Technical Brief: IO Performance and Configuration Benchmarking on High Performance Cluster and SMP Computing Systems at the National Center for Supercomputing Applications

April 12, 2005

This is a technical summary brief of IO performance benchmarking performed on NCSA HPC systems on behalf of the United States National Archives and Records Administration. The purpose of this study was to examine various SMP and clustered filesystem environments to determine stability and performance curves over different types of access and storage patterns. One core question to be answered through this project was the ability of cheap commodity clusters and clustered filesystems to achieve similar sustained performance levels for large-scale document repositories as possible through SMP "monolithic" storage systems.

Benchmarks are provided for three systems:

- Cobalt, a 1000-CPU SGI Altix with 3TB of memory, a 2000-disk storage system, and fiber connectivity. Tested with 24 CPUs. Filesystem is cXFS running on a 2000-disk storage array with multiple fiber connections.
- Tungsten, a 1450-node PC cluster with each node having 3GB RAM, dual 3.2Ghz Xeon CPUs. Tested with a single node. Filesystem is Lustre running across GigE TCP/IP to a dedicated disk array.
- Mercury FastIO Partition, a 128-node PC cluster. Tests were done on the head node with 4 CPUs and 8GB of memory. Filesystem is GPFS running across fibre to NCSA SAN.

Cobalt is a fully-outfitted and fully-tuned SGI system, so its performance numbers are fairly indicative of the maximum performance potential of that class of SGI system. Cobalt uses multiple fibre connections to a 2000-disk dedicated disk array.

On serial IO, Cobalt was able to scale smoothly from 1 to 80 threads using buffer sizes ranging from 65K to 4MB. As expected, with 80 threads and a 4MB buffer size, the system achieved peak performance of the configurations tested, and the constant scaling to that point suggests that the system has not yet reached the limits of its scalability. In this configuration a sustained rate of 1.6GB/s was achieved. This seemed to be far too high for disk access and so a second data file was loaded onto the cXFS filesystem from the login node (a physically separate SGI system) to eliminate the possibility of the filesystem cache playing a role in this performance. The second file yielded the same performance numbers, suggesting that the disk array was indeed able to sustain this IO performance.

On parallel IO, Cobalt's performance degraded substantially as expected. The application being used to test random IO performance failed in 60 and 80 thread configurations when the buffer size was greater than 1MB, so these configurations were unable to be tested. Of the tested configurations, the peak bandwidth was achieved with

80 threads and a 65K block size, for 84.4MB/s. This is ideal, because this blocksize closely mimics the average filesize in many document repositories.

Tungsten is a PC cluster representative of the type of system widely deployed at HPC data centers today. It uses a Lustre clustered filesystem running across GigE over TCP/IP which manages a dedicated disk array.

Tungsten's Lustre deployment was unable to sustain more than approximately 10 concurrent threads for serial IO, so the serial IO tests were only able to measure 1, 5, and 10 thread configurations. The system illustrated clear saturation above 5 threads, where it maxed at 123MB/s.

Tungsten's parallel performance was nothing short of abysmal. It was only able to run 1, 5, and 10 threads with a 65K buffer size. All other configurations resulted in the filesystem hanging. The system saturated at 5 threads.

Mercury illustrates the performance potential of a cluster with direct fibre SAN access. Tests on Mercury were run on the quad-CPU head node, which has a direct fibre link to the NCSA SAN.

Serial IO was not able to be measured on Mercury due to an apparent incompatibility with the serial IO benchmarking application.

This system illustrated by far the most impressive parallel IO performance of any of the systems measured. The system was very easily able to eclipse Cobalt's performance. Of particular note was that optimal performance was achieved with a 1MB block size on this system, and the 4MB block size actually led to decreased performance, unlike the other systems. Peak performance was reached with 60 threads with a 1MB buffer size, which was able to sustain an astounding 517MB/s.

It should be noted that for the two cluster systems, all threads were run on a single dual or quad-CPU node, while on the SGI SMP system, each thread was assigned its own CPU. This might suggest that the system itself was the limiting factor for bandwidth in the two cluster systems, in that these benchmarks did not measure the ability of the filesystem to withstand multiple simultaneous applications handing it a heavy load. However, in this case, the purpose of the benchmarks were to measure the ability of a single application (such as a single document repository file server) to tax the filesystem with both streaming (sequential) and single-file (random) requests.

