On enhancing GridFTP and GPFS performances

To cite this article: A Cavalli et al 2010 J. Phys.: Conf. Ser. 219 052024

View the article online for updates and enhancements.

Related content
- StoRM-GPFS-TSM: A new approach to hierarchical storage management for the LHC experiments
  A Cavalli, L dell’Agnello, A Ghiselli et al.
- Identifying Gaps in Grid Middleware on Fast Networks with the Advanced Networking Initiative
  Dave Dykstra, Gabriele Garzoglio, Hyunwoo Kim et al.
- A comparison of data-access platforms for BaBar and ALICE analysis computing model at the Italian Tier1
  A Fella, F Furano, L Li Gioi et al.
On Enhancing GridFTP and GPFS performances

A. Cavalli\textsuperscript{1}, C. Ciocca\textsuperscript{1}, L. dell’Agnello\textsuperscript{1}, T. Ferrari\textsuperscript{1}, D. Gregori\textsuperscript{1}, B. Martelli\textsuperscript{1}, A. Prosperini\textsuperscript{1}, P. Ricci\textsuperscript{1}, E. Ronchieri\textsuperscript{1}, V. Sapunenko\textsuperscript{1}, A. Sartirana\textsuperscript{2}, D. Vitlacil\textsuperscript{1} and S. Zani\textsuperscript{1}

\textsuperscript{1}INFN CNAF, Bologna, Italy
\textsuperscript{2}LLR - École Polytechnique, Paris, France

E-mail: \textsuperscript{1}elisabetta.ronchieri@cnaf.infn.it, vladimir.sapunenko@cnaf.infn.it
E-mail: \textsuperscript{2}sartiran@llr.in2p3.fr

Abstract. One of the most demanding tasks which Computing in High Energy Physics has to deal with is reliable and high throughput transfer of large data volumes. Maximization and optimization of the data throughput are therefore key issues which have to be addressed by detailed investigations of the involved infrastructures and services.

In this note, we present some transfer performance tests carried out at the INFN-CNAF Tier-1 center, using SLC4 64-bit Grid File Transfer Protocol (GridFTP) servers and a disk storage system based on the General Parallel File System (GPFS) from IBM. We describe the testbed setup and report the measurements of throughput performances in function of some fundamental variables, such as number of parallel file and number of streams per transfer, concurrent read and write activity and size of data blocks transferred.

During this activity, we have verified that a significant improvement in performances of the GridFTP server can be obtained using 64bit version of Operating System and GPFS from IBM.

1. Introduction

The optimization of Wide Area Network (WAN) data distribution in the framework of Grid Computing is one of the key issues in the deployment, commissioning and operation of the computing models of the current generation of High Energy Physics (HEP) experiments. The computing infrastructures of these experiments are required to handle huge amounts of data that have to be stored and analysed. Moreover, such computing systems are built by resources which are worldwide distributed and connected by the Grid middleware services.

The Grid infrastructure manages data transfers by a set of layered services: the File Transfer Service [1], the Storage Resource Managers [2], and Grid File Transfer Protocol (GridFTP) servers [3]. On the lowest level, the GridFTP protocol is used to efficiently manage high-performance, reliable and secure point to point data transfer. GridFTP servers are standard components of the Globus middleware suite [4] and directly interface with the different file system technologies at fabric level.

In this paper, we present the results of some tests focused on the interaction and performance issues in a storage setup which combines GridFTP servers with the IBM General Parallel File System (GPFS) [5]. The aim of such tests is to measure the evolution of performances in function of some fundamental variables: number of parallel transfers, number of streams per transfer, concurrent read and write activity, and size of data blocks transferred.
The testbed was deployed at the INFN-CNAF Tier-1 center, located in Bologna, Italy. We employed a GPFS file system built over a Storage Area Network (SAN) infrastructure. GridFTP servers were configured on two SAN-attached GPFS running Scientific Linux CERN 4 (SLC4) 64-bit Operating System. During the testing activity, all the main metrics were measured: CPU load, memory usage, network utilization, data read and write rates.

