From *SIAM News*, Volume 41, Number 6, July/August 2008

Los Alamos National Lab and IBM Bring Computing into the Petascale Era

On learning that Roadrunner had broken the petaflop/s barrier, SIAM News *asked Los Alamos scientists for details about the unique hybrid supercomputer,the IBM/LANL collaboration that leveraged video game technology to produce it, and early results obtained on it by lab scientists.*

By John Turner and Andy White

At 3:30 A.M. on May 26, 2008, Los Alamos National Laboratory and IBM ushered in the era of petascale computing when the supercomputer Roadrunner achieved sustained performance of more than one quadrillion (10^{15}) floating-point operations per second (1 petaflop/s) on the industry-standard LINPACK benchmark for high-performance computing. Within days of the benchmark run, LANL researchers were running scientific applications on the supercomputer, such as molecular dynamics, plasma physics, and models of the human visual system, at unprecedented levels of performance.

These achievements are the result of a multi-year collaboration between Los Alamos and IBM to build the first large-scale hybrid supercomputer (details can be found at http://www.lanl.gov/roadrunner/).

Roadrunner: A Natural Next Step for LANL

The collaboration officially began in the fall of 2006, when a contract was awarded to IBM to build NNSA's next supercomputer. One design goal for the system was a sustained petaflop/s on the LINPACK benchmark, which solves a large dense system of linear equations by Gaussian elimination with partial pivoting. Word of the system was greeted with both skepticism and excitement, largely because of the unique design, which couples processors of three types in a hybrid, accelerated architecture.

This accelerated design was a natural extension of work that had begun in earnest in 2002 to examine a variety of computational accelerators. One example of such work was an internally funded Laboratory-Directed Research and Development project. This project investigated the use of accelerators, such as video cards or graphical processing units (GPUs) and Field-Programmable Gate Arrays (FPGAs), to improve the performance of key numerical algorithms—for the solution of linear systems of equations, radiative heat transfer, and the shallow-water equations, among others. Although results were mixed, overall they were encouraging enough to demonstrate that an accelerated design could work.

More importantly, work on the project was good preparation for research staff who would be involved in Roadrunner.

Roadrunner Design

In essence, Roadrunner is a standard, albeit large, Linux cluster consisting of a modest number of nodes made up of commodity processors running Linux interconnected by InfiniBand. Each node is accelerated, from approximately 14 gigaflop/s peak to 450 gigaflop/s, by Cell processors, which for certain tasks perform at considerably higher levels. Each Roadrunner node consists of two dual-core AMD Opteron processors in an IBM LS21 blade connected to two IBM QS22 blades, each containing two IBM PowerXCell 8i processors—a total of four Opteron cores and four IBM PowerXCell 8i processors per node.

The IBM PowerXCell 8i processor (http://www.ibm.com/technology/cell/) is an improved version of the Cell Broadband Engine (CBE) used in the Sony PlayStation 3 video game console. The CBE, which was jointly developed by STI, a consortium of Sony, Toshiba, and IBM, is a hybrid parallel processor on a chip. It consists of a PowerPC core, which runs the Linux operating system and performs such housekeeping tasks as thread management, augmented by eight Synergistic Processing Elements (SPEs). Each SPE is a 128-bit vector engine with 256 kilobytes of local memory and its own memory controller. The nine processing elements, along with memory and I/O, are connected to a very-high-bandwidth on-chip bus. Independent memory controllers for each SPE result in both high performance (in that data movement and computation can be overlapped) and flexibility (in that the SPEs can operate on similar data in synchrony, independently on completely different tasks or in a pipelined, streaming mode).

This leveraging of a (uniquely capable) commodity processor reflects a trend that has been apparent for some time. In the early days of high-performance computing, national laboratories and other large institutions could influence the technology significantly, but these days we are dwarfed by such markets as games and game consoles. A recurring question for us now is, which commodity advances can we take advantage of for HPC? For Roadrunner, IBM was able to leverage the investment in design and fabrication of the original CBE chip and extend it to produce a version more suited to HPC. The primary differences between the CBE and the IBM PowerXCell 8i are a 65-nm (rather than 90-nm) manufacturing process, dramatically improved double-precision (64-bit) arithmetic, and use of DDR2 memory rather than Rambus (XDR) memory.

In aggregate, each Roadrunner node is theoretically capable of more than 400 gigaflop/s of double-precision performance. One hundred eighty of these nodes, along with 12 I/O nodes, are connected by InfiniBand 4x DDR in one Connected Unit (CU). The full Roadrunner system consists of 17 CUs, again connected by InfiniBand, for a total of just over 6000 dual-core Opterons and just over 12,000 IBM PowerXCell 8i processors.

