Multi-Path File System over iSCSI (MPFSi) Applied Technology

2008 EMC Proven™ Professional Knowledge Sharing

Ayyaswamy Thangavel
IP Storage Practice Leader
EMC South Asia
Thangavel_Ayyaswamy@emc.com

Table of Contents

Disclaimer: The views, processes or methodologies published in this compilation are those of the authors. They do not necessarily reflect EMC Corporation's views, processes, or methodologies.

MPFSi: Applied Example

Introduction

Continuous pressure to expand product portfolios brings ever-shortening product design cycle time. To meet these challenges, Computer Aided Engineering (CAE) and Finite Element Analysis (FEA) is becoming increasingly essential to help evaluate various design options and, at the same time, reduce cost by limiting or eliminating the need to produce prototypes for destructive testing. However, FEA does require significant computation power and fast input/output operations per second (IOPS) to the storage device.

Objective

This article examines the deployment of EMC® Multi-Path File System over iSCSI (MPFSi) technology at Leading Consumer Product Manufacturer's Singapore facility. MPFSi technology was used to help overcome storage data access performance issues often evident in large cluster deployments involving 200 or more CPUs/Cores performing simultaneous access to a central storage.

Performance, scalability and availability are key requirements for this solution to support the grid environment which the customer built using 50+ SuSE Linux Servers, including EMC Solutions consisting of CLARiiON®, Celerra®, MPFSi technology and LAN/SAN switches with IP Storage Services, and FC modules.

The customer provided the applications and jobs that are executed on the High Performance Computing farm to study the performance.

What is CAD/CAM/CAE?

CAD: Computer Aided Design
CAM: Computer Aided Manufacturing
CAE: Computer Aided Engineering Analysis

Typical Process Flow: Concept > Detailed Design > Analysis > Manufacturing
CAD/CAM/CAE is a set of tools running on graphical workstations used by product manufacturers from concept to product.

In today's highly competitive market, the challenge for an aggressive product development company is to improve efficiency and productivity, while maintaining or improving quality. That challenge translates into a search for improved process flows and methodologies that realize greater concurrency between product design and manufacturing process design, while simultaneously compressing and optimizing cycle times.

The Finite Element Analysis solution allows design optimization within short development schedules. Efficient design of multiple configurations of the same product and automatic application of changes throughout the design enable users to bring products to market quickly. Many load regimes, including mechanical (with shock/drop), thermal, and Multiphysics capabilities can be considered early in the design phase. This enables troubleshooting potential design problems before beginning to make prototypes, saving substantial time and money.

In short, Finite Element Analysis is about:

- Unified modeling and simulation
- Simulation methods to allow robust analyses
    - Multiple load types can be applied to a single model
    - Techniques available to efficiently handle the problems of different size scales typically found in electronic assemblies

Benefits of using CAD/CAM/CAE tools are:

- Lower cost
- Faster time to release/market
- Shorter design cycles
- Increased lead times
- Increased productivity
- Greater innovation

Solution Capabilities

- Drop testing of diverse products.
- Design analysis of consumer appliances
- Modeling and simulation

CAD Model of Cell Phone used for Testing

The job used for this testing consisted of a simplified model of a cell phone impacting a fixed rigid floor. For confidentiality, the real model of the mobile phone tested by the customer is not displayed here.

The cell phone components were meshed using a variety of element types. The material behavior was modeled using linear elasticity, etc. The components were assembled using surface-based mesh ties and placed into a general contact domain that also included the floor. The initial velocity and orientation of the cell phone was defined such that a severe oblique impact occurred.

What is EMC Celerra MPFSi Technology?

EMC Celerra Multi-Path File System over iSCSI (MPFSi) is a network file system (NFS)-based file storage interconnect topology that increases aggregate I/O performance to data shared in a grid computing architecture by up to four times that of conventional NFS.

MPFSi is also a scalable high performance data sharing solution for high performance computing environments such as that being discussed in this article. Other solutions include native NFS, use disks in external storage, JBOD in the compute Nodes, etc. Their purpose is to deliver very high throughput at lower latency. MPFS of MPFSi allows clients to send the request through the IP network and get the response over the faster FC or ISCSI network.

