Software Roadmap to Plug and Play Petaflop/s

William T. C. Kramer,* Jonathan Carter, David Skinner, Lenny Oliker, Parry Husbands, Paul Hargrove, John Shalf, Osni Marques, Esmond Ng, Tony Drummond, Kathy Yelick, Lawrence Berkeley National Laboratory

Summary

In the report "Software Roadmap to Plug and Play Petaflop/s," a group of Berkeley Lab's high performance computing experts outline the potential problems that need to be addressed in advance to ensure the scientific productivity of the forthcoming petascale systems.

In the next five years, the DOE expects to field systems that reach a petaflop of computing power in scale. In the near term (two years), DOE will have several "near-petaflops" systems that are 10% to 25% of a peraflop-scale system. A common feature of these precursors to petaflop systems (such as the Cray XT3 or the IBM BlueGene/L) is that they rely on an unprecedented degree of concurrency, which puts stress on every aspect of HPC system design. Such complex systems will likely break current "best practices" for fault resilience, I/O scaling, and debugging, and even raise fundamental questions about programming languages and application models. It is important that potential problems are anticipated far enough in advance that they can be addressed in time to prepare the way for petaflop-scale systems.

DOE asked the NERSC and Computational Research Divisions at Lawrence Berkeley National Laboratory to address these issues by considering the following four questions:

1. What software is on a critical path to make the systems work?
2. What are the strengths/weaknesses of the vendors and of existing vendor solutions?
3. What are the local strengths at the labs?
4. Who are other key players who will play a role and can help?

Berkeley Lab responded to these questions in the report "Software Roadmap to Plug and Play Petaflop/s" (https://www.nersc.gov/news/reports/LBNL-59999.pdf). The report is organized as follows.

Section 1 provides a high-level answer to question #1, "What software is on the critical path to make the systems work?" We broadened the response to include both hardware and software issues because the two are so intricately entwined on systems of this scale. We also differentiate near-term (2007) challenges from those that we anticipate in the long term (2008 and beyond).

Section 2 addresses question #2, "Describe the strengths and weaknesses of the vendors and existing vendor solutions," using data collected from the NERSC-5 procurement.

* 510-486-7577, WTKramer@lbl.gov

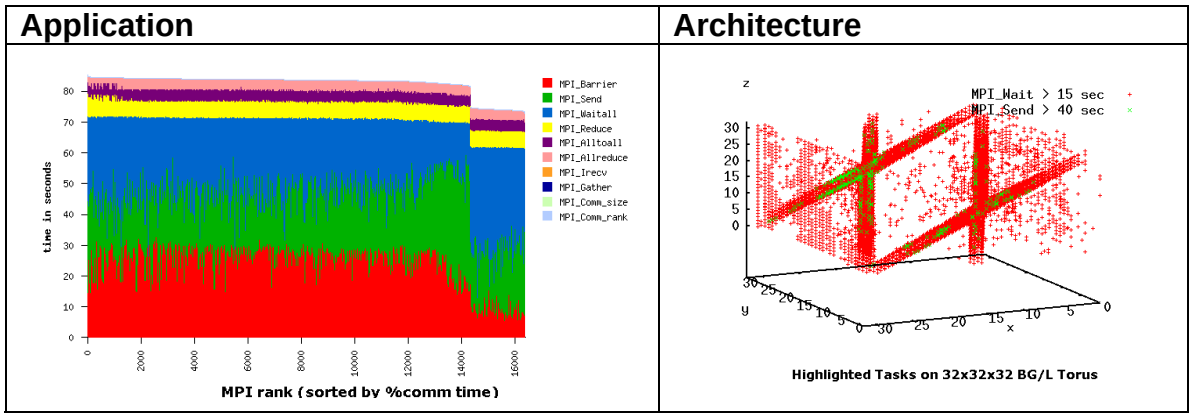

Figure 1. Two views of the same application reveals performance is more complex than a single number. Examining how performance varies within a parallel application (in MPI rank space) or within the architectural space of the torus reveals important performance characteristics and makes bottlenecks to performance clear. In application space (left), we see that most of the tasks spend more time communicating than a particular small fraction of tasks. This performance discontinuity is a likely indication of load imbalance. By moving to the architectural space (right) of machine coordinates within the torus, the bottlenecks that make up the discontinuity become clear. Given the massive concurrency involved, care must be taken to retrieve and render the relevant data but not all performance data.

Section 3 addresses question #3, "What are the local strengths at the labs?" by describing the local strengths at NERSC and LBNL for responding to the challenges of petascale computing described in the earlier sections.

Section 4 addresses question #4, "Who are other key players who will play a role and can help?" by identifying key players at other institutions who can be considered key partners for addressing the problems posed in earlier sections.

Section 5 provides supplemental information regarding NERSC's effort to use non-invasive workload profiling to identify application requirements for future systems (Figure 1). The data collected by NERSC may be valuable for proactively identifying bottlenecks in current systems and anticipating future application requirements.

Section 6 describes a set of codes that provide good representation of the application requirements of the broader DOE scientific community. The success of these codes is a bellwether for the overall success of these computing platforms for the range of DOE scientific applications.

Section 7 is a comprehensive production software requirements checklist that was derived from the experience of the NERSC-3, NERSC-4, and NERSC-5 procurement teams. It presents an extremely detailed view of the software requirements for a fully functional petaflop-scale system environment. It also includes an assessment of how emerging near-petaflop systems (XT3, BG/L, Power SP) conform or fail to conform to these requirements.

For further information on this subject contact:
Barbara Helland, Program Manager
Mathematical, Information, and Computational Sciences Division
Office of Advanced Scientific Computing Research
Phone: 301-903-3127
Barbara.Helland@science.doe.gov