Why Accelerators? Why Cell?

As shown in Figure 1, Roadrunner represents a path to petascale dramatically different from, say, the approach taken by the designers of IBM's BlueGene, or from a simple scaling up of standard clusters. A diversity of architectures is necessary at this point in time, when the silicon foundations of high-performance computing are changing radically. A machine that could achieve a sustained petaflop/s on LINPACK, for example, would require roughly 32,000 quad-core processors, depending on clock rate and efficiency. Even at four processors per node, this hypothetical machine would require 8000 nodes—significantly more than the number in Roadrunner.

The modest node count of Roadrunner has significant advantages, including reduced power consumption: Along with the petaflop/s record, QS22 and Roadrunner are setting new records in power efficiency. The long history of supercomputing at LANL illustrates how dramatic these improvements are.

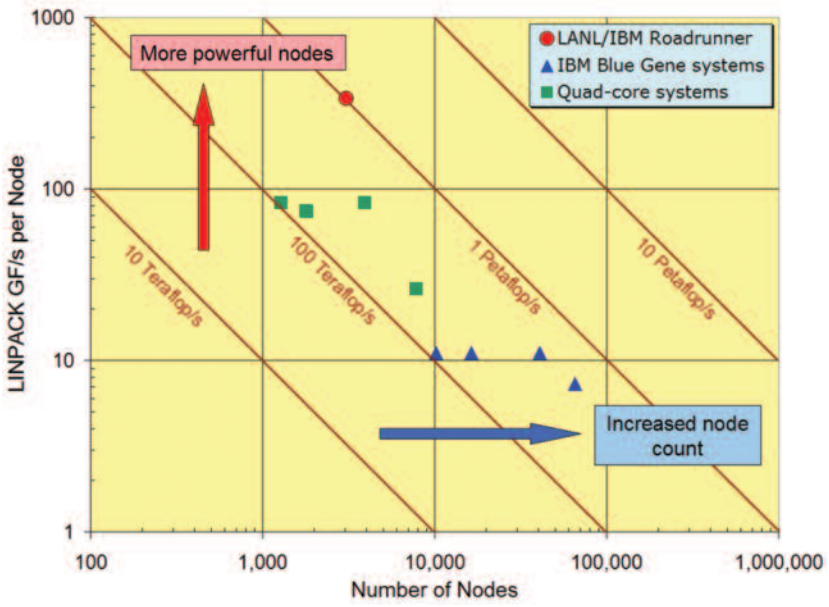

Figure 1. *With fewer, more powerful hybrid nodes, Roadrunner represents a different path to petascale computing.*

The first vector supercomputer installed at LANL (in 1976) needed 115 kilowatts to deliver about 100 megaflop/s on highly optimized matrix operations hand-coded in assembly language. The 10 million such machines that would be required to match Roadrunner's performance would require 1.15 terawatts—roughly equivalent to the entire electricity-generating capacity in the U.S.

Fortunately, over the past 32 years, the power efficiency of the fastest computers in the world has increased more than 500,000-fold. The Roadrunner system delivered its record-breaking 1.026 petaflop/s using only 2.345 megawatts, within the power limits of a large computer facility. Roadrunner's power efficiency rating of 437 megaflop/s/watt places it at number 3 on the Green500 list (http://green500.org/). Two German systems based on QS22 blades were numbers 1 and 2.

Programming Roadrunner

There's no doubt that the complexity of Roadrunner's hybrid design can appear daunting to application developers. The challenges include:

- the need for three separate compilers (for the Opteron, PowerPC, and SPE instruction sets),
- the need to explicitly manage data movement between the Opteron and IBM PowerXCell 8i memory spaces, and
- the need to explicitly manage movement of both program and data into and out of the local store for each SPE within the IBM PowerXCell 8i.

The byte ordering differences between Opteron and PowerPC (little and big endian, respectively) were not the challenge we anticipated.

A critical task was to ensure that applications of importance to national security can be modified for Roadrunner with significant performance gains. LANL applications range from single-physics science codes used for detailed investigation of physical phenomena to multiphysics codes that simulate complex physical interactions taking place over a range of spatial and temporal scales. The applications utilize a wide range of numerical algorithms, including Eulerian, Lagrangian, and hybrid techniques on meshes, particle methods (both with and without underlying meshes), and sparse linear algebra techniques.

