Storage Devices, Local File System and Crossbar Network File System Characteristics, and 1 Terabyte File IO Benchmark on the "Numerical Simulator III"*

Naoyuki FUJITA *Hirofumi OOKAWA*
fujita@nal.go.jp *ookawa@nal.go.jp*
National Aerospace Laboratory of Japan

Abstract

We benchmarked a mass storage system named "CeMSS" on the "Numerical Simulator III" System. It has eighty (80) RAID-5 disk arrays and forty (40) LTO tape drives, as a storage devices, and has an HSM based local file system and crossbar network file system. We also described CeMSS design outline. In order to clear our benchmark perspective we defined "Standard IO Characteristic" and "User IO Pattern". We recognized that the disk and tape devices of CeMSS are optimized at 2[MB] and 128+[KB] IO size. Using 16-way disks, user application programs can use over 1[GB/s] IO throughput on the NS-III. And under 80-way disk condition, CeMSS could operate a 1[TB] file within 10 minutes.

1. Introduction

In 2002, the National Aerospace Laboratory of Japan introduced the *Numerical Simulator III System*. The *Numerical Simulator III System* has a mass storage system named *CeMSS (Central Mass Storage System)*, which has eighty (80) RAID-5 disk arrays and forty (40) LTO tape drives. *CeMSS* is connected with *CeNSS (Central Numerical Simulation System)* via a four (4) Gigabyte bi-directional crossbar network. In the field of computational fluid dynamics, huge scale numerical simulations -- e.g.: simulations on combustion flow with chemical reactions -- are possible recently. However, these simulations require high-speed file IO systems that can operate files at a rate of about one (1) Gigabyte per second.

We realized the throughput requirements of *CeMSS* with the methodology listed below. In order to inspect the throughput and understand *CeMSS*'s characteristics in detail we measured the raw device IO characteristics, local file system IO characteristics, and crossbar network file system IO characteristics of the *Numerical Simulator III System*. Furthermore, we benchmarked a one (1) Terabyte file IO. From measuring the characteristics, we recognized that the RAID device on *CeMSS* is optimized at two (2) Megabytes IO size, the LTO drive on *CeMSS* is optimized at 128+ Kilobytes IO size, the local file system performs at about three to four (3 to 4) Gigabytes per second with eighty-way disks, and the crossbar network file system performs at about one point six (1.6) Gigabytes per second with eighty-way disks. Set to the above conditions *CeMSS* can write/read a one (1) Terabyte file within 10 minuets.

2. "Numerical Simulator III" Mass Storage System

At the National Aerospace Laboratory of Japan (NAL), we introduced the *Numerical Simulator I System* which utilized the compute server *VP400* and promoted the Navier-Storks equation based on numerical simulations in the 1980s. In the 1990s, we built the *Numerical Simulator II system* which utilized the compute server *Numerical Wind Tunnel* and promoted parametric study calculations of complete aircraft aerodynamic simulations. Then in 2002, we introduce the *Numerical Simulator III System* (NS-III). NS-III has nine (9) [TFLOPS] compute servers (*CeNSS: Central Numerical Simulation System*), a 3D-visualization server (*CeViS: Central Visualization System*), and a high-speed mass storage system (*CeMSS: Central Mass Storage System*). We are going to do multidisciplinary numerical simulations, unsteady flow analysis, and so forth, on the NS-III. Figure 1 shows an overview of NS-III.

According to our estimation of requirements on *CeMSS* [1], *CeMSS* has to have about one (1) Gigabyte per second throughput, while single storage devices have several Megabytes per second throughput. So *CeMSS* should be a parallel IO system. When we make a parallel IO system on multi-node parallel computers (where "node" is a computing node and *CeNSS* is a multi-node parallel computer), we can consider two typical IO models; one is the nodes-wide-parallel-IO model while the other is the IO-node model (Figure 2). By using the

Proceedings of the 20 th IEEE/11 th NASA Goddard Conference on Mass Storage Systems and Technologies (MSS'03)
0-7695-1914-8/03 $17.00 © 2003 IEEE

