A data Grid testbed environment in Gigabit WAN with HPSS

Atsushi Manabe, Setsuya Kawabata, Youhei Morita, Takashi Sasaki, Hiroyuki Sato, Yoshiyuki Watase and Shigeo Yashiro
High Energy Accelerator Research Organization,KEK, Tsukuba,Japan
Tetsuro Mashimo,Hiroshi Matsumoto, Hiroshi Sakamoto,Junichi Tanaka, Ikuo Ueda
International Center for Elementary Particle Physics University of Tokyo (ICEPP, Tokyo, Japan)
Kohki Ishikawa, Yoshihiko Itoh, Satomi Yamamoto
IBM Japan, Ltd.
Tsutomu Miyashita
KSK-ALPA co. ltd.

For data analysis of large-scale experiments such as LHC Atlas and other Japanese high energy and nuclear physics projects, we have constructed a Grid test bed at ICEPP and KEK. These institutes are connected to national scientific gigabit network backbone called SuperSINET. In our test bed, we have installed NorduGrid middleware based on Globus, and connected 120TB HPSS at KEK as a large scale data store. Atlas simulation data at ICEPP has been transferred and accessed using SuperSINET. We have tested various performances and characteristics of HPSS through this high speed WAN. The measurement includes comparison between computing and storage resources are tightly coupled with low latency LAN and long distant WAN.

1. Introduction

In the Atlas Japan collaboration, International Center for Elementary Particle Physics University of Tokyo(ICEPP) will build a "Tier-1" regional center and High Energy Accelerator Research Organization(KEK) will build a "Tier-2" regional center for the Atlas experiment of the Large Hadron Collider (LHC) project at CERN. The two institutes are connected by the Super Sinet network which is an ultrahigh-speed Japanese academic researches Internet backbone. On the network with the Grid technologies a test bed was constructed to study requisite functionality and performance issues for the tiered regional centers.

High Performance Storage System (HPSS) with high density digital tape libraries could be a key component to handle petabytes of data produced by Atlas experiment and to share such data among the regional collaborators. HPSS parallel and concurrency data transfer mechanisms, which support disk, tape and tape libraries, are effective and scale to support huge data archives. This paper describes about integration of HPSS into a Grid architecture and the performance measurement of HPSS in use over a high-speed WAN.

2. Test bed system

The computer resources for the test bed were installed to ICEPP and KEK site. One Grid server in each site and HPSS servers in KEK were connected to the Super Sinet. The Internet backbone, Super Sinet connects research institutes at 10 Gbps with operation of Optical Cross Connect for fiber/ wavelength switching and the two are directly connected at 1 Gbps. All resources including network were isolated from other users and dedicated for the test. Figure 1 and Table I shows our hardware setup.

Three storage system components were employed. One disk storage server each at KEK and ICEPP shared its host with the Grid server. The remaining HPSS software components used some part of the KEK central computer system. The HPSS data flow depicted in Fig. 2. The HPSS Servers includes core servers, disk movers, and tape movers tightly coupled by an IBM SP2 cluster network switch.

In the case of original (kerberos) pftp server performance measurement, pftpd was run in the core HPSS server. In the case of GSI-enabled HPSS server which will be mentioned in 4, pftpd was run in the same processors as the disk mover. The disk movers were directly connected to the test bed LAN through their network interface cards. HPSS disk movers were dedicated only to the test.

NorduGrid middleware ran on the Grid servers. Other computing elements (CE) acted as a Portable Batch System (PBS) [1] that was not required to install the NorduGrid middleware.

NorduGrid middleware ran on the Grid servers. Other computing elements (CE) acted as a Portable Batch System that was not required to install the NorduGrid middleware.

The NorduGrid[5] is a pioneer Grid project in Scandinavia that added upper layer functionality, which is necessary to HEP computing, on the Globus tool kit. The middleware was simple to understand and offered functionality sufficient for our test bed study.

Table II shows the versions of middleware used in the test bed.

Table I Test bed Hardware

ICEPP	**Grid and PBS server**
	1 × Athlon 1.7GHz 2CPU
	Computing Element
	4 × pentium III 1.4GHz 2CPU
KEK	**Grid and PBS server**
	1 × Pentium III 1GHz 2CPU
	Computing Element
	50 × pentium III 1GHz 2CPU
	HPSS disk mover
	2 × Power3 375MHz
	HPSS tape mover and Library
	19 × Power3 375MHz, IBM 3590