Concept of MPFSi

- Integrates FC SAN or IP SAN and IP NAS
- Celerra MPFSi allows UNIX, Linux, and Windows clients to access file system data over direct storage area network (SAN)
- Supports Standard File Sharing with improved performance for some applications
  - NFS or CIFS Request via IP
  - Data delivery via IP or FC SAN
- Greater client scalability through meta offload
- Basic components to accomplish connectivity:
  - Symmetrix or CLARiiON Storage system
  - Celerra Network Server with FMP Protocol
  - EMC HighRoad client FC HBA , TOE, or GigE Card
  - IP network and SAN connectivity for clients

Celerra MPFS Protocol

- File Mapping Protocol (FMP): an additional file sharing protocol:
  - File Server support for FMP complements NFS/CIFS protocols
- There are actually two MPFS protocols:
  - File Mapping Protocol (FMP): used by the client to lock file extents and their locations on the disks
  - FMP/Notify Protocol: used by the server to inform clients about lock revocations and mapping changes
- Supports parallel access of multiple clients to same file using range locking
  - Read/write locking scheme is used to ensure data consistency across clients
- The server is also responsible for other functions such as security, concurrency control, and cache invalidation. MPFS protocols are used in addition to either NFS or CIFS.

Celerra MPFSi Write Operation

- MPFSi Client process a write I/O request to the MPFS server
- MPFSi server allocates and returns free file system blocks for the client write
- MPFSi Client writes the data directly over the iSCSI path (IP SAN)
- MPFSi delivers data in the form of disk blocks
  - Disk blocks are fixed-size file chunks that use storage protocol to read/write data in a SAN environment using protocols such as SCSI, Fibre Channel, or iSCSI

Celerra MPFSi Read Operation

- Client sends a mapping request to FMP server on the Data Mover to read the data
- FMP server retrieves the physical locations of the data for the specific read
- MPFSi Client receive the file extent map for the specific read (a list of 8 KB file system blocks in run-length format)
- MPFSi Client accomplish the read directly over the iSCSI path (IP SAN)

HPC Architecture using MPFSi Technology at Leading Consumer Product Manufacturer, Singapore

In the customer environment, MPFSi software executes on 56 Linux clients, enabling the client to use the Gigabit Ethernet LAN to request access to a file and then perform the read or write across a 1Gbps iSCSI (IP-SAN) connection directly between client and storage. This capability decreased overall Data Mover CPU utilization, allowing improved scalability.

In one of the applications that processed large data objects sequentially, MPFSi enabled as much as a two- to three-times performance improvement versus NFS. For another application that processed small data objects, the performance was equal to NFS in several runs. MPFS clients access file data from CLARiiON systems over iSCSI connections, which increases the speed at which the data files can be delivered to Linux clients compared to NFS connections for both request and response.

Components used in the HPC environment are:

Network: Gigabit LAN, SAN, iSCSI, Myrinet
Servers: 54 x Compute Nodes, 2 x Head Nodes
Operating System: SuSE 9, Linux Enterprise
Storage: NS40G, CX3-20 with 45 FC Disks
Finite Element Analysis, Application: ME and RF simulation applications
Management Application: Scali

CLARiiON Disk Layout and Celerra File Systems

In the CLARiiON array RAID 5, 4+1 RAID Groups were created across all 60 FC drives, with some drives configured as Hot Spares. Two LUNs were created on RAID Group and they are presented to alternate Storage Processors to balance the back-end load. Celerra File Systems were created without using AVM. The size of each file system is documented below.

