

Performance Comparison of GPFS 1.3 and GPFS 1.4 For POSIX and MPI -IO

Richard M. Hedges and William E. Loewe

July 26, 2001

U.S. Department of Energy

Lawrence
Livermore
National
Laboratory

DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.

This report has been reproduced
directly from the best available copy.

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information
P.O. Box 62, Oak Ridge, TN 37831
Prices available from (423) 576-8401
<http://apollo.osti.gov/bridge/>

Available to the public from the
National Technical Information Service
U.S. Department of Commerce
5285 Port Royal Rd.,
Springfield, VA 22161
<http://www.ntis.gov/>

OR

Lawrence Livermore National Laboratory
Technical Information Department's Digital Library
<http://www.llnl.gov/tid/Library.html>

Performance Comparison of GPFS 1.3 and GPFS 1.4 For POSIX and MPI -IO

This report by the SIOP observes the effects of recent hardware and software changes for parallel I/O performance to the GPFS parallel filesystem. The IBM SP machine (frost) has been upgraded from Mohonk with GPFS 1.3 to Mohonk2 with GPFS 1.4. In addition, the Colonyswitch adapters have been upgraded from Single/Single to Double/Single.

The tests discussed here were performed on frost using 60 compute nodes and the GPFS filesystem using 2 dedicated I/O nodes (servers).

The tests performed utilize the POSIX and MPI -IO interface to GPFS. The noted system changes to frost have improved both POSIX and MPI -IO peak read performance and have not diminished peak write performance. We note that as the bandwidth of mounted disks is near fully utilized, there was no expectation of significant performance improvement.

For POSIX, the best write rate did not change from 550 MB/sec. The read rates improved from 500 MB/sec to 600 MB/sec, however.

For MPI -IO, the best write rates did not change from 550 MB/sec. The read rates improved from 470 MB/sec to 570 MB/sec, in line with the improvements observed.

The MPI -IO discontinuous test results show that improvement is significant (nearly a factor of 2 beyond 40 nodes). The performance of this particular test is more sensitive to the improved switch performance characteristics because the data passes across the switch twice: once in the MPI -IO data shipping phase to assemble large block; and then to write the data out to the disks as this requires two passes across the switch.

MACHINE CHARACTERISTICS

Frost, the IBM SP/6000 used for these tests, has 64 compute nodes, 2 dedicated GPFS I/O server nodes, and 1 login node.

As of May 2001, frost nodes were connected to a Colony switch by single/single adapters. This configuration on frost had the two server nodes each serving 12 logic disks through 4 Cambex PC1000 fibre channel HBAs (2 each in 2 RIO drawers). The disk server system is a Data Direct Next SDD300 using a dual hstd (high speed traffic director) system with each hstd containing 4 fibre channel ports connected to the server node. The backend of the SDD (San Data Director) had 10 fibre channel loops connected to 8 data disks, 1 parity disk, and 1 spare. With 36 tiers in this system, 24 disks were created at about 860 GB each. The total transfer rate of these disks is approximately 700 MB/sec. This hardware configuration ran the PSSP3.2 (Mohonk) with GPFS 1.3.

In late June 2001, frost was upgraded to double/single switch adapters and PSSP3.3 (Mohonk2) with GPFS 1.4. Except for the new adapters, this hardware configuration on frost is the same as the earlier configuration.

TESTS & RESULTS

A maximum of 60 compute nodes were used due to a few nodes being unavailable during test times. The parallel codes `ior_posix.c` and `ior_mpiio.c` were used for various access patterns for POSIX and MPI -IO. For discontinuous testing under MPI -IO, `mpiio_discontig.c` was used.

Three general data layout patterns have been used: 1) a *segmented* pattern, where all data from a single process will be written to a contiguous portion of the file; 2) *strided*, where data from the processes is interleaved in the file, and 3) *discontiguous*, where the file is divided into 1k blocks, which are then randomly assigned to the processes.

POSIX

The first round of testing using GPFS 1.3 used the POSIX interface with strided and segmented access patterns, varying the node count and transfer size. The comparable tests were performed on GPFS 1.4.

