

## Overview

The IBM System BlueGene/P comes from the factory with IBM's Compute Node Kernel, which is well-suited to running supercomputer applications. But it is also possible to run Linux in the compute fabric. 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 BlueGene/P, and presents the software you need if you have a BlueGene/P and want to get started with Linux.

## Supercomputers and Cloud Computers

Having your own BlueGene is rather 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 5 years' time would be simple; we will most likely have oil reservoir models, airline seat pricing models, gas turbine flow visualisations, and the like to show off; the market will be mature. However, today is today, we're in at the 'ground floor' of a new and growing business, so we're adapting IBM's General Parallel File System for the purpose.

## IBM General Parallel File System

This IBM program product 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 centres.

A typical GPFS installation will consist of maybe 10 server machines, each with up to a few hundred disk spindles. These server machines provide POSIX file system services for hundreds-to-thousands of network-connected client machines.

GPFS provides for data replication, volume management, backup/restore, continuous operation in respect 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 NFS.

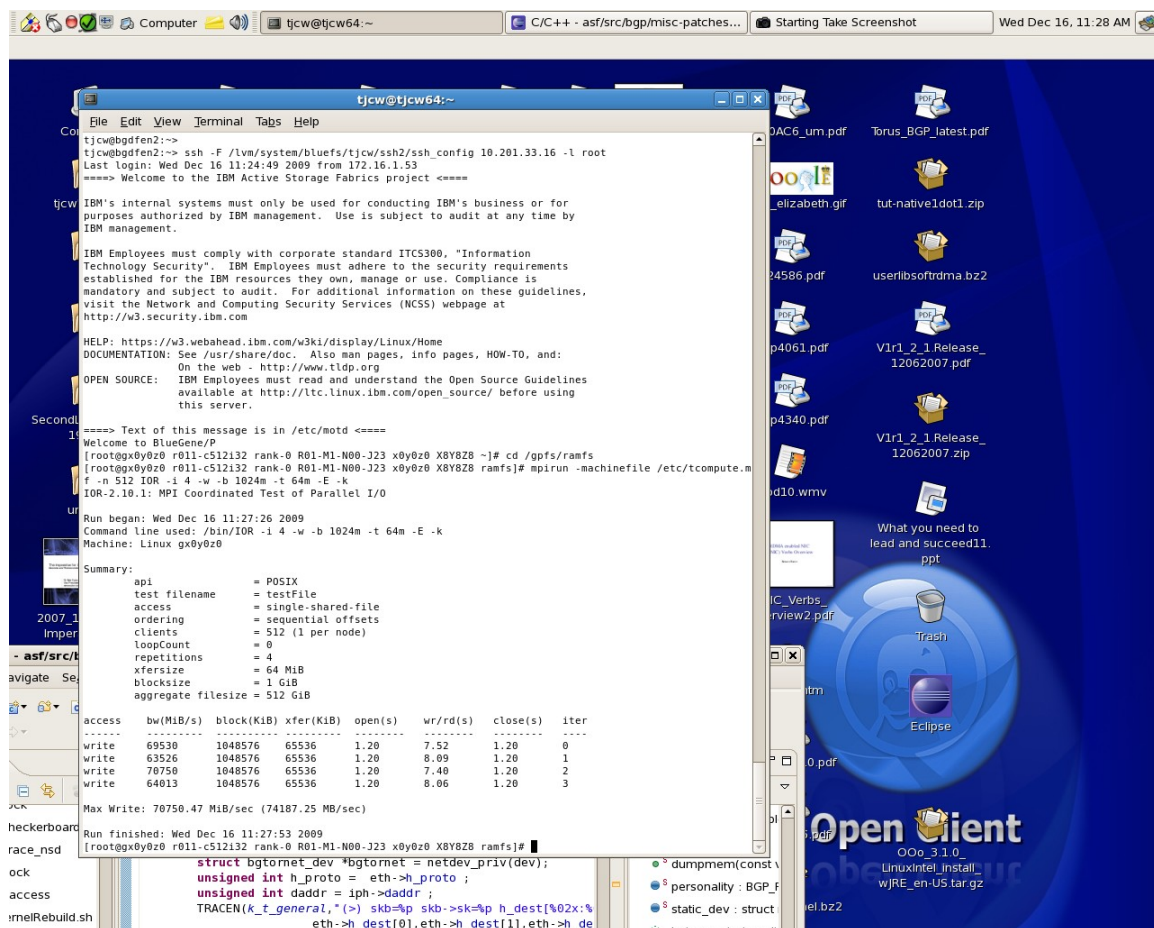
As part of IBM Research, my team has privileged access to GPFS. This article does not constitute a commitment that this version of GPFS will be available with standard pricing and warranty terms in your country; however your IBM salesman is always willing to discuss your requirements.

