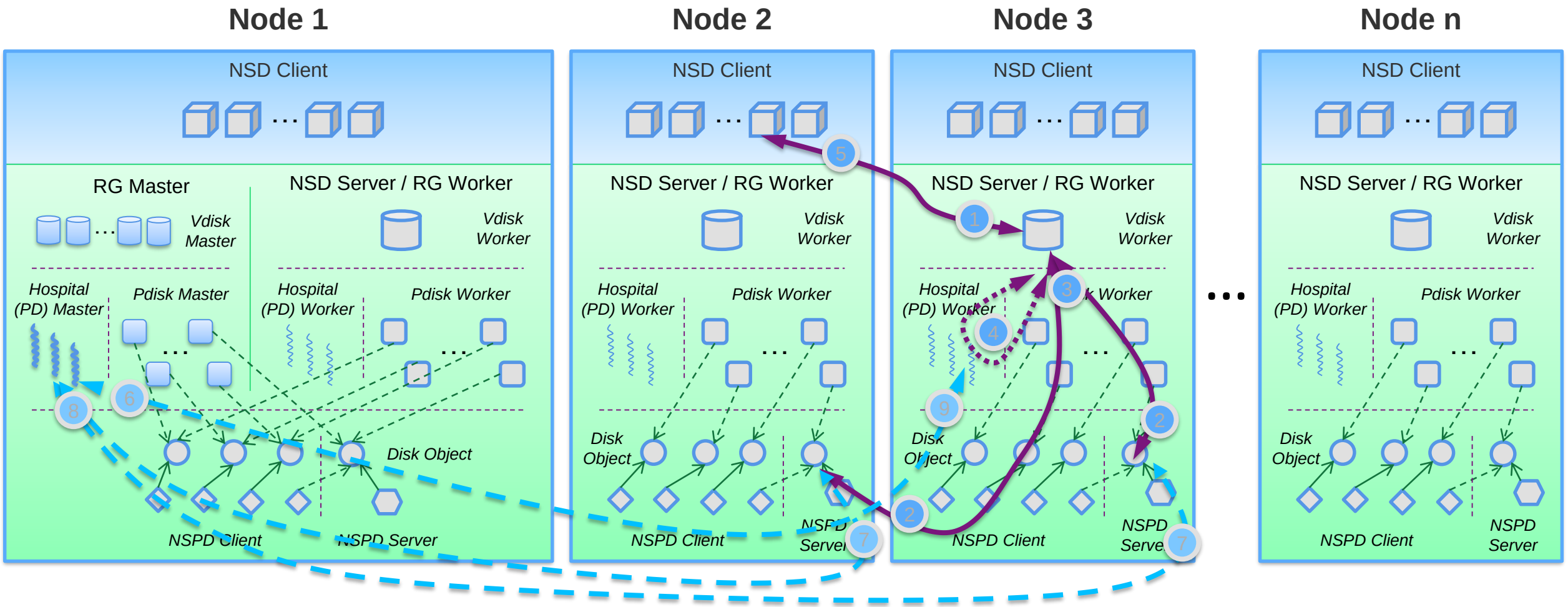
- **Will the FPO feature in Scale DAE and DME continue to be supported, and if so for how long?**
- IBM does not currently plan to deprecate or remove this capability in a subsequent release of the product; it will remain supported and current with updates to the operating systems. Customers do not need to change any of their existing applications and scripts that use FPO at this time. They should not expect significant new functionality or enhancements to FPO.

- **Can ECE (a) be fully managed and accessed from containers through SEC/CSI, (b) be run itself in container mode?**
- ECE will support containerized workloads today and containerization in the future in the same way as other editions. See the Knowledge Center for details.

- **For more information, on Slack, join the #scale-ece channel in the IBM Storage Systems workspace**

Backup

I/O & Pdisk Diagnosis Path Illustrated



GNR Definitions

mmvdisk: The command suite for simplified GNR/vdisk administration.

Disk: A block storage device. Examples include SSDs and HDDs.

Pdisk: An abstraction of a disk that encompasses all the physical paths and properties of the disk.

NSD: The abstraction of a filesystem disk used by GPFS.

Filesystem: A GPFS filesystem is striped across a collection of NSDs.

Server: A GPFS cluster node that has a number of disks available to it, and serves abstractions based on those disks, both to other servers as pdisks and to GPFS as filesystem NSDs.

Recovery group: A recovery group is a collection of pdisks and servers. Filesystem NSDs called *vdisks* may be created within a recovery group, and may be configured to have various levels of data protection, including tolerance and correction of disk errors, and tolerance and recovery of disk and server failures.

Server set: All the servers within a recovery group.

Declustered array: A declustered array is a subset of the pdisks within a recovery group that all share similar characteristics, such as size and speed. A recovery group may contain multiple declustered arrays, which may not overlap (that is, a pdisk must belong to exactly one declustered array).

Vdisk: An erasure-code-protected virtual NSD partitioned among the pdisks of a declustered array of a recovery group, and served by one of the recovery group servers.

Log group: A log group is a subset of the vdisks within a recovery group that all share the same transaction log. A recovery group may contain multiple log groups, which may not overlap (that is, a vdisk must belong to exactly one log group). All of the vdisks in a log group are served by one of the recovery group servers, which then is known as the log group server.

Vdisk set: A vdisk set is a collection of vdisks with identical sizes and attributes, one in each log group across one or more recovery groups. Vdisk sets are externally managed according to the conventions of the **mmvdisk** command. With vdisk sets, the **mmvdisk** command creates GPFS filesystems that are striped uniformly across all log groups.

https://tuna.pok.stglabs.ibm.com/perseuswiki/Vlestor_mmvdisk_command

Disclaimer

IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at IBM’s sole discretion.

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user’s job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.