

PAPER • OPEN ACCESS

## Performance of the AMS Offline software on the IBM Blue Gene/Q architecture

To cite this article: V Choutko *et al* 2017 *J. Phys.: Conf. Ser.* **898** 072002

View the [article online](#) for updates and enhancements.

### You may also like

- [Tuneable gradient Helmholtz-resonator-based acoustic metasurface for acoustic focusing](#)

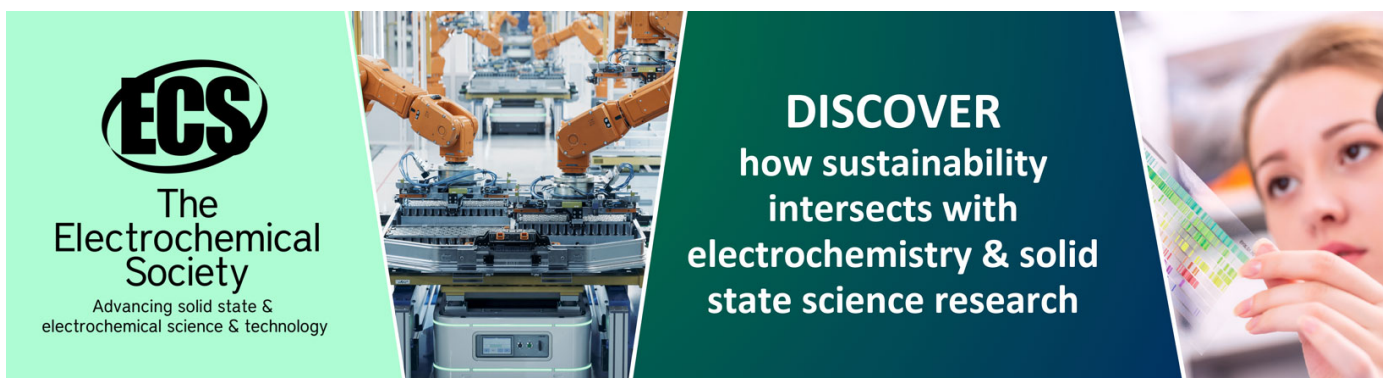
Kemeng Gong, Xiaofan Wang, Huajiang Ouyang et al.

- [Coupling rare event algorithms with data-based learned committor functions using the analogue Markov chain](#)

Dario Lucente, Joran Rolland, Corentin Herbert et al.

- [Experimental realization for abnormal reflection caused by an acoustic metasurface with subwavelength apertures](#)

Xuanjun Liu, Xinwu Zeng, Dongbao Gao et al.



**ECS**  
The  
Electrochemical  
Society  
Advancing solid state &  
electrochemical science & technology

**DISCOVER**  
how sustainability  
intersects with  
electrochemistry & solid  
state science research

# Performance of the AMS Offline software on the IBM Blue Gene/Q architecture

V Choutko<sup>1</sup>, A Egorov<sup>1</sup> and B S Shan<sup>2</sup>

<sup>1</sup>Massachusetts Institute of Technology, Cambridge, MA

<sup>2</sup>Beihang University, Beijing, China

E-mail: [vitali.choutko@cern.ch](mailto:vitali.choutko@cern.ch)

**Abstract.** The Alpha Magnetic Spectrometer (AMS) is a high energy physics experiment installed and operating on board the International Space Station (ISS) since May 2011 and expected to last beyond 2024. The details of porting the AMS software to the IBM Blue Gene/Q Architecture are discussed. The performance of the AMS reconstruction and simulation software in that architecture is evaluated and compared to the performance obtained on Intel based architecture.