In this use with BlueGene, we allocate 1GB of the RAM of each BlueGene node as if it was a disk spindle. The whole GPFS system consists of 512 server nodes, each with 1 'disk' of size 1GB, providing a coherent POSIX file system image to client application running on the 512 server nodes. This is an unusual 'geometry' for a GPFS cluster; but it is viable.

Normally you would configure GPFS with the supplied administrative commands; however this is unwieldy with as many as 512 server nodes, so I configured GPFS with scripts which write the configuration files. These scripts are available for download [here](#).

## Interleave Or Random

IOR is a file system benchmark from University of California. Here is a screenshot of it running on an 8x8x8 block --- 512 nodes --- of BlueGene.



```
tjcw@tjcw64:~  
tjcw@bgdfen2:~$ ssh -F /lvm/system/bluefs/tjcw/ssh2/ssh_config 10.201.33.16 -l root  
Last login: Wed Dec 16 11:24:49 2009 from 172.16.1.53  
===== Welcome to the IBM Active Storage Fabrics project =====  
  
IBM's internal systems must only be used for conducting IBM's business or for  
purposes authorized by IBM management. Use is subject to audit at any time by  
IBM management.  
  
IBM Employees must comply with corporate standard ITC5300, "Information  
Technology Security". IBM Employees must adhere to the security requirements  
established for the IBM resources they own, manage or use. Compliance is  
mandatory and subject to audit. For additional information on these guidelines,  
visit the Network and Computing Security Services (NCSS) webpage at  
http://w3.security.ibm.com  
  
HELP: https://w3.webhead.ibm.com/w3ki/display/Linux/Home  
DOCUMENTATION: See /usr/share/doc. Also man pages, info pages, HOW-TO, and:  
On the web - http://www.tldp.org  
OPEN SOURCE: IBM Employees must read and understand the Open Source Guidelines  
available at http://lrc.linux.ibm.com/open_source/ before using  
this server.  
  
Second:  
1) ===== Text of this message is in /etc/motd =====  
[root@qx0y0z0 r011-c512132 rank-0 R01-M1-N00-J23 x0y0z0 X8Y8Z8 ~]# cd /gpfs/ramfs  
[root@qx0y0z0 r011-c512132 rank-0 R01-M1-N00-J23 x0y0z0 X8Y8Z8 ramfs]# mpirun -machinefile /etc/tcompute.m  
f -n 512 IOR -i 4 -w -b 1024m -t 64m -E -k  
IOR-2.10.1: MPI Coordinated Test of Parallel I/O  
  
Run began: Wed Dec 16 11:27:26 2009  
Command line used: /bin/IOR -i 4 -w -b 1024m -t 64m -E -k  
Machine: Linux qx0y0z0  
  
Summary:  
api = POSIX  
test filename = testFile  
access = single-shared-file  
ordering = sequential offsets  
clients = 512 (1 per node)  
loopCount = 0  
repetitions = 4  
xfersize = 64 MiB  
blocksize = 1 GiB  
aggregate filesize = 512 GiB  
  
access bw(MiB/s) block(KiB) xfer(KiB) open(s) wr/rd(s) close(s) iter  
write 69530 1048576 65536 1.20 7.52 1.20 0  
write 63526 1048576 65536 1.20 8.09 1.20 1  
write 70750 1048576 65536 1.20 7.40 1.20 2  
write 64013 1048576 65536 1.20 8.06 1.20 3  
  
Max Write: 70750.47 MiB/sec (74187.25 MB/sec)  
  
Run finished: Wed Dec 16 11:27:53 2009  
[root@qx0y0z0 r011-c512132 rank-0 R01-M1-N00-J23 x0y0z0 X8Y8Z8 ramfs]#  
  
struct bgdtormet_dev *bgdtormet = netdev_priv(dev);  
unsigned int h_proto = eth->h_proto ;  
unsigned int daddr = iph->daddr ;  
TRACEN(K_f_general, "(>) skb=<np skb->sk=<np h_dest[%02x:%  
eth->h_dest[0],eth->h_dest[1],eth->h_de
```