Note that these are preliminary benchmarks. We are working on a more robust benchmarking tool to verify these numbers, as well as allow us to do additional types of experiments.

Machine	Access Type	FS Type	Num Threads	Blocksize	Bandwidth (MB/s)
Tungsten	Random	Lustre	1	65K	0.84
Tungsten	Random	Lustre	5	65K	1.87

Tungsten	Random	Lustre	10	65K		1.86	
Tungsten	Random	Lustre	20	65K	FAIL		
Tungsten	Random	Lustre	40	65K	FAIL		
Tungsten	Random	Lustre	80	65K	FAIL		
Tungsten	Random	Lustre	1	512K	FAIL		
Tungsten	Random	Lustre	5	512K	FAIL		
Tungsten	Random	Lustre	10	512K	FAIL		
Tungsten	Random	Lustre	20	512K	FAIL		
Tungsten	Random	Lustre	40	512K	FAIL		
Tungsten	Random	Lustre	80	512K	FAIL		
Tungsten	Random	Lustre	1	1MB	FAIL		
Tungsten	Random	Lustre	5	1MB	FAIL		
Tungsten	Random	Lustre	10	1MB	FAIL		
Tungsten	Random	Lustre	20	1MB	FAIL		
Tungsten	Random	Lustre	40	1MB	FAIL		
Tungsten	Random	Lustre	80	1MB	FAIL		
Tungsten	Random	Lustre	1	4MB	FAIL		
Tungsten	Random	Lustre	5	4MB	FAIL		
Tungsten	Random	Lustre	10	4MB	FAIL		
Tungsten	Random	Lustre	20	4MB	FAIL		
Tungsten	Random	Lustre	40	4MB	FAIL		
Tungsten	Random	Lustre	80	4MB	FAIL		
Tungsten	Serial	Lustre	1	65K		47	
Tungsten	Serial	Lustre	5	65K		97	
Tungsten	Serial	Lustre	10	65K		122	
Tungsten	Serial	Lustre	20	65K	FAIL		
Tungsten	Serial	Lustre	40	65K	FAIL		
Tungsten	Serial	Lustre	80	65K	FAIL		
Tungsten	Serial	Lustre	1	512K		96	
Tungsten	Serial	Lustre	5	512K		123	
Tungsten	Serial	Lustre	10	512K		123	
Tungsten	Serial	Lustre	20	512K	FAIL		
Tungsten	Serial	Lustre	40	512K	FAIL		
Tungsten	Serial	Lustre	80	512K	FAIL		
Tungsten	Serial	Lustre	1	1MB		98	
Tungsten	Serial	Lustre	5	1MB		124	
Tungsten	Serial	Lustre	10	1MB		123	
Tungsten	Serial	Lustre	20	1MB	FAIL		
Tungsten	Serial	Lustre	40	1MB	FAIL		
Tungsten	Serial	Lustre	80	1MB	FAIL		
Tungsten	Serial	Lustre	1	4MB		98	
Tungsten	Serial	Lustre	5	4MB		123	
Tungsten	Serial	Lustre	10	4MB		123	
Tungsten	Serial	Lustre	20	4MB	FAIL		
Tungsten	Serial	Lustre	40	4MB	FAIL		
Tungsten	Serial	Lustre	80	4MB	FAIL		
Mercury	Serial	GPFS/SAN	1	65K			85
Mercury	Serial	GPFS/SAN	5	65K			28

