Explore True Performance Using Application Benchmark for the Next Generation HPC Systems

Robert Henschel https://orcid.org/0000-0003-2289-9398
Junjie Li
Rudolf Eigenmann
Sunita Chandrasekaran https://orcid.org/0000-0002-3560-9428

Indiana University

PTI Technical Report PTI-TR20-003

April 15, 2020

Please cite as:

Henschel, R., Li, Junjie, Eigenmann, R., & Chandrasekaran, S. (2020). *Explore True Performance Using Application Benchmark for the Next Generation HPC Systems.* (PTI Technical Report PTI-TR20-003). Indiana University. Retrieved from http://hdl.handle.net/2022/25344

NSF EAGER Grant No. OAC-1842623

Explore True Performance Using Application Benchmark for the Next Generation HPC Systems

First NSF EAGER SPEC HPG Workshop Report

September 12-13, 2019 Alexandria, VA, USA

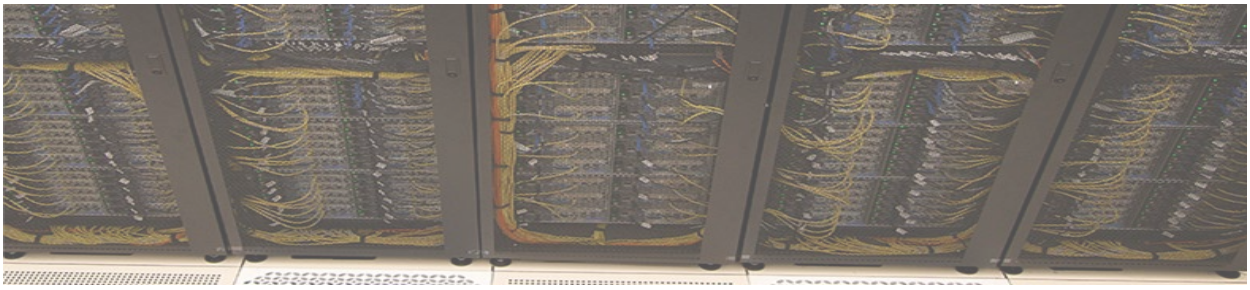
Organized by:

Research Technologies, Indiana University
Robert Henschel, Junjie Li

Department of Computer Science, University of Delaware
Rudolf Eigenmann, Sunita Chandrasekaran

This work is supported (in part) by the National Science Foundation under EArly-concept Grants for Exploratory Research (EAGER) Grant No. OAC-1842623.

DOI: https://doi.org/10.5967/jmkd-6p64



EXECUTIVE SUMMARY

Large scale computing systems in national labs, institutes and centers look for applications and benchmark suites that can stress test their hardware and software environments. In order to push the limits of the computing system, we need suites that go beyond mini-applications and kernels. To that end, SPEC HPG (Standard Performance Evaluation Corporation, High Performance Group) is building a benchmark suite called HPC2020 that will comprise of real-world scientific applications facilitating measurement and stress testing of homogeneous and heterogeneous multi-node systems equipped with traditional and/or specialized cores. A solicitation [4] was released in 2017 for the same and at the time of writing this report, the benchmark suite is in development, with a release date in late 2020.

In order to gather feedback and suggestions from experts on this important topic of application benchmarking, PIs of NSF EAGER award OAC-1842623 Robert Henchel (SPEC HPG Chair) Indiana University, Rudolf Eigenmann and Sunita Chandrasekaran from the University of Delaware organized a workshop on September 12 and 13, 2019 along with the HPG Secretary Junjie Li also from Indiana University. Experts from large supercomputing centers and national labs along with pioneers and leaders in benchmark creation were invited to attend this workshop.

The purpose of this workshop was to gather feedback and suggestions from experts for the HPC2020 benchmark suite and understand the requirements and needs of the community to measure the computing power of large systems.

This report summarizes a number of insights and suggestions from the presenters at the workshop. We plan to look into incorporating these into the HPC2020 suite. The report spans four categories (1) Key Takeaways, (2) Points to consider to build a robust suite and (3) thoughts and reflections from the presenters. We aim for broad readership and hope that the material and the discussions gathered are useful for both research and development in contexts beyond benchmarking.