CLARiiON CX3-20 Disk Layout

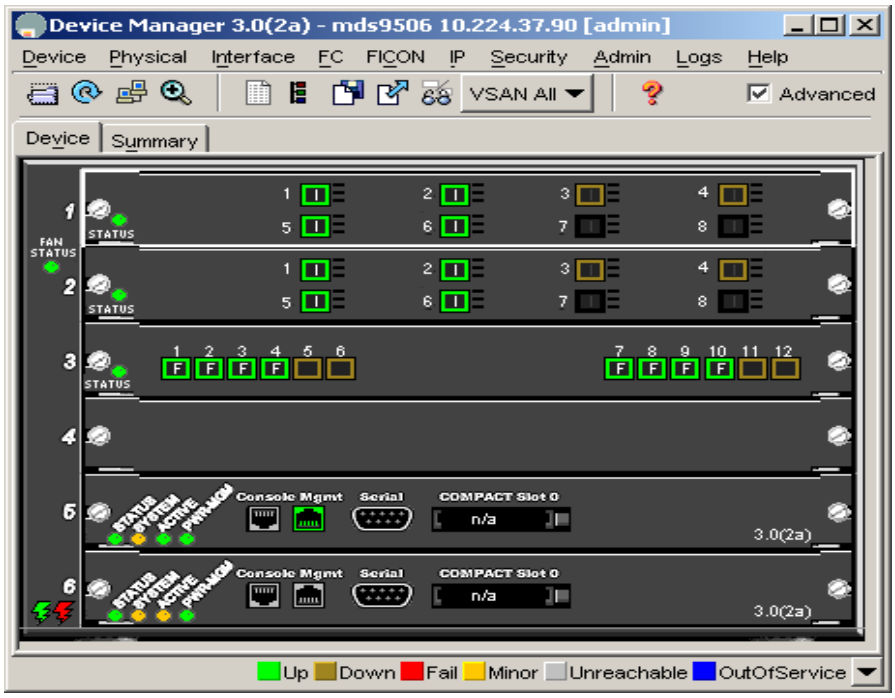
Encl 3	(4+1 RAID5) RG8		(ALU 22, HLU 32) – SPA			(4+1 RAID5) RG9		(ALU 24, HLU 34) - SPA			(4+1 RAID5) RG10		(ALU 26, HLU 36) - SPA		
			(ALU 23, HLU 33) - SPB					(ALU 25, HLU 35) - SPB					(ALU 27, HLU 37) - SPB		
Encl 2	(4+1 RAID5) RG5		(ALU 16, HLU 26) – SPA			(4+1 RAID5) RG6		(ALU 18, HLU 28) - SPA			(4+1 RAID5) RG7		(ALU 20, HLU 30) - SPA		
			(ALU 17, HLU 27) - SPB					(ALU 19, HLU 29) - SPB					(ALU 21, HLU 31) - SPB		
Encl 1	(4+1 RAID5) RG2		(ALU 10, HLU 20) - SPA			(4+1 RAID5) RG3		(ALU 12, HLU 22) - SPA			(4+1 RAID5) RG4		(ALU 14, HLU 24) - SPA		
			(ALU 11, HLU 21) - SPB					(ALU 13, HLU 23) - SPB					(ALU 15, HLU 25) - SPB		
Encl 0	Flare Operating System NAS Control LUN (4+1 RAID5) RG0					Hot	(4+1 RAID5)RG1 (ALU 8,HLU18) – SPA				Hot	Hot		Hot	Hot
	(ALU 0-5, HLU 0-5) – SPA (ALU 6, HLU16) – SPA (ALU 7, HLU17) – SPB					Spare	(ALU 9,HLU19) - SPB				Spare	Spare		Spare	Spare
Slots	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Celerra File Systems

Data Mover or VDM	File System Name	File System Type	TimeFinder Candidate	AVM	Storage Pool	Size (GB)
server_2	sim00_fs	UXFS	NO	NO	mpfsi_stv1	960
server_2	simua_fs	UXFS	NO	NO	mpfsi_stv2	780
server_2	sim01_fs	UXFS	NO	NO	mpfsi_stv3	360
server_2	sim02_fs	UXFS	NO	NO	mpfsi_stv4	360
server_2	sim03_fs	UXFS	NO	NO	mpfsi_stv5	240
server_2	sim04_fs	UXFS	NO	NO	mpfsi_stv5	240
server_2	sim05_fs	UXFS	NO	NO	mpfsi_stv3	120
server_2	pbs_fs	UXFS	NO	NO	mpfsi_stv4	75
server_2	app_fs	UXFS	NO	NO	mpfsi_stv1	50
server_2	home_fs	UXFS	NO	NO	mpfsi_stv2	21
server_2	scali_fs	UXFS	NO	NO	mpfsi_stv2	50
server_2	simadmin_fs	UXFS	NO	NO	mpfsi_stv3	50
server_2	scratch_fs	UXFS	NO	NO	scratch_stv	372

