Designing and optimizing a cluster file system solution

Kalyana Chadalavada, Sanjay Lalwani, Amina Saify
Dell Inc.
{kalyana_chadalavada@dell.com, sanjay_lalwani@dell.com, amina_saify@dell.com}

Introduction:
With the ever increasing size of cluster solutions for High performance computing, apart from the system area network, file system has also been an area of interest in scaling the capability across a large system. There are multiple solutions available for providing a scalable file system solution. But the challenge lies in architecting the solution that will optimally utilize all the components of the system. This will include the back end storage, the IO nodes, the network, file system and the operating system parameters.

We would like to share with the users and solution architects of high performance computing clusters and file systems our experiences with designing a highly available multi-server scalable file system for medium to large clusters. For our exercise, we chose the IBRIX Fusion file system on the Intel based servers. The objective was to design a solution to meet general purpose criteria, which is described in detail below.

Objectives:
Design a highly available multi-server scalable cluster file system.

Test bed:
Intel Xeon based dual core dual socket servers
4Gbps Fibre Channel storage
Non-blocking enterprise class gigabit Ethernet switch
IBRIX Fusion Cluster File System
RedHat Enterprise Linux 4

Every Parallel File System has its own design advantage. Apart from the design, all the other components including storage, operating system and network, though standards based, also affect the various parameters like scalability and performance of the file system. For getting the desired results, we need to not only understand the design and tuning parameters of any file system, but also focus on configuring all the participating components to yield their best.

For this exercise, our goal was to maximize "**file system throughput**" for a given computational power of an HPC cluster. The measure this parameter, we used standard file system benchmark i.e. IOZONE. We did our experiments with IBRIX Fusion file system and 4Gbps Fibre Channel storage. Towards realizing our goal, we experimented with the following parameters:

- Storage level tuning
  - RAID Type
  - Number of disks per RAID group

    - Number of LUNs per RAID group
    - Number of LUNs per Storage Processor
    - Number of add-on disk array enclosures per Storage Processor
    - Cache, prefetch, page size optimizations
- IO Node OS tuning
    - Read ahead
    - HBA module optimizations
    - TCP/IP parameter tuning
- Network tuning
    - Multiple modes of Ethernet channel bonding (NIC teaming)
- IBRIX file system tuning

With experimental results from all these studies, we were able to arrive at an optimal design to uniformly saturate every component in the solution and meet the objectives.

Some deductions were:

- Enclosures per storage processor
- Disks per RAID volume
- Number of clients per IO node

We will talk about our experiments, deductions and derived best practices not specific to any particular file system.