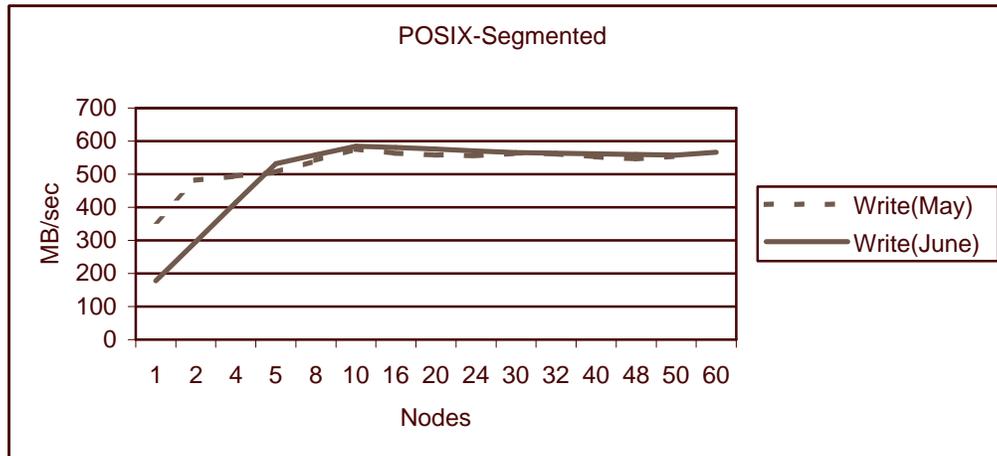
The hardware/software changes in moving from GPFS 1.3 to GPFS 1.4 has led to improvement in the POSIX read performance and has not diminished write performance. As the disks were already saturated, there was not an expectation of performance improvement beyond 700 MB/sec.

So, despite read rates increasing, it was no surprise that write rates have not changed. For POSIX, the best write rates did not change from 550 MB/sec.

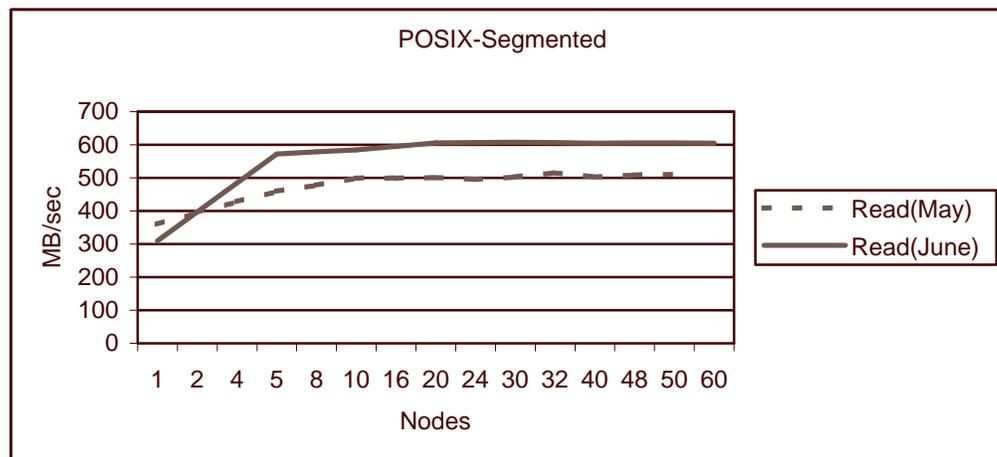
Segmented

Testing for scalability with a segmented pattern, test 1.4 is as follows:

Clients/Node:1 Filesize:.5GB -30GB Code:ior_posix.c May:GPFS1.3
Nodes:1 -60 Transfersize:512KB Machine:frost.llnl.gov June:GPFS:1.4



