Performance of Popular HPC Applications on the Intel Knights Landing Platform

Antonio Gómez-Iglesias, Feng Chen, Lei Huang, Hang Liu,
Si Liu, Antia Lamas-Linares, John Cazes, Carlos Rosales
Texas Advanced Computing Center
The University of Texas at Austin
Austin, Texas
Email: {agomez,chenk,huang,hliu,siliu,alamas,cazes,carlos}@tacc.utexas.edu

Abstract—**We present direct performance measurements for six popular HPC applications in the Knights Landing (KNL) platform. Performance numbers for Sandy Bridge and Haswell processors are provided for contrast. The applications (NAMD, Gromacs, FLASH4, WRF, Quantum Espresso and NCBI BLAST) were selected from among the ten most used in the Stampede supercomputer at the Texas Advanced Computing Center as well as good representative of workloads used by large number of users and, given their diversity, should be representative of typical HPC workloads. All runs were performed with publicly available codes without modification – except a single line added to FLASH4 to enable threading in a given code section – and so results should be expected to improve as developers gain access to KNL. Current results are promising, with execution on a single KNL processor showing speedups up to 2.7X with respect to a dual socket Sandy Bridge and up to 1.7x with respect to a dual socket Haswell.**

I. INTRODUCTION

The relentless pursuit of increase in computational power has lead to a common trend in the World of HPC: increased core counts and more specialized CPU cores. Accelerators such as GPUs or the first generation Intel Xeon Phi – Knights Corner or KNC – are good examples of this trend, providing high peak performance within a moderate power envelope. The common presence of accelerated systems in the TOP500 list is testament to the popularity of this choice in the HPC community. Accelerated systems are not without challenges though. The presence of a PCIe interface between host CPU and accelerator often creates a data motion bottleneck that reduces the achieved sustained performance of scientific applications. minimizing this bottleneck in heterogeneous computing infrastructures requires complex implementations and increases both time of development and code maintenance cost.

The second generation of the Intel Xeon Phi processor, the Knights Landing (KNL) platform, attempts to address some of this issues by providing a many-core architecture on a socket. On paper the KNL architecture represents a qualitative leap over the previous generation (KNC). It includes features that make programming simpler and more effective such as support for out-of-order and unaligned fast instructions, as well as a completely redesigned connectivity mesh across cores and attached high speed memory. The single most important change in KNL is the fact that it is a self-hosted processor running a regular Linux OS, making it amenable to homogeneous cluster configurations and eliminating the PCIe bottleneck present in heterogeneous accelerated solutions.

Int his work we present an evaluation of the performance of several well known HPC applications, which have been chosen based on the overall utilization of the Stampede supercomputer at the Texas Advanced Computing Center. Stampede hosts thousands of users from all backgrounds and research fields, and thus these applications should be representative of the wider range of workloads used in across the scientific computing community. The utilization data was provided by XALT.

II. CASE STUDIES

All applications were executed in three different systems at the Texas Advanced Computing Center: Stampede (Sandy Bridge, dual socket Xeon E5-2680, 32 GB DDR3); Lonestar 5 (Haswell, dual socket Xeon E5-2690v3, 64 GB DDR4); and Stampede KNL Upgrade (Knights Landing, single socket Xeon Phi 7250, 16GB MCDRAM + 96 GB DDR4). Intel compiler version 15.0.2 was used in Stampede, while versions 16.0.1 and 16.0.3 were used in Lonestar 5 and the Stampede KNL Upgrade respectively.

Previous work has shown that of the many possible memory configuration modes for the Knights Landing platform the Cache-Quadrant combination provides the highest performance for many scientific applications, so all KNL results in this study were obtained using that mode [7].

A. NAMD

NAMD[5] is a parallel molecular dynamics code for high-performance simulation of large molecular systems used by many research teams around the world and the 2002 Gordon Bell Prize winner. NAMD is also the most used application on Stampede. This study uses the APOA1 NAMD benchmark with modified input. The APOA1 benchmark is a 92-thousand atom molecular dynamics simulation of Apolipoprotein A-1 which is the primary component of the high-density lipoprotein cholesterol molecule.

B. WRF

The Weather Research Forecasting (WRF) model is a mesoscale weather model widely used for research and operational weather forecasting [4]. It is among the ten most used application on Stampede. For this investigation the benchmark case is a 12km simulation over the Continental U.S. (CONUS) domain on October 24, 2001 with a time step of 72 seconds.