## 1. Introduction

The Alpha Magnetic Spectrometer (AMS) [1] is a high energy physics experiment operating on board of the International Space Station (ISS). The detector has a large geometrical acceptance of  $0.5m^2 \cdot sr$ . It is equipped with a permanent magnet, a Time of Flight hodoscope, a precision nine layer silicon tracker, a gaseous Transition Radiation Detector, a Ring Imaging Cerenkov Detector, and an 3D Electromagnetic Calorimeter. In five years of operation more than 85 billion of cosmic ray events were triggered, recorded and transferred to the ground.

## 2. IBM Blue Gene/Q Architecture overview

The IBM Blue Gene/Q architecture is described in detail in Ref. [2]. It includes the login nodes, equipped with the POWER7 3.55 GHz processors and 128 GB of memory which run a Linux 2.6 operation system; and the compute nodes, each equipped with the 16+1 core PowerPC A2 1.6 GHz processor running a lightweight proprietary kernel (CNK), with 16 GB of memory. The PowerPC A2 processor features four-way hyperthreading, so up to 64 threads per node can be run. All performance studies were done on the compute nodes, while ported software equally works well on the login nodes.

There are a few distinct features of this architecture which were found to be essential for the software porting:

- A 64-bit address space;
- Big-endian data format;
- Limited support for Linux system calls; in particular, no support for *fork()* and *system()* calls on the compute nodes;
- Massive parallelization beyond SMP. Open MPI [3] is used usually to synchronize threads running on different nodes.



### 2.1. IBM Blue Gene/Q Architecture Compilers

The IBM compilers xLC 12.1 and xLF 14.1 were used to compile and link all the software. These compilers support the OpenMP [4] directives, with the major exception of not supporting the *omp threadprivate* pragma for any STL container (vector, map, etc). The xLC 12.1 compiler supports a subset of C++11 directives, including thread local storage (TLS) via the *\_\_thread* directive with the same exception for STL containers.

## 3. Software Porting

The actual porting of the software has been done on the JUQUEEN computer of the Juelich SuperComputing Center [5], which consists of 28,672 computing nodes. The minimal job configuration in the batch system includes 32 nodes, i.e., up to 2048 threads, while a typical job configuration consists of 128 to 512 nodes, i.e., up to 32,768 threads.

### 3.1. Porting of the ROOT software

The ROOT 5.34 [6] software was not available on this platform (codenamed here as *linuxppcbgxl*) owing to the incompatibility between the ROOT CINT interpreter and 64-bit addressing space coupled with the big-endian data format of PowerPC processors [7]. This was fixed by changing a single line in the *cint/cint/src/value.h* file like:

```
< if (buftype == 'i') return (T) buf->obj.i;
---
> if (buftype == 'i') return (T) buf->obj.in;
```

Another minor issue was an xLC compiler internal error during compilation of the RooFit dictionary. This was fixed by division of the dictionary file into several parts.

After this the successful build of the root executable and all shared and static libraries became possible, see figure 1.

### 3.2. Porting of the GEANT4 software

The IBM Blue Gene/Q architecture was not supported by the GEANT4.10.1 package [8]. To support it, the following architecture file *Linux-ppc-mt.gmk* was added:

```
..
CXX      := bgxlc_r
CXXFLAGS := -q64 -qmaxmem=-1 -D__PPC64
ifdef G4USE_STD11
    CXXFLAGS += -qlanglvl=extc1x
endif
ifdef G4MULTITHREADED
    CXXFLAGS += -qthreaded -qsmp=omp -qtls
endif
..
```

