IBM Spectrum LSF and IBM Spectrum Scale User Group Erasure Code Edition

Christopher D. Maestas
cdmaestas@us.ibm.com

Please Note

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

Notices and disclaimers

Notices and disclaimers continued

- Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.**
- The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.
- IBM, the IBM logo, ibm.com and [names of other referenced IBM products and services used in the presentation] are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml.

Demand for storage-rich servers is growing rapidly

Supplier mandates

- “We buy from Dell, HPE, 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 perception

- “We want the economic benefits of commodity hardware”
- “We don’t want to pay for high-end or even mid-range storage”

But conventional servers are no substitute for storage arrays

High failure rates of commodity hardware means poor durability

Long recovery time of traditional RAID impacts availability

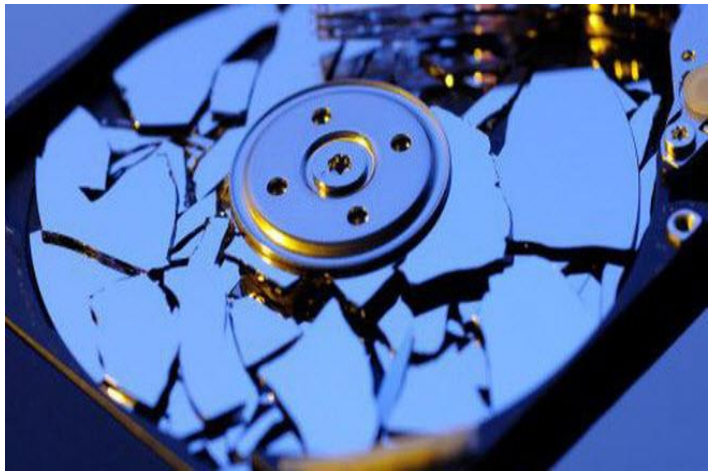
Enterprise storage features are missing
(life cycle management, snapshots, etc.)

Limited scalability – how to pool storage from a large number of servers?

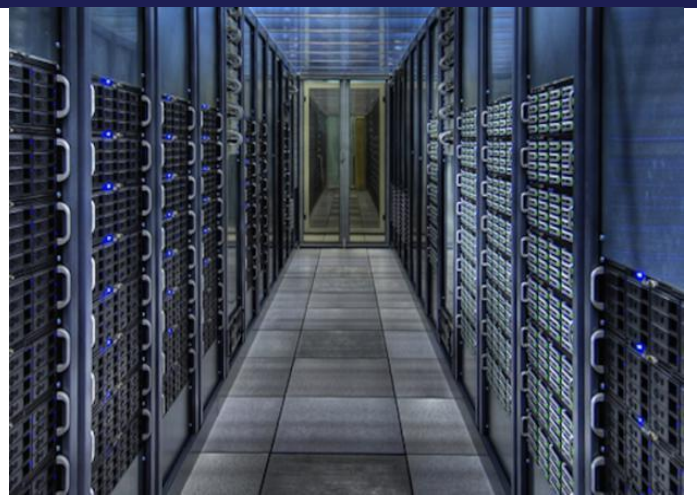
Limited performance of traditional file systems, often gated by the speed of a single server

Replication sacrifices storage utilization and increases costs
(just 33% to 50% available storage)

Very large server farms are tough to manage with constant break/fix



IBM Spectrum Scale Erasure Code Edition (ECE)



Delivers all the capability of
Spectrum Scale Data Management Edition

- Enormous scalability
- Very high performance
- Enterprise manageability

Plus: Durable, robust, and storage-efficient

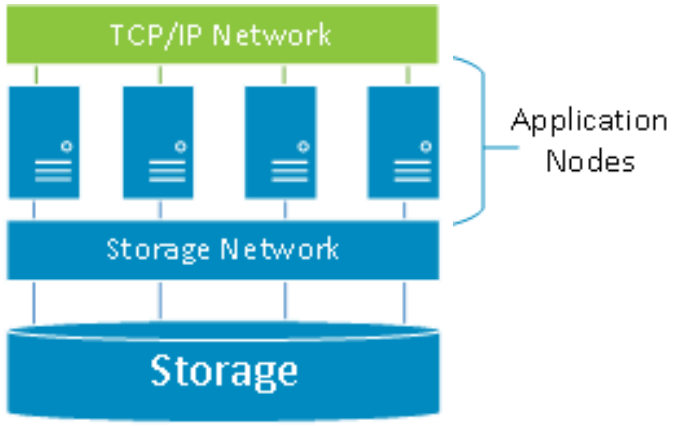
- Distributes data across nodes and drives for higher durability *without* the cost of replication
- End to end checksum identifies and corrects errors introduced by network or media
- Rapid recovery and rebuild after hardware failure

