Petascale Storage Using Ceph And The BackBlaze Storage Pod

Tim Wickberg, Christopher D. Carothers, R. Lindsay Todd
Rensselaer Polytechnic Institute

Design Goals

- Inexpensive: $500,000
- Large capacity: Over 1 Petabyte Usable
- Fast - over 5 GB/sec write, 10GB/sec read
- Reliable - can loose any two storage servers and their disks

RAID Kills Performance, Use Replication Instead

A RAID array using 3TB hard drives will have at least one uncorrectable bit error during reconstruction. This drives the adoption of RAID-6.

But RAID kills performance. A single SATA disk provides 75 random 4k IOPS. A 4+2P RAID array, due to parity and stripe requirements, is limited to the same 75 IOPS write, and 300 IOPS read.

If you add replication to guard against the loss of any storage servers or arrays, you now have 12 disks as a replicated pair of RAID6 4+2P arrays providing only 75 IOPS write and 600 IOPS read, for 4/6 (RAID overhead) * 1/2 (replication overhead) = one third of the raw disk space.

Instead you can replicate data twice (2x replication). Each file written in is stored on three separate disk drives. Same one-third of the raw disks space available as usable space, but the write IOPS is no longer constrained by RAID. The same 12 disks under 2x replication instead can provide 300 IOPS write, and 900 IOPS read.

This is 300% better write performance, and 50% better read performance with the same hardware.

Ceph

Ceph is the only mainstream open-source parallel filesystem that can currently support the replication model required.

The CRUSH layer in Ceph allows us to control the data placement, such that replicas are spread out between storage pods and racks, to prevent any single pod failure from causing data loss.

BackBlaze Storage Pod

The BackBlaze Storage Pod is a commodity hardware approach to large scale storage systems.

There is no redundancy within the pod. But by aggregating a large number of these pods together with Ceph's replication mechaisms we can provide high availability at the system level.

Our pod design varies from the stock BackBlaze Pod:

- Different motherboard, with additional PCIe slots.
- 40Gbps Infiniband adapter, rather than using the Gigabit Ethernet.
- Reduced capacity, but better performance. Changing the SATA controllers from 3 4-port models to 4 2-port models gives us an additional 100MB/s path to the disks, at the loss of 1 SATA port (5 disks).
- Intel SSD added in for metadata storage.

Rensselaer
CCNI COMPUTATIONAL CENTER for NANOTECHNOLOGY INNOVATIONS
SCOREC SCIENTIFIC COMPUTATION RESEARCH CENTER

Proposed System

The proposed system includes 64 storage pods, each with 24x 3TB Hitachi Deskstar hard drives, and an Intel 300GB SSD.

Quantity	Component	Price	Total
64	Modified BackBlaze Pods	$4,500	$288,000
64 * 24 = 1536	3TB Hitachi Hard Drives	$120	$184,320
64	300GB Intel SSDs	$500	$32,000
			$504,320

This results in 4.7-Petabytes of raw disk storage, and 19.2-Terabytes of SSD storage for metadata.

After replication on both the filesystem (2x) and metadata (3x), the resulting system is expected to have 1.5-Petabytes of usable storage, 4.5-Terabytes of metadata storage.

Expected performance from the system including replication overhead is:

- 8.5 GB/sec write, 25GB/sec read
- 21k IOPS write, 64k IOPS read
- Metadata operations at 300k IOPS write, 2.5million IOPS read.

BackBlaze Pod - Photo by Chris Dag - http://www.flickr.com/photos/chrisdag/6074480108/