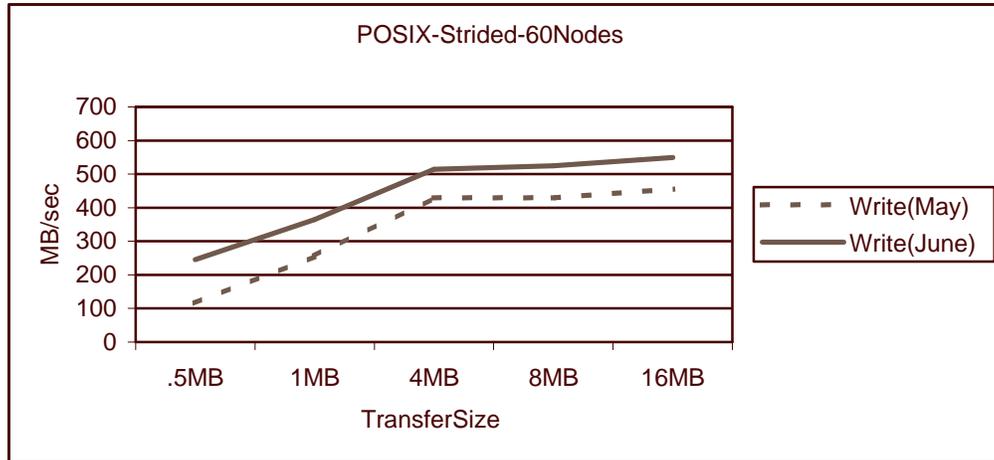
The read rates improved from 500 MB/sec to 600 MB/sec, from GPFS 1.3 to GPFS 1.4.



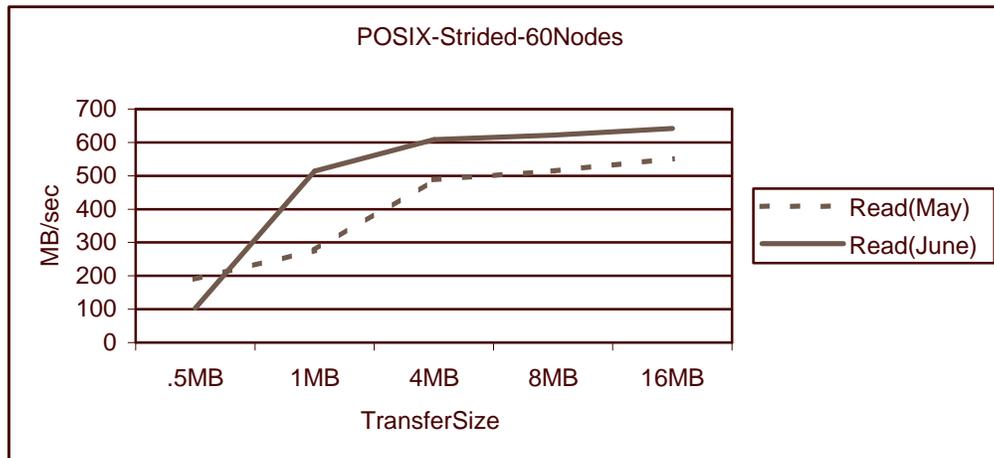
Strided

For writes to a strided access pattern, the performance increased nearly 100 MB/sec for all stride sizes.

Clients/Node:1 Filesize:30GB Code:ior_posix.c May:GPFS1.3
Nodes:60 Stride size:.5MB –16MB Machine:frost.llnl.gov June:GPFS:1.4



For reads from a strided access pattern, the performance increased by approximately 100 MB/sec for all stride sizes.



Further, it is interesting to note that for the strided pattern in test 1.1 with transfer sizes set to 16MB, the maximum read is 549 MB/sec on GPFS 1.3. Though with only 3 repetitions of each test, it appears that this is nearly 7% faster than any of the segmented 1.4 tests for read. The same is true on GPFS 1.4 for the strided pattern in test 1.1 with transfer sizes set to 16MB. This read rate of 642 MB/sec, too, is better than any of the segmented test on GPFS 1.4 by nearly 6%.

