Performance of the AMS Offline Software at National Energy Research Scientific Computing Centre and Argonne Leadership Computing Facility

Vitali Choutko[1], *Alexander* Egorov[1], *Alexandre* Eline[1], and *Baosong* Shan[2,*]

[1]Massachusetts Institute of Technology, Laboratory for Nuclear Science, MA-02139, United States
[2]Beihang University, School of Mathematics and System Science, Beijing 100191, China

Abstract. The Alpha Magnetic Spectrometer [1] (AMS) is a high energy physics experiment installed and operating on board of the International Space Station (ISS) from May 2011 and expected to last through year 2024 and beyond. More than 50 million of CPU hours has been delivered for AMS Monte Carlo simulations using NERSC and ALCF facilities in 2017. The details of porting of the AMS software to the 2nd Generation Intel Xeon Phi Knights Landing architecture are discussed, including the MPI emulation module to allow the AMS offline software to be run as multiple-node batch jobs. The performance of the AMS simulation software at NERSC Cori (KNL 7250), ALCF Theta (KNL 7230), and Mira (IBM BG/Q) farms is also discussed.

1 Introduction

1.1 Intel Xeon Phi Knights Landing architecture

The Intel Xeon Phi Knights Landing architecture is described in details in Ref. [2]. Knights Landing is the second generation Many Integrated Core (MIC) architecture product of Intel. It is available in two forms, as a coprocessor or a host processor (CPU), based on INTEL's 14 nm process technology, and it also includes integrated on-package memory for significantly higher memory bandwidth.

Knights Landing contains up to 72 Airmont (Atom) cores with four-way hyper-threading, supporting up to 384 GB of "far" DDR4 2133 RAM and 8–16 GB of stacked "near" 3D MCDRAM [3]. Each core has two 512-bit vector units and supports AVX-512 SIMD instructions.

1.2 National Energy Research Scientific Computing Centre

The National Energy Research Scientific Computing Centre (NERSC) [4] is a high performance computing facility operated by Lawrence Berkeley National Laboratory for the United States Department of Energy Office of Science. As the mission computing centre for the Office of Science, NERSC houses high performance computing and data systems used by 7,000 scientists at national laboratories and universities around the country. NERSC is located on the main Berkeley Lab campus in Berkeley, California.

NERSC installed the second phase of its supercomputing system "Cori" with 9,668 compute nodes based on Knights Landing architecture in the second half of 2016. It features:

[*] e-mail: baosong.shan@cern.ch

- Each node contains an Intel Xeon Phi Processor 7250 @ 1.40GHz.
- 68 cores per node with support for 4 hardware threads each (272 threads total).
- 96 GB DDR4 2400 MHz memory per node using six 16 GB DIMMs (115.2 GB/s peak bandwidth). The total aggregated memory (combined with MCDRAM) is 1 PB.
- 16 GB of on-package, high-bandwidth memory with bandwidth projected to be 5X the bandwidth of DDR4 DRAM memory, (>460 GB/sec); over 5x energy efficiency vs. GDDR52; over 3x density vs. GDDR52.

After the upgrade, Cori was ranked 5th on the TOP500 list of world's fastest supercomputers in November 2016. [5]

1.3 Argonne Leadership Computing Facility (ALCF)

Argonne National Laboratory is a scientific and engineering research national laboratory operated by the University of Chicago Argonne LLC for the United States Department of Energy located near Lemont, Illinois, outside Chicago. Belonging to Argonne National Laboratory, Argonne Leadership Computing Facility is a national scientific user facility that provides supercomputing resources, including computing time, resources and data storage, and expertise to the scientific and engineering community in order to accelerate the pace of discovery and innovation in a broad range of disciplines.

In 2017, the Theta supercomputing system installation finished and it entered production mode. Theta includes 3,624 Knights Landing nodes:

- Each node contains an Intel Xeon Phi Processor 7230 @ 1.30GHz.
- 68 cores per node with support for 4 hardware threads each (272 threads total).
- 192 GB DDR4 and 16 GB MCDRAM memory per node.
- 128 GB SSD per node.

2 AMS Offline Software Practices in NERSC and ALCF

AMS is using NERSC (Edison and Cori) and ALCF (Theta and Mira) facilities for simulation.

2.1 Software porting

Mira is based on IBM Blue Gene/Q architecture, and thanks to our experiences [6] during using JuQueen [7], we are able to run the ported binaries without any issue.

Knights Landing (KNL) architecture has an important improvement for end users compared with its predecessor, the Knights Corner (KNC): the KNL is a self-booting, standalone processor and it is binary compatible with the standard Xeon instruction set, which means that it can run legacy software, compilers, tools and profilers without recompilation. This feature saved us from building another separate distribution of our offline software.

