



ARTICLE

Running Linux and IBM Spectrum Scale on IBM supercomputers

Resources to facilitate standard operating systems and network protocols on IBM's leadership supercomputing hardware

By T. Christopher Ward | Published November 27, 2018 - Updated November 27, 2018

Infrastructure Linux Systems

Overview

Almost all of the world's top 500 supercomputers today run Linux®. Mostly they have batch job submission systems, which partition the supercomputer as required for the applications, and run the applications in sequence in their allocated partitions in an attempt to keep the expensive supercomputer at maximum utilization.

It is also possible to run Linux in the compute fabric as a multiuser operating system. This standard programming environment broadens the set of applications which can run on the leadership hardware and makes it easy to put the supercompute capability in the hands of scientists, engineers, and other business personnel who need it.

This article shows a Linux application running on an IBM® POWER9™ (model 8335-GTW) supercomputer cluster, and presents the software you need if you have a machine like this and want to get started with Linux.

Supercomputers and cloud computers

Having your own supercomputer is like having your own Amazon Elastic Compute Cloud. The benchmarks and test cases that you use to measure previous generations of computers (mainframes, PCs, games consoles, cellphones) don't really apply in this *new world*.

Fortunately, some of the software developed for those other types of computers can be pressed in to service to make some basic measurements, to showcase these new computers, and to illustrate who in a modern competitive business needs to have access to these facilities.

Writing this article in five years' time would be simple; we might most likely have oil reservoir models, airline seat pricing models, gas turbine flow visualizations, and similar techniques to show off; the market would be mature. However, today is today, we're in at the *ground floor* of new and growing business, so we're adapting IBM General Parallel File System (IBM GPFS™) for the purpose.

IBM General Parallel File System is now IBM Spectrum Scale

IBM GPFS, now IBM Spectrum Scale™, started life as the *multimedia file system*, intended for streaming video at predictable bandwidth from server farms. It is now actively marketed for data management in enterprise data centers.

A typical IBM Spectrum Scale installation consists of maybe 10 servers, each with up to a few hundred disk spindles. These servers provide POSIX file system services for hundreds to thousands of network-connected client systems.

IBM Spectrum Scale provides data replication, volume management, backup/restore, continuous operation in case of disk and server failures, improved serviceability, and scalability. These are features needed by enterprises and are what distinguish this IBM technology from open technology such as Network File System (NFS).

In our scenario with the POWER9 cluster, we allocate a solid-state disk of 1.5 TB on each POWER9 node as if it was a disk spindle. The whole IBM Spectrum Scale system consists of 16 server nodes, each with one disk of size 1.5 TB, providing a coherent POSIX file system image to client applications running on the 16 server nodes. This is an unusual geometry for an IBM Spectrum Scale cluster; but it is viable.