LIST OF ATTENDEES

- Alan Sussman, National Science Foundation (NSF), USA
- Alejandro Suarez, National Science Foundation (NSF), USA
- Almadena Chtchelkanova, National Science Foundation (NSF), USA
- Bob Chadduck, National Science Foundation (NSF), USA
- Bronson Messer, Oak Ridge National Laboratory (ORNL), USA
- David Bailey, Lawrence Berkeley National Laboratory (LBNL), USA
- David Richards, Lawrence Livermore National Laboratory (LLNL), USA
- Edward Walker, National Science Foundation (NSF), USA
- Henry Jin, National Aeronautics and Space Administration (NASA), USA
- Jack Wells, Oak Ridge National Laboratory (ORNL), USA
- Junjie Li, Indiana University (IU), USA
- Piotr Luszczek, University of Tennessee Knoxville (UTK), USA
- Rich Vuduc, Georgia Institute of Technology (GaTech), USA
- Robert Henschel, Indiana University (IU), USA
- Rudolf Eigenmann, University of Delaware (UD), USA
- Stefan A Robila, National Science Foundation (NSF), USA
- Sunita Chandrasekaran, University of Delaware (UD), USA
- William Kramer, NCSA/University of Illinois, USA
- Zhao Liu, National Supercomputing Center in Wuxi, China

WORKSHOP PROGRAM

Thursday, September 12, 2019

1:00 pm	Welcome and Introductions by Robert Henschel download1, download2
1:15 pm	Overview of SPEC High Performance Group and HPC2020 Benchmark Suite by Sunita Chandrasekaran, Robert Henschel download1, download2
3:00 pm	Break
3:15 pm	Invited Talk I: Perspectives on Application Benchmarking at OLCF by Jack Wells, Bronson Messer download
4:15 pm	Invited Talk II: All Benchmarks are Proxy Apps, but not all Proxy Apps Make Good Benchmarks by David Richards download
5:15 pm	Group Dinner

Friday, September 13, 2019

8:00 am	Invited Talk III: Holistic Performance Assessment for Complex Computational and Data Analysis Systems by William Kramer download
9:00 am	Invited Talk IV: Benchmarking HPC Systems for Running NASA Workloads by Henry Jin download
10:00 am	Break
10:15 am	Invited Talk V: The System and Applications of Sunway TaihuLight by Zhao Liu download
11:15 am	Invited Talk VI: Benchmarking: Doomed to Succeed or Simply Succeeding by Piotr Luszczek download
12:15 pm	Group Lunch
1:30 pm	Forming SPEC High Performance Group Advisory Board by Sunita Chandrasekaran, Rudolf Eigenmann, Robert Henschel, Junjie Li
3:30 pm	Closing Remarks

REPORT CONTENT

The following sections will highlight some of the critical takeaways that help define the purpose of a benchmark suite and the common expectations from such a suite. We then discuss important points to consider when building a robust benchmark suite. Some of these points are easy to incorporate and will be included in SPEC HPC2020 right away while other points will influence the development of version 2 of the benchmark suite. To that end, SPEC HPC2020 will be iteratively improved based on feedback from the community, leveraging the SPEC infrastructure to sustain benchmark maintenance over a long period of time. The remainder of the document discusses the different presentations given by the workshop attendees. Presenters from national labs, supercomputing centers, and academia delivered a number of different perspectives on application benchmarking and we aim to capture them below.

Key Takeaways

The workshop resulted in the following key recommendations for the SPEC HPC2020 Benchmark suite. The suite must be able to:

- Cover different languages and programming models/framework
- Cover a range of algorithms by spanning multiple scientific domains
- Drive the maturity of compilers
- Drive collaboration between the scientists, hardware, and software teams so that the goal is beyond just measuring FOM (Figure of Merit - a metric to judge the value of a system)
- Provide a public database for peer-reviewed benchmark results that serve as a reference for performance studies or procurements
- Evaluate if the system performance stays as expected throughout the system's lifetime (e.g. after upgrades, changes, and regular use)

Additional recommendations for HPC2020

- Proxy applications do not always represent the physics of the full applications, and it varies from proxy application to proxy application
- Strongly recommend application developers should provide solutions and correctness tests with the application, including an acceptable range of error
- Maintain a central repository to collect all run results
- Publish results so that they can be used while making machine purchasing decisions by universities/labs/institutes/centers
- Design a performance model per application and compare the same against standard roofline model
- Make performance analysis available with the benchmark narrative
- Determine how many cores were idle and set a threshold for how many idle cores are OK while calculating SPEC score

- Allow code alterations - SPEC does so for peak results - in order to demonstrate the best performance for architectures under consideration
- Allow for the inclusion of libraries and packages that are required by the applications, as long as there are few external dependencies

Thoughts and reflections from presenters

- Robert Henschel (Chair of SPEC HPG) from Indiana University, USA along with Rudolf Eigenmann and Sunita Chandrasekaran from University of Delaware, USA presented on "Overview of SPEC High Performance Group and HPC2020 Benchmark Suite" sharing an overview of the goals of SPEC HPG, introducing HPC2020 benchmark suite to the workshop attendees, using the SPEC HPG benchmark suite for teaching and training purposes and efforts undertaken to disseminate SPEC ACCEL V1.3 to the HPC community via tutorials at ISC, SC, ICS and PEARC conferences spanning 2015 till date. Since 2018, SPEC has been offering HPG benchmarks free of charge to qualified non-profit organizations worldwide. Read more