2.2 Time Divided Variables deployment

CernVM File System (CVMFS) [8] had been missing in both facilities. Very recently it started to be deployed at NERSC, but our repository (/cvmfs/ams.cern.ch) has not been included yet. At NERSC we build a Docker image to provide our database of Time Divided Variables to achieve the best performance of simultaneous starting of MPI jobs. In ALCF the database is extracted to local computing nodes before real simulation starting.

2.3 Job management

As described in Ref. [9], a light-weight production platform was designed to automate the processes of reconstruction and simulation production in AMS computing centres. The platform manages all the production stages, including job acquisition, submission, monitoring, validation, transferring, and optional scratching. The platform is based on script languages (Perl [10] and Python [11]) and sqlite3 [12] database, and it is easy to deploy and customise, according to the needs of different batch systems, storage, and transferring method. This platform is used in both NERSC and ALCF.

2.4 MPI emulation

Large scaled jobs are preferred/allowed at NERSC and ALCF. The MPI emulator cite-bib:bluegene developed for JuQueen platform is used to emulate the required features of Open MPI messaging.

3 Results

The AMS simulation software uses memory and startup time optimised GEANT-4.10 package which allows it to run on modern processors with large number of cores and limited amount of memory per core, such as Intel Xeon, IBM BlueGene/Q and Intel KNL [13].

In particular, jobs with up to 3,400 KNL nodes and 700,000 threads were successfully running in NERSC Cori facility; jobs with up to 600 KNL nodes and ~100,000 threads on ALCF Theta facility; and jobs with up to 4,096 PowerPC nodes and ~250,000 threads on ALCF Mira facility.

Figure 1 shows the measured performance of the AMS software on single node of Intel KNL at ALCF Theta (a), Intel KNL at NERSC Cori (b), and Intel Xeon at NERSC Edison (c) facilities. The AMS software performance on ALCF Mira hardware is similar to that shown in Figure 2 of Ref. [6].

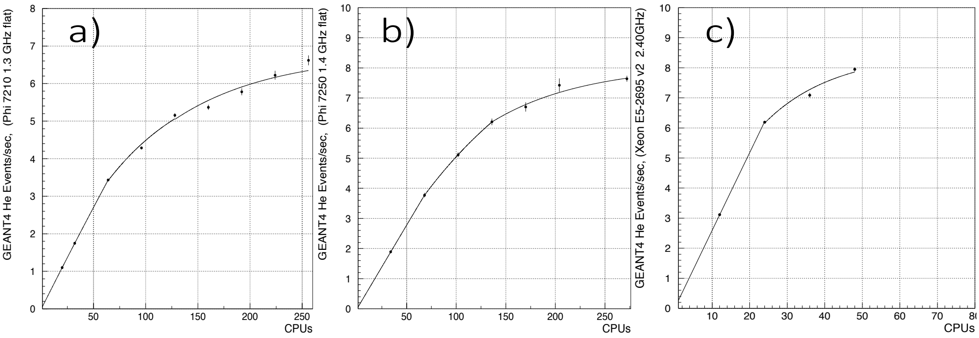

Figure 1. The measured performance of AMS software on Intel KNL hardware at ALCF Theta (a), NERSC Cori (b), and Intel Xeon hardware at NERSC Edison (c) facilities. The linear scaling versus number of threads used in application is seen up to the number of physical cores in the processors.

Figure 2 shows the AMS software's large scale performance for jobs with up to 3,400 KNL nodes and 700,000 threads at NERSC facility. As shown, the AMS software performance scales well with number of nodes and/or threads.

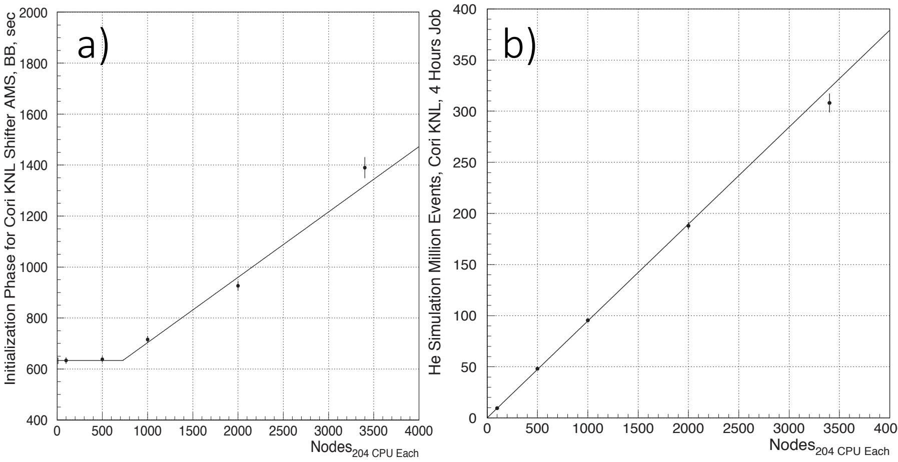

Figure 2. The AMS software large scale performance at NERSC facility. (a) Job starting time using Shifter and Burst Buffer technologies (b) Number of events simulates as function of number of KNL nodes used.

