

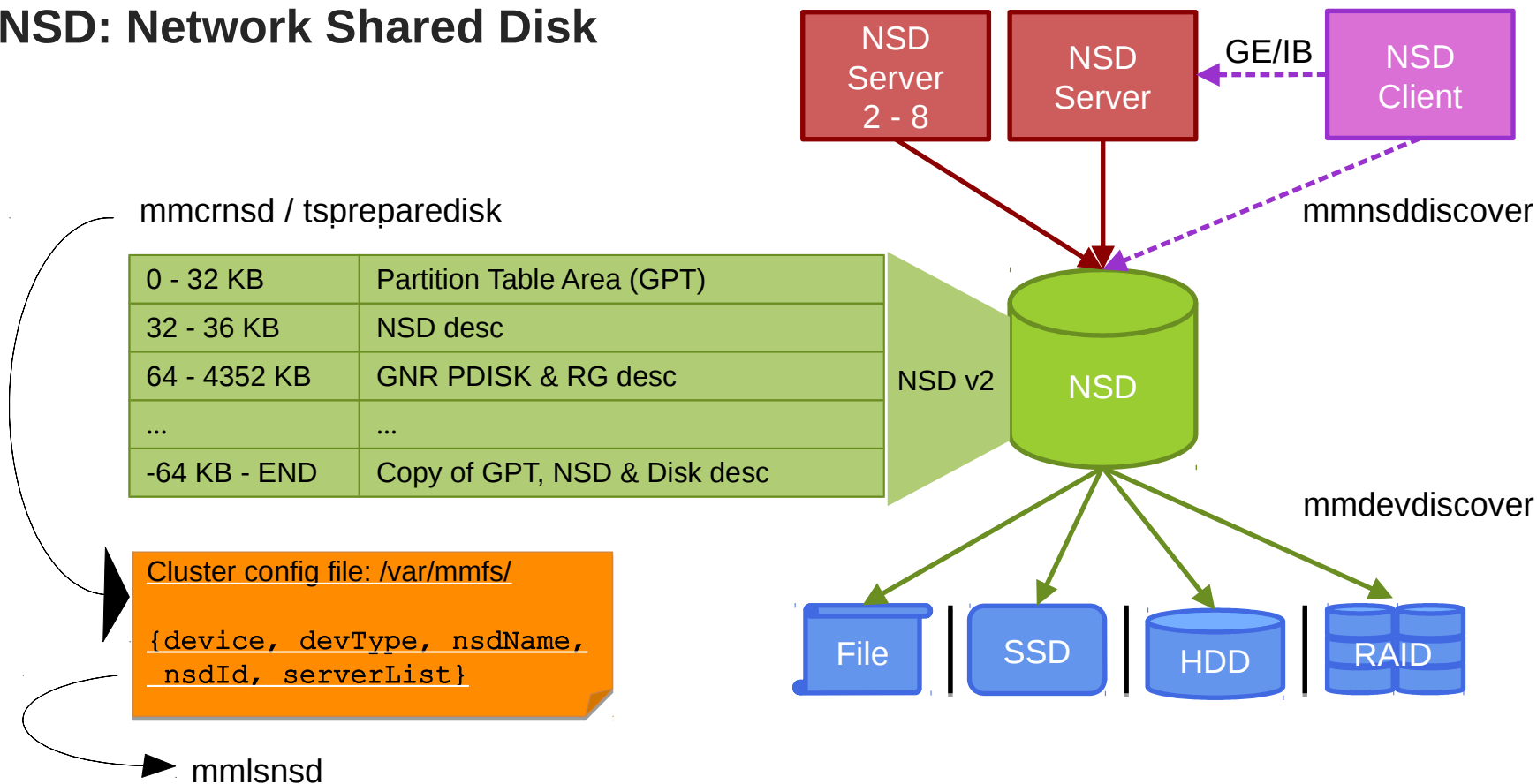


IBM Spectrum Scale

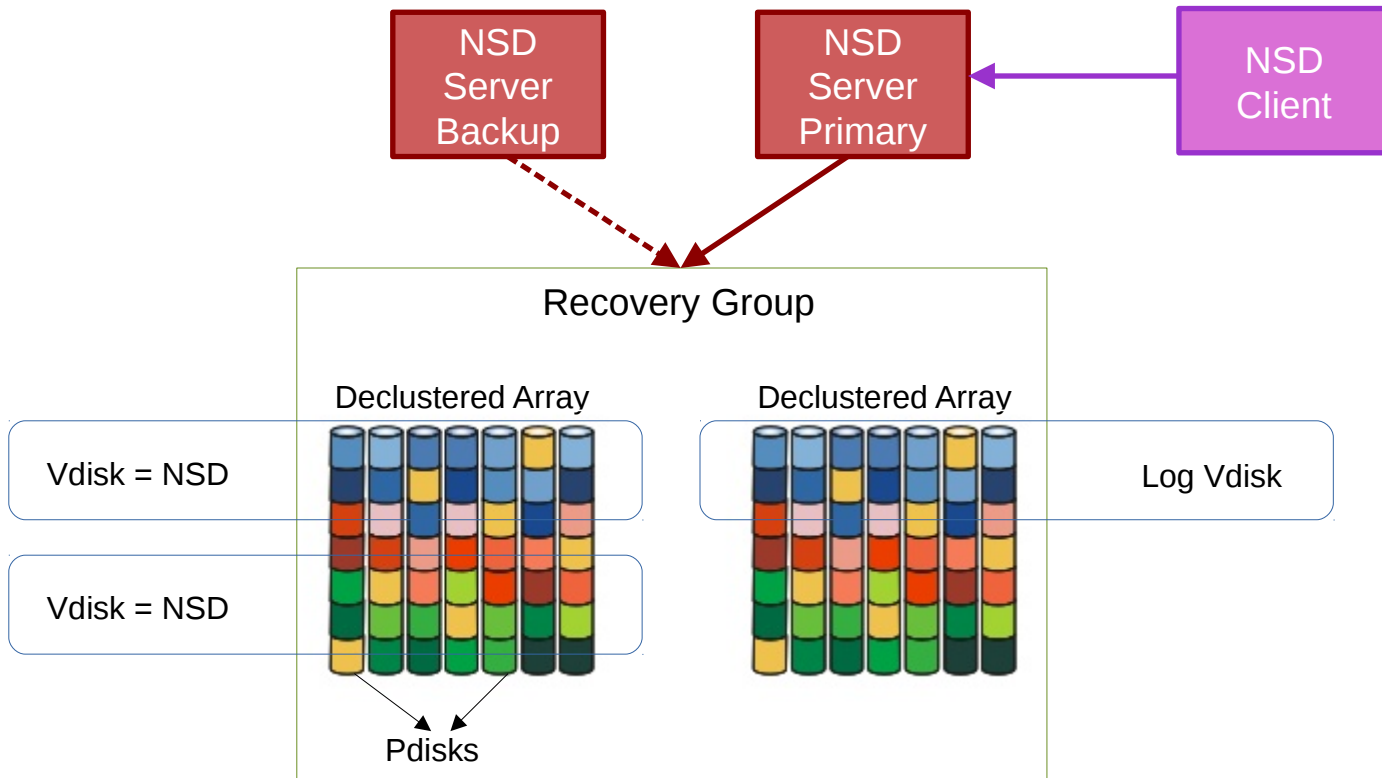
Problem Determination : Disks

Karthik Iyer
Spectrum Scale Development

NSD: Network Shared Disk



ESS: Elastic Storage Server



Disk Problems

- NSD Server Down
- Disk Down
- Corruption
- Performance
- Out of Space



NSD Server Down

- Symptoms
 - File system panic, Node Assert
 - E_IO, E_NODEV, E_DISK_UNVAIL, E_ALL_UNAVAIL
 - Forced unmount
 - Log recovery failure
- Detection
 - mmgetstate -N <nsd_servers>
 - Use mmhealth event notification
 - IBM Spectrum Scale GUI : Monitoring > Events page
 - mmhealth node eventlog command

