Optimizing network software to advance scientific discovery

April 16 2019, by Ariana Tantillo

Brookhaven Lab collaborated with Columbia University, University of Edinburgh, and Intel to optimize the performance of a 144-node parallel computer built from Intel's Xeon Phi processors and Omni-Path high-speed communication network. The computer is installed at Brookhaven's Scientific Data and Computing Center, as seen above with technology engineer Costin Caramarcu. Credit: Brookhaven National Laboratory

High-performance computing (HPC)—the use of supercomputers and parallel processing techniques to solve large computational problems—is of great use in the scientific community. For example, scientists at the U.S. Department of Energy's (DOE) Brookhaven National Laboratory rely on HPC to analyze the data they collect at the large-scale experimental facilities on site and to model complex processes that would be too expensive or impossible to demonstrate experimentally.

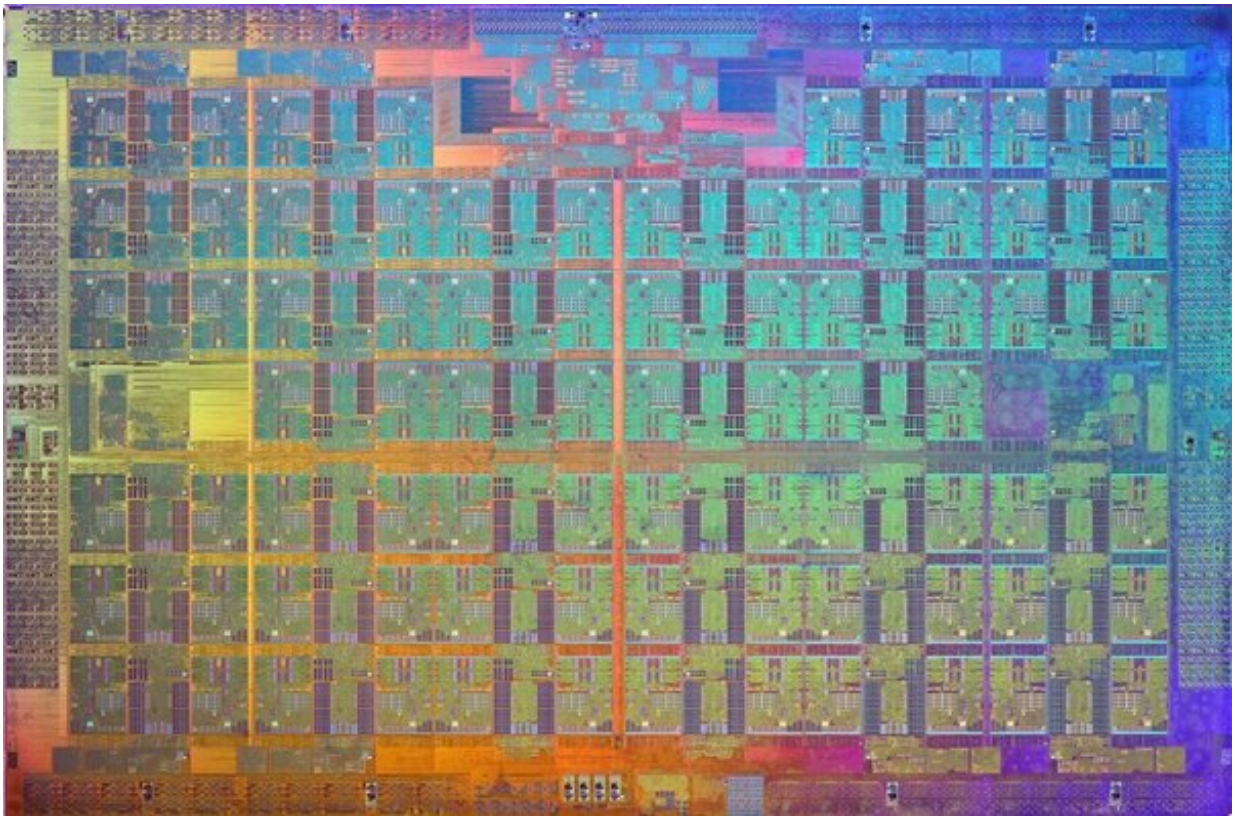
Modern science applications, such as simulating particle interactions, often require a combination of aggregated computing power, high-speed networks for data transfer, large amounts of memory, and high-capacity storage capabilities. Advances in HPC hardware and software are needed to meet these requirements. Computer and computational scientists and mathematicians in Brookhaven Lab's Computational Science Initiative (CSI) are collaborating with physicists, biologists, and other domain scientists to understand their data analysis needs and provide solutions to accelerate the scientific discovery process.