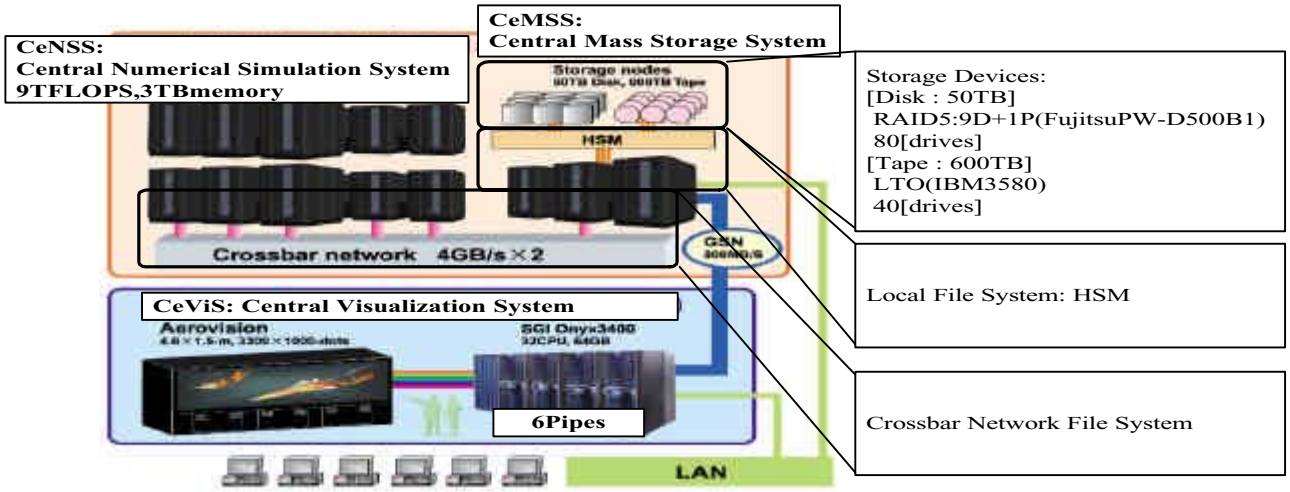


Figure 1. Numerical SimulatorIII Mass Storage System Overview

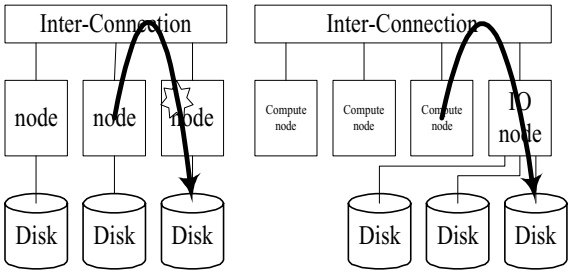


Figure 2. Two typical parallel IO models on multi-node parallel computer

Table 1. *CeMSS*'s Outline Design Items

Items		Selected design
Individual storage system or Local file system		Local file system
HSM[*1] or not		HSM
Reliability/Redundancy	Disk	RAID5
	Tape	(Original + Copy) tape media
Cooperation with 3D-visualization system		GSN[*2] + Original library
Cooperation with workstation		NFS[*3]

[*1] Hierarchical Storage Management [*2] Gigabyte System Network [*3] Network File System

nodes-wide-parallel-IO model, we can easily build a parallel IO infrastructure because each node has its own IO port, so that we do not have to prepare special resources for building a parallel IO. However, the nodes-wide-parallel-IO may cause a collision between IO operations and computing operations. Namely, whenever one node needs data that is stored in its neighbor's storage, the neighboring node is obliged to do the IO operations while doing computing operations. This situation makes it difficult for us to estimate IO operation times as well as computing operation times. On the other hand, the IO-node model will give us more steady and higher IO performance although an extra IO node is required. We adopted the IO-node model on NS-III.

As mentioned above, we chose the IO-node model for *CeMSS*. There are further items that should be considered in order to build an efficient storage system. Table 1 summarizes these items. Figure 3 shows the resultant system design of *CeMSS*.

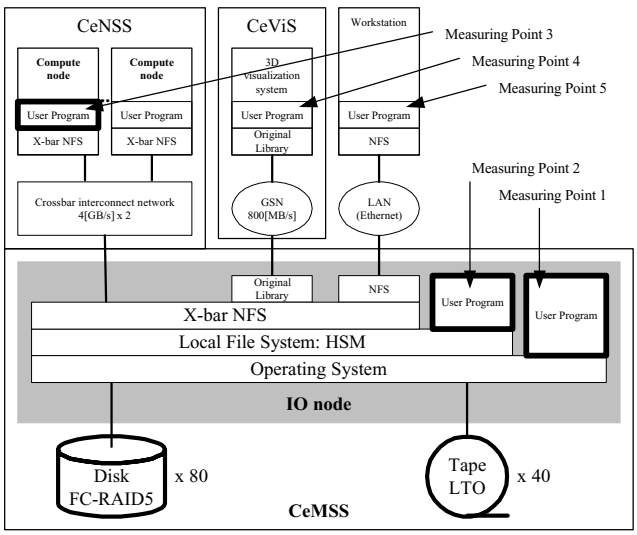


Figure 3. System Design of CeMSS

3. Mass Storage Benchmark Perspective

Estimating IO performance on *CeMSS* with an HSM as a local file system we must take some steps. The detailed estimation steps are as follows;