Cisco MDS and Catalyst Switch Modules and connections

The Cisco MDS Switch has two IP Storage Services modules, each populated with four iSCSI interfaces, and there is one FC module. The connections are described below (see next page):



Cisco MDS Switch IP Storage Services Module connections to Catalyst Switch Gigabit Ethernet Module

Module/Port

IPS1/1 --- Gi4/1
IPS1/2 --- Gi4/2
IPS1/5 --- Gi4/9
IPS1/6 --- Gi4/10
IPS2/1 --- Gi4/17
IPS2/2 --- Gi4/18
IPS2/5 --- Gi4/25
IPS2/6 --- Gi4/26

CLARiiON Storage Processor connections to FC Switch

Module/Port: CLARiiON, Celerra Ports
3/1: SPA Port 0
3/2: SPB Port 0
3/3: Data Mover 3 Port 0
3/4: Data Mover 2 Port 0

3/7: SPA Port 1
3/8: SPB port 1
3/9: Data Mover 3 Port 1
3/10: Data Mover 2 Port 1

More configuration details are documented in Appendix A.

Pre-production Test Setup

The test environment consisted of four Compute Nodes and two Head Nodes, Celerra NS40 with dual Data Movers, Catalyst and MDS Switch with IP Services Blade.

The plan was to run several iterations to insure the solution is working. The customer developed a synthetic benchmark that ran several scenarios from small to large block I/O. The performance results are tabulated under “Results” section.

The output of compute farm is a file which is then read by the post processor of a File Element Analysis application. The post-processing module read the output of analysis and displayed the results graphically on the screen. The study also included performance tests using different communication protocols, such as NFS, SAN, and iSCSI.

