ON ENHANCING GridFTP AND GPFS PERFORMANCES

A. Cavalli, C. Ciocca, L. dell'Agnello, T. Ferrari, D. Gregori, B. Martelli, A. Prosperini, P. Ricci, E. Ronchieri[1], V. Sapunenko[1], A. Sartirana[2], D. Vitlacil, S. Zani

INFN CNAF, Bologna, Italy, {elisabetta.ronchieri,vladimir.sapunenko}@cnaf.infn.it[1]

GRIF, Paris, France, sartiran@llr.in2p3.fr[2]

Abstract

Many High Energy Physics experiments must transfer large volumes of data. Therefore, the maximization of data throughput is a key issue, requiring detailed analysis and setup optimization of the underlying infrastructure and services. In Grid computing, the data transfer protocol called GridFTP is widely used for efficiently transferring data in conjunction with various types of file systems.

We focus on the interaction and performance issues in a setup, which combines GridFTP server with the IBM General Parallel File System (GPFS), adopted for providing storage management and capable of handling petabytes of data and billions of files. A typical issue is the size of the data blocks read from disk used by the GridFTP server version 2.3, which can potentially impair the data transfer threshold achievable with an IBM GPFS data block.

We propose an experimental deployment of GridFTP server characterized by being on a Scientific Linux Cern 4 (SLC4) 64-bit platform, having GridFTP server and IBM GPFS over a Storage Area Network (SAN) infrastructure aimed to improve data throughput and to serve distributed remote Grid sites. We present the results of data-transfer measurements, such as CPU load, network utilization, data read and write rates, obtained performing several tests at INFN Tier1 where the described deployment has been setup. During this activity, we have verified a significant improvement of the GridFTP performances (of almost 50%) on SLC4 64-bit over SAN saturating the Gigabit with a very low CPU load.

Test Descriptions

Several tests were performed with the aim to independently evaluate the performances of the different layer of a typical GPFS-based storage system. All test have been performed varying the values for the relevant parameters, as the number of parallel files transferred and the number of streams per file. Tests can be divided into three groups:

Group 1: GPFS bare performance on a SAN node. This test simply measures the performance for a simple file copy on a node included in the SAN. Files can be read, written or contemporaneously read and written on the GPFS file system. Up to 5 parallel file copies were tested.

Group 2: GridFTP performances on a SAN enabled FTP server.
This test measures read (or write) performances of FTP transfers from (or to) GPFS storage on a FTP server included in the SAN. Up to 20 parallel files and 10 stream per file were tested.

Group 3: FTP transfers among 2 SAN enabled FTP servers. This test measures the performances for FTP transfers on the 1Gb + 1Gb link between a couple of servers both included in the SAN. Both unidirectional and bidirectional transfers were tested, up to 20 parallel files and 10 stream per file.

Tests were performed on a 32TB GPFS file system. Servers were SAN enabled SLC4 64-bit GridFTP (globus 2.3), 2 CPUs quad-core, 16GB of memory.

Conclusions

The tests 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 FTP. On the left part of the poster you may find plots regarding some of the most relevant tests performed.

Group 1: Tests from group 1 measured the bare read/write performance from a node included in the SAN. The GPFS showed unidirectional read/write performances up to 500-550MB/s. Contemporaneous reading and writing from the file system can be sustained at ~300 [MB/s]. The latter performance seems to smoothly decrease to 150 [MB/s] as the number of parallel files increase up to 5.

Group 2: Tests from group 2 measured the read/write performance to/from a single FTP server included in the SAN. Performances vary from 250-300 [MB/s] read/write rate with 1-2 parallel transfers down to 150-100 [MB/s] with 5-10 parallel transfers. This seems to be fairly independent from the number of streams used in a single FTP transfer.

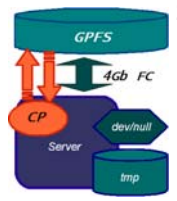
Group 3: Tests from group 3 measured the transfers via LAN between two SAN nodes FTP servers (both reading and writing on the same GPFS file system). Unidirectional transfers from between the 2 servers can be sustained saturating the 1 Gb Ethernet link. This is independent from the number of parallel transfers and if stream per file.

Bidirectional transfer among the two servers showed as well to be able to saturate the two 1 Gb network interfaces with a ~120 [Mb/s] read/write performance. The saturation actually takes place for 5 or more parallel transfers. With a single transfer the overall read/write rate is ~80 [MB/s].

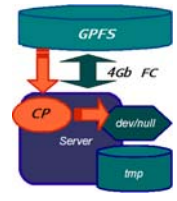
The performance dependency on the number of parallel files can be explained by the usage of the operative system buffer: this needs further investigation.