1. Estimate storage device characteristics
1.1. Disk device characteristics
1.2. Tape device characteristics
2. Estimate basic file system performance
2.1. One file disk IO performance
2.2. One file migrating performance
2.3. One file staging performance
3. Estimate file system performance of actual condition
3.1. Multiple files disk IO performance
3.2. Multiple files migrating performance
3.3. Multiple files staging performance

NS-III has measuring points for IO throughput (See Figure 3; Measuring Point (MP) 1-5). In this paper we report results at three measuring points. The following are the selected measuring points;

MP 1:Raw device IO characteristics on *CeMSS*
MP 2:Local file system IO characteristics on *CeMSS*
MP 3:Crossbar network file system IO on *CeNSS*

Before starting the discussion, we would like to define some terms for measurement.

Disk Device --- One RAID disk array unit in this paper. One RAID has 9 data disks and 1 parity disk

Raw IO Size --- Number of bytes of an IO unit read from or written to a device, which is assigned to disk device in advance

User IO Size --- Total IO number of bytes per one IO user operation

File Size --- Total number of bytes in a file

Number of Device Parallelization --- Total number of disk devices, which are used when one user IO block is read or written.

Benchmarks and studies on benchmark strategy were researched [2],[3],[4],[5]. As those papers said, when we benchmark and/or measure characteristics, there are a lot of parameters to determine. In order to clear our benchmark perspective, we defined the "Standard Characteristic" and "User IO pattern".

3.1. "Standard Characteristic"

The "Standard Characteristic" is throughput profile via raw IO size. This characteristic shows the raw device characteristic. When we see "Standard Characteristic", we can recognize the number of maximum actual throughput of a certain device, the range of raw IO sizes on which a certain device works effectively, and so on.

3.2. "User IO Pattern"

We assumed that a user writes and/or reads 2[MB] data per one user IO operation and the total number of bytes per one file is 2[GB]. That is to say, one file is created by about one thousand write-operations. Of course, user IO patterns are tightly related to each application, so this pattern is just an example.

3.3. Parameters

As we defined the "Standard Characteristic" and "User IO patterns", we can define parameters for measuring characteristics and benchmarking. Here are some parameters; File Size, User IO Size, Raw IO Size, Number of Device Parallelization. We measure the characteristics via IO sizes and number of parallelizations in this paper.

4. Characteristics and 1 Terabyte File IO Benchmark

4.1. Raw Device IO Characteristics on *CeMSS*

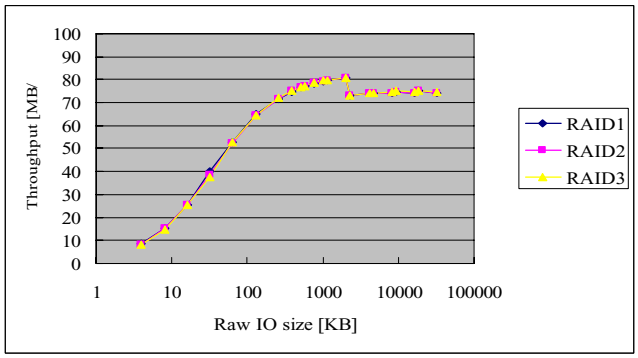
On *CeMSS*, the disk device is a RAID disk array unit. So seizing disk device characteristics, we measured IO throughput corresponding to raw IO size. Throughput was calculated from the time stamp just before/after an IO operation and file size. And with this measurement, the IO operation is "write" and/or "read" low-level file IO function call, and the IO operation was performed on a raw device directly. Figure 4 shows some of the results.