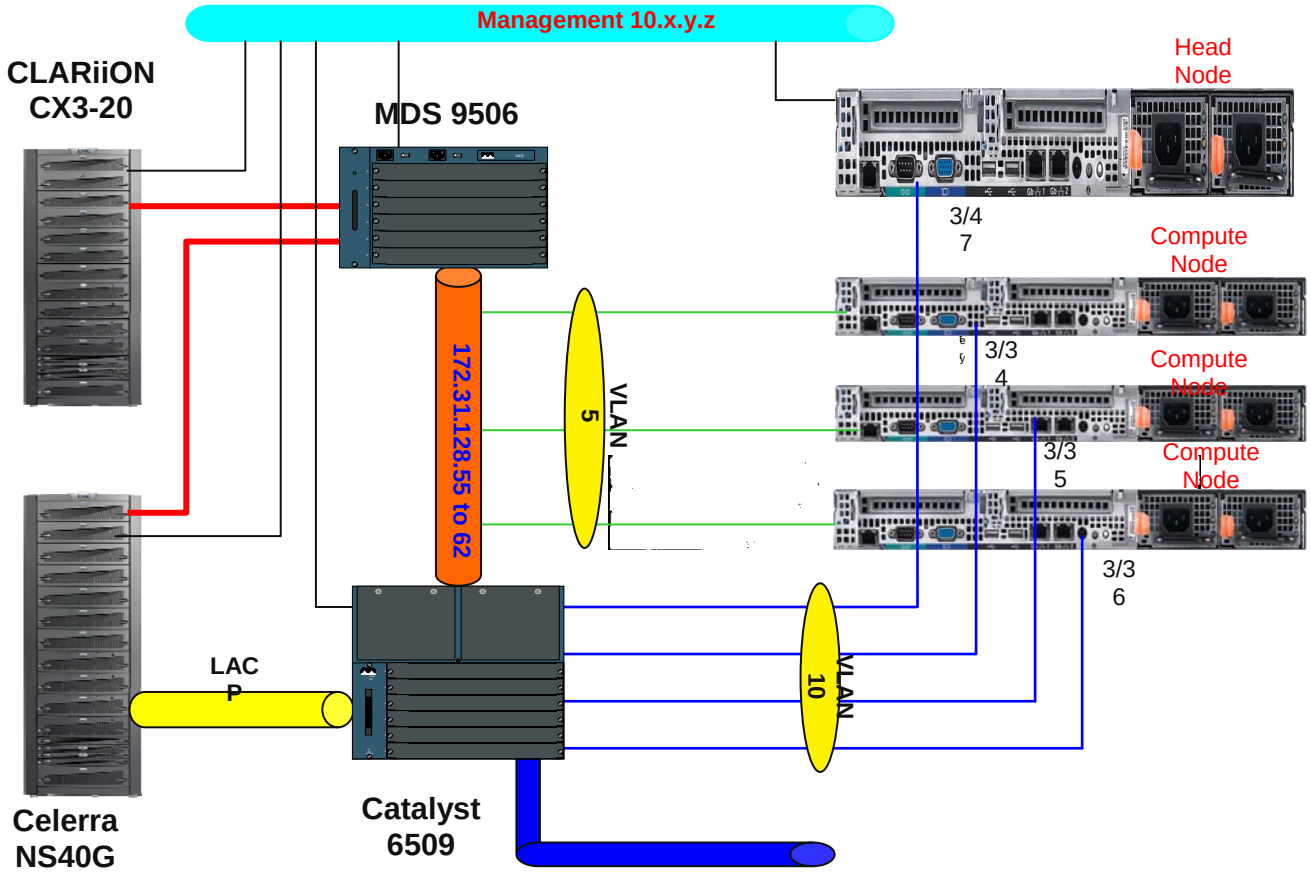


Fig 1: MPFSi Setup: Pre-Production

Production Setup

Figure 2 illustrates the MPFSi architecture currently under production. Several jobs were being executed and the results were stored for a few days in the Celerra. There was no Backup and Recovery requested by the customer as part of this solution because the output data can be regenerated again if required.

The production environment consists of 52 Compute Nodes and two Head Nodes, Celerra NS40 with dual Data Movers, LAN and SAN Switch with IP Services Blade. The production network is Gigabit Ethernet LAN, Management LAN, Fibre Channel SAN, and iSCSI Network. There are 8 x iSCSI connections between Catalyst and MDS switches and these are connected using two Gigabit Ethernet modules and IP Services modules, respectively.

Each Compute and Head node had at least three network connections, i.e., NFS, iSCSI Management, and optional Myrinet on some of the nodes. MPFSi Clients were installed in all the nodes, and the required operating system and MPFSi client parameter settings followed the recommendations documented in the EMC Best Practices document referred to in this article. The Scali management application was used by the customer to copy the configuration across multiple nodes.

During production, we tested the Bandwidth Utilization in this path, which was much lower than theoretical bandwidth available, i.e., 8Gbps. Some of the settings from the Best Practices document are highlighted in red below.