This shows a session from a 'desktop' machine to the supercomputer. 'r011-c512i32' is a half-rack, consisting of 512 processing nodes, each with 4x '450' PowerPC processing cores and 4 GB of memory, for a total of 2048 processors and 2TB of RAM. The half-rack can also drive about 128 Gbit/second of TCP/IP data to and from external Internet.

I log on to the compute node at location x0y0z0 out of X8Y8Z8, and issue the 'mpirun' command to ask for all the processing nodes to be joined up over TCP/IP as an MPI job. MPI runs 'IOR', a distributed file system benchmark which you could run over NFS amongst a cluster of workstations. In this case IOR is running over the IBM General Parallel File System with its data in ramdisk, and it achieves a peak data rate of 70 Gigabytes/second over TCP/IP amongst the 512 nodes.

## Conclusion

IOR is a synthetic benchmark, not a real application, of course. But if you want to compete in the world of high-performance computing, you need to be able to run it well. And it shows how easy and quick it is with IBM System BlueGene/P 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 this Linux; the capability to receive and transmit 256Gbit/second per rack, 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, one rack of IBM System BlueGene/P may be able to drive 30000 domestic broadband connections at full rate, continuously. We may be at the beginning of a new opportunity for those who will provide and consume this type of connected data infrastructure.

## Resources

<http://git.anl-external.org/bg-linux/repos/linux-2.6.29.1-BGP.git/> . Source code for the version of Linux used on the IBM System BlueGene/P for this test is hosted here, courtesy of Argonne National Laboratory. You may need the version control system <http://git-scm.com/> to unpack the source code ready for building.

<http://kernel.org/> is the public Linux repository.

<http://www.mcs.anl.gov/research/projects/zeptoos/> , another version of Linux for IBM System BlueGene/P, has support for 256MB memory pages and for the [http://dcmf.anl-external.org/wiki/index.php/Main\\_Page](http://dcmf.anl-external.org/wiki/index.php/Main_Page) Deep Computing Messaging Framework standard.

<https://asc.llnl.gov/sequoia/benchmarks/#ior> , the home of the IOR benchmark.

<http://www.mcs.anl.gov/research/projects/mpich2/> MPICH2, a high-performance, widely-portable (and Free) implementation of the Message Passing Interface standard.

<http://www-03.ibm.com/systems/clusters/software/gpfs/index.html> IBM General Parallel File System, one of the IBM software products that is useful as part of a BlueGene solution.

<http://www-01.ibm.com/software/lotus/openclient/> is the IBM-marketed software running on the desktop computer in the screenshot.

<http://www.cygwin.com/> contains an implementation of 'ssh' for Microsoft Windows. This key enabling technology allows those who have not yet 'made the transition' to access their infrastructure servers from their desktop systems. It is open source technology, collected and distributed by Red Hat, an IBM business partner.

<http://www.alcf.anl.gov/>, the Argonne Leadership Computing Facility. Until you get your own IBM System BlueGene/P, you can submit a proposal to use theirs. So far, ALCF has committed 400 million processor-hours of IBM System BlueGene/P time to proposals that have been accepted. Some of the projects and accomplishments at ALCF are chronicled here <http://www.alcf.anl.gov/collaborations/projects/index.php>.

<http://drbl.sourceforge.net/>, Diskless Remote Boot Linux from <http://www.nchc.org.tw/en/> National Center for High-Performance Computing Taiwan, is a way to construct an entry-level 'Cloud Computer' from a classroom-ful of traditional Personal Computers.