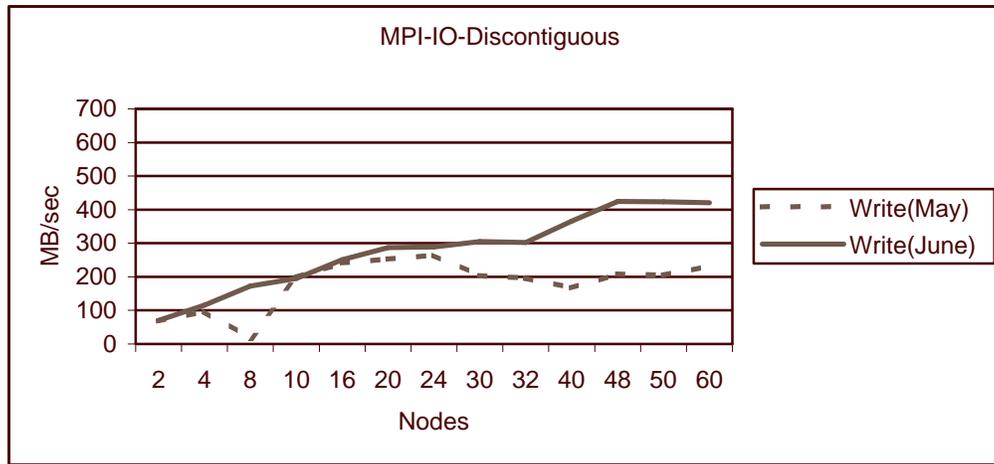
MPI -IO

The MPI -IO read performance, too, has improved and the write performance has maintained the maximum writerateset by the disks.

For MPI -IO, the best writerates did not change from 550 MB/sec. The read rates improved from 470 MB/sec to 570 MB/sec.

Discontiguous

Clients/Node:1	Filesize:30GB	Code:mpiio_discontig.c	May:GPF1.3
Nodes:2 -60	Blocksize:1KB	Machine:frost.llnl.gov	June:GPF1.4



The MPI -IO Discontiguous test results show that the improvement for data shipping is significant (nearly a factor of 2 beyond 40 nodes). It is not worthy as this write pattern requires two passes across the switch. This suggests that with reducing the bottleneck at the disks, there might be an improvement in performance since the switch could handle more traffic faster.

Segmented

Clients/Node:1

Filesize:30GB

Code:ior_mpiio.c

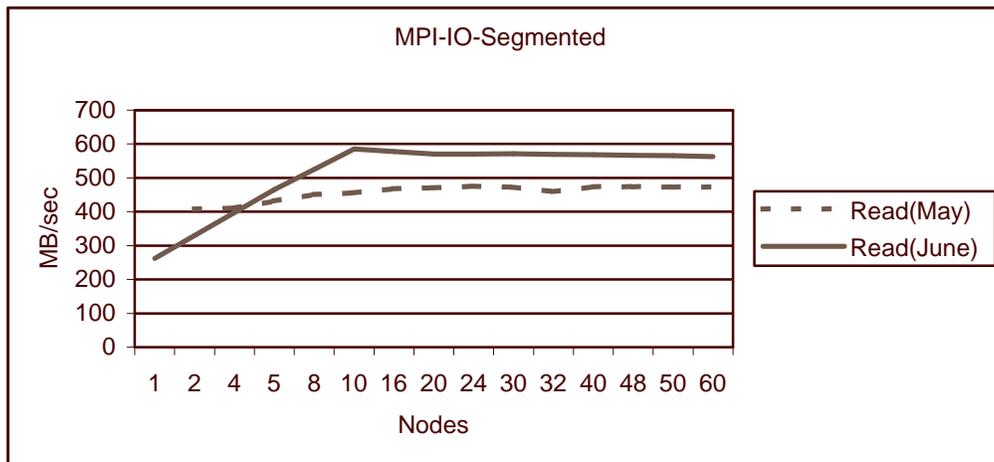
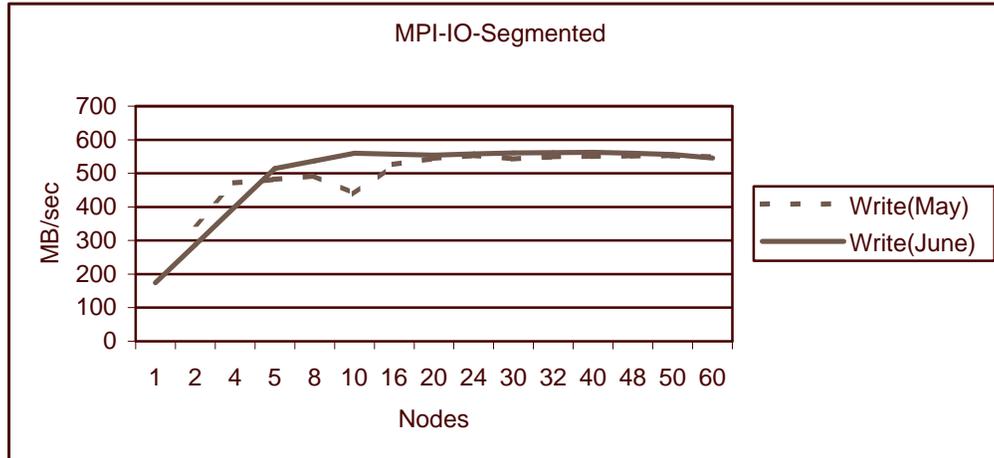
May:GPFS1.3