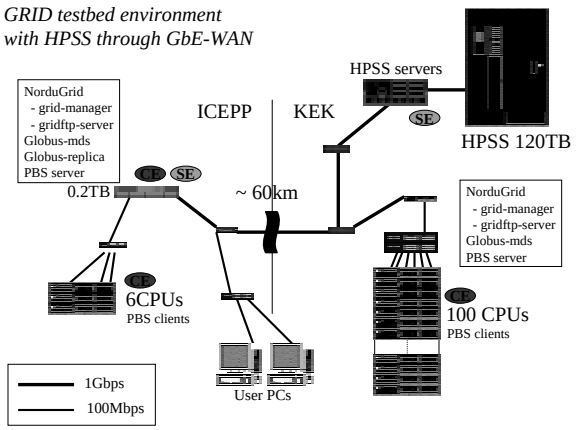

Figure 1: Layout of the test bed hardware

3. HPSS performance over high-speed WAN

3.1. basic network performance

Before end to end measurement, basic Gigabit Ethernet performance between IBM HPSS servers at KEK and a host at ICEPP through the WAN and a host on the KEK LAN was measured using netperf [2] and is shown in figure 3. Round Trip Time (RTT) averaged 3 to 4 ms. The network quality of service was quite good and free from packet loss (¡0.1HPSS server was 256kB (the size was fixed to optimize IBM SP2 switching network performance) and was 64MB in clients at both KEK (over LAN) and ICEPP (over WAN). The processors running the HPSS servers limited the maximum raw TCP transfer performance, as seen in the graph the network performance varied with socket buffer size. Beyond 0.5MB, network access performance through both LAN and WAN became almost equivalent and saturated.

Figure 4 shows the network performance with a

Table II Test bed Software

software	version
Globus	2.2.2
NorduGrid	0.3.12
PBS	2.3.16
HPSS	4.3

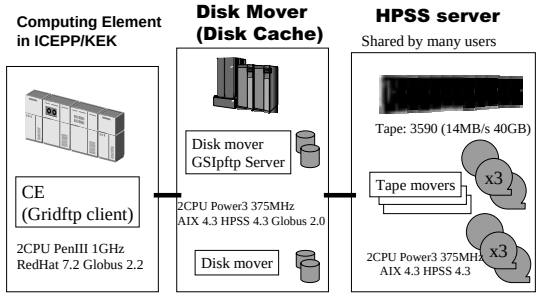

Figure 2: HPSS players.

number of simultaneous transfer sessions through the WAN and the LAN. In the situation where socket buffer size was 100KB, up to 4 parallel simultaneous stream sessions improved network throughput. Using greater buffer size than 1MB, multiple stream sessions did not improve the aggregate network transfer speed. And network utilization was limited by the performance of the processors running the HPSS servers.

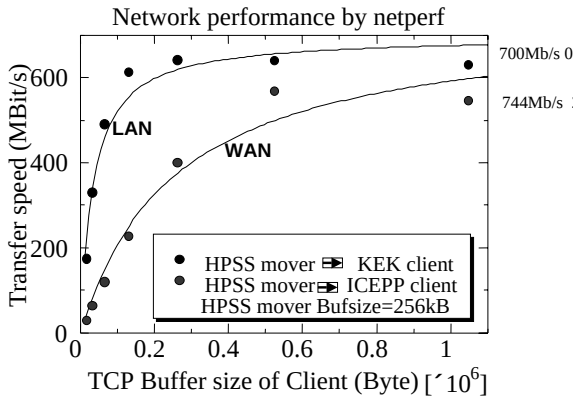

Figure 3: Basic GbE network transfer speed.

3.2. HPSS client API performance

Figure 5 shows data transfer speed by using the HPSS client API and comparison between access from LAN and over WAN. The transfer was from/to HPSS

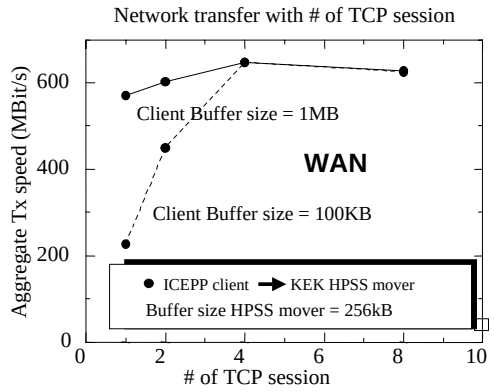

Figure 4: Network performance with no. of TCP stream sessions.

disk mover disk to/from client host memory. The transferred file size was 2GB in all case. Disk access speed in the disk mover was 80MB/s. It shows that even with a larger API buffer, WAN access speed was about a half of LAN access both for reading and writing from/to HPSS server.

To increase HPSS WAN performance in future tests, the newer pdata protocol provided in HPSS 4.3 can be employed. This will improve pget performance. To get the same effect on pputs, the pdata-push protocol provided in HPSS 5.1 is required.