A few source files needed to be changed, to add the thread specification for this architecture and also to cope with xLC compiler template specialization initialization limitations. The following files were modified:

```
global/management/include/tls.hh
global/management/include/G4TWorkspacePool.hh,
particles/management/src/G4ParticlesWorkspace.cc
geometry/navigation/src/G4VIntersectionLocator.cc
```

```

[vsk1006@juqueen2 install]$ $R00TSYS/bin/root -b
*****
*
*           W E L C O M E   t o   R O O T           *
*
*   Version    5.34/09           26 June 2013      *
*
*   You are welcome to visit our Web site          *
*           http://root.cern.ch                    *
*
*****

ROOT 5.34/09 (v5-34-09@v5-34-09, May 31 2016, 23:21:19 on linuxppcbgxlc)

CINT/ROOT C/C++ Interpreter version 5.18.00, July 2, 2010
Type ? for help. Commands must be C++ statements.
Enclose multiple statements between { }.
root [0]   gSystem->Load("$AMSWD/lib/linuxppcbgxlc5.34/ntuple_slc4_PG.so");
AMSCoreCommonsI-I-HardwareIdentifiedAs Linux 2.6.32-642.3.1.el6.ppc64 ppc64 5
AMSCoreCommonsI-W-AMSDataDir variable is not defined.
AMSCoreCommonsI-W-Default value /afs/cern.ch/exp/ams/Offline/AMSDataDir will be used
.
AMSCoreCommonsI-I-Identified as BigEndian 64 bit machine. 4294967295
AMSCoreCommonsI-I-amsdatabase /afs/cern.ch/exp/ams/Offline/AMSDataDir/DataBase/
AMSCoreCommonsI-I-MipsFromCPUInfo: 3550
AMSCoreCommonsI-I-SystemIdentified as POWER7 (architected), altivec supported
AMSCoreCommonsI-I-ComputerEvaluatedAsMips 2378
AMS Software version v5.01/1074/15 build Thu Jun 23 16:13:39 2016

AMSaPool::SetLastResort-I-ResortSet 2000000
AMSaPool::SetLastResort-I-ResortSet 2000000
root [1]  TH1D *p=new TH1D("1d","1d",100,0,1)
root [2]  sizeof(p)
(const int)8

```

**Figure 1.** The 64-bit ROOT starting screen in a JUQUEEN login node.

Following the changes, the GEANT4.10.1 libraries were built and the test examples successfully ran in multi-threaded mode.

### 3.3. Porting of the CERNLIB software

The port of CERNLIB [9] software was needed, as AMS software depends on it, to ensure that the FORTRAN local variables being initialized in the stack to allow thread safe processing. The MINUIT package also needed to be adapted to the thread safe mode using the OpenMP technique.

### 3.4. Porting of the AMS software

**3.4.1. Simulation of Linux system() calls** Due to absence of the Linux *system()* directive support on the CNK kernel, the following system calls were rewritten using C++ language I/O constructions: *mkdir*, *rm*, *rmdir*, *cat*, *grep*, *ln -s*, ... .

**3.4.2. Memory Management** Due to a lack of support for Linux system routines like *getrlimit()* ..., proprietary routines were used to estimate the amount of free memory available for job execution.

**3.4.3. C++ features** Portions of the AMS software which contained threadprivate STL vector and/or maps were rewritten. In one particular case the threadprivate map was replaced by an array of maps with explicit thread addressing, using the OpenMP and GEANT4.10.1 methods.

*3.4.4. FORTRAN features* The AMS software uses DPMJET2.5 FORTRAN code to simulate the nucleus-nucleus interactions. To provide the thread safe usage of this library OpenMP was used. Incompatibility of the thread local storage usage between xLF 14.1 and xLC 12.1 compilers was found which prevented using the default TLS option of the xLF 14.1 compiler. The combination of `-qsmg=omp:noostls -g5` FORTRAN compiler flags was found to work correctly.

*3.4.5. ROOT dictionary* Due to a platform specific Table of Content (TOC) 24-bit size issue, which results in R\_PPC\_REL24 errors during linking, (as generated code was too large for 24-bit relative jumps to reach from some branch sites [10]), the single ROOT dictionary file had to be divided into many (10) smaller files. These files were compiled separately and successfully linked together.

*3.4.6. Linking* For the AMS software version with the Open MPI support, no static linking was possible, because of another TOC size issue. Dynamic linking to those libraries did not show the problem. Finally, a statically bound executable with no Open MPI support and with MPI emulation was used during the software performance tests.