An HPC industry leader

For decades, Intel Corporation has been one of the leaders in developing HPC technologies. In 2016, the company released the Intel Xeon PhiTM processors (formerly code-named "Knights Landing"), its second-generation HPC architecture that integrates many processing units (cores) per chip. The same year, Intel released the Intel Omni-Path Architecture high-speed communication network. In order for the 5,000 to 100,000 individual computers, or nodes, in modern supercomputers to work together to solve a problem, they must be able to quickly communicate with each other while minimizing network delays.

Soon after these releases, Brookhaven Lab and RIKEN, Japan's largest comprehensive research institution, pooled their resources to purchase a small 144-node parallel computer built from Xeon Phi processors and

two independent network connections, or rails, using Intel's Omni-Path Architecture. The computer was installed at Brookhaven Lab's Scientific Data and Computing Center, which is part of CSI.



An image of the Xeon Phi Knights Landing processor die. A die is a pattern on a wafer of semiconducting material that contains the electronic circuitry to perform a particular function. Credit: Intel

With the installation completed, physicist Chulwoo Jung and CSI computational scientist Meifeng Lin of Brookhaven Lab; theoretical physicist Christoph Lehner, a joint appointee at Brookhaven Lab and the University of Regensburg in Germany; Norman Christ, the Ephraim Gildor Professor of Computational Theoretical Physics at Columbia

University; and theoretical particle physicist Peter Boyle of the University of Edinburgh worked in close collaboration with software engineers at Intel to optimize the network software for two science applications: particle physics and machine learning.

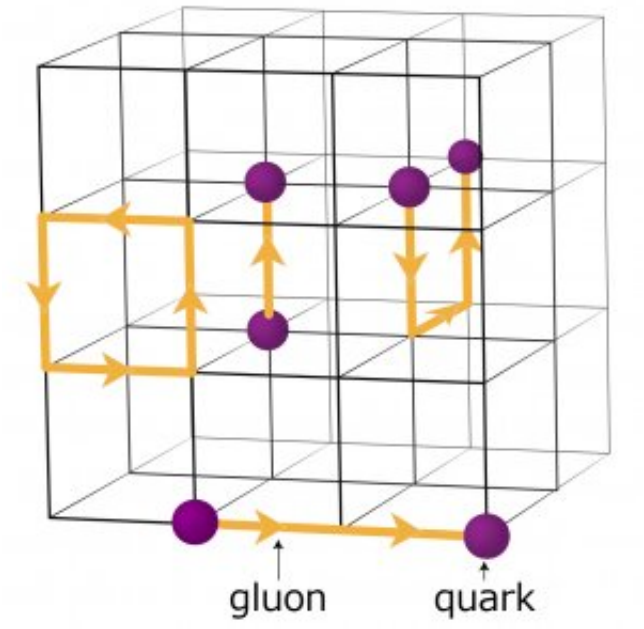
"CSI had been very interested in the Intel Omni-Path Architecture since it was announced in 2015," said Lin. "The expertise of Intel engineers was critical to implementing the software optimizations that allowed us to fully take advantage of this high-performance communication network for our specific application needs."

Network requirements for scientific applications

For many scientific applications, running one rank (a value that distinguishes one process from another) or possibly a few ranks per node on a parallel computer is much more efficient than running several ranks per node. Each rank typically executes as an independent process that communicates with the other ranks by using a standard protocol known as Message Passing Interface (MPI).

For example, physicists seeking to understand how the early universe formed run complex numerical simulations of particle interactions based on the theory of quantum chromodynamics (QCD). This theory explains how elementary particles called quarks and gluons interact to form the particles we directly observe, such as protons and neutrons. Physicists model these interactions by using supercomputers that represent the three dimensions of space and the dimension of time in a four-dimensional (4-D) lattice of equally spaced points, similar to that of a crystal. The lattice is split into smaller identical sub-volumes. For lattice QCD calculations, data need to be exchanged at the boundaries between the different sub-volumes. If there are multiple ranks per node, each rank hosts a different 4-D sub-volume. Thus, splitting up the sub-volumes creates more boundaries where data need to be exchanged and

therefore unnecessary data transfers that slow down the calculations.



