Tyrone®

Case Study:

Accomplishing redundancy on Lustre based PFS with DRBD

A premier research institute wanted a HPC along with Parallel File System (PFS) to be deployed which would have complete redundancy. The PFS was based on Lustre file system, which didn't have any redundancy features; this turned a simple HPC deployment to a challenging one for Netweb Technologies.

CHALLENGE:

- The institute sought a 48 node HPC with 4.5 Tera Flop performance along with a PFS of 30 TB storage with 100% redundancy and a throughput of 1.5 GB/s.
- The PFS being on Lustre file system was a challenge to build as Lustre didn't supported redundancy.

SOLUTION:

- High Performance Computing setup installed.
- Object Storage Servers used for setting up Parallel file System.

BUSINESS RESULTS:

- By incorporating DRBD and High Availability, the PFS based on a Lustre file system delivered by Netweb Technologies provides full redundancy.
- The 48 node cluster can simultaneously access the storage (PFS) with 1.5 GB/s throughput

DEPLOYMENT ENVIRONMENT:

- Tyrone storage array.
- Six object storage servers of 12 TB each and each having two volumes of 6TB configured on RAID 5.
- Infiniband switch between Meta data servers and object storage servers.

Being one of India's premier research institutes it is well-known across the world. The research institute provides a vibrant academic ambience hosting more than 200 Researchers. The research institute is funded by the Department of Science and Technology, Government of India and is a deemed university.

EXECUTIVE SUMMARY

Customer Name:

India's premier Scientific Research Institute

Industry:

Education and Research

Corporate Office:

Bangalore, India

Accomplishing Redundancy on Lustre Based PFS with DRBD

Challenges:

Being one of the renowned research institutes in India; the institute wanted to deploy a High Performance Computing (HPC) for their research and training purposes. Netweb Technologies was deputed to make this deployment of HPC at the research institute. The requirements set by the research institute for the HPC environment were straight forward; but to meet those requirements turned out to be a challenge, mainly for the Parallel File System deployment.

The research institute had put forth a requirement of 48 node HPC that would have a theoretical peak performance of 4.5 Tera flops and should have a Parallel File System (PFS) of 30 TB storage which would have 100% redundancy and a throughput of 1.5 GB/s. The setting up of the HPC 48 node cluster was straight forward for Netweb Technologies, but to provide the corresponding PFS storage posed a challenge.

Lustre, having its name derived from Linux and cluster, is the most popular file system which is used in HPC community and has the highest share in top Parallel File System (PFS) deployments worldwide. Therefore, Lustre by default was the choice of file system for the PFS deployment at the research institute. The only hindrance was that Lustre doesn't have any native support for redundancy. And since, redundancy was the primary requirement laid for PFS by the research institute, it posed the biggest challenge for Netweb Technologies to provide the solution with redundancy.

Solution:

A Parallel File System mainly constitutes of Meta Data servers and the connected storage nodes called as Object Storage servers. Meta Data servers store the location of the data which is actually residing on the physical storage nodes i.e. Object Storage servers. A PFS is usually used for large file storage, which is stored distributed across all the storage nodes. If suppose, a disk fails in any of the Object storage server, then not only the portion of data in that disk is lost, but also across all the nodes the complete data file becomes corrupted. To provide a redundancy to such a setup on Lustre file system required to look at possibilities beyond the traditional approach.

The setup planned by Netweb Technologies was to have their "twin servers" to act as Meta data servers. The "twin server" is approach where everything is on dual configuration inside one enclosure. This approach by Netweb Technologies is a novel way for space saving as well as to provide full redundancy within one box. And likewise for Object Storage servers, Netweb Technologies planned to have each node having two volumes; whereby one volume will be active and other in passive state. The passive state volume would be the mirrored image of active volume of the corresponding node. Thus, in a round-robin fashion each active volume in one node of the PFS will have a passive mirrored volume residing in another node of the PFS. This, way even if one volume in any node fails, the passive volume residing in another node could provide redundancy. But it was easier to plan than to accomplish, as Lustre didn't had any support for redundancy. To accomplish redundancy through RAID within a server/machine is easy, but across two different machines it is not possible with RAID. Therefore, Netweb Technologies were forced to find a technique through which redundancy could be achieved across cluster nodes.

DRBD to rescue:

To achieve redundancy, the technicians of Netweb Technologies used Distributed Replicated Block Devices (DRBD) which is normally used in High Availability (HA) clusters. DRBD layers logical block devices over existing local block devices on participating cluster nodes. DRBD ensures that the writes happening on block device on

Accomplishing Redundancy on Lustre Based PFS with DRBD

primary node are propagated simultaneously to the block device on secondary node. If for reasons, primary node fails; then the secondary node is promoted to the primary state. With this technique, redundancy could be made possible for the research institute's PFS solution.

Netweb Technologies was building 6 node PFS solution for the research institute, with each node of 12 TB, where each node had two volumes of 6 TB configured on RAID 5; where one volume was active while other was to be the mirrored image of another volume in the cluster node. By using DRBD, they were able to ascertain that, the active volume of first node is getting mirrored in the passive volume on second node and so on in a round robin fashion across all the 6 cluster nodes. Now, if any primary/active volume fails, then with DRBD the passive volume is made active and the data continuity continues without any hindrance. Thus, redundancy on the storage nodes in the PFS solution was achieved. Likewise, redundancy was provided on Meta Data servers, so that if one fails the other Meta data server which had the mirrored image could take over the functioning of the PFS solution.

Now, if any primary/active volume fails, then with DRBD the passive volume is made active and the data continuity continues without any hindrance. Thus, redundancy on the storage nodes in the PFS solution was achieved. Likewise, redundancy was provided on Meta Data servers, so that if one fails the other Meta data server which had the mirrored image could take over the functioning of the PFS solution.

Benefits:

The non-traditional approach to accomplish redundancy on PFS solution has helped the research institute to have a High Performance Computing cluster environment where they can achieve 4.5 Tera flops of performance and can access the storage system simultaneously from 48 nodes at 1.5 GB/s throughput.

- **100% Redundancy:** By using DRBD along with High Availability concepts, the PFS solution is designed to provide full redundancy which could occur either from disk failure or from a complete node failure. Even if a cluster node fails, its backup is in another node which gets promoted to the active state, without hindering any performance of the PFS and maintaining the data consistency. Even the hardware is designed to provide redundancy feature.
- **High Performance:** With Infiniband as the backbone, high throughputs can be achieved. So that even if all 48 nodes of the HPC access the storage system the performance doesn't get disrupted and a sustained 1.5 GB/s throughput is maintained.

* Disclaimer for Case Study

The case study is intended for informational purpose only pertaining to Netweb Technologies solutions. The cases cited here are real and customer names have been withheld, however to get detailed further information about the case kindly contact Netweb Technologies. Links to this case study from external sources are allowed, however any other re-distribution of this content for commercial purposes is strictly prohibited.