The New Supercomputer Revolution

Pete Beckman

Senior Scientist, Argonne National Laboratory

Co-Director, Northwestern University / Argonne Laboratory Institute of Science and Engineering

ABSTRACT

Forty years ago only a handful of scientists were using supercomputers for advanced calculations. Today, large scientific computing clusters are everywhere -- from small colleges to large universities, global businesses, and research laboratories. Two main factors explain why supercomputers have become so widespread: a stable message-passing programming model and a largely open source software stack that effectively leverages community development. The stable programming model and the constantly improving software stack have allowed for sustained investments in complex simulation codes. Code teams have devised new algorithms and increased the performance and scalability of their applications, leading to realistic modeling and prediction of complex phenomena. As we look forward into the next 20 years, we see dramatic changes on the horizon.

Disruptive new computing technology has already begun to change the scientific computing landscape. Hybrid CPUs, manycore systems, and low-power system-on-a-chip designs are being used in today's most powerful HPC systems. New memory architectures, such as Micron's 3D Hybrid Memory Cube, will open the door for new system designs. However, these examples represent technology shifts that are here today. As we follow the trends embodied by the DOE FastForward program and look toward exascale, we see even more specialized designs supporting ultra-low-power cores, dynamic software-managed power limits, embedded network controllers and message routers, increased failure rates, massive node parallelism, and reduced per-core memory bandwidth and capacity. For the next few years, novel designs will flourish as new technologies are explored.

Furthermore, architectures seem to be diverging, dividing the once cohesive software stack. Power efficiency is now a key concern. All these changes mean that we will need to develop new programming models, new architectures, and a new software stack that can respond dynamically to system power.

It is not just computer architecture that is rapidly changing. Computational scientists are adapting their programming models and their computational methods. For exascale data sets, some initial analysis and visualization will likely be done in situ, while the data is still in memory. Uncertainty quantification has also launched new, highly parallel workflows. Furthermore, in order to tackle massive parallelism, new programming models will over-decompose the computational problem into lightweight tasks or threads that can increase concurrency, hide latency, and improve resilience. Decades-old software and computer architectures will be replaced with a new supercomputer revolution.