4.2. Local File System IO Characteristics on *CeMSS*

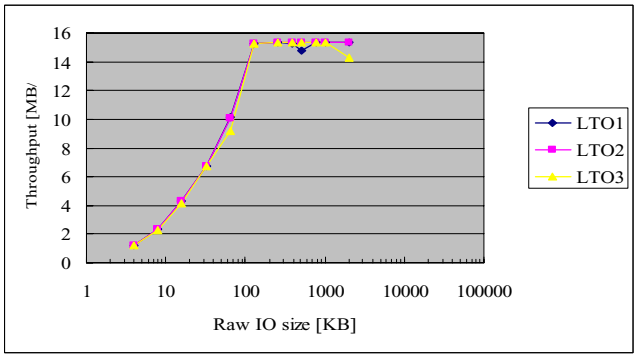
We show two characteristics; one is throughput via user IO size, another is throughput via the number of device parallelizations. Throughput was calculated the same as section 4.1. Figure 5 shows the results.

4.3. Crossbar Network File System IO Characteristics on *CeNSS*

We show one characteristic; throughput via number of device parallelizations. Throughput was calculated the same as section 4.1. Figure 6 shows the results.

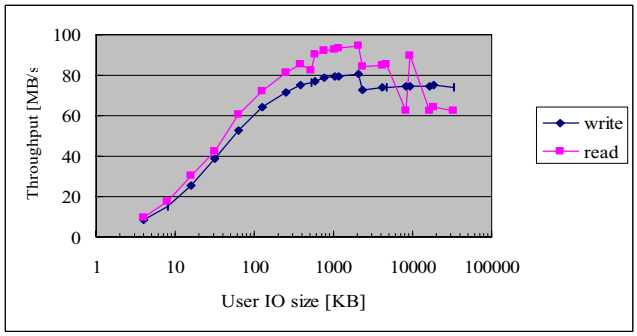


(a) Disk write characteristic

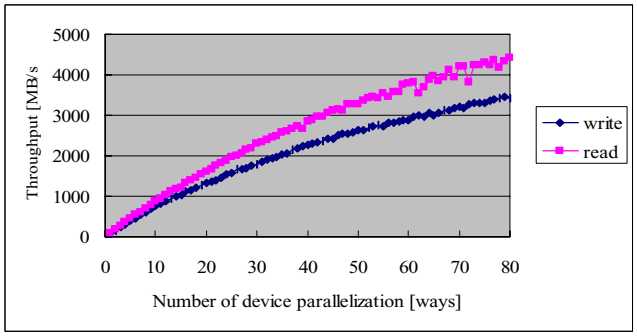


(b) Tape write characteristic

Figure 4. Raw Device IO Characteristics



(a) User IO size



(b) Number of device parallelization

Figure 5. Local File System IO Characteristics

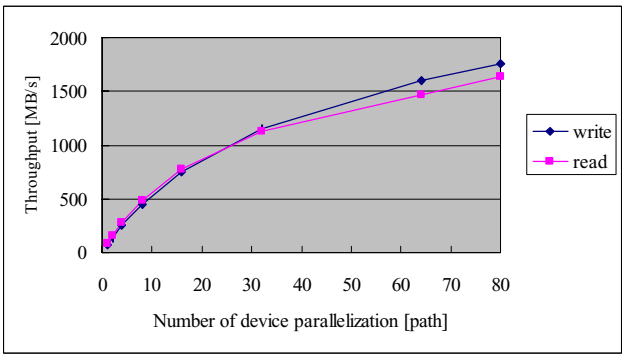


Figure 6. Crossbar Network File System IO Characteristics

4.4. One Terabyte File IO Benchmark

According to our estimation [1], when we calculate complete aircraft simulations and/or direct numerical simulations of turbulence, we need from several gigabytes to one terabyte file IO. But, at the present moment, we do not have storage systems which can handle such large scale files at high speeds. As we understood the characteristics of *CeMSS* shown above, we did a one Terabyte file IO benchmark in order to check if *CeMSS* can be a solution for the high-speed large file IO. Table2 shows the result.

Table 2. 1 Terabyte File IO Throughput

File size: 1120[GB], IO size: 160[MB]

Measuring Point	Throughput [MB/s]	
	Write	Read
Local File System on *CeMSS* (80 ways RAID)	3407	4284
Crossbar NFS on *CeNSS* (80 ways RAID)	1856	1724

5. Discussion

From the results of the raw device IO characteristics (Figure 5), we learned that this RAID is optimized at 2[MB] IO size, and from the datasheet (not shown), the actual throughput for write is 81[MB/s] and for read is 95[MB/s]. LTO tape drive is optimized at 128+[KB] IO size, and from the datasheet (not shown), the actual throughput for write is 15[MB/s] and for read is 14[MB/s].

