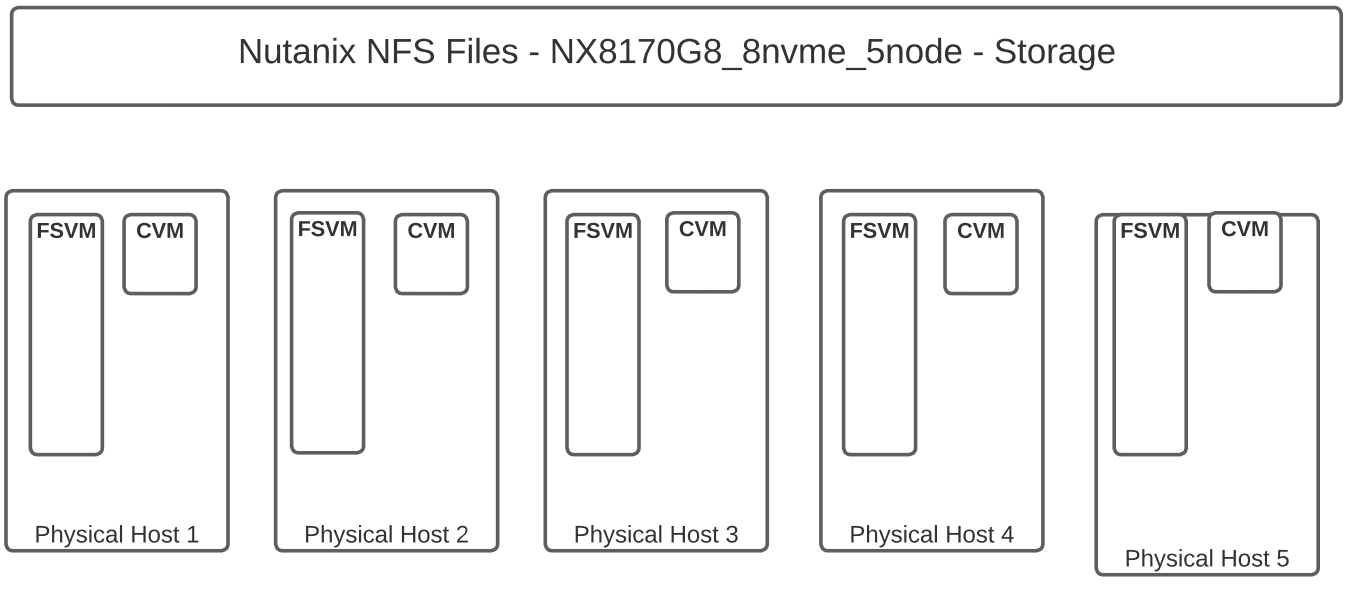

Workload Optimizations to achieve 13 GPUs per FSVM at around 92% utilization

CVM

- 32 vcpus
- 64 GB

FSVM

- 24 vcpus
- 256GB memory
- Numa pinned
- Two virtual NICs each with 8 queues each
- Interrupt sharding using custom script
- Files Version: 4.3.0

Share & Data layout

- Distributed share
- One top level directories per FSVM
- 14 vdisks Per FSVM for the Share
- Sequential workload type for share

Clients

- Ubuntu OS - Ubuntu 22.04.2 LTS - Kernel: 5.15.0-76-generic
- 64 vcpus
- 180GB memory
- NUMA pinned
- One clients per node
- 8 NIC queues
- swap off
- NFS mount with v4.0, nconnect=16
- sunrpc.tcp_slot_table_entries=32

Example Benchmark Run Command:

./benchmark.sh run --workload unet3d --num-accelerators 13 --param dataset.data_folder=/mnt/NTNX_NFS_Files/client1fs52 --param dataset.num_files_train=25000 --param checkpoint.checkpoint_folder=/mnt/NTNX_NFS_Files/client1fs52 --param reader.read_threads=4 --param dataset.num_subfolders_train=20

Example NFS mount command from Linux client:

sudo **mount**-t nfs 10.57.116.51:/s1 /mnt/NTNX_NFS_Files -o nconnect=16 -o vers=4.0