Plus: Delivered at hyperscale

- 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

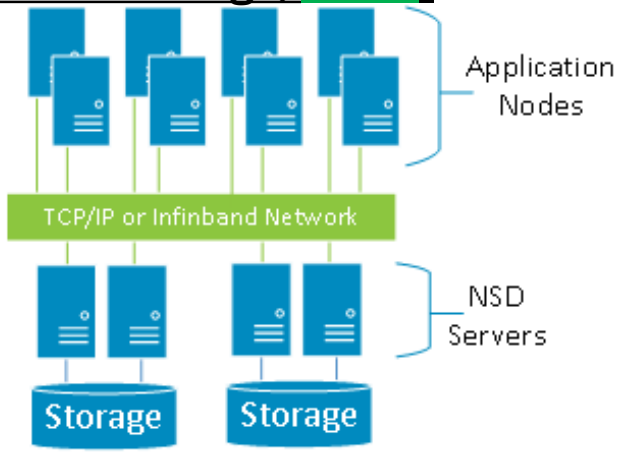
Spectrum Scale deployment models

Enterprise Integrated Model (SAN, NVMeoF, iSCSI)



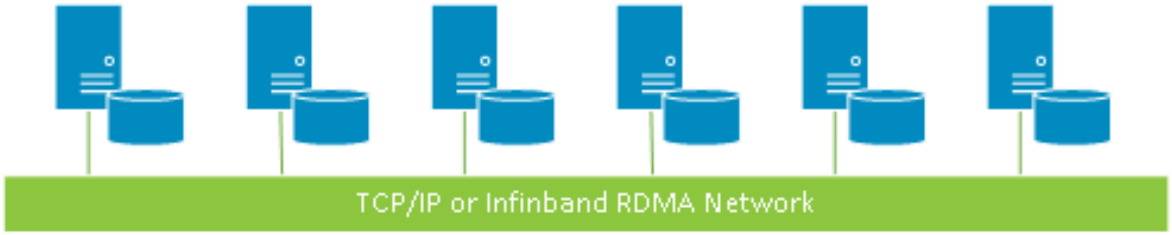
Unify and parallelize storage silos

Network Shared Disk (NSD) Model (twin tailed storage, ESS)



Modular High-Performance Scaling

Shared Nothing Cluster (SNC) Model (Storage Rich Servers (replication, erasure code edition))



Span storage rich servers for converged architecture or HDFS deployment

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

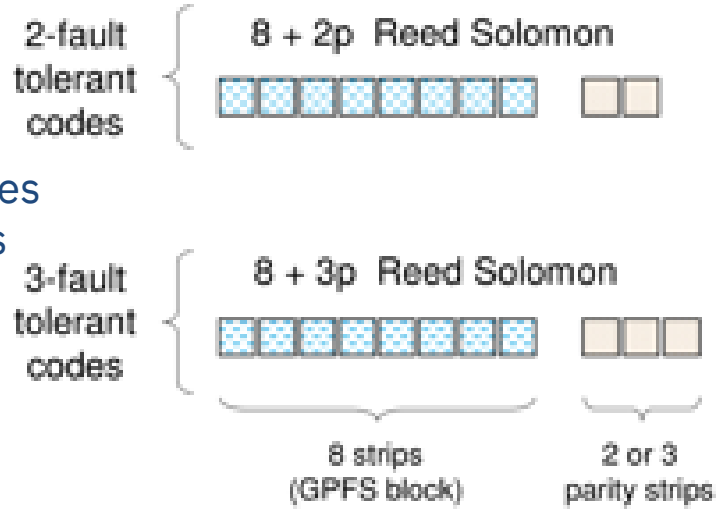
Fundamentals of erasure coding

Reed-Solomon erasure coding delivers efficient error detection and correction

- K strips of data plus N parity strips calculated using Reed-Solomon encoding functions
- Allows data to survive loss (erasure) or data error (correction) in up to N strips
- Reed-Solomon widely used where most likely error case is lost data, for example, a burst of interference in a satellite transmission, or lost/damaged storage media