Mercury	Serial	GPFS/SAN	10	65K		32	
Mercury	Serial	GPFS/SAN	20	65K	FAIL		
Mercury	Serial	GPFS/SAN	40	65K	FAIL		
Mercury	Serial	GPFS/SAN	80	65K	FAIL		
Mercury	Serial	GPFS/SAN	1	512K		84	
Mercury	Serial	GPFS/SAN	5	512K		71	
Mercury	Serial	GPFS/SAN	10	512K		121	
Mercury	Serial	GPFS/SAN	20	512K	FAIL		
Mercury	Serial	GPFS/SAN	40	512K	FAIL		
Mercury	Serial	GPFS/SAN	80	512K	FAIL		
Mercury	Serial	GPFS/SAN	1	1MB		82	
Mercury	Serial	GPFS/SAN	5	1MB		124	
Mercury	Serial	GPFS/SAN	10	1MB		176	
Mercury	Serial	GPFS/SAN	20	1MB	FAIL		
Mercury	Serial	GPFS/SAN	40	1MB	FAIL		
Mercury	Serial	GPFS/SAN	80	1MB	FAIL		
Mercury	Serial	GPFS/SAN	1	4MB		86	
Mercury	Serial	GPFS/SAN	5	4MB		138	
Mercury	Serial	GPFS/SAN	10	4MB		185	
Mercury	Serial	GPFS/SAN	20	4MB	FAIL		
Mercury	Serial	GPFS/SAN	40	4MB	FAIL		
Mercury	Serial	GPFS/SAN	80	4MB	FAIL		
Mercury	Random	GPFS/SAN	1	65K		2.99	
Mercury	Random	GPFS/SAN	5	65K		4.79	
Mercury	Random	GPFS/SAN	10	65K		36	
Mercury	Random	GPFS/SAN	20	65K		77	
Mercury	Random	GPFS/SAN	40	65K		91.5	
Mercury	Random	GPFS/SAN	80	65K		102.9	
Mercury	Random	GPFS/SAN	1	512K		12.8	
Mercury	Random	GPFS/SAN	5	512K		29.7	
Mercury	Random	GPFS/SAN	10	512K		140.3	
Mercury	Random	GPFS/SAN	20	512K		157.7	
Mercury	Random	GPFS/SAN	40	512K		288.9	
Mercury	Random	GPFS/SAN	80	512K		228.3	
Mercury	Random	GPFS/SAN	1	1MB		17.3	
Mercury	Random	GPFS/SAN	5	1MB		102.2	
Mercury	Random	GPFS/SAN	10	1MB		145.7	
Mercury	Random	GPFS/SAN	20	1MB		188.3	
Mercury	Random	GPFS/SAN	40	1MB		402.5	
Mercury	Random	GPFS/SAN	60	1MB		517.9	NOTE - 80 failed, so this is 60
Mercury	Random	GPFS/SAN	1	4MB		61.6	
Mercury	Random	GPFS/SAN	5	4MB		87.4	
Mercury	Random	GPFS/SAN	10	4MB		95.2	
Mercury	Random	GPFS/SAN	20	4MB		87.2	
Mercury	Random	GPFS/SAN	40	4MB		83.3	
Mercury	Random	GPFS/SAN	50	4MB		98.9	NOTE- 60 and 80 failed, so this is 50

Cobalt	Serial	cXFS	1	65K	79
Cobalt	Serial	cXFS	5	65K	151
Cobalt	Serial	cXFS	10	65K	192
Cobalt	Serial	cXFS	20	65K	234
Cobalt	Serial	cXFS	40	65K	243
Cobalt	Serial	cXFS	80	65K	436
Cobalt	Serial	cXFS	1	512K	165
Cobalt	Serial	cXFS	5	512K	316
Cobalt	Serial	cXFS	10	512K	430
Cobalt	Serial	cXFS	20	512K	489
Cobalt	Serial	cXFS	40	512K	633
Cobalt	Serial	cXFS	80	512K	1107
Cobalt	Serial	cXFS	1	1MB	170
Cobalt	Serial	cXFS	5	1MB	438
Cobalt	Serial	cXFS	10	1MB	462
Cobalt	Serial	cXFS	20	1MB	707
Cobalt	Serial	cXFS	40	1MB	1377
Cobalt	Serial	cXFS	80	1MB	1535
Cobalt	Serial	cXFS	1	4MB	168
Cobalt	Serial	cXFS	5	4MB	586
Cobalt	Serial	cXFS	10	4MB	986
Cobalt	Serial	cXFS	20	4MB	1703
Cobalt	Serial	cXFS	40	4MB	1452
Cobalt	Serial	cXFS	80	4MB	1615

Cobalt	Random	cXFS	1	65K	1.1