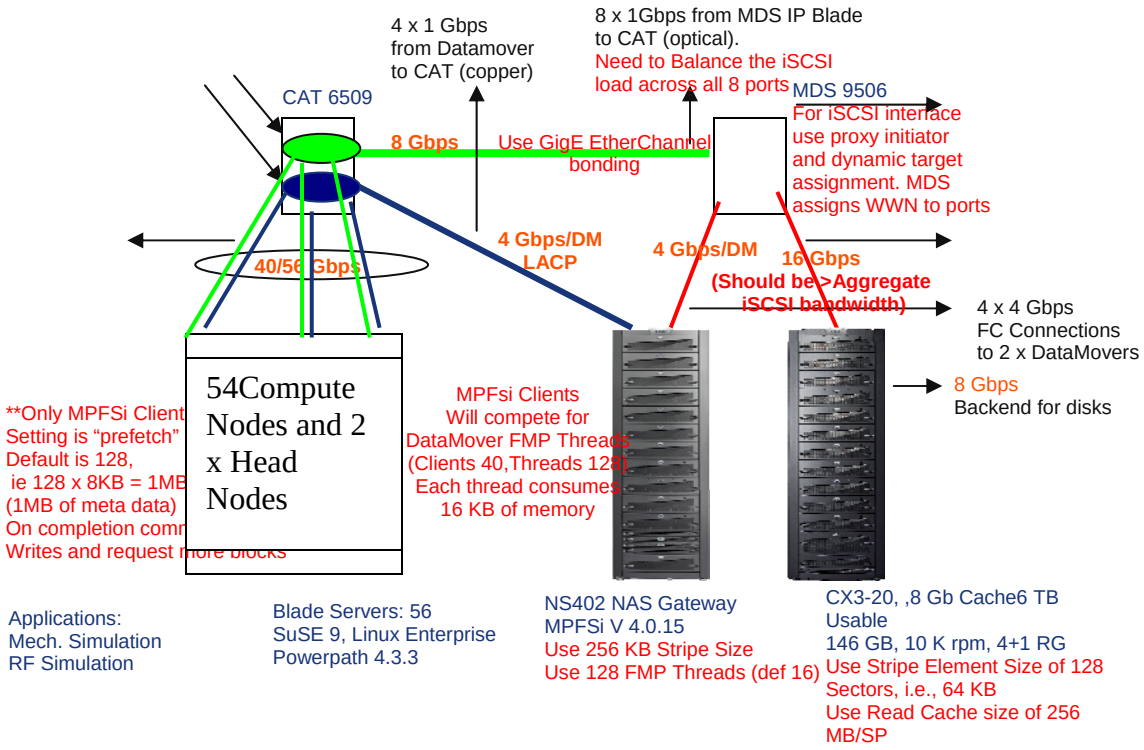


Fig 2: Production Setup

Jobs used for Testing

The customer ran two types of jobs. One job had characteristics like large block, sequential workload and the second job had small block, random workload characteristics.

During our tests, the jobs were executed to run for short duration and longer duration. Initially, the jobs which ran for shorter durations had issues in opening the output file using the Post-processor module, and jobs executed for longer duration had no issues. The issue was escalated to EMC and was solved by using the previous version of MPFSi client which is supported on Linux Kernel versions used by the customer. The results of the jobs are reported in the following section.

Results

At the end of the job run, a file was written through iSCSI connection. This output file was read by the post-processing application and graphical results displayed on the computer screen. To ensure customer confidentiality, the real model is not shown here.

Performance

The results of two simulation jobs run using NFS and MPFSi tests are tabulated below:

Application	NFS	MPFSi	Comments
ME Simulation	16min 35sec	8min 10sec	Large Blocks
RF Simulation	1 hr 31min	1hr 15min	Small Blocks

Customer Benefits

As the MPFSi client requests are sent via NFS and the results are delivered via commodity IP SAN, the customer realizes the following benefits:

- Standard file sharing with faster MB/s
- Greater client scalability at lower cost than FC
- Deliver to the full throughput potential of the SAN-attached storage arrays connected to SAN and IP-SAN grid nodes via iSCSI
  - Cheaper alternative to FC for 100's of grid compute nodes

MPFSi Best Practices

The EMC Best Practices guide captures experiences of NAS solutions engineering and others in development, test, performance, and field support in regard to setup, use, and scaling of an MPFSi architecture in a grid computing environment. iSCSI throughput calculation and Linux, MPFSi parameter settings are as below.