When Spectrum Scale writes data blocks

- Calculate N parity strips and store K + N total strips
- Distribute data and parity strips as widely as possible across racks, servers and drives in order to minimize impact of any failure: can survive loss of any N servers or drives
- Failure domains provide for high hardware failure tolerance



When Spectrum Scale reads data blocks

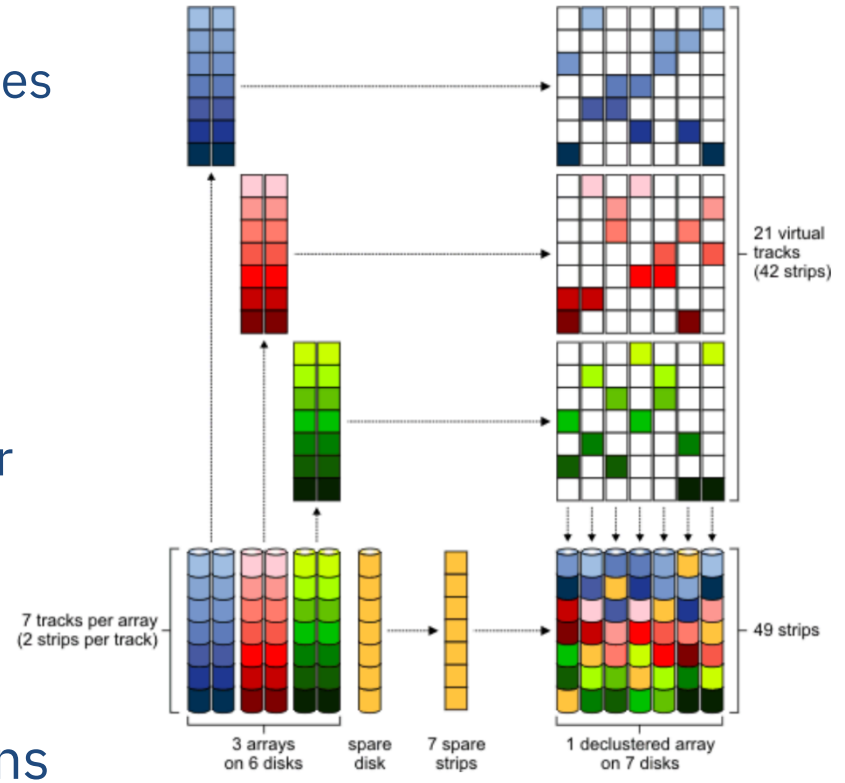
- Normal case is to read and aggregate the K data strips adding no extra overhead
- Only read parity strips and rebuild data when a lost or corrupt strip is detected

Spectrum Scale erasure code advantages

Faster and smarter rebuild operations compared to RAID arrays
- Uses many drives in parallel, distributes work across many nodes
- Normal rebuilds have minimal impact on system performance
- Critical rebuilds complete in minutes
- Rebuilds can be deferred with sufficient protection

Improved storage efficiency compared to replication
- 8+2P and 8+3P offer 25% - 38% overhead vs 100% - 200% for replication
- 4+2P and 4+3P also supported
- Spare capacity is also distributed across all drives and nodes

Higher performance than traditional erasure code implementations
- Patented strategies optimize IO data paths, read and multi-layer write caching
- Suitable for analytics, AI, and demanding read/write workloads, not just read-heavy workloads or cool data



Disk Hospital delivers hardware manageability at hyperscale

Identifies device problems before hard drive failure
- Dead disks
- Connectivity issues
- Media errors
- Slow drives

Attempts corrective action to revive sick or failing devices
- Power cycle non-responsive drives
- Recompute and rewrite corrupted data
- Rediscover disk connectivity

Maintains “health record” for each device
- If device is accumulating too many errors, remove from service
- If device is persistently slow, remove from service

Data Integrity is automatically managed with end-to-end checksums

Every IO has a checksum added to data trailer

For writes, verifies checksum when data passes from

- Client (compute node) to storage node
- Storage node to storage media
- Write also include a sequence number in metadata to detect dropped/skipped writes