### *3.5. MPI Emulation*

For the massively parallel jobs which can be run on the JUQUEEN computer, inter nodes communication can be done via the Open MPI libraries, provided by IBM. As the AMS software parallelization is limited inside one node by using OpenMP (in the case of reconstruction) and the mixture of POSIX threads and OpenMP (in the case of simulation), see Ref. [11], the only communications needed are the proper ranking of the jobs and the synchronization of the job finishing phases. Despite Open MPI was able to provide the desired features, custom software was written to emulate the required features of Open MPI messaging. It includes:

- MPI-like initialization, to create the job rank and properly reroute the input and output files;
- A special thread to govern the job execution, to calculate CPU limits, to send and receive messages from other jobs, and to intercept and reinterpret all Linux signals, including SIGSEGV;
- MPI-like termination routine, called during the job static destructor execution, to ensure simultaneous termination of all the jobs.

## **4. Results**

The efficient use of the JUQUEEN batch system for simulation jobs containing up to 2048 nodes and 131K threads was possible. This allows us to use JUQUEEN for massive simulated AMS data production.

The reconstruction jobs, due to their well defined amount of input data, could not be efficiently used in the MPI environment. This prevented us from using JUQUEEN for massive AMS data reruns.

### *4.1. Simulation Software performance*

We were able to run up to 62 threads per compute node. This became possible due to our customized GEANT4 memory manager [11]. Figure 2 shows the performance of the IBM Blue Gene/Q versus the number of threads per node. As seen, no degradation due to 4-way hyperthreading can be noticed. The multithread overhead is limited to about 0.7%.

In terms of absolute values, the Intel(R) Xeon(R) CPU E5-2699 v3 @ 2.30GHz (18 cores, 36 threads, 145 Watt) outperforms the Blue Gene/Q node (16 cores, 64 threads, 60 Watt) by

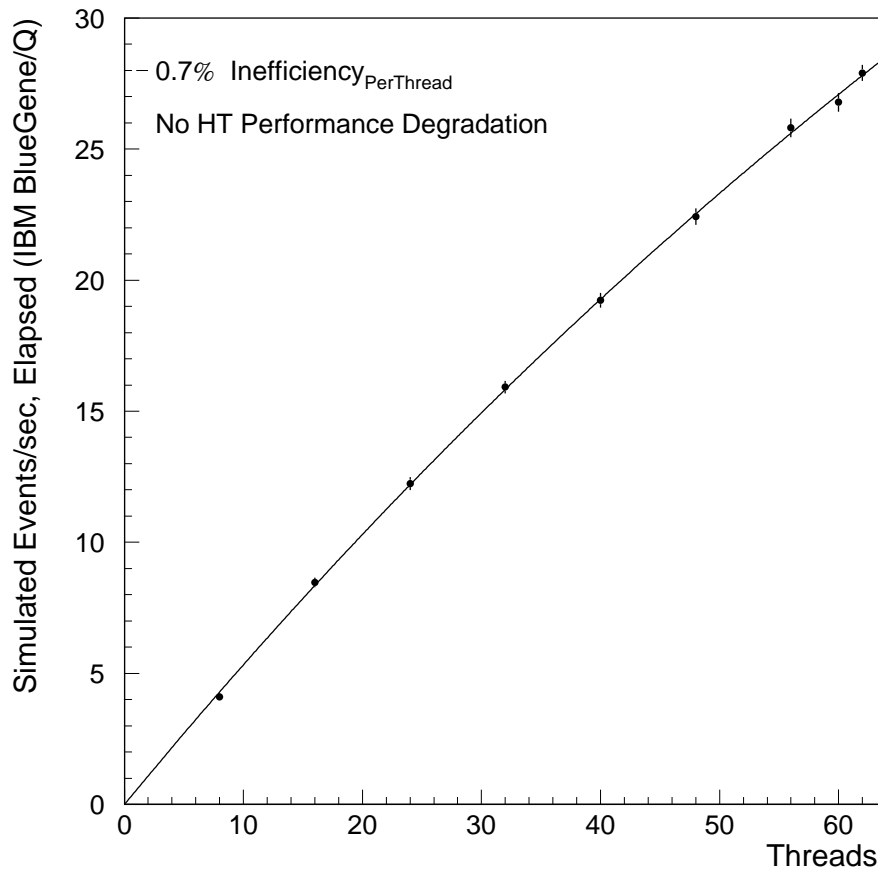
a factor of five, which gives the Intel architecture an advantage of a factor of two in terms of computing power per watt. However, due to the large number of cores available, the Blue Gene/Q architecture is still quite attractive for large scale computations.