The rest of this paper is organized as follows. Section 2 presents details of the deployed GridFTP and GPFS testbed. Section 3 summarizes all the selected case studies. Section 4 briefly describes how we ran tests and measured the metrics. Section 5 is devoted to the presentation of the results. Conclusions and outlook are given in Section 6.

2. Experimental Deployment of GridFTP and GPFS

In this section we briefly describe the facilities deployed to perform the tests. Figure 1 shows the schema of the SAN connecting the storage, the IBM GPFS servers and the GridFTP servers.

![Figure 1. Experimental Deployment of GridFTP and GPFS.](image)

**Storage** - We have used four Logical Unit Numbers (LUNs) each of 8 TB. The LUNs are served by an EMC CLARiiON CX4-960 [6] subsystem.

**Blade Servers** - There are six M600 Blades from Dell PowerEdge M1000e Modular Blade Enclosure [7]. Each blade server has been equipped with:

- 2 quad core Intel(R) Xeon(R) CPU E5410 at 2.33 GHz;
- 16 GB of RAM;
- 2 SAS Hard Drives of 130 GB each configured in RAID 1 using on-board LSI Logic / Symbios Logic SAS1068E controller and used as local disk.

The six blade servers have been divided in two groups: one of them is composed by four blades that run the GPFS servers, Network Shared Disk (NSD), and the other one by two
blades used as GridFTP servers with direct access to GPFS file system by Fibre Channel (FC) connection on external storage. This configuration allows for a complete separation between GridFTP and GPFS data flow as it shown on Figure 1.

**Network** - There are dual-channel QLogic Corporation QLA2432 4 Gb/s FC adapter and two onboard Gigabit Ethernet ports (only one has been used). The interconnection architecture with external storage is supported by 4 Gb/s optical FC links connecting each blade with two Brocade M4424 switches [8] installed inside the Dell M1000e Blade Enclosures, while the Ethernet interconnection, we have used, has 48-port Extreme X450 [9] switch with 10 Gb/s uplink.

### 3. Case Studies

The tests aim to independently evaluate the performances of the different layers of a typical GPFS-based storage system. In order to do so we have defined three different case studies, each one testing a different layer of the storage system. In this section we give a brief description of the different case studies:

**Case I** - *GPFS bare performance on a SAN node.* For preliminary tests we have used simple copy (i.e., cp) command to check transfer rates between memory, local and SAN-attached disks.

---

**Figure 2.** *GPFS bare performance on a SAN node - CP from-to GPFS.*

**Figure 3.** *GPFS bare performance on a SAN node - CP from GPFS to /dev/null.*

**Figure 4.** *GPFS bare performance on a SAN node - CP from /tmp to GPFS.*

**Figure 5.** *GPFS bare performance on a SAN node - CP from /tmp to GPFS.*
This test aims to measure the performances for a simple file copy on a SAN node. Files can be contemporaneously read from and written to the GPFS file system as described in Figure 2. They can be as well read from the GPFS file system and written either to the null device (i.e., /dev/null) or to a temporary directory (i.e., /tmp) where the local disk is used) as shown in Figures 3 and 4 respectively. Finally, files can be also read from /tmp and written to the GPFS file system as described in Figure 5.

In UNIX, the null device is a special file, which discards all data written to it reporting that the write operation succeeded, and provides no data to any process that reads from it returning end-of-file immediately [10].

It is important to evaluate which disk I/O devices may be considered a bottle neck having FC disks, i.e. GPFS partition, as the input stream and local disk /tmp partition as the output data stream. The intention was also to evaluate which factors introduce a performance worsening reading from GPFS and writing to /dev/null and then reading from GPFS and writing to /tmp: if performances were the same, read from GPFS would be the problem, considering that local disk access should be slower than FC disk.

Case II - GPFS performances on a SAN enabled GridFTP server.