For reads, verifies checksum when data passes from

- Storage media to storage node
- Storage node to client

Background scrub task also 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

Testing and preparing for ECE https://github.com/IBM/SpectrumScaleTools

```
# ./mor.py -h
usage: mor.py [-h] --ip IPv4_ADDRESS [--path PATH/] [--no-cpu-check]
              [--no-md5-check] [--no-mem-check] [--no-os-check]
              [--no-packages-check] [--no-net-check] [--no-storage-check]
              [--no-sysctl-check] [--toolkit] [-v]

optional arguments:
  -h, --help            show this help message and exit
  --ip IPv4_ADDRESS     IP address linked to device used for NSD
  --path PATH/          Path ending with / where JSON files are located.
                        Defaults to local directory
  --no-cpu-check        Does not run CPU checks
  --no-md5-check        Does not check MD5 of JSON files
  --no-mem-check        Does not run memory checks
  --no-os-check         Does not run OS checks
  --no-packages-check   Does not run packages checks
  --no-net-check        Does not run network checks
  --no-storage-check    Does not run storage checks
  --no-sysctl-check     Does not run sysctl checks
  --toolkit             To indicate this is being run from Spectrum Scale
                        install toolkit
  -v, --version         show program's version number and exit
```

This tool assesses the readiness of a single node to run IBM Spectrum Scale Erasure Code Edition (ECE). This tool only checks for requirement of a system that run ECE, no other software or middleware on top in the same server.

This tool is run when installing ECE with the Spectrum Scale toolkit, it is used by the toolkit to do a more comprehensive inter node checking from a cluster perspective, this tool does only check at node level. Each run it generates a JSON file with name IP_ADDRESS.json where some data is saved, on standalone mode this file is only for reference.

The tool requires the packages that are listed on on packages.json, in addition to those would need nvme-cli if NVME are installed and storcli if SAS card is installed.

The tool requires one parameter (--ip) to be passed, it has to be the IP where RAID traffic is going to happen. It does not allow names of a node it must be an IPv4 address

Testing and preparing for ECE https://github.com/IBM/SpectrumScaleTools

```
# ./koet.py -h
usage: koet.py [-h] [-l KPI_LATENCY] [-c FPING_COUNT] [-m KPI_THROUGHPUT]
               [-p PERF_RUNTIME] [-v]

optional arguments:
  -h, --help            show this help message and exit
  -l KPI_LATENCY, --latency KPI_LATENCY
                        The KPI latency value as float. The maximum required
                        value for certification is 1.0 msec
  -c FPING_COUNT, --fping_count FPING_COUNT
                        The number of fping counts to run per node and test.
                        The minimum required value for certification is 500
  -m KPI_THROUGHPUT, --min_throughput KPI_THROUGHPUT
                        The minimum MB/sec required to pass the test. The value
                        has to be at least 10 seconds. The minimum required
                        value for certification is 2000
  -p PERF_RUNTIME, --perf_runtime PERF_RUNTIME
                        The seconds of nsdperf runtime per test. The value has
                        to be at least 30 seconds. The minimum required value
                        for certification is 1200
  -v, --version         show program's version number and exit
```

This tool will run a network test across multiple nodes and compare the results against IBM Spectrum Scale Key Performance Indicators (KPI). This tool attempts to hide much of the complexity of running network measurement tools, and present the results in an easy to interpret way.

Note:

You need to first populate the hosts.json file with the IP addresses of the nodes to participate in the test. Node names are not allowed.

This test can require a long time to execute, depending on the number of nodes. This tool will display an estimated runtime at startup.

What Hardware should I use?

Processor:	**x86_64 – 8 cores min**
Memory:	**Greater than 64 GB**
Drives per node:	**24** **– For NVMe Drive Types 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 requirements**
Min OS:	**RHEL 7.5 or 7.6**
Servers system disk:	**Recommend 100 GB configured as RAID1**
SAS Host Bus Adapter:	**LSI SAS HBA** **models SAS3108, SAS3216, or SAS3516**
SAS Data Drives:	**SAS or NL-SAS HDD or SSDs**
NVMe Data Drives:	**Enterprise class NVMe drives with U.2 form factor**
Fast Drive Requirement:	**At least one additional SSD or NVMe drive for IBM ECE logging per server**
Network:	**25 Gb Ethernet min or Infiniband with RDMA**

ECE notes 1

- Do you have extra nodes outside of ECE servers if you are using them for CES, AFM, TCT, GUI, Protect, Archive?
- Do you have extra nodes for helper nodes?
- Do you or your IBM/BP partner have experience installing and managing Spectrum Scale?
- Are you prepared to run separate networks for NSD Clients, NSD Storage communication (Erasure Coding) and CES (if using protocols)?
- Can you use jumbo frames (MTU 9000) or RDMA?
- Will you connect to an existing Scale (5.0.3.1 min) or ESS (5.3.4 min)?

Thank You.

IBM Storage & SDI