As the physics analysis of the AMS experiment is moving to the nuclei particles with higher mass and higher energy, the computing power requirement is growing. Figure 3 shows the CPU time spent on simulations from 2014 to 2018, and as shown in Figure 4, NERSC and ALCF altogether contributed 38% of the total simulation CPU time for the AMS experiment in 2018.

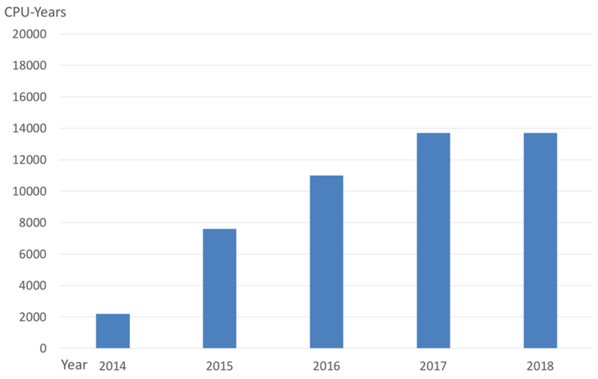

Figure 3. The amount of CPU years spent on AMS simulations from 2014 to 2018.

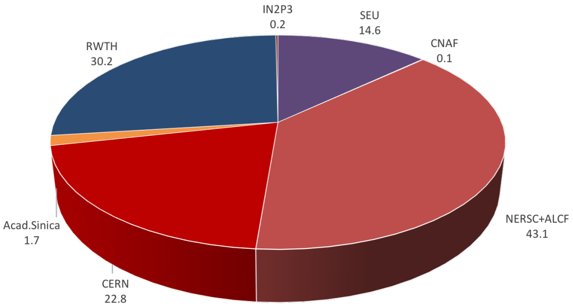

Figure 4. The simulation CPU time (in million CPU hours) contribution of AMS computing centres in 2018.

4 Conclusions

The AMS offline software has been deployed and tested at NERSC and ALCF computing centres. The measured performance on Intel KNL shows linear scaling versus number of threads up to the number of physical cores. Large scale jobs requiring up to 3400 KNL nodes and around 700,000 threads have been run on Cori and scale well with the number of nodes/threads. In 2017 and 2018 NERSC and ALCF have contributed over one third of the total CPU hours for AMS simulation, and we expect both centres will make more contributions in future.

5 Acknowledgement

This work has been completed using resources from the National Energy Research Scientific Computing Centre under Contract No. DE-AC02-05CH11231 and the Argonne Leadership Computing Facility under Contract No. DE-AC02-06CH11357.

References

[1] S. Ting, Nuclear Physics B-Proceedings Supplements **243**, 12 (2013)

[2] A. Sodani, R. Gramunt, J. Corbal, H.S. Kim, K. Vinod, S. Chinthamani, S. Hutsell, R. Agarwal, Y.C. Liu, Ieee micro **36**, 34 (2016)

[3] *Xeon phi*, https://en.wikipedia.org/wiki/Xeon_Phi

[4] *National energy research scientific computing center*, https://en.wikipedia.org/wiki/National_Energy_Research_Scientific_Computing_Center

[5] H.W. Meuer, E. Strohmaier, J. Dongarra, H. Simon, M. Meuer, *November 2016 top500 supercomputer sites* (2016)

[6] V. Choutko, A. Egorov, B. Shan, *Performance of the AMS Offline software on the IBM Blue Gene/Q architecture*, in *Journal of Physics: Conference Series* (IOP Publishing, 2017), Vol. 898, p. 072002

[7] M. Stephan, J. Docter, Journal of large-scale research facilities JLSRF **1**, 1 (2015)

[8] C. Aguado Sanchez, J. Bloomer, P. Buncic, L. Franco, S. Klemer, P. Mato, *CVMFS-a file system for the CernVM virtual appliance*, in *Proceedings of XII Advanced Computing and Analysis Techniques in Physics Research* (2008), Vol. 1, p. 52

[9] V. Choutko, O. Demakov, A. Egorov, A. Eline, B. Shan, R. Shi, *Production Management System for AMS Computing Centres*, in *Journal of Physics: Conference Series* (IOP Publishing, 2017), Vol. 898, p. 092034

[10] L. Wall et al., *The perl programming language* (1994)

[11] G. Van Rossum et al., *Python Programming Language.*, in *USENIX Annual Technical Conference* (2007), Vol. 41

[12] M. Owens, G. Allen, *SQLite* (Springer, 2010)

[13] V. Choutko, A. Egorov, A. Eline, B. Shan, *Computing Strategy of the AMS Experiment*, in *Journal of Physics: Conference Series* (IOP Publishing, 2015), Vol. 664, p. 032029