C. Gromacs

GROMACS is a versatile package to perform molecular dynamics, i.e. simulate the Newtonian equations of motion for systems with hundreds to millions of particles [2]. It is primarily designed for biochemical molecules like proteins, lipids and nucleic acids that have a lot of complicated bonded interactions. Hen egg white lysozyme (PDB code 1AKI) and water (SPC model) solutions were simulated using version 5.1.2. The simulated systems consist a total of 140124 atoms. All simulations were performed in the isothermal isobaric (NpT) ensemble at 300 K and 1 atm.

D. Quantum Espresso

Quantum Espresso [3] is an integrated suite of Open-source codes for electronic-structure calculations and materials modeling at the nanoscale. It is based on density-functional theory, plane waves, and pseudopotentials. A medium size benchmark input for PWscf, AUSURF112, is selected and computed by Quantum Espresso V5.4.0.

E. FLASH

FLASH is a hydrodynamics code that presents a modular and extensible design. It uses a collection of legacy codes and inter-operable modules to generate different architectures. In this case, we have used the StirTurb setup with a small problem size (64^3) [6].

F. NCBI BLAST

One of the most popular tools for sequence similarity searches is the Basic Local Alignment Search Tool (BLAST) by NCBI [1]. We consider a threaded implementation of NCBI BLAST since the MPI version has not been actively maintained in several years.

III. RESULTS

The single node performance for each one of the applications on Sandy Bridge, Haswell and KNL are displayed in Fig. **??**, normalized to the Sandy Bridge result in each case. In general, it can be seen how KNL either outperforms or, at least, provides similar results to those achieved when using Haswell. Only in the case of Gromacs and NCBI BLAST, Haswell significantly improves the performance achieved with KNL. Overall, it is possible to see how the many-core approach is, in this case, comparable if not better than the used multi-core alternatives provided by Intel. This fact, combined with some of the characteristics of the KNL (self-hosted system, running a full operating system), makes this new processor an optimal alternative for some of the most popular HPC applications.

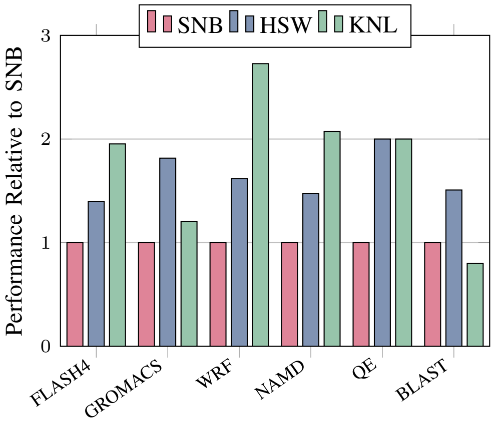

Fig. 1: Results for the applications of interest on Sandy Bridge, Haswell and KNL processors

Finally, it is important to remark that no effort was put into optimizing any of these codes to target any of the capabilities of the processors analyzed. Furthermore, because of the maturity of Sandy Bridge and Haswell processors, it is possible that some of these applications have been previously optimized for these targets, while no optimization has been yet added to support KNL.

REFERENCES

[1] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403 – 410, 1990.

[2] H.J.C. Berendsen, D. van der Spoel, and R. van Drunen. GROMACS: A message-passing parallel molecular dynamics implementation. *Computer Physics Communications*, 91(1):43 – 56, 1995.

[3] Paolo Giannozzi et al. Quantum espresso: a modular and open-source software project for quantum simulations of materials. *Journal of Physics: Condensed Matter*, 21(39):395502, 2009.

[4] W. C. Skamarock et al. A description of the advanced research WRF version 3. Technical report, National Center for Atmospheric Research, 2008.

[5] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kalé, and Klaus Schulten. Scalable molecular dynamics with namd. *Journal of Computational Chemistry*, 26(16):1781–1802, 2005.

[6] David Radice, Sean M. Couch, and Christian D. Ott. Implicit large eddy simulations of anisotropic weakly compressible turbulence with application to core-collapse supernovae. *Computational Astrophysics and Cosmology*, 2(1):1–17, 2015.

[7] A. Sodani, R. Gramunt, J. Corbal, H. S. Kim, K. Vinod, S. Chinthamani, S. Hutsell, R. Agarwal, and Y. C. Liu. Knights landing: Second-generation intel xeon phi product. *IEEE Micro*, 36(2):34–46, Mar 2016.