Sep 2012 ASC Newsletter Items – Sandia National Laboratories SAND2012-####

SAND2012-7153P

The Survey Says…

As the high-performance computing community looks toward developing exascale systems, power consumption is considered the most challenging obstacle. Researchers and practitioners from every area of system architecture are coming together to examine component and subsystem power use, as well as future trends. In this spirit of this examination, Sandia, LANL, and Clemson University researchers performed a survey of three supercomputers during normal operation: Cielo, hosted at LANL; Red Sky, hosted at Sandia; and Palmetto, a commodity cluster hosted at Clemson University. Each institution gathered rack-level power statistics, enabling the power budget to be partitioned between compute and storage resources.

The survey results offer a reassuring perspective on storage system efficiency. Of the three machines surveyed, none used more than six percent of their power on disk systems. Further, an aggregate survey of the entire LANL secure computing environment, which includes Cielo, Roadrunner, capacity clusters, and twenty petabytes of data storage, found that it used less than 2.5% of its power on all storage infrastructure, including disks, storage networking, and servers. Because 94% or more of the power per machine was dedicated to computation, efficiencies gained in compute-related subsystems will have the largest impact on future exascale systems.

The data collected also allowed the researchers to project how future systems will consume power, and how system design must change to remain sustainable. According to estimates, simply scaling the size of the storage system to meet bandwidth demands will not be possible. An exascale-class storage system in 2020 would include more than 100,000 disks and consume 66% of the 20 MW exascale power budget. However, incorporating burst buffers into an exascale-class system is estimated to reduce power use by 90% (to 6.6% of the power budget) while meeting performance requirements.

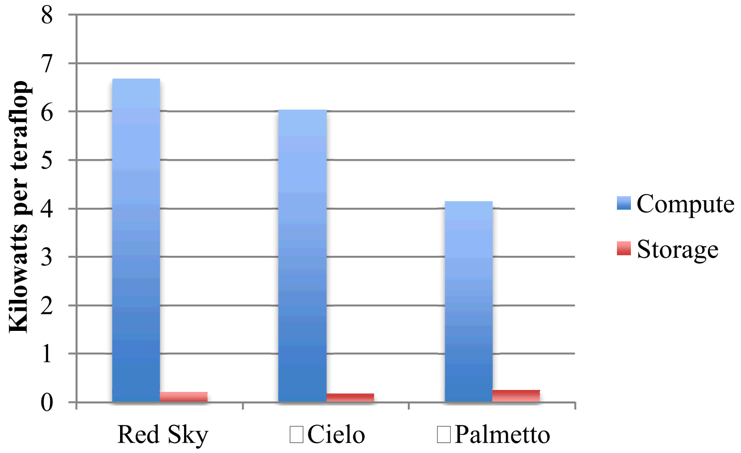

Figure 1: The amount of power dedicated to computation and storage, scaled to system size.

Reference Implementation Released for Updated Network Protocol Specification

Sandia recently released a reference implementation of the Portals 4.0 interconnect programming interface specification designed to enable scalable, high-performance network communication for massively parallel computing systems. Portals has evolved from a component of early lightweight compute node operating systems to provide scalable interconnect performance when deployed on production systems to an important vehicle for enabling interconnect research and software/hardware co-design. Previous versions of Portals ran on several successful vendor-supported systems, including the Intel ASCI Red machine and the Cray XT series.

Unlike other user-level network programming interfaces, Portals employs a building block approach that encapsulates the semantic requirements of a broad range of upper-level protocols needed to support high-performance computing applications and services. For example, Portals provides benefits like scalable buffering for MPI, but also enables functionality needed for system services like remote procedure calls and parallel file system network communication. This building block approach has also enabled hardware designers to focus on developing components that accelerate key functions in Portals, facilitating the application/architecture co-design process.

The most recent version of the Portals specification is the result of a close collaboration between Sandia and researchers at Intel working on advanced network interface hardware. This collaboration has led to two CRADAs between Sandia and Intel over the last two years. In addition, the ASC collaboration with CEA/DAM, the military applications division of the French Atomic Energy and Alternative Energies Commission, has led to a partnership between Sandia and CEA. CEA researchers added support for Portals 4.0 to their MultiProcessor Computing (MPC) software stack and plan to explore more advanced capabilities in future implementations.

The reference implementation of Portals 4.0 was developed in collaboration with System Fabric Works. It is layered on top of the OpenFabrics Verbs interface, allowing applications to be developed and tested using InfiniBand network hardware. Sandia gave invited talks about Portals 4.0 and this reference implementation at the OpenFabrics Alliance Annual Workshop at the end of March and at the IEEE Symposium on High Performance Interconnects at the end of August. Several research papers about Portals 4.0 have been published in the last year, and a paper entitled “A Low Impact Flow Control Implementation for Offload Communication Interfaces” that describes how Portals 4.0 supports scalable receiver-based resource exhaustion recovery for MPI will be presented at the upcoming European MPI Users’ Group Conference. The following graph shows simulation data from Sandia’s Structural Simulation Toolkit that illustrates the benefit of Portals 4.0 triggered operations in supporting a non-blocking Allreduce operation on several thousand nodes. Such non-blocking collective operations will soon be available in the MPI 3.0 Standard.