This test aims to measure the GridFTP transfers performances over the SAN infrastructure. Files can be read from the GPFS file system and written to /tmp, and vice versa as shown in Figures 6 and 7. They can be contemporaneously read from and written to the GPFS file system as described in Figure 8. In each example a GridFTP server is used.

In Case I a given file is copied from a block device to another block device or memory, while in this case the file is transferred reading from a block device and then sending packets to one or more TCP sockets, therefore GridFTP introduces overheads that may affect final performances. However, having configured network parameters in order to adopt loopback device (i.e., a block device that can be used as a disk but really points to an ordinary file somewhere) during the usage of GridFTP with IP address or with localhost, no network layer is introduced.

Case III - GPFS performances with GridFTP transfers among two SANs enabled GridFTP servers.

This last test measures the performances of GridFTP transfers among two SANs enabled servers connected by a 1 Gb plus 1 Gb Ethernet link (i.e., up to 1 Gb for each direction). Both unidirectional and bidirectional transfers have been tested. Files can be contemporaneously read from and written to the GPFS file system by using two different GridFTP servers as described in Figures 9 and 10. Figure 9 shows that each GridFTP can read from and write to GPFS, while Figure 10 details that one GridFTP is used to read from GPFS and the other one to write to GPFS.

In this case a given file is read from a block device via a GridFTP server to 1 Gb link and written to the same block device using another GridFTP server to 1 Gb link. In Case III, therefore, the 1 Gb LAN connection limits the performances to ~ 25% of those observed in Case II during transfers from-to GPFS.

4. Testbed Setup

In this section we briefly describe the testbed used for running the different tests and measure the metrics. The setup was very simple and relied on basic UNIX and Grid middleware command wrapped in a perl script. The latter just allowed for constant submission of transfers, keeping a stable load, and managed the scheduling of various tests in sequence. In particular, we submitted GridFTP transfers using the globus-url-copy GridFTP client available with the Globus Toolkit middleware distribution, version 4.

The measurement of the metrics on the servers was provided by dstat monitoring tool. This is a python based command line tool which allows for measurement of the usage of CPU, the
network activity, the load average, and others. Through a plugin mechanism \texttt{dstat} can interface and measure the performance of various devices. External plugins are enabled by using a single
The used command line
\[ \texttt{dstat} -M \text{time,epoch,cpu,mem,load,gpfs,net 30} \]  
provides logging of the following information: time stamp, both in hour and epoch format; CPU usage; memory usage; overall load; gpfs input and output rates; inbound and outbound network rates. All values are averaged over thirty second intervals.

Table 1 summarises all the statistics provided by the \texttt{dstat} command. The details of CPU usage are averaged over all the processors in the system.

| Statistic | Event | Description |
|-----------|-------|-------------|
| CPU       | usr   | It shows the percentage of time taken by user processes. |
|           | sys   | It identifies the system processor usage by the kernel. |
|           | idl   | It is the percentage of time the CPU is sitting idle. |
|           | wai   | It shows the time tasks are waiting for I/O to complete. |
|           | hiq   | It is the percentage of time spent processing hardware interrupts. |
|           | siq   | It is the percentage of time spent processing software interrupts. |
| Load      | 1 min | It shows an average of the number of runnable processes waiting for the CPU. |
|           | 5 min | It shows an average of the number of runnable processes waiting for the CPU. |
|           | 15 min| It shows an average of the number of runnable processes waiting for the CPU. |
| Mem       | used  | It is the used memory in bytes. |
|           | buff  | It is the buffered memory in bytes. |
|           | cach  | It is the cached memory in bytes. |
|           | free  | It is the free physical memory in bytes. |
| GPFS      | read  | It is the number of byte read. |
|           | write | It is the number of byte written. |
| Network   | receive | It is the number of byte received. |
|           | send  | It is the number of byte sent. |

5. Simulation Results
In this section we summarize the tests results. Tables 2 and 3 show respectively testbed information for software, and set up characteristics used for performing experiments.