IBM Innovation Centers <http://www-304.ibm.com/jct01005c/isv/iic/> can arrange for demonstrations of this and other IBM technology, planet-wide. Approach via your IBM account team.

[http://www-03.ibm.com/press/us/en/attachment/28424.wss?fileId=ATTACH\\_FILE2&fileName=IBM\\_Blue\\_Gene\\_Supercomputer.jpg](http://www-03.ibm.com/press/us/en/attachment/28424.wss?fileId=ATTACH_FILE2&fileName=IBM_Blue_Gene_Supercomputer.jpg) – a picture of an installed IBM System BlueGene/P system.

“Using the Active Storage Fabrics model to address petascale storage challenges”  
<http://portal.acm.org/citation.cfm?id=1713072.1713086&coll=DL&dl=GUIDE&CFID=4932808&CFTOKEN=98431226>  
documents more of my team's work with this version of Linux on the IBM System BlueGene/P.

<http://lofar.org/> is a radiotelescope the size of Europe, powered by an IBM System BlueGene/P at the centre. Visit their web site and see the breakthroughs in astronomy, geophysics, and agriculture that have been facilitated by their investment in IBM technology.

<http://www.hpc-colony.org/> is project led by the US Government, to explore scalable services on computing systems with very large numbers of processors. If you ever wanted to know “Why Linux?”, they explain it well. (Note to editor: This is a 'live' customer project, before mentioning this please get approval from Jose Moreira , [jmoreira@us.ibm.com](mailto:jmoreira@us.ibm.com) , to confirm that the customer is content to be used as reference)

<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 <http://www.pbs.org/> the Public Broadcasting TV service.

<http://www.tryscience.org/> and <http://www.tryengineering.org/> are some encouragement from IBM and others for the next generation of scientists and engineers.

[https://www-01.ibm.com/chips/techlib/techlib.nsf/products/PowerPC\\_460S\\_Embedded\\_Core](https://www-01.ibm.com/chips/techlib/techlib.nsf/products/PowerPC_460S_Embedded_Core) will lead you to the user manual for the '460' processor core, the closest currently-marketed product to the '450' used in BlueGene/P.

[http://www-03.ibm.com/technology/manufacturing/technology\\_Tour\\_300mm\\_Foundry.html](http://www-03.ibm.com/technology/manufacturing/technology_Tour_300mm_Foundry.html) is a tour of one of IBM's two silicon chip manufacturing plants.

<http://www.mosis.com/IBM/> is a good starting point for businesses that would like IBM to manufacture prototype quantities of 'system-on-a-chip' microprocessors similar to those in the BlueGene/P. Nintendo have sold 50,000,000 such IBM chips in their 'Wii' games consoles, but most clients start smaller.

<http://www-03.ibm.com/linux/services.html> IBM Global Services is available as your professional innovation partner for all uses you choose to make of your IBM System BlueGene/P

<http://www-03.ibm.com/systems/deepcomputing/> IBM Systems and Technology Group is available to discuss your requirements for IBM System BlueGene hardware now and in the future.

# developerWorks article template using OpenOffice.org Writer

**Important:** Please ensure that your input in this information form is inside the fields. To enter data in a very narrow field , press ctrl-shift-F9 and then use the **Next** button to move through fields.

**Type of Submission:** Article

**Title:** Running Linux on the IBM System BlueGene

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

**Keywords:** BlueGene, Supercomputing, Cloud Computing, Linux, Open Standards, TCP/IP

**Prefix:** Mr

**Given:** T.

**Middle:** Christopher

**Family:** Ward

**Suffix:**

**Job Title:** Senior Software Engineer

**Email:** tjcw@uk.ibm.com

**Bio:** Chris Ward joined the IBM UK Development Laboratories in Hursley, England, in 1982 with a degree in Engineering from Cambridge University. He has worked on the development of many products for IBM, from Disk Files to Branded Middleware. He is privileged to be working on technology which is as valuable to IBM's future customers as IBM Websphere Software and IBM Lotus Software are to IBM's customers of today.

**Company:** IBM

**Photo filename:** headshot.jpg