TeraGrid: Past, Present, and Future

Phil Andrews, Director, National Institute for Computational Sciences, Oak Ridge, Tennessee
TeraGrid Abstract:

- The TeraGrid is the amalgamation of the major U.S. National Science Foundation HPC centers, currently with approximately a dozen Resource Provider sites and a total computational capacity of over 1 Petaflop. This is scheduled to increase to about 2 PF in 2009 and 3 PF in 2010. The TeraGrid itself will experience a major transformation in 2010, with a new organizational model expected to replace the current approach. This reorganization is presently in the proposal stage. We will discuss the history of the TeraGrid, the past changes in the funding and strategic plans, and any available information on future directions, including technology, standards, and international collaboration.
TeraGrid: Beginnings

- 2001: NSF sent “Dear Colleagues” letter to SDSC, NCSA to provide coordinated systems
- For complex reasons, CACR (CalTech) and Argonne National Laboratory (ANL) added
- Significant funding earmarked for connecting network: initially intended as a mesh, then a more flexible “Grid”
- Named: Distributed TeraScale Facility
Technical Details (2001, DTF)

- Intended to be highly homogenous (IA64)
- Every node GbE connected
- 40 Gb/s backbone, 30 Gb/s to each site
- Standardized on Globus middleware
- Initial award: $53M
- Initial capability: 13.6 TF
- High point of standardization for TeraGrid
- Intended use: distributed MPI system!
TeraGrid: 13.6 TF, 6.8 TB memory, 79 TB internal disk, 576 network disk
Extended TeraScale Facility (ETF)

- Added Pittsburgh Supercomputing Center, no longer homogenous
- Site specialization (compute, data, visualization, etc.)
- Distributed independent sub-tasks, not a single task
- Began to relax standardization
The TeraGrid (including ETF)

**Caltech**: Data collection analysis
- 0.4 TF Intel IA-64
- 80 TB Online Disk
- IA-32 Datawulf
- Sun Storage Server

**ANL**: Visualization
- 1.25 TF Intel IA-64
- 20 TB Online Disk
- 96 Visualization nodes

**SDSC**: Data-Intensive
- 4 TF Intel IA-64
- 1.1 TF IBM Power4
- 500 TB Online Disk
- Sun Disk Server
- IBM Database Server

**NCSA**: Compute-Intensive
- 10 TF Intel IA-64 (128 Large-Memory nodes)
- 230 TB Online Disk

**PSC**: Heterogeneity
- 6.0 TF Compaq EV68
- 0.3 TF EV7
- 221 TB Online Disk
- Disk Cache Server

**LEGEND**
- Clusters
- Storage Server
- Online Disk Storage
- Visualization Cluster
- Shared Memory

**Clusters**
- LA Hub
- Chicago Hub

**Network**
- 40Gb/sec Extensible Backplane Network
TeraGrid: more sites, less network

- Added Purdue, Indiana, TACC
- Introduced Coordinated TeraGrid Software and Services (CTSS)
- Network “minimum” reduced to 10 Gb/s
- Began work on parallel global file systems
TeraGrid Network
Tracks 1 and 2, 2005: Big Iron gets Bigger

- 4 yearly Track 2 RFPs announced: $30M for capital, operations funded separately, production in 2007, 8, 9, 10
- Single Track 1 RFP, $200M for “sustained PetaScale system”, production in 2011
- TeraGrid connectivity and consistency important
Track2A award

- Recommended to TACC (U.T. Austin) in 2006, awarded in 2007
- Sun System, AMD processors, CBBW network interconnect, custom (Magnum) switch (2)
- Over 60,000 cores, 500+ TF, ~125TB, ~3.4MW
- Began production Q1, 2008
Track2A, Texas Advanced Computing Center, Austin, Texas
Track2B: Kraken Cray XT5 at NICS
Kraken XT5 Overview

- Processor: Quad-core 2.3 GHz AMD Opteron
- Number of Nodes/Cores: 8256 nodes / 66,048 cores
- Peak Performance: 607.64 TF
- Interconnect: Seastar 2 (3D Torus)
- Memory: 100 TB
- Disk: 3.3 PB 30 GB/s I/O
- Upgrade to ~100,000 cores, ~1PF later this year
- Providing over 400M CPU hours/year
Current TeraGrid

- Very heterogeneous, both size, architecture
- CTSS now in “kits”; sites can opt in or out
- Added NCAR, LONI; lost CACR
- Split services into “common”, provided by TeraGrid, “specific” provided by RP’s
- Common allocation process, accounting
- Almost 1.5 PF aggregate
Future TeraGrid

- Track2C: ~1PF SGI system at PSC (2009)
- Track2D: data intensive, experimental, throughput, and test-bed systems (2010)
- Track1: ~10PF IBM system at NCSA (2011)
- Current TeraGrid expires in 2011
- Follow-on: “eXtreme Digital” (XD) proposals; two competing partnerships
Early XT5 Users
Mathieu Luisier – Purdue University

