4 applications sustain 1 PF on Blue Waters

Four large-scale science applications (VPIC, PPM, QMCPACK and SPECFEM3DGLOBE) have run above 1 PF/s of sustained performance, and the Weather Research & Forecasting (WRF) run on Blue Waters is the largest WRF simulation ever documented. These applications are part of the NCSA Blue Waters Sustained Petascale Performance (SPP) suite and represent valid scientific workloads. The CUG 2012 paper “Analyses and Modeling of Applications Used to Demonstrate Sustained Petascale Performance on Blue Waters” details many of the computation aspects of these SPP applications.

VPIC

VPIC integrates the relativistic Maxwell-Boltzmann system in a linear background medium for multiple particle species, in time with an explicit-implicit mixture of velocity Verlet, leapfrog, Boris rotation and exponential differencing based on a reversible phase-space volume conserving second order Trotter factorization. The Petascale Computing Resource Allocation (PRAC) team led by Homayoun Karimabadi is using VPIC in for kinetic simulations of magnetic reconnection of high temperature plasmas (H+ and e-). Magnetic reconnection is an energy conversion process that occurs within high-temperature plasmas and often produces an explosive release of energy as magnetic fields are reconfigured and destroyed. Simulated here are force-free current sheets, thought to be relevant to the solar atmosphere and many astrophysical problems. NCSA and Cray worked to improve compiler optimization of loops not already using optimized vector compiler intrinsic functions, optimizations to eliminate extra data copies, added FMA4 compiler intrinsic functions to improve computer performance and used Cray I/O buffering functionality. The three-dimensional, nearest-neighbor communication maps well to the Gemini network. The VPIC science problem had a 3072x3072x2464 cell domain with 7.44103E+12 particles. The science problem was run on 22,528 nodes with 180,224 MPI ranks with 4 OMP threads/rank, and achieved 1.25 PFLOPS sustained over 2.5 hrs.

PPM

The PRAC science team led by Paul Woodward is using their PPM code to address problems in Inertial Confinement Fusion (ICF). PPM was developed by the team to address stellar processes. The ICF problem, like the giant star problem the team solved on the Blue Waters Early Science System (ESS), involves turbulent mixing due to instabilities at multi-fluid interfaces, and in both problems the details of this mixing affect the combustion that results. The ICF problem exercises all the features of the PPM codes, including strong shocks and very elaborate treatments of unstable multi-fluid interfaces, while at the same time performing an important scientific problem that has a highly transient character. PPM is designed to be a high-performing application. The Woodward PRAC team has taken significant steps to take advantage of many processor architecture features, including aggressive pipelining, compiler generated SIMD vectorization, minimizing working set size, and shared L3 cache optimization. Communication is primarily halo exchanges on a 3D Cartesian mesh which are overlapped by computations, as is the I/O. The Cray Fortran compiler generated highly vectorized code and provided high performance math functions. Cray implemented a rank reordering scheme to interleave the I/O server tasks in the MPI rank list. Lustre striping was set as appropriate for the larger output files. The test case uses a 7040^3 zone mesh. It was run across 21,417 XE nodes with 85,668 MPI ranks and 8 threads per rank. 121 nodes were used as I/O servers. For just under an hour of runtime a sustained performance of 1.23 PF/s was achieved and 14 TB of data were written.

QMCPACK

The Shiwei Zhang PRAC team is using the QMCPACK application on Blue Waters. QMCPACK uses quantum Monte Carlo methods to solve the many-body Schrödinger equation by stochastically sampling the configuration domain. First, a Variational Monte Carlo (VMC) algorithm is used to quickly find a “ballpark” estimate of the solution, and then a Diffusion Monte Carlo (DMC) algorithm is used to refine the estimate from the VMC phase. A number of VMC walkers (each a complete state representation) that randomly move
through the energy domain each time step are sampled to create walkers for the DMC phase. The output of the DMC phase is the lowest energy state within a statistical uncertainty, which can be reduced by taking more samples (i.e., using more DMC walkers). Recent predictions using density functional theory (DFT) suggest that hydrogen may be in a superconducting state that could extend from a zero temperature solid state phase to a room temperature metallic liquid. For the test problem the input is for a hydrogen structure under high pressure with VMC only. Cray, NCSA and the QMCPACK team worked on several improvements to an already well-performing code. To fully benefit from the AMD Interlagos processor, FMA4 compiler intrinsics were added to key routines and the AMD version of the math library (libm) was used to replace the GNU release. QMCPACK was run with a 432-atom high-pressure hydrogen problem on 22,500 XE nodes with 4 MPI ranks per node and 8 OpenMP threads per rank. The run achieved sustained performance of 1.037 PF/s for < 1 hour of execution.

**SPECFEM3D_GLOBE**

The SPECFEM3D series of codes is a simulation of seismic wave propagation in the earth, modeling the globe as a finite element mesh. While some of the codes model a region of the earth in isolation, SPECFEM3D_GLOBE models propagation of waves from earthquakes through the entire earth and is designed to scale to systems with hundreds of thousands of processor threads. The code takes seismographs and inverts them to synthesize what the displacement was at the epicenter of the earthquake. This is used, among other things, to model the displacement of a close-by earthquake to understand how buildings need to be constructed to withstand earthquakes. The goal for the model problem is to run at a high enough resolution that the shortest period of accurately simulated seismic waves is below the "two second barrier," or ideally down to a shortest period of 1 second. The runtime process involves two phases: the mesh generation and the solver. The mesh for a SPECFEM3D_GLOBE simulation is based upon a mapping from the cube to the sphere called the cubed sphere that breaks the globe into six chunks, each of which is further subdivided. The spectral-element method solver is a continuous Galerkin technique with optimized efficiency because of its tensorized basis functions and has very good accuracy and convergence properties. NCSA improved the mesh performance by using system memory to store the mesh files for each MPI rank and MPI task reordering reduced communication overhead for both the mesh generator and the solver. Optimizations applied to the solver consisted of a small number of compiler directives to improve the generally excellent vectorization done by the Fortran compiler. These compiler directives inhibited overall aggressive loop unrolling and reordering in a few places. Higher-level optimizations by subroutine inlining were also helpful. The run was done on 21,675 XE nodes with 693,600 MPI ranks and sustained over 1 PF/s.

**WRF**

The Weather Research and Forecasting (WRF) Model is a next-generation mesoscale numerical weather prediction system designed to serve both operational forecasting and atmospheric research needs. It features multiple dynamical cores, a three-dimensional variational (3DVAR) data assimilation system, and a software architecture allowing for computational parallelism and system extensibility. WRF is suitable for a broad spectrum of applications across scales ranging from meters to thousands of kilometers and is the major mesoscale weather forecasting model used by thousands of registered users across the globe. The model problem computes an 18-hour forecast for a region centered on Alaska with a 1km horizontal resolution (6075 x 6075 x 28 =~1billion points) and is believed to be the largest full forecast WRF domain ever run, reaching a new level of sustained performance. Cray and NCSA worked to improve WRF performance by implementing a MPI rank ordering that maximizes on-node halo exchange, improved compiler vectorization of vertical columns, implemented an alternative parallel I/O strategy to exploit Lustre filesystem performance. A total of 8,192 XE nodes were used in the run with 262,144 MPI tasks with 2 OpenMP threads per task. A performance of 166 TFLOP/s sustained was achieved, completing the forecast in under 30 minutes. The previous best was ~50 TFLOP/s on a Cray XT5.