NSD Server Down: Reason & Resolution

- Node expel
 - Backup NSD servers
 - Transparent failover and failback
 - mmchdisk suspend
 - FPO
 - Auto recovery suspends the disks if it fails to change the availability of the disk to Up
- Client nodes start before NSD Server
 - mmchconfig nsdServerWaitTimeForMount (seconds to wait)

Disk Down

- Symptoms
 - File system panic, Node Assert
 - E_IO, E_NODEV, E_DISK_UNVAIL, E_ALL_UNAVAIL
 - Forced unmount
 - Log recovery failure
 - Data loss depends on replication
- Detection
 - mmlsdisk -e
 - Status { Ready, ToBeEmptied, Emptied }
 - Availability { Up, Down, Recovering, Unrecovered }
 - SCSI errors in syslog
 - MMFS_DISKFAIL, MMFS_SYSTEM_UNMOUNT

Disk Down: Reason & Resolution

- Disk path issue
 - Use multipathed device
 - multipath -ll
 - mmnsddiscover
 - Backup NSD servers
 - Transparent failover and failback
 - Never use local path
 - mmmount -o useNSDserver=always

Disk Down: Reason & Resolution (2)

- Media failure
 - Use Spectrum Scale RAID for Disk Hospital
or
 - FPO
 - mmchconfig restripeOnDiskFailure
 - Auto recovery suspends the disks and restripes the data off the suspended disks
or
 - mmchdisk suspend
 - mmdeldisk -p / Offline mmfsck
 - Interesting inodes file
or
 - mmmount -o ro
 - Copy out what is possible and recreate file system

Disk Down: Reason & Resolution (3)

- NSD descriptor overwritten
 - `mmlsnsd -X #` get NSD ID of bad disk
 - `tspreparedisk -F -n <device> -p NSDId #` patch in the NSD ID into NSD descriptor
 - `mmfsadm test readdescraw <device> #` verify descriptors
 - `mmdevdiscover`
 - or
 - `# NSD V2 supports backup (-B) and restore (-R)`
 - `tspreparedisk -[B|R] /tmp/desc.copy <device>`
 - `mmdevdiscover`

Corruption

- Symptoms

- File system panic, Node Assert
- E_IO
- Forced unmount
- Log recovery failure
- Data loss depends on replication

- Detection

- Sector Error: disk surface check tool
- MMFS_FSSTRUCT
- NSD checksum
 - mmchconfig nsdCksumTraditional : client initiated
 - mmchconfig nsdDumpBuffersOnCksumError : dump data related to NSD checksum errors
 - mmfsadm dump nsdcksum : statistics on client and server nodes
 - mmfsadm dump cfgmgr : for read and write checksum errors per node

Corruption: Reason & Resolution

- Bad sectors
 - Use Spectrum Scale RAID for Disk Scrubbing
or
 - mmchdisk suspend
 - mmfileid
 - Identifies files that are on areas of a disk that are damaged or suspect
 - Outputs {InodeNumber LogicalDiskAddress SnapshotId Filename}
 - mmchdisk start
 - dd conv=sync,noerror
 - Delete files / (tsdbfs & mmfsck)
 - mmdeldisk
 - mmdeldisk -c
 - Interesting inodes file

Corruption: Reason & Resolution (2)

- Sector overwrite, lost write, network corruption
 - Use Spectrum Scale RAID for End to End Checksum
or
 - `mmchconfig nsdRAIDAssertOnNsdChecksumError`
 - Trigger an assert when a checksum error occursor
 - Offline `mmfsck`

Performance

- Symptoms

- NSD RPC timeout
- I/O hang, MMFS_LONGDISKIO

- Detection

- Monitoring > Statistics page in the IBM Spectrum Scale GUI
- mmperfmon query; mmdumpperfdata
- mmdiag --waiters
- iostat, mmdiag -iohist
- mmfsadm dump threads
- Test Sequential Performance with gpfsperf