- Jack Wells and Bronson Messer from the Oak Ridge National Lab (ORNL), Tennessee, USA presented on "Perspectives on Application Benchmarking at OLCF" that elaborated about the lab's efforts with respect to preparing a vanguard of applications for accelerated computing as part of the CAAR-OLCF5 effort [1]. The lab already uses the SPEC HPG ACCEL benchmark for compiler development and performance analysis. The lab is interested in a benchmark that is (a) beyond a suite of mini-applications but a scalable set of applications that can cover a variety of programming models 'X' in MPI +X in order to drive compiler development and measure its maturity (b) allows source code modifications to get the best linear time performance when possible and (c) foster collaborations between the application developer, hardware and software teams. The speakers also discussed ISCM metric - INCITE System Capability Metric - that aims to take a balanced view of the system behavior that are critical to application performance. The metric is not expected to be a static value but expresses the capability of the system at a given time stamp. (Once the paper on ISCM is made available publicly, we will update this report with a citation)

- David Richards from Lawrence Livermore National Laboratory (LLNL), CA, USA presented on "All Benchmarks are Proxy Apps, but not all Proxy Apps Make Good Benchmarks" cautioning that often mini-applications do not fully represent the main application. Simplified physics aren't always faithful to the functionality of the main algorithm. There is a difference between proxy applications and a benchmark suite. Proxy applications can become benchmarks only when appropriate run rules and figures of merit are defined. Benchmark curation over a period of time is important.

- William Kramer who leads Pittsburgh Supercomputing Center, Philadelphia, USA presented on "Holistic Performance Assessment for Complex Computational and Data Analysis Systems" sharing his vast experiences with benchmarking effort. Benchmark data would be more valuable with information that includes the cost of systems and not just the performance. Partnering with intersect360 could supply "street price cost" [2]. A benchmark metric also needs to address that the performance stays as expected throughout the system's lifetime (e.g. after upgrades, changes, and regular use). Good attributes of a benchmark include proportionality, reliability, consistency, independence, ease of use, and repeatability.

- Henry Jin from NASA, Ames, CA, USA presented on "Benchmarking HPC Systems for Running NASA Workloads" enforcing the purpose of a benchmark suite to be able to evaluate pathfinding architecture. The suite should not only evaluate for good performance, but also assess worst-case scenario performance, such as MPI network contention. It is often the case that the benchmark does not pay attention to convergence. Different platforms or hardware architectures seek different algorithms to arrive at optimal performance. Need strategies to arrive at a common solution with respect to both CPUs and accelerators.

- Zhao Liu from the National Supercomputing Center, Wuxi, China presented on "The System and Applications of Sunway TaihuLight" sharing experiences on evaluating SPEC, HPL, NPB results on their system, Sunway Taihulight. Zhao also highlighted the challenges to designing a benchmark from applications. From the application standpoint, the challenges include (1) an application consisting of a large number of lines of code thus making it a challenge to decipher the functionalities needed to create a benchmark, (2) creating a benchmark that is a misfit between its design and hardware architecture, (3) an application containing too many hotspots to no hotspots,(4) lack of personnel with interdisciplinary knowledge and experience, as a result the benchmark lacks a real insight into the science itself.

- Piotr Luszczek from University of Tennessee, Knoxville, Tennessee USA presented on "Benchmarking: Doomed to Succeed or Simply Succeeding" reviewing the history of LINPACK benchmarks [5], features and highlights of HPCG benchmark suite, and the new half-precision LINPACK for Accelerator Introspection (HPL-AI) benchmark that is capable of doing reduced precision math.

- Some useful references include:
    - An article by David Bailey on 12 ways to fool the masses https://www.researchgate.net/publication/24293656_Twelve_Ways_to_Fool_the_Masses_When_Giving_Performance_Results_on_Parallel_Computers

- Workload Analysis of Blue Waters https://www.osti.gov/servlets/purl/1365199
- The System Sustained Performance (SSP) benchmark developed by NERSC https://www.nersc.gov/users/computational-systems/cori/nersc-8-procurement/trinity-nersc-8-rfp/nersc-8-trinity-benchmarks/ssp/

References

[1] CAAR OLCF5 https://www.olcf.ornl.gov/caar/frontier-caar/
[2] Intersect360 http://www.intersect360.com/
[3] Allan, Benjamin A. Tue . "Figures of merit for production HPC.". United States. doi:10.2172/1571365. https://www.osti.gov/servlets/purl/1571365.
[4] SPEC HPC2020 Search Program https://www.spec.org/hpg/search/
[5] LINPACK https://www.netlib.org/benchmark/hpl/