From the results of the local file system IO characteristics, we learned that there is little overhead on the local file system, because the characteristic curve profiles (Figure 4(a) and Figure 5(a) write) are almost the same. But when we compared local file systems and crossbar network file systems, we recognized the performance declined. From Figure 5(b) and Figure 6,

maximum read throughput on local file system is about 4300[MB/s], but maximum read throughput on crossbar network file system is about 1600[MB/s], and on the crossbar network file systems, read operation is slower than write operation.

From Table 3, *CeMSS* has over 1[GB/s] of throughput, so we can write/read 1 [TB] file within 10 minutes. It is useful for large file handling.

As shown in Figure 3, *CeMSS* has three types of interfaces; a crossbar network file system interface, a GSN interface, and a LAN (Ethernet) interface. Using these interfaces, user application programs can access storage devices uniformly via *CeMSS*. Research and development for uniform storage access are carried out actively [6,7]. The Storage Resource Broker is middleware that has the ability to manage archives, file systems, and databases uniformly. Using the SRB client API, user application programs can access very various storage managed by SRB. *CeMSS* has only three access interfaces mentioned above, but it provides over 1[GB/s] of throughput for user application programs. High-speed IO is important to scientific computing users, especially CFD researchers.

6. Conclusion

We built a mass storage system named *CeMSS* on the *Numerical Simulator III System*. It has eighty RAID5 disk arrays and forty LTO tape drives, as a storage devices, and has an HSM based local file system and crossbar network file system. *CeMSS* has crossbar interface, GSN interface, and Ethernet interface to connect to other systems that want to use storage devices. We recognized that the disk and tape devices of *CeMSS* are optimized at 2[MB] and 128+[KB] IO size. Using 80-way disks, user application programs can use over 1[GB/s] IO throughput on the NS-III. And under this condition, *CeMSS* can operate a 1[TB] file within 10 minutes. In addition we defined "Standard Characteristics" and "User IO Pattern" in order to clear our benchmark perspective.

In future work, we are going to analyze crossbar network file system performance. And we would like to measure multiple file IO access throughput on *CeMSS* and *CeNSS* and also measure throughput on *CeViS*: 3D-visualization systems and workstations on LAN, and make NS-III a total numerical simulation facility.

Acknowledgment

We are grateful to Katsumi Yazawa and Kazuhiro Kanaya for their work measuring the performance of *CeMSS*.

References

[1] Naoyuki Fujita and Yuichi Matsuo, "High-Speed Mass Storage System of Numerical Simulator III and its Basic I/O Performance Benchmark",A collection of abstract, Parallel CFD 2002, 2002, Kyoto Institute of Technology and JAERI

[2] Phil Andrews, "Tom Sherwin and Victor Hazlewood, High-Speed Data Transfer via HPSS using Striped Gigabit Ethernet Communications", Proceedings, 18th IEEE Symposium on Mass Storage Systems and Technologies / 9th NASA Goddard Conference on Mass Storage Systems and Technologies,2001

[3] Emin Gabrielyan and Roger D.Hersch, "SFIO a Striped file I/O library for MPI", Proceedings, 18th IEEE Symposium on Mass Storage Systems and Technologies / 9th NASA Goddard Conference on Mass Storage Systems and Technologies,2001

[4] Marti Bancroft, Phillip L.(Rocky) Snyder and Mark Woodyard, "Two Case Studies of the Application of Dynamic Modeling Techniques in Performance Assessment and Prediction of Complex Shared Storage Architectures", Proceedings, 18th IEEE Symposium on Mass Storage Systems and Technologies / 9th NASA Goddard Conference on Mass Storage Systems and Technologies,2001

[5] Thomas M. Ruwart, "File System Benchmarks, Then, Now, and Tomorrow", Proceedings, 18th IEEE Symposium on Mass Storage Systems and Technologies / 9th NASA Goddard Conference on Mass Storage Systems and Technologies,2001

[6] Reagan W. Moor, "Knowledge-based Grids", 18th IEEE Symposium on Mass Storage Systems and Technologies / 9th NASA Goddard Conference on Mass Storage Systems and Technologies,2001

[7] Baru,C., R,Moor, A.Rajasekar, M.Wan, "The SDSC Storage Resource Broker", Proc. CASCON'98 Conference, 1998