#### 4.2. Reconstruction Software performance

Up to 64 threads per node were run without problem, as the memory requirements are less challenging for the AMS reconstruction software. However due to the fact that AMS reconstruction jobs are somewhat I/O bound and the relatively low I/O bandwidth for the IBM Blue Gene/Q nodes, the wall clock performance is saturated above 32 threads per job.

### 5. Conclusions

AMS and other (ROOT, GEANT, CERNLIB) software was successfully ported to the IBM Blue Gene/Q architecture. Massively parallel jobs, up to 2048 nodes and 131K threads successfully ran on the JUQUEEN computer for 24 hours by the wall clock, which is the maximum amount



**Figure 2.** Performance of the AMS simulation software on the IBM Blue Gene/Q node vs the number of threads. The line shows the fit of the  $\frac{\text{Events}_{\text{PerThread}}}{\text{Sec}} \cdot \frac{1-(1-\xi)^{\text{Threads}}}{\xi}$  to the measured performance. The value of the  $\xi$  parameter, which measures the per thread inefficiency, was found to be about 0.007.

of time allowed by the batch job scheduler. The massive production of AMS simulated data is expected to start in the 2017 and to deliver up to 15% of the AMS simulated data using 10% of the capacity of the JUQUEEN supercomputer.

## 6. Acknowledgements

The authors thank the Juelich Supercomputer Center for hospitality and for providing access to the JUQUEEN computer.

## References

- [1] Ting S C 2013 *Nuclear Physics B, Proc. Suppl.* **243-244**
- [2] Blue Gene/Q Application Development URL <http://www.redbooks.ibm.com/redbooks/pdfs/sg247948.pdf>
- [3] Open MPI: Open source high performance computing URL <https://www.open-mpi.org/>
- [4] Dagum L and Menon R 1998 *Computational Science & Engineering, IEEE* **5** 46–55
- [5] JUQUEEN Juelich Blue Gene/Q  
URL [http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUQUEEN/JUQUEEN\\_node.html](http://www.fz-juelich.de/ias/jsc/EN/Expertise/Supercomputers/JUQUEEN/JUQUEEN_node.html)
- [6] Antcheva I, Ballintijn M, Bellenot B, Biskup M, Brun R, Buncic N, Canal P, Casadei D, Couet O, Fine V *et al.* 2011 *Computer Physics Communications* **182** 1384–1385
- [7] URL [http://savannah.web.cern.ch/savannah/HEP\\_Applications/savroot/bugs/70542.html](http://savannah.web.cern.ch/savannah/HEP_Applications/savroot/bugs/70542.html)
- [8] Agostinelli S, Allison J, Amako K a, Apostolakis J, Araujo H, Arce P, Asai M, Axen D, Banerjee S, Barrand G *et al.* 2003 *Nuclear instruments and methods in physics research section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **506** 250–303
- [9] Shiers J *et al.* 1996 *CERN Geneva*
- [10] Argonne leadership computing facility  
URL <https://www.alcf.anl.gov/user-guides/compiling-and-linking-faq>
- [11] Choutko V *et al.* 2015 *Journal of Physics: Conference Series* vol 664 (IOP Publishing) p 032029