The existing mover and pdata protocols are driven by the HPSS mover with the mover requesting each data packet by sending a pdata header to the client mover. The client mover then sends the data. This exchange creates latency on a WAN. The pdata-push protocol allows the client mover to determine the HPSS movers that will be the target of all data packets when the data transfer is set up. This protocol eliminates the pdata header interchange and allows the client to just flush data buffers to the appropriate mover. The result is that the data is streamed to the HPSS mover by TCP at whatever rates it can be delivered by the client side mover and written to the HPSS mover devices.

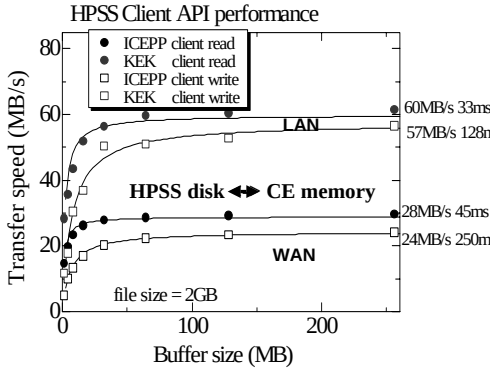

Figure 5: HPSS client API performance

3.3. pftp-pftpd transfer speed

Figure 6 shows data transfer speed by using HPSS pftp from HPSS disk mover to client /dev/null dummy device. Again as in the previous HPSS client API transfer, even with a pftp buffer size of 64MB, access speed from WAN was about a half of LAN access. In addition, enabling single file transfer with multiple TCP stream by using the pftp 'pwidth' option was not effective in our situation. In our server layout, two disk mover hosts each had two RAID disks. Therefore, up to 4 concurrent file transfers could effect higher network utilization and overall throughput, and was so seen in WAN and LAN access case. In the same figure (Fig. 6) data transfer speed was shown from HPSS disk mover to client disks which had writing performance of 35-45MB/s. Though disks both in server and client hosts had exceeding 30MB/s access speed and also network transfer speed exceeded 80MB/s, overall transfer speed dropped into 20MB/s. It is because these three resources access was not executed in parallel but done in series.

Figure 7 shows elapsed time for access of data in tape library. Thanks to HPSS functionality and an adequate number of tape movers and tape drives, the system data throughput performance scaled with the number of concurrent file transfers. If all the tapes are off drives, since the library had only two accessors, performance scaled up to two concurrent transfers.

Comparison (Fig. 8) of writing to HPSS disk mover from client over WAN and LAN is rather complicated. In our setup, HPSS server had 4 independent disks but client had only one. Reading multiple files in parallel (N files → N files; reading N files simultaneously at client and writing to N files to the server) from a single disk slows down the aggregate access performance by contention of disk heads.

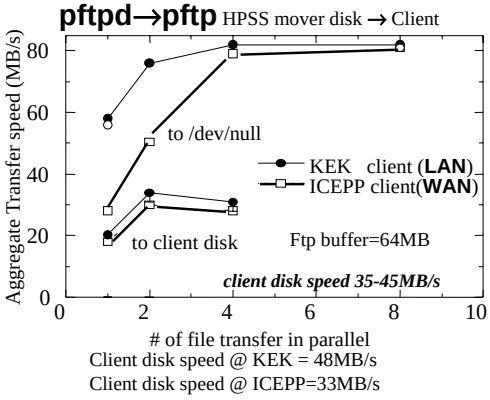

Figure 6: performance pftpd-pftp read to client /dev/null and disk

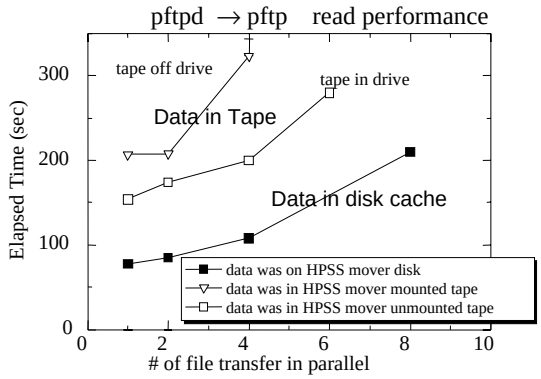

Figure 7: pftpd-pftp read to client disk from tape archive performance

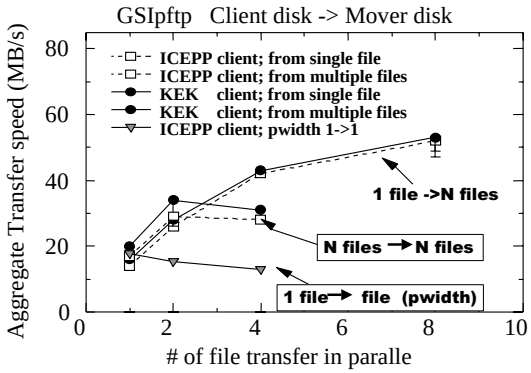

Figure 8: pftpd-pftp write to server cachedisk performance

4. GSI-enabled pftp

GridFTP[3] is a standard protocol for building data GRID and supports the featues of Grid Security Infrastructure (GSI), Multiple data channels for parallel transfers, partial file transfers, third-party transfer and reusable and authenticated data channels.