All of them have been executed varying the values for the relevant parameters such as the number of transferred parallel files and the number of streams per file expressed with M and N respectively. The tests have allowed to collect a lot of useful information on the behavior and performances in accessing a typical size file on a GPFS storage by direct POSIX access or by
Table 2. Testbed software details.

| Type            | Version       |
|-----------------|---------------|
| OS              | SLC4 64-bit   |
| dstat           | 0.6.6         |
| GPFS            | 3.2.1-9       |
| globus gridFTP  | 2.3           |
| globus-url-copy | 3.2.1         |
| cp              | 5.2.1         |

Table 3. Set up characteristics.

| Parameter       | Value       |
|-----------------|-------------|
| GPFS page pool  | 8 GB        |
| default GPFS cache | 2.4 GB   |
| default TCP buffer | 262 KB   |
| max TCP buffer  | 2 GB        |

GridFTP. We have verified a low CPU load looking at case simulations. Plots regarding some of the most relevant performed tests are reported in this section together with summary results.

**Case I - GPFS bare performances on a SAN node.**

Up to five parallel file copies have been tested. In this case network does not have any role. The GPFS file system has shown unidirectional read (i.e., GPFS-/dev/null where disk is not used) and write (i.e., /tmp - GPFS) performances up to $\sim 0.50$ GB/s as shown in Figures 11 and 12.

![Figure 11](image1.png) **Figure 11.** Bare reading rates from GPFS as function of M.

![Figure 12](image2.png) **Figure 12.** Bare writing rates on GPFS as function of M.

Maximum sustained rate of contemporaneous read and write operations reaches $\sim 0.3$ GB/s as shown in Figure 13. The performance seems to smoothly decrease to 0.15 GB/s as the number of parallel files increase up to five. Figure 14 highlights this behaviour up to twenty parallel file copies.

**Case II - GPFS performances on a SAN node enabled GridFTP server.**

Up to twenty parallel files and ten streams per file have been tested. In this case GridFTP is directly SAN-attached, therefore network throughput is almost zero. Performances vary from 0.25-0.3 GB/s read and write rate with 1-2 parallel transfers down to 0.15-0.1 GB/s with 5-10 parallel transfers. This seems to be fairly independent from the number of streams used in a single GridFTP transfer. Figures 15 and 16 show the GPFS...
throughput in read and write respectively, given an $i$ stream and a $j$ parallel file with $i \in N$ and $j \in M$.

Table 4 reports the average of GPFS throughput in read and write, transferring a file from GPFS to \texttt{/tmp} and vice versa specifying the number of streams $N$ and the number of parallel files $M$.

**Case III - GPFS performances with GridFTP transfers among 2 SANs enabled GridFTP servers.**

Up to twenty parallel files and ten streams per file have been tested. In this case GridFTP resources are directly SAN-attached.

Unidirectional transfers between the two GridFTP servers can be sustained saturating the 1 Gb Ethernet link. This is independent from the number of parallel transfers and the
Table 4. Average of GPFS throughput GB/s.

| Test Type  | N  | M  | Write | Read | N  | M  | Write | Read |
|------------|----|----|-------|------|----|----|-------|------|
| /tmp-GPFS  | 1  | 1  | 0.314 | 0    | 10 | 1  | 0.323 | 0    |
| /tmp-GPFS  | 1  | 5  | 0.495 | 0    | 10 | 5  | 0.489 | 0    |
| /tmp-GPFS  | 1  | 10 | 0.485 | 0    | 10 | 10 | 0.463 | 0    |
| GPFS-/tmp  | 1  | 1  | 0    | 0.171| 10 | 1  | 0    | 0.155|
| GPFS-/tmp  | 1  | 5  | 0    | 0.081| 10 | 5  | 0    | 0.093|
| GPFS-/tmp  | 1  | 10 | 0    | 0.071| 10 | 10 | 0    | 0.072|

stream per file. Figures 17 and 18 highlight GPFS performances with two GridFTP servers used to read from GPFS and write to GPFS respectively.