Tests Group 1: results

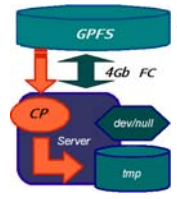
1.1



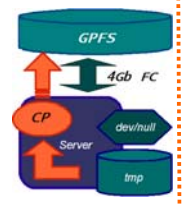
1.4



1.2



1.3



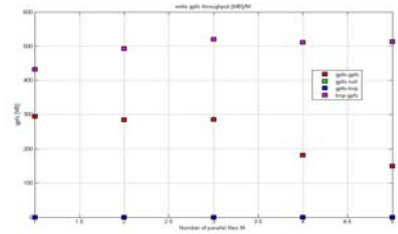


Fig. 1 – Average of GPFS throughput performance for "cp" write operations

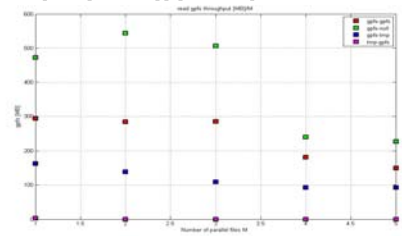
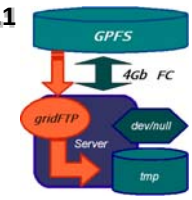


Fig. 2 – Average of GPFS throughput performance for "cp" read operations

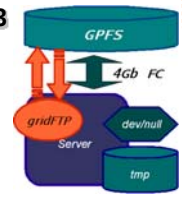
Network Throughput ≈ 0 [B/s]

Tests Group 2: results

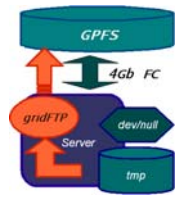
2.1



2.3



2.2



Type Tests	N	M	Av. of gpfs thr. [MB] in w	Av. of gpfs thr [MB] in r
gpfs -> gpfs	1	1	275	248
		5	136	136
		10	153	153
	10	1	214	214
		5	133	133
		10	151	151
	20	1	210	210
		5	134	134
		10	152	152
gpfs -> /tmp	1	1	0	171
		5	0	81
		10	0	71
	10	1	0	155
		5	0	93
		10	0	72
	20	1	0	148
		5	0	86
		10	0	73
/tmp -> gpfs	1	1	314	0
		5	495	0
		10	485	0
	10	1	323	0
		5	489	0
		10	463	0
	20	1	317	0
		5	477	0
		10	469	0

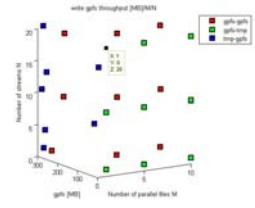


Fig. 3 – Average of GPFS throughput performance for globus-url-copy write operations

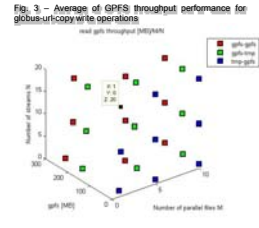
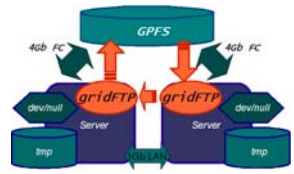


Fig. 4 – Average of GPFS throughput performance for globus-url-copy read operations

Network Throughput ≈ 0 [B/s]
(GridFTP directly linked to FC)

Tests Group 3: results

3.1



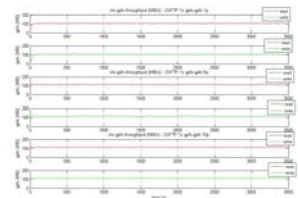


Fig. 5 – read/write GPFS throughput with 1,5,10 parallel transfers and 1 stream per transfer

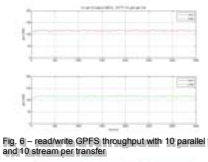


Fig. 6 – read/write GPFS throughput with 10 parallel transfers and 10 stream per transfer

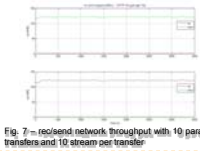
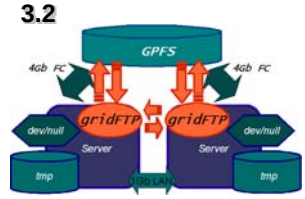


Fig. 7 – rec/send network throughput with 10 parallel transfers and 10 stream per transfer

3.2



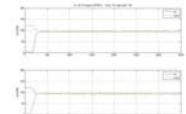


Fig. 8 – rec/send network throughput with 10 parallel transfers and 10 stream per transfer from/to the 2 GridFTP servers

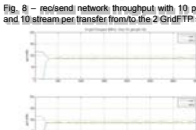


Fig. 9 – read/write GPFS throughput with 10 parallel transfers and 10 stream per transfer from/to the 2 GridFTP servers

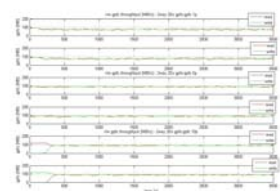


Fig. 10 – read/write GPFS throughput with 1,5,10 parallel transfers and 20 streams per transfer from/to the 2 GridFTP servers