The project specifically included design and implementation of libraries that would facilitate hybrid programming, as required for Roadrunner. IBM and Los Alamos collaborated on library design, and IBM implemented the Data and Communication Synchronization (DaCS) library and the Application Library Framework (ALF). The DaCS library passed an early test with flying colors when the Roadrunner node was re-designed to use PCIe rather than InfiniBand with no significant alteration of the software's interface design. DaCS and ALF are now part of IBM's Software Development Kit for Multicore Acceleration. The generalization of these libraries is significant, as programming multicore, many-core, and heterogeneous chips, as well as future hybrid architectures, will be a fundamental challenge over the next decade.

Assessing Roadrunner's Performance

A spectrum of representative algorithms were implemented on testbed hardware to determine whether Roadrunner would effectively support its intended workload. Our focus was on the molecular dynamics application SPaSM, the particle-in-cell code VPIC, the Milagro package for implicit Monte Carlo simulation of radiation transport, and Sweep3D, a kernel extracted from the PARTISN deterministic neutron transport package. By using the IBM PowerXCell 8i processors to accelerate portions of the codes, the team was able in each case to restructure the application to achieve performance improvements.

Performance modeling was essential to the assessment process, because only the PowerXCell 8i was actually available for testing. Models of each of the four codes were constructed, based on their hybrid implementations for the full Roadrunner system, including the Triblade and two

levels of InfiniBand interconnect. This effort provided accurate forecasts of performance of both the hybrid node and the Roadrunner system scaling to full size, demonstrating that Roadrunner would yield significant speedup relative to the unaccelerated system, as well as considerable performance advantage over other state-of-the-art systems.

Figure 2 shows what the full Roadrunner system has achieved on SPaSM, one of the targeted science applications: near-perfect scaling and sustained double-precision performance of 361 teraflop/s.

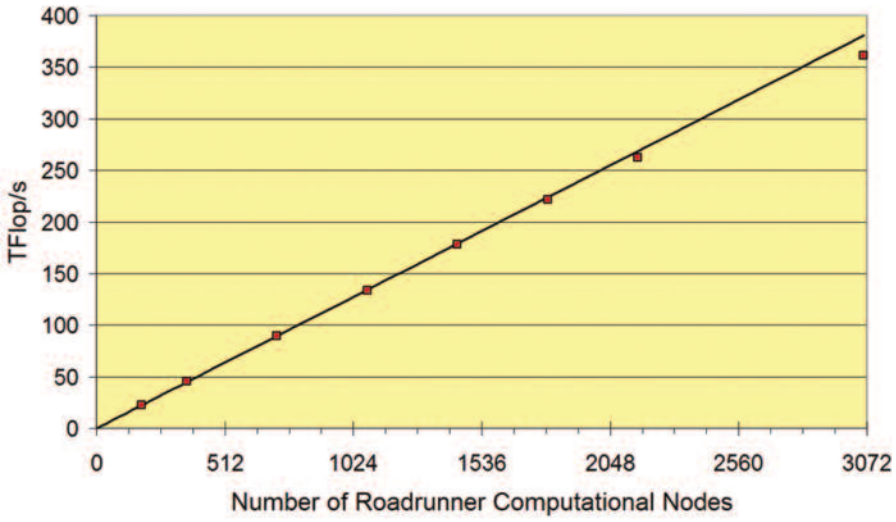

Figure 2. *Near-ideal scaling and sustained double-precision performance of 361 teraflop/s have already been achieved for the molecular dynamics code SPaSM on Roadrunner.*

Next Steps for Roadrunner

Roadrunner presents intriguing opportunities for computational science. As part of our plans to stabilize and integrate the system into our environment, we will run a suite of unclassified science applications on Roadrunner. These applications range from astrophysics and cosmology to viral phylogenetics and nanotechnology, and work is under way to prepare these applications for hybrid computing. Los Alamos also plans to acquire additional connected units in 2008 for its Institutional Computing program.

Along with the applications, a profound change is now taking place in computing: All programs, from the world shaking to word processing, must soon be fundamentally parallel at the chip level. Roadrunner provides a first-of-a-kind look at this future of high-performance computing. Building on Roadrunner, computer science efforts at Los Alamos will focus on the effective programming and utilization of a new generation of high-performance computers as we move along an as yet unknown path from petaflop/s to exaflop/s.

Both authors are at Los Alamos National Laboratory. John Turner is Group Leader for Computational Physics (CCS-2) and Andy White is Deputy Associate Director for Theory, Simulation and Computation.