- Accelerating Nanoscale Transistor Innovation through Petascale Simulation
- ~3M hours allocated on Kraken (with Gerhard Klimeck)
- Objectives:
  - Test Nanoelectronics Simulator OMEN – full-band, atomistic quantum transport simulator for 1-D, 2-D, 3-D nanoelectronics devices
  - Full-machine scaling runs
  - Achieve new science (model larger devices)
- Results:
  - Almost perfect scaling up to 64K cores
  - Tuning of structure dimensions – 210 Tflops
  - Realistic circular nanowire transistor structures with diameters up to 10nm
- Impacts:
  - Impressively short startup time
  - High reliability on 64K cores: ~15 jobs started, no failure
  - Straightforward migration from XT4 to XT5
  - Strong support from NICS through website and help@teragrid.org

- 173 TFLOPs for Matrices of Size N=55,440
- 208 TFLOPs for N=129,360
Early XT5 Users
Diego Donzis – University of Maryland

- Direct Numerical Simulations (DNS): compute turbulent fluctuations at all scales according to exact, unsteady, 3D conservation equations
- ~15M hours allocated on Kraken (with PK Yeung)
- Computationally very demanding: simple geometry, efficient codes
- Computational power today (Kraken XT4, XT5): simulations at conditions not possible earlier
  - 4096^3 grid (current record by Kaneda 03, but short simulations and no passive scalars)
  - Understanding of mixing, resolve long standing issues
- Strong scaling from 32K to 64K cores:
  - 80% for 4096^3; 85% for 8192^3
- Alltoalls: bottleneck
  - Better performance with 4 cores/node (solid blue symbol)
  - Strong scaling ~ 95% 32k -> 64k
- Science: production 4096^3 simulation reached stationary state; ready to include Lagrangian particles
  - study dispersion at record Reynolds number

\[
\text{Time step (secs)} \quad \frac{1}{M} \\
\text{Circles: 4096^3} \\
\text{Squares: 8192^3}
\]

\[
\text{Number of cores} \\
\text{3D FFTs: } N^3 \log_2(N) \text{ for } N^3 \text{ grid points} \\
\text{2D domain decomposition: best proc grid} \\
\text{Perfect scaling: } N^3 \log_2(N) / M \\
(M: \# \text{ of cores})
\]
Early XT5 Users
Yifeng Cui
Southern California Earthquake Center

- Largest and most detailed earthquake simulation on San Andreas fault to improve public awareness and readiness for the ‘Big One’
- ~8M hours allocated on Kraken (with Tom Jordan)
- 600 x 300 x 80 km domain, 100m resolution, 14.4 billion grids, upper frequency limit to 1-Hz, 3 minutes, 50k time steps, min surface velocity 500m/s, dynamic source, velocity properties SCEC CVM4.0, 1 terabyte inputs, 5 terabytes output
- Structured 3D with $4^{th}$ order staggered-grid finite differences for velocity and stress, Fortran 90, message passing done with MPI using domain decomposition, I/O with MPI-IO
- Simulation completed within 5 hours on Kraken XT5 during Friendly User week, a milestone in earthquake simulation

PetaShake: a petascale cyberfacility for physics-based seismic hazard analysis funded by NSF
Early XT5 Users
Robert Harkness – UC San Diego

- ~6M hours allocated on Kraken (with Mike Norman)
- ENZO Cosmology Simulations
  - Lyman Alpha Forest
  - 100 Mpc/h box, 2048^3 mesh, 8 billion dark matter particles
    H, He ionization, heating & cooling. 15 Baryon fields,
    8 particle properties (position, velocity, mass and index).
  - Re-started* from Kraken-XT4 simulation at redshift Z=1.3.
  - Hybrid MPI/OpenMP code, 2048 MPI tasks, 4 and 8
    OpenMP threads.
    - Input dataset ~ 1 TByte. Output datasets ~1 TByte per dump.
      Over 100 TB written by Feb 10th - with NO ERRORS.
    - Results archived in NICS HPSS - rates above 300 MB/sec.
- Strong Scaling & Hybrid Scaling Test
  - 1024^3 mesh, 1 billion dark matter particles.
  - 128, 256, 512 & 1024 MPI tasks and 1, 2, 4 or 8 OpenMP
    threads.
- AMR code testing
  - 1024^3 L10 restart 512 MPI tasks, 4 threads, 443,802
    subgrids!
Intermittency in Interstellar Turbulence
Alexei Kritsuk, UC, San Diego

• Developed and tested a new numerical method for ideal MHD.
  – Compared its performance with that of ZEUS, FLASH, RAMSES, and other ideal MHD codes
  – Grid resolutions for the decaying turbulence problem: 256^3 and 512^3

• Carried out a parameter survey of supersonic MHD turbulence at resolution of 512^3 with different degrees of magnetization

• Largest runs used 1024 cores for simulations on 512^3 grids, up to 75,000 integration time steps per run

Using computer simulations to investigate conformational dynamics of HIV-1 protease
Carlos Simmerling, SUNY Stony Brook

- HIV-1 protease
  - This enzyme is responsible for an essential step in the viral maturation and lifecycle of HIV-1. Inhibition of this enzyme results in noninfectious immature virus particles.
  - However, this enzyme mutates to elude drugs which bind to its active site and inhibit its action.
  - How do mutations of HIV-1 protease affect inhibitor binding?

- Conformational Dynamics of HIV-1 protease
  - Mutations occur both inside and outside of the active site. It is not well understood how mutations outside the active site affect inhibitor binding.
  - The mutations exterior to the active site are thought to affect the mobility of two gate-keeping molecular flaps near the active site.
  - Both experimental and computational methods were utilized and compared to examine the affect of mutations on HIV-1 protease.