A schematic of the lattice for quantum chromodynamics calculations. The intersection points on the grid represent quark values, while the lines between them represent gluon values. Credit: Brookhaven National Laboratory

Software optimizations to advance science

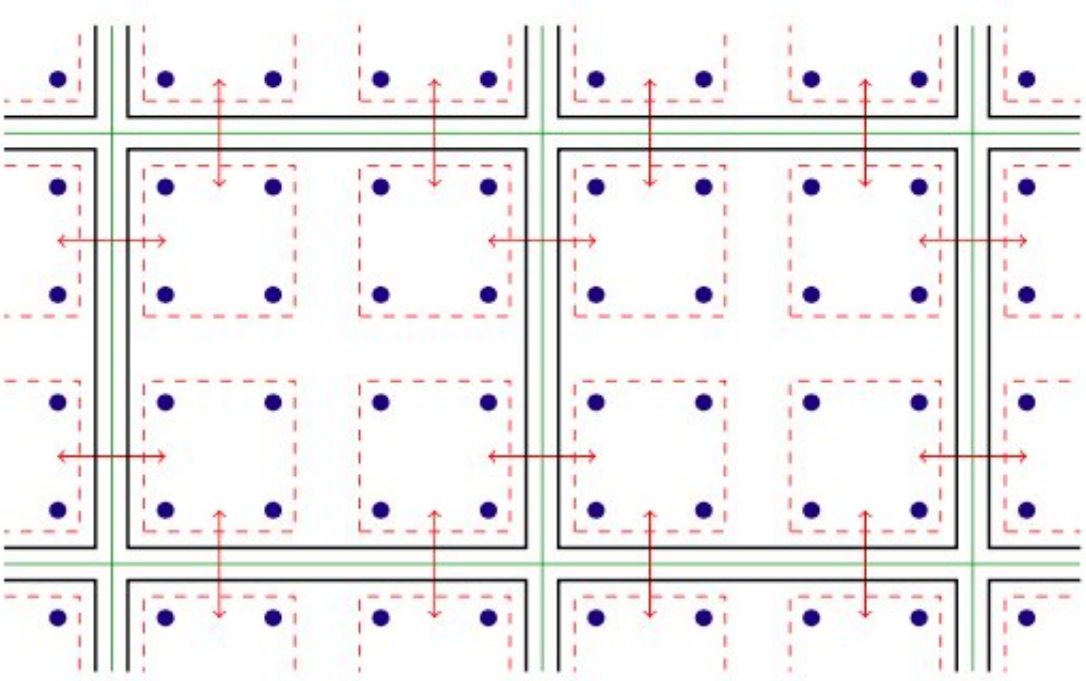
To optimize the network software for such a computationally intensive scientific application, the team focused on enhancing the speed of a single rank.

"We made the code for a single MPI rank run faster so that a proliferation of MPI ranks would not be needed to handle the large communication load present for each node," explained Christ.

The software within the MPI rank exploits the threaded parallelism

available on Xeon Phi nodes. Threaded parallelism refers to the simultaneous execution of multiple processes, or threads, that follow the same instructions while sharing some computing resources. With the optimized software, the team was able to create multiple communication channels on a single rank and to drive these channels using different threads.

The MPI software was now set up for the scientific applications to run more quickly and to take full advantage of the Intel Omni-Path communications hardware. But after implementing the software, the team members encountered another challenge: in each run, a few nodes would inevitably communicate slowly and hold the others back.



Two-dimensional illustration of threaded parallelism. Key: green lines separate physical compute nodes; black lines separate MPI ranks; red lines are the communication contexts, with the arrows denoting communication between nodes or memory copy within a node via the Intel Omni-Path hardware. Credit: Brookhaven National Laboratory

They traced this problem to the way that Linux—the operating system used by the majority of HPC platforms—manages memory. In its default mode, Linux divides memory into small chunks called pages. By reconfiguring Linux to use large ("huge") memory pages, they resolved the issue. Increasing the page size means that fewer pages are needed to map the virtual address space that an application uses. As a result, memory can be accessed much more quickly.

With the software enhancements, the team members analyzed the performance of the Intel Omni-Path Architecture and Intel Xeon Phi processor compute nodes installed on Intel's dual-rail "Diamond" cluster and the Distributed Research Using Advanced Computing (DiRAC) single-rail cluster in the United Kingdom. For their analysis, they used two different classes of scientific applications: particle physics and machine learning. For both application codes, they achieved near-wirespeed performance—the theoretical maximum rate of data transfer. This improvement represents an increase in network performance that is between four and ten times that of the original codes.

"Because of the close collaboration between Brookhaven, Edinburgh, and Intel, these optimizations were made available worldwide in a new version of the Intel Omni-Path MPI implementation and a best-practice protocol to configure Linux memory management," said Christ. "The factor of five speedup in the execution of the physics code on the Xeon Phi computer at Brookhaven Lab—and on the University of Edinburgh's new, even larger 800-node Hewlett Packard Enterprise "hypercube" computer—is now being put to good use in ongoing studies of fundamental questions in particle physics."

More information: Accelerating HPC codes on Intel(R) Omni-Path Architecture networks: From particle physics to machine learning.