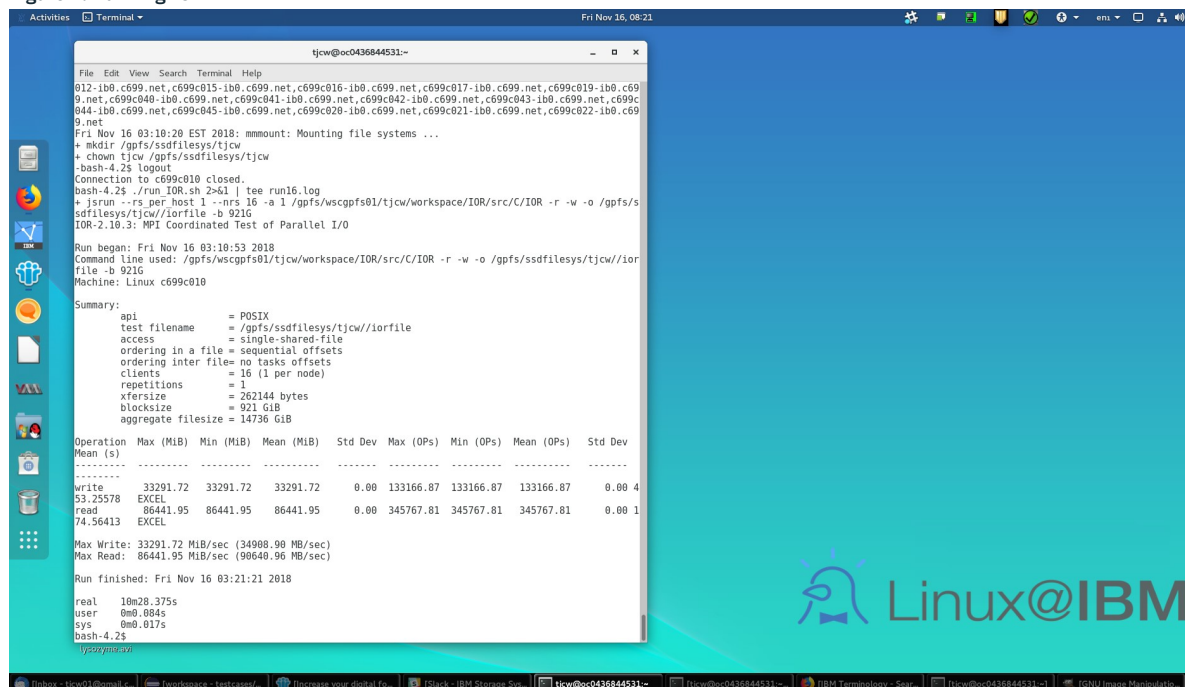
I had access to 16 server nodes; cluster sizes vary from two nodes all the way up to several thousand nodes depending on the intended application.

The scripts I used to configure and run this benchmark are available for download in the [Resources](#) section.

Interleaved or Random

Interleaved or Random (IOR) is a file system benchmark from the University of California. Figure 1 shows a screen capture of it running on 16 nodes of the POWER9.

Figure 1. Running IOR



```
File Edit View Search Terminal Help
tjcw@oc0436844531:~
012-ib0.c699.net,c699c015-ib0.c699.net,c699c016-ib0.c699.net,c699c017-ib0.c699.net,c699c019-ib0.c69
9.net,c699c040-ib0.c699.net,c699c041-ib0.c699.net,c699c042-ib0.c699.net,c699c043-ib0.c699.net,c699c
044-ib0.c699.net,c699c045-ib0.c699.net,c699c020-ib0.c699.net,c699c021-ib0.c699.net,c699c022-ib0.c69
9.net
Fri Nov 16 03:10:20 EST 2018: mmount: Mounting file systems ...
+ mkdir /gpfs/ssdfilesys/tjcw
+ chown tjcw /gpfs/ssdfilesys/tjcw
-bash-4.2$ logout
Connection to c699c010 closed.
bash-4.2$ ./run_IOR.sh 2>&1 | tee run16.log
+ jsrun --rs_per_host 1 --nrs 16 -a 1 /gpfs/wscgpf01/tjcw/workspace/IOI/src/C/IOI -r -w -o /gpfs/s
sdfilesys/tjcw//iorfile -b 921G
IOR-2.10.3: MPI Coordinated Test of Parallel I/O

Run began: Fri Nov 16 03:10:53 2018
Command line used: /gpfs/wscgpf01/tjcw/workspace/IOI/src/C/IOI -r -w -o /gpfs/ssdfilesys/tjcw//ior
file -b 921G
Machine: Linux c699c010

Summary:
api = POSIX
test filename = /gpfs/ssdfilesys/tjcw//iorfile
access = single-shared-file
ordering in a file = sequential offsets
ordering inter file = no tasks offsets
clients = 16 (1 per node)
repetitions = 1
rferysize = 262144 bytes
blocksize = 921 GiB
aggregate filesize = 14736 GiB

Operation Max (MiB) Min (MiB) Mean (MiB) Std Dev Max (OPs) Min (OPs) Mean (OPs) Std Dev
Mean (s)
-----
write 33291.72 33291.72 33291.72 0.00 133166.87 133166.87 133166.87 0.00 4
33.2578 EXCEL
read 86441.95 86441.95 86441.95 0.00 345767.81 345767.81 345767.81 0.00 1
74.56413 EXCEL

Max Write: 33291.72 MiB/sec (34988.98 MB/sec)
Max Read: 86441.95 MiB/sec (90640.96 MB/sec)

Run finished: Fri Nov 16 03:21:21 2018

real 10m28.375s
user 0m0.084s
sys 0m0.017s
bash-4.2$
lyss@mesiv:
```

