Book of Basics - Drive Breakdown

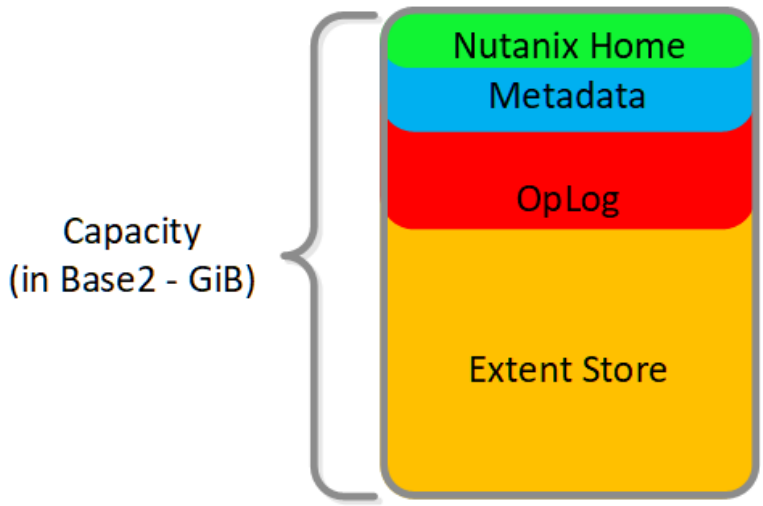
In this section, I'll cover how the various storage devices (Performance (NVMe/SSD) / Capacity (SSD/HDD) are broken down, partitioned, and utilized by the Nutanix platform. NOTE: All of the capacities used are in Base2 Gibibyte (GiB) instead of the Base10 Gigabyte (GB). Formatting of the drives with a filesystem and associated overheads has also been taken into account.

Performance Disk Devices

Performance devices are the highest performance device in a node. These can be NVMe or a mix of NVMe and SSD devices. They store a few key items, as explained below

- Nutanix Home (CVM core)
- Metadata (Cassandra / AES storage)
- OpLog (persistent write buffer)
- Extent Store (persistent storage)

The following figure shows an example of the storage breakdown for a Nutanix node's performance device:



Performance Drive Breakdown

Graphics and proportions aren't drawn to scale. When evaluating the Remaining GiB capacities, do so from the top down. For example, the Remaining GiB to be used for the OpLog calculation would be after Nutanix Home and Cassandra have been subtracted from the formatted SSD capacity.

Nutanix Home is mirrored across the first two SSDs to ensure availability and has a 60GiB reservation for two devices.

As of 5.0 Cassandra is sharded across multiple SSDs in the node (currently up to 4) with an initial reservation of 15GiB per SSD (can leverage some Stargate SSD if metadata usage increases). In dual SSD systems, metadata will be mirrored between the SSDs. The metadata reservation per SSD is 15 GiB (30GiB for dual SSD, 60GiB for 4+ SSD).

Prior to 5.0, Cassandra was on the first SSD by default, if that SSD fails the CVM will be restarted and Cassandra storage will then be on the 2nd. In this case the metadata reservation per SSD is 30 GiB for the first two devices.

The OpLog is distributed among all SSD devices up to a max of 12 per node (Gflag: max*ssds*for_oplog). If NVMe devices are available, OpLog will be placed on those devices instead of SATA SSD.

The OpLog reservation per disk can be calculated using the following formula: MIN(((Max cluster RF/2)*400 GiB)/ numDevForOplog), ((Max cluster RF/2)*25%) x Remaining GiB). NOTE: The sizing for OpLog is done dynamically as of release 4.0.1 which will allow the extent store portion to grow dynamically. The values used are assuming a completely utilized OpLog.

For example, in a RF2 (FT1) cluster with 8 SSD devices that are 1TB the result would be:

- MIN(((2/2)*400 GiB)/ 8), ((2/2)*25%) x ~900GiB) == MIN(50, 225) == 50 GiB reserved for Oplog per device.

For a RF3 (FT2) cluster this would be:

- MIN(((3/2)*400 GiB)/ 8), ((3/2)*25%) x ~900GiB) == MIN(75, 337) == 75 GiB reserved for Oplog per device.

For a RF2 (FT1) cluster with 4 NVMe and 8 SSD devices that are 1TB the result would be:

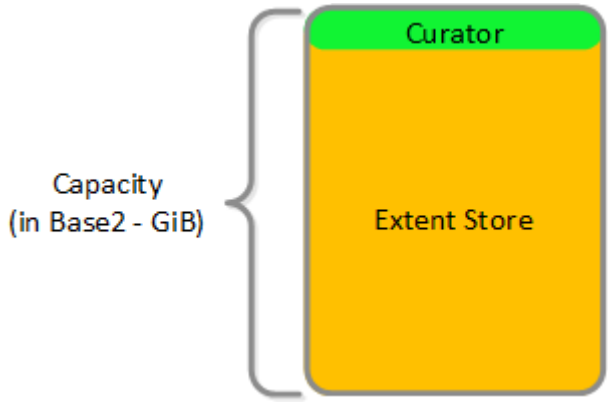
- MIN(((2/2)*400 GiB)/ 4), ((2/2)*25%) x ~900GiB) == MIN(100, 225) == 100 GiB reserved for Oplog per device.

The Extent Store capacity would be the remaining capacity after all other reservations are accounted for.

HDD Devices

Since HDD devices are primarily used for bulk storage, their breakdown is much simpler:

- Curator Reservation (Curator storage)
- Extent Store (persistent storage)



HDD Drive Breakdown