Reference: EMC Celerra MPFSi for Linux HPC/Grid Computing Best Practices Guide.

a) **Throughput:** The following calculation was used to calculate the maximum iSCSI throughput possible using MDS 9506 with two IP Services modules each populated with four interfaces only.

```
4-port IP Storage Services module x 2 = (4 * 1 Gb/s )x 2 = 8 Gb/s
4-port on 2 storage processors = 4x 4Gb/s = 16 Gb/s.
Maximum iSCSI traffic = smaller of 8 Gb/s (Ethernet) and 16 Gb/s (Fibre Channel) = 8 Gb/s.
In Bytes that's 8 Gb/s / 8 bits per byte = 1 GB/sec.
```

b) **Prefetch** - The MPFSi client software had one parameter that was adjusted. Prefetch is a parameter specifying the amount of 8 KB blocks of metadata to prefetch.

mpfsctl prefetch 2048

c) There was one parameter for 2.4 Linux kernels. For this parameter to be saved across **reboots, a line** was added to /etc/sysctl.conf that looks like the following:

vm.max-readahead = 8192

Other recommendations to consider:

- Monitor all systems in the environment (Linux MPFSi clients, switches, Data Movers and disk arrays)
- Monitor Data Mover statistics for CPU utilization, memory utilization, network
- Watch for Fallthroughs that occur in MPFSi.
- Disable logging on the Storage Processors to increase performance when not actively monitoring statistics.

Performance Analysis

A recommendation to collect the metrics below to do a performance analysis in case there are bottlenecks during the test or preproduction and production phases proved unnecessary as the customer noticed the performance difference over MPFSi compared to NFS job runs. The Navisphere Analyzer data collection was disabled as requested by the customer—and also as per the best practices document—as the system was expected to go into production immediately.

Monitor from server running test (host perspective):

- I/O stat or SAR
  - I/O Q depth
  - CPU busy

Navisphere Analyzer Data Collection:

- Total I/O sec
- Read/write ratio
- Read hit rates
- I/O Blocksize
- KBytes transferred

Celerra Monitor

- Data Mover CPU busy
- Data Mover network throughput

Lessons Learned

1. IP Services module has only Optical Gigabit Interface, so the customer's Catalyst Switch used for Gigabit LAN should have Optical Gigabit interface as well. At first, the customer had only Copper Gigabit Interface. This was noticed before installation and reduced the delay.
2. The customer's Post Processor application was unable to read the output file written using the latest version of MPFSi Client, while it was able to read with jobs run using NFS connections. The issue was escalated and a similar setup was replicated at EMC Hopkinton Lab where issues with MPFSi client were identified. The issue was with NFS changes with different Linux Kernel version, MPFSi Client version.
3. Dedicated interfaces on servers for NFS, iSCSI, and management were used in the customer environment. We recommend planning each of these network configurations and requesting an exact number of IP Addresses, Subnet Masks, Default Gateways, etc. for servers, switches, storage, etc.

4. Manual volume management was used instead of using Celerra AVM to optimize the disk throughput for different application and other NFS, CIFS Shares.
5. While the application writing large block sequentially had shown an increase in performance compared to NFS, the other application writing small block did not show any performance improvement, as expected. The recommendation is to understand the customer's application write sizes and read sizes in the early stages of discussion to avoid any surprises during proof of concept or production.
6. Follow the guidelines of the Best Practices document for operating system, network and storage configuration, and settings unless there are special requirements, if any.

Opportunities for EMC Celerra MPFSi Solutions

EMC Celerra MPFSi technology delivers maximum performance when data is transferred in large blocks and sequentially streamed. This solution can be applied to several industries, though this article discussed only one of the applied examples where product quality and time-to-market pressure are evident.

Some of the target industries are:
Aerospace
Automotive
Consumer Products
Weather Forecasting
Grid Computing
Oil and Gas

Conclusion

The two objectives below were met using the EMC Celerra MPFSi Technology

a) Large block application with performance exceeding current implementation.
b) Small block application with performance equal to current implementation.

The customer collected all the performance data and compared the results during several job runs, satisfying both of the above objectives.

About the Author

1. Ayyaswamy Thangavel,
   IP Storage Practice Leader
   EMC South Asia

   Twenty years of working experience in the IT Field. Experience includes CAD/CAM Teaching and Consulting, TCP/IP Network Management Consulting and Deployment, and IP Storage Solutions Presales Support.

   .