Refer to Listing 1 for the text from Figure 1.

Listing 1. Running IOR

```
+ jsrun --rs_per_host 1 --nrs 16 -a 1 /gpfs/wscgpf01/tjcw/workspace/IOI/src/C/IOI -r -w -o /gpfs/ssdfilesys/tjcw//iorfile -b 921G
IOR-2.10.3: MPI Coordinated Test of Parallel I/O

Run began: Fri Nov 16 03:10:53 2018
Command line used: /gpfs/wscgpf01/tjcw/workspace/IOI/src/C/IOI -r -w -o /gpfs/ssdfilesys/tjcw//iorfile -b 921G
Machine: Linux c699c010

Summary:
api = POSIX
test filename = /gpfs/ssdfilesys/tjcw//iorfile
access = single-shared-file
ordering in a file = sequential offsets
ordering inter file = no tasks offsets
clients = 16 (1 per node)
```

[Show more](#)

This shows a session from a desktop to the supercomputer. *c699c010* is one of the 16 nodes allocated to this job, each with 44 POWER9 processors, six NVIDIA Tesla GPUs; a 1.5 TB solid-state disk and 605 GB of RAM, for a total of 704 POWER9 processors, 96 GPUs, 24 TB of solid-state disk, and 9.6 TB of RAM.

Log on to the launch node named *c699launch01*, and issue the `jsrun` command to ask for one processor core on each processing node to be joined up over TCP/IP as a Message Passing Interface (MPI) job.

```
jsrun --rs_per_host 1 --nrs 16 -a 1
/gpfs/wscgpf01/tjcw/workspace/IOI/src/C/IOI -r -w -o
/gpfs/ssdfilesys/tjcw//iorfile -b 921G
```

MPI runs IOR, a distributed file system benchmark which you could run over NFS among a cluster of workstations. In this case, IOR is run over the IBM Spectrum Scale File System with its data in solid-state disk, and it achieves an average write data rate of 33.3 GBps and an average read data rate of 86.4 GBps over Mellanox InfiniBand among the 16 nodes. These data rates are limited by the transfer speeds to and from the solid-state disks.

It would be possible to ask `jsrun` to run the MPI job over all 704 processor cores on the 16 nodes by specifying `--rs_per_host 44 -nrs 704`, but one core per node is sufficient in this benchmark to use the whole capability of the solid-state disks.

Conclusion

IOR is a synthetic benchmark and not a real application. But if you want to compete in the world of high performance computing, you need be able to run it well. And it shows how easy and quick it is with this IBM POWER9 cluster and Linux, to leap from a standard desktop environment to a standard cloud computing environment.

This article has not attempted to explore another significant feature of the IBM POWER9 model 8335-GTW; the capability to receive and transmit 10 Gbps per node of data traffic over the standard TCP/IP protocol. This feature may have the most long-term significance, as it enables IBM's customers to construct web-scale architectures for serving data to and analyzing the needs of their customers. For example these 16 IBM POWER9 nodes may be able to drive 8000 domestic broadband connections at the rate required for video streaming at DVD quality, continuously. We may be at the beginning of a new opportunity for those who might provide and consume this type of connected d: infrastructure.