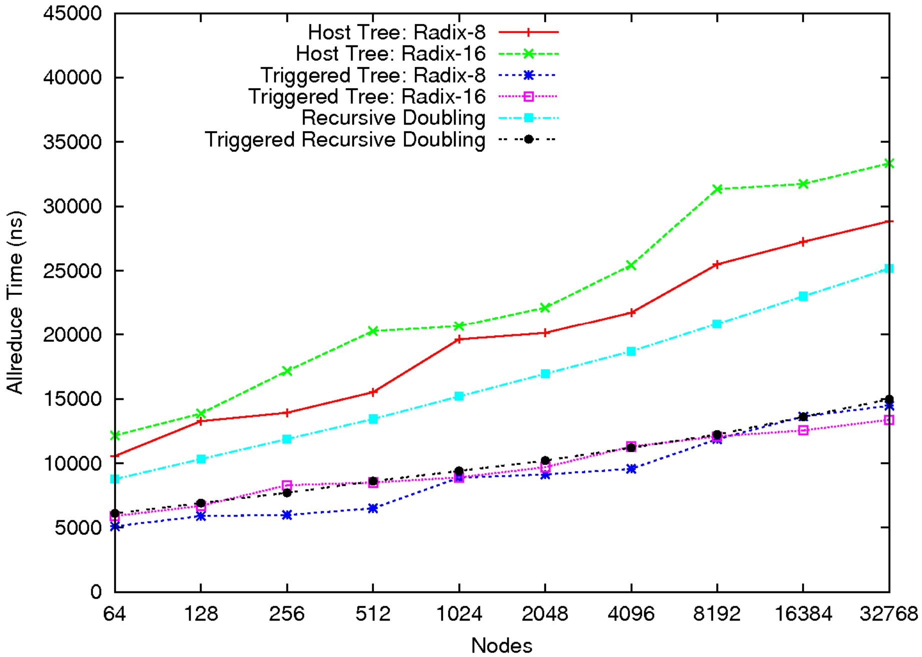

Figure 2: Structural Simulation Toolkit (SST) data show the benefit of Portals 4.0 triggered operations.

Performance-based Code Assessment for Low Mach Large Eddy Simulations (LES)
Sandia has completed a performance-based assessment of fluid dynamics simulation capabilities within the Sierra code base. The improved performance of an acoustically incompressible LES capability did not sacrifice the generality needed to address key needs of the B61 Life Extension Plan (LEP) and W88 ALT programs. Flexibility in software design is necessary for development of new capabilities that will support these programs, while performance is necessary to ensure that new and existing capabilities have a timely impact on qualification and design activities.

Conducted on Cielo, code performance and scaling simulations used up to 65,536 cores. Near optimal algorithmic scaling for linear system solves was demonstrated, and improvements of factors of 3 to 4 were achieved in CPU performance. Future work will address remaining scaling bottlenecks and performance of the matrix assembly.

The simulations used unstructured hexahedral mesh element counts ranging from 17.5 million to 1.12 billion elements. These mesh sizes and core counts are among the largest simulations within the unstructured low Mach community. In addition to software-related performance and scalability improvements, algorithmic advances were realized. Collectively, these activities and advances represent a path forward to exascale simulations in Sierra.

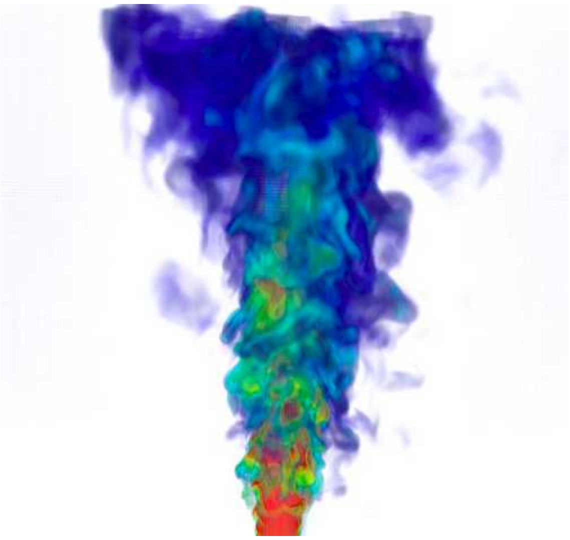

Figure 1: Volume rendering of a conserved scalar mixture fraction field in a turbulent open jet (Re = 6,600)

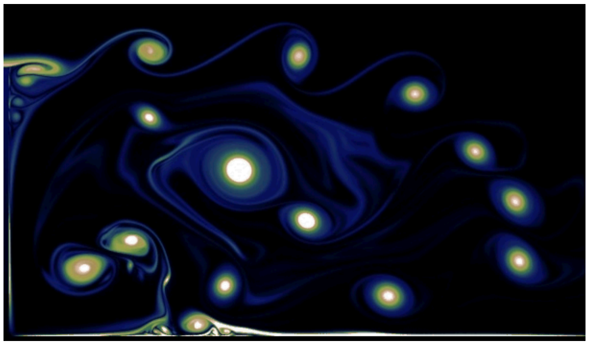

Figure 2: Vorticity contours for turbulent flow (Re = 45,000) past a backward facing step.

LES treatment of fluid turbulence is required for qualification efforts for aerodynamics, fire environments, and captive-carry loading. The unsteady nature of flows related to Abnormal Thermal and Normal Delivery environments requires LES for accurate environment prediction. Other less expensive techniques, such as Reynolds-Averaged Navier-Stokes (RANS), have proven to be inadequate. The characterization of fire environments requires sub-centimeter resolution to capture Rayleigh/Taylor instabilities leading to large-scale plume core collapse in pool fires of 5-10 meters. Many lessons learned for acoustically incompressible LES are also applicable for compressible LES, which is necessary for aerodynamic simulations. Resolution of vortex/fin interactions will require over 200 million element meshes for design calculations, and even more for qualification. Recent gains in performance and scalability will make these large LES simulations practical.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.