Nodes:1 -60

Transfersize:512KB

Machine:frost.llnl.gov

June:GPFS:1.4



OUTPUTDATA

Test: Clients/Node:1 Filesize:30GB Code:ior_posix.c May:GPFS1.3
 1.1 Nodes:60 Transfersize:.5 -16MB Machine:frost.llnl.gov June:GPFS:1.4

Interface Pattern Filesize Nodes Reps
 POSIX Strided 30G 60 3

[Note:AllRatesinMB/sec]

Blocksize	Transfersize	29-May	Write	Read
.5MB	.5MB		116	245
1MB	1MB		256	365
4MB	4MB		429	514
8MB	8MB		429	525
16MB	16MB		456	549

		28-Jun	Write	Read
.5MB	.5MB		190	103
1MB	1MB		277	514
4MB	4MB		488	608
8MB	8MB		514	622
16MB	16MB		551	642

Test: Clients/Node:1 Filesize:30GB Code:ior_posix.c May:GPFS1.3
 1.4 Nodes:60 Transfersize: 512KB Machine:frost.llnl.gov June:GPFS:1.4

Interface Pattern Transfersize Nodes Reps
 POSIX Segmented 512KB 60 3

[Note:AllRatesinMB/sec]

Nodes	29-May	Write	Read
2		359	360
4		482	395
8		506	460
10		541	478
16		576	499
20		563	499
24		558	501
30		556	495
32		563	503
40		562	515
48		553	503
50		546	509
60		555	510

Test:1.4(cont.)

	28-Jun	Write	Read
1		178	310
5		532	572
10		585	585
20		576	606
30		565	608
40		562	605
50		558	606
60		566	605

Test: Clients/Node:1
2.1 Nodes:2 -60

Filesize:30GB
Blocksize:1KB

Code:mpio_discontig.c
Machine:frost.llnl.gov

May:GPFS1.3
June:GPFS:1.4

Interface	Pattern	Reps	Blocksize	
MPI-IO	Discontiguous	5	1KB	
		[Note:AllRatesinMB/sec]		
Filesize	Nodes	29-May	Write	28-Jun
1GB	2		68	70
2GB	4		98	116
4GB	8		17	172
5GB	10		198	195
8GB	16		242	251
10GB	20		253	287
12GB	24		264	289
15GB	30		204	305
16GB	32		196	302
20GB	40		167	366
24GB	48		208	424
25GB	50		205	423
30GB	60		232	421

Test: Clients/Node:1 Filesize:30GB Code:ior_posix.c May:GPFS1.3
 2.3 Nodes:60 Transfersize:.5 -16MB Machine:snow.llnl.gov June:GPFS:1.4

Interface	Pattern	Transfersize	Reps
MPI-IO	Segmented	512KB	3
[Note:AllRatesinMB/sec]			
Nodes	29-May	Write(MB/sec)	Read(MB/sec)
2		350	408
4		472	411
8		493	451
10		439	456
16		527	468
20		545	471
24		553	475
30		543	473
32		549	459
40		550	474
48		551	474
50		552	473
60		549	473
	28-Jun	Write	Read
1		174	262
5		514	465
10		560	585
20		554	570
30		561	571
40		563	568
50		556	566
60		546	563