The pftp and ftp provided with HPSS software was not required or designed to support data Grid infrastructure. For future releases, HPSS Collaboration Members have introduced data Grid pftp requirements and the HPSS Technical Committee (TC) has convened a Grid Working Group to propose a development plan. As an interim and partial HPSS data Grid interface solution, the HPSS Collaboration is distributing the GSI-enabled pftp solution developed by Lawrence Berkeley National Laboratory (LBL). The HPSS TC is also working with the GridFTP development project underway at Argonne National Laboratory.

To acquire an HPSS data Grid interface necessary for our test bed, we requested and received a copy of Lawrence Berkeley National Laboratory's recently developed GSI-enabled pftp. The protocol itself is pftp but it supports GSI-enabled AUTH and ADAT ftp-command.

Table III commands in FTP protocol

GridFTP	GSI-enbled pftp
SPAS,SPOR,ETET	PBSZ,PCLO,PORPN,
ESTO,SBUF,DCAU	PPOR,PROT,PRTR,PSTO
AUTH,ADAT	
RFC959 commands	

As shown in table III which lists commands in each FTP protocol. while GSI-enabled pftp and GridFTP have different command set for parallel transfer, buffer management and Data Channel Authentication (DCA), the base command set is common. Fortunately unique functions to each protocol are optional and the two protocols are able to communicate. Installing and testing the GSI-enabled pftp proved that the GSI-enabled pftp daemon form LBL could be successfully accessed from gsinftp and url-copy (standard globus clients).

```
&(executable=gsim1)
(arguments=''-d'')
(inputfiles=
("Bdata.in"
"gsiftp://dt05s.cc:2811/hpss/manabe/data2"))
(stdout=datafiles.out)
(join=true)
(maxcputime="36000")
(middleware="nordugrid")
(jobname="HPSS access test")
(stdlog="grid_debug")%
(ftpThreads=1)
```

sample XRSL

As for performance measurement of 2GB file being accessed from HPSS, GSI-enabled pftp and normal kerberos pftp had equivalent elapsed time. Figure 9 shows aggregate transfer speed over the number of independent simultaneous file transfer. However, in the case where GSI enabled-pftpd server does not run on HPSS disk mover where accessed data resides, transfer speed halved. In original pftp where pftpd running in HPSS core server, data path is directly established between pftp client and disk mover. On the other hand, GSI-enabled pftp, data flow was from disk mover, via pftpd to client host. When the disk mover and pftpd server do not reside on the same host, two successive network transfer are incurred.

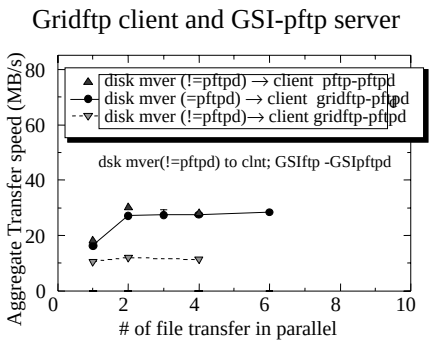

Figure 9: from GSI-enabled pftpd to Gridftp read performance

5. summary

ICEPP and KEK configured NorduGrid test bed with HPSS storage server over High speed GbE domestic WAN. Performance was measured several times for comparison between LAN and WAN access. From that, we found that network latency affected HPSS pftp and client API data transfer speed. The "GSI-enabled pftpd" developed by LBL was successfully adapted to the interface between Grid infrastructure and HPSS.

Our paper is a report on work-in-progress. Final results require that the questions relative to raw TCP performance, server/client protocol traffic, and pftp a protocol be further evaluated; that any necessary modifications or parametric changes be acquired form our HPSS team members; and that measurements be taken again. Further understanding of the scalability and the limits of multi-disk mover configurations would be gained from measuring HPSS network utilization and performance using higher performance network interfaces adapters, system software and infrastructure, and processor configurations.

References

[1] http://www.openpbs.org
[2] http://www.netperf.org
[3] http://www.globus.org/datagrid/gridftp.html
[4] http://www.sdsc.edu/hpss/
[5] http://www.nordugrid.org, You can find NorduGrid papers in this proceedings too.
[6] S.Yashiro et. al., "Data transfer using buffered I/O API with HPSS", CHEP'01, Beijing, Jul.2001