Resources

- <https://github.com/tjcw/dW-spectrum-scale> – the scripts used to configure and run this benchmark.
- <http://kernel.org/> is the public Linux repository.
- [ZeptoOS](#) is an early version of Linux for IBM System BlueGene/P and has support for 256 MB memory pages.
- <https://github.com/LLNL/ior> is the home of the IOR benchmark.
- [Open MPI](#) is a high performance, widely-portable (and free) implementation of the Message Passing Interface standard. This open source component is the basis for the [IBM Spectrum MPI product](#)
- [IBM Spectrum Scale](#) is the new name for IBM General Parallel File System, one of the IBM software products that is useful as part of a supercomputer solution from IBM or other vendors.
- [OpenSSH in Windows 10](#) – Windows 10 now contains an implementation of ssh, so even if you use a desktop operating system from t “other” vendor you can still access your supercomputing infrastructure servers from your desktop systems.
- [The Argonne Leadership Computing Facility](#) – Until you get your own IBM supercomputer, you can submit a proposal to use theirs.
- [Diskless Remote Boot Linux](#), from [National Center for High-Performance Computing Taiwan](#), is a way to construct an entry-level “Cloud Computer” from a classroom-full of traditional personal computers.
- [IBM Innovation Centers](#) can arrange for demonstrations of this and other IBM technology, planet-wide. Approach through your IBM account team.
- [Using the Active Storage Fabrics model to address petascale storage challenges](#) documents some of my team’s earlier work with Linux on the IBM System BlueGene/P.
- [LOFAR](#) is a radiotelescope the size of Europe, powered by an IBM System BlueGene/P at the center. Visit their website and see the breakthroughs in astronomy, geophysics, and agriculture that have been facilitated by their investment in IBM technology.
- [HPC-Colony Project](#) is a project led by the US Government to explore scalable services on computing systems with very large number: of processors. If you ever wanted to know *Why Linux?*, they explain it well.
- <http://sdf.lonestar.org/index.cgi?telnet> might be your fastest entry into the world of Cloud Computing. It gives access to a Unix system somewhere in the Internet, funded on the same basis as [the Public Broadcasting TV service](#).
- [Teachers try science](#) and [Try engineering](#) provide some encouragement from IBM and others for the next generation of scientists and engineers.

- [POWER9 Processor User's Manual](#) is the user manual for the POWER9 processor, manufactured by IBM and others, thanks to the OpenPOWER Foundation.
- [IBM Linux servers](#) – Learn about the IBM Linux servers and mainframes.
- [IBM IT infrastructure](#) IBM is available to discuss your requirements for IBM supercomputer hardware now and in the future.
- [top500.org](#) – The current top 500 supercomputers run Linux.

COMPONENTS IBM POWER SYSTEMS

SOCIAL



CONTENTS

- Overview
- Supercomputers and cloud computers
- IBM General Parallel File System is now IBM Spectrum Scale
- Interleaved or Random
- Conclusion
- Resources

Related content

<p>ARTICLE OCT 09, 2018</p> <p>Enhancing QEMU virtio-scsi with Block Limits vital product data (VPD) emulation</p> <hr/> <p>Databases Infrastructure +</p>	<p>ARTICLE SEP 25, 2018</p> <p>Custom PHP script to monitor the availability of server resources</p> <hr/> <p>Databases Infrastructure +</p>	<p>TUTORIAL SEP 18, 2018</p> <p>Connecting Python and Node.js applications to an Oracle database</p> <hr/> <p>Databases IBM Power Systems +</p>
---	---	--

IBM Developer

- About
- Site Feedback & FAQ
- Submit content
- Report abuse
- Third-party notice

Follow us

Select a language

- English
- 中文
- 日本語
- Русский
- Português
- Español

Code Patterns

- Articles
- Tutorials
- Recipes
- Open Source Projects

Videos

- Newsletters
- Events
- Cities
- Developer Answers



한글



[Contact](#) [Privacy](#) [Terms of use](#) [Accessibility](#) [Feedback](#) [Cookie Preferences](#)