[IBM Spectrum Scale Wiki > Performance > Testing File System Performance](#)

- Test network performance with nsdperf

[IBM Spectrum Scale Wiki > Network Configuration > Testing network performance with nsdperf](#)

- mmcallback diskIOHang (> 5 mins)

Performance: Reason & Resolution

- NSD client nodes not using VERBS RDMA
 - `mmlsconfig | grep verbsRdma`
 - `mmfsadm test verbs status`
- Uneven NSD server workload
 - Ensure client I/O is equally distributed across all the NSD servers
- Sub-optimal block allocation method
 - Cluster allocation method diminish as number of nodes or disks increases
 - Scatter allocation method provides more consistent file system performance by averaging out performance variations
- Network needs tuning

[IBM Spectrum Scale Wiki > Network Configuration > Best Practices Network Tuning](#)

Performance: Reason & Resolution (2)

- Incorrect NSD server settings
 - $\text{pagepool} = 3 * \# \text{LUNS} * \text{maxBlockSize} * (100 / \text{nsdBufSpace})$
 - NSD server does not cache any of the data
 - $\text{nsdMaxWorkerThreads}$ = scale with number of LUNs
 - Maximum number of NSD threads that will be concurrently transferring data.
 - Manages flow control on the network between the clients and the servers.
 - $\text{nsdSmallThreadRatio}$ = determined by workload I/O size
 - Ratio of NSD server queues for small vs large I/Os
 - $\text{nsdThreadsPerQueue} = \text{nsdMaxWorkerThreads} / (\text{number of queues needed})$
 - Number of threads assigned to process each NSD server I/O queue
- Incorrect NSD client settings
 - maxMBpS = twice the IO throughput expected for the node
 - Hint to calculate how many prefetch/writebehind threads should be scheduled

Performance: Reason & Resolution (3)

- Hung I/O
 - `mmdiag -deadlock`
 - Disk lease timeout
 - Dead man switch panics the node kernel if there is any outstanding I/O requests
 - `mmchconfig panicOnIOHang`
 - Panics kernel when a local I/O request pending in the kernel for more than five minutes.

Out of Space

- Symptoms
 - ENOSPC
 - Recovery failure
 - Slow allocation
- Detection
 - Use Spectrum control for Capacity Management
 - Identify when a file system runs out of space
 - View size and age of snapshots
 - View file system relationships, quotas, NSDs, compare pools across cluster

Out of Space: Reason & Resolution

- Quota information not enabled/missing/incorrect
 - mmchfs -Q
 - mmlsquota (in-doubt space)
 - mmcheckquota
- Too many snapshots
 - mmlssnapshot -d
 - mmdelsnapshot (oldest first)
 - Snapshot ILM
 - RULE 'MigToCap' MIGRATE FROM POOL 'SSD' TO POOL 'Capacity'
- No ILM
 - mmdf
 - Migrate to tape/cloud with IBM Spectrum Protect
- Disk deletion
 - Add disk before delete

Out of Space : Reason & Resolution (2)

- Storage pool full
 - mmlspool
 - mmadddisk
 - Change placement policy
- Fragments
 - mmdf
 - mmdefragfs
- Lost blocks
 - mmfsck -o
- Max inode limit reached
 - mmchfs --inode-limit <new_limit>
- Inode expansion
 - Delete independent filesets

References

- Product Documentation
 - Planning for GPFS: Disk considerations, Recoverability considerations
 - Troubleshooting: Disk issues, File system issues, Performance issues
 - Administering: Managing disks
 - IBM Spectrum Scale RAID: Administration
- IBM Developer Works
 - Forum: General Parallel File System (GPFS)
 - IBM Spectrum Scale Wiki



IBM Spectrum Scale

Thank You

karthik.iyer@in.ibm.com