![Figure 17](image1.png)  
**Figure 17.** Reading rates from GPFS as function of N and M by the 1st GridFTP.

![Figure 18](image2.png)  
**Figure 18.** Writing rates from GPFS as function of N and M by the 2nd GridFTP.

Bidirectional transfers among the two servers have shown as well to be able to saturate 1Gb/s Ethernet link with a $\sim 0.24$ GB/s read (write) performance (1 Gbit = $\sim 0.12$ GB/s). Figures 19 show GPFS throughput with read by the 1st GridFTP and write by the 2nd GridFTP (i.e., the GPFS throughput in the other sense is almost the same to this one). The saturation actually takes place for five or more parallel transfers. Figures 19 also detail with a single transfer (i.e., M=1) the overall read/write rate is $\sim 0.08$ GB/s.

Figures 17, 18, 19(a), and 19(b) show the GPFS throughput in read and write respectively, given an $i$ stream and a $j$ parallel file with $i \in N$ and $j \in M$.

The performance dependency on the number of parallel files can be explained by the usage of the operative system buffer: this needs further investigation.

6. Conclusions

In this paper, we presented some tests which have been performed at INFN-CNAF Tier-1 with GridFTP servers on a GPFS storage system. In order to evaluate independently the
performances of the different layers involved in a typical data transfer task we have identified three case studies: GPFS bare performance on a SAN node, GPFS performances on a SAN enabled GridFTP server and GPFS performances with GridFTP transfers among 2 SAN enabled GridFTP servers. The testbed setup was described. We reported the measurements of throughput performances in function of the number of parallel transferred files and the number of streams per transfer. The results show a strong dependence of the transfer rate on the number of parallel files. We need further investigations in order to fully understand such dependence.

In addition it would be interesting to also verify this layout in interaction with the SRM layer in order to optimize both the number of streams per transfer and the number of parallel files, and see how system scales when the rate of transfer requests increases. This activity is part of the planned future tests.

Figure 19. GPFS rates as function of N and M, reading by the 1st GridFTP and writing by the 2nd GridFTP.

References
[1] E. Laure, S. M. Fisher, A. Frohner, C. Grandi, P. Kunszt, A. Krenek, O. Mulmo, F. Pacini, F. Prelz, J. White, M. Barroso, P. Buncic, F. Hemmer, A. Di Meglio, and A. Edlund, "Programming the Grid with gLite," In Journal of Computational Methods in Science and Technology, Vol. 12, No. 1, pp. 33–45, 2006.
[2] A. Shoshani, A. Sim, and J. Gu, "Storage Resource Managers: Middleware Components for Grid Storage," In Proceedings of NASA CONFERENCE PUBLICATION, pp. 209–224, 2002, http://romulus.gsfc.nasa.gov/msst/conf2002/papers/d02ap-ash.pdf.
[3] W. Allock, "GridFTP: Protocol Extensions to FTP for the Grid," Global Grid Forum GFD-R-P.020, 2003.
[4] Globus middleware, http://www.globus.org/.
[5] F. Schmuck, and R. Haskin, "GPFS: A shared–disk file system for large computing clusters," In Proceedings of the 2002 Conference on File and Storage Technologies (FAST), 2002.
[6] CLARIION CX4 Model 960, http://www.emc.com/products/detail/hardware/clarion-cx4-model-960.htm.
[7] PowerEdge M1000e Modular Blade Enclosure, http://www.dell.com/content/products/productdetails.aspx?c=us&cs=555&l=en&s=biz.
[8] Brocade M4424 4Gb Fibre Channel Switch for M1000e-Series Blade Enclosures, http://www.dell.com/content/products/productdetails.aspx?c=us&cs=555&l=en&s=biz.
[9] SUMMIT X450e, http://www.extremenetworks.com/products/switch-x450e.aspx.
[10] The Open Group Base Specifications, Issue 6, IEEE Std 1003.1, 2004 Edition, Copyright 2001-2004 The IEEE and The Open Group, http://www.opengroup.org/onlinepubs/009695399/.