Trends in High Performance Technical Computing and Their Impact on Scientific Simulation

March 14, 2002
Workshop on Advanced Computational Methods for Solving the Nuclear Many-Body Problem

Dr. Mark K. Seager
Asst. Dept Head for Tera-scale Systems
LLNL Integrated Computing and Communications Department

Simulation of Flow in Gas Turbine Engine
Stanford: Level 1 Alliance
Three remarkable trends have changed science as we know it

- Scientific simulation is now on par with Theory and Experiment as fundamental elements of the scientific method

Requirements for predictive scientific simulations with large economic benefit are changing the state of the art

- ASCI Program as example HPTC market driver
- Japanese ESP second example

Disruptive technologies in commodity clusters and open source software put HPTC on a different cost performance curve

Very large clusters require

- New facilities
  - Single site, low latency coms
- or
- Grid computing paradigm
  - Multiple sites, high latency coms

Improved determination of the size, age and composition of the universe is a goal of the Djehuty project, using the world’s only operating code that can model a complete star in 3D. Visualization produced by Chromium distributed visualization.
Challenge is to maintain confidence in a smaller stockpile without nuclear testing and modernization

Past

- Underground nuclear testing
- Large numbers of warheads
- Many different system designs
- Frequent modernization
- Numerical simulations
- Simulation of the high-explosive pre-nuclear phase

Present and Future

- No nuclear testing
- Modest numbers of warheads
- Few system designs
- No modernization
- AGEX
- Simulation
- Simulation of the high-explosive pre-nuclear phase
Primary & Secondary Assessments Drive Required Computing Capabilities

- **Platforms Mat’ls & Physics**
  - FY00
    - 1.5 TB memory, 76 TB disk
  - FY01
    - 6 TB memory, 150 TB disk
  - FY02
    - 30 TB memory, 600 TB disk
  - FY03
    - 12 TB memory, 600 TB disk
  - FY04
    - 12 TB memory, 600 TB disk
  - FY05
    - 60 TB memory, 1500 TB disk

- **SSP**
  - FY00
    - 3-D burn simulation
      - Simple engineering features
      - O(10M) zones, O(2w) runtime, O(200GB) memory
  - FY01
    - Full-scale advanced HE models
    - Transport capability
    - Full-scale 2-D transport
    - 3-D burn simulation
      - O(100M) zones, O(8TB) memory, O(2w) runtime
      - O(1B) zone safety simulation
  - FY02
    - Full-scale advanced material models
    - Full-scale 2-D transport
    - 3-D burn simulation
      - O(300M) zones, O(10TB) memory, O(2w) runtime
      - O(1B) zone safety simulation
  - FY03
    - Full-scale advanced material models
    - Full-scale 2-D transport
    - 3-D burn simulation
      - O(500M) zones, O(10TB) memory, O(2w) runtime
      - O(1B) zone safety simulation
  - FY04
    - Full-scale advanced material models
    - Full-scale 2-D transport
    - 3-D burn simulation
      - O(1B) zones, O(30TB) memory, O(2w) runtime
      - O(1B) zone safety simulation
  - FY05
    - Full-scale advanced material models
    - Full-scale 2-D transport
    - 3-D burn simulation
      - O(1B) zones, O(30TB) memory, O(2w) runtime
      - O(1B) zone safety simulation

- **Validation**
  - Continued Use
  - Continued Use
  - Continued Use

SSP and ASCI have recently revised these milepost definitions for the out-years.
ASCI timescales go from Femto-sec to years with grand challenge computations at every level.

<table>
<thead>
<tr>
<th>Time Unit</th>
<th>Level Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Femtoseconds</td>
<td>Level 1: Atomic physics opacity</td>
<td>-</td>
</tr>
<tr>
<td>Nanoseconds</td>
<td>Level 2: Molecular/atomic level simulations</td>
<td>Fires, Aging, Explosions</td>
</tr>
<tr>
<td>Milliseconds</td>
<td>Level 3: Turbulence and mix simulations</td>
<td></td>
</tr>
<tr>
<td>Minutes</td>
<td>Level 4: Continuum models</td>
<td></td>
</tr>
<tr>
<td>Years</td>
<td>Level 5: System validation</td>
<td>Aging, Fires</td>
</tr>
</tbody>
</table>

Distance (m):
- $10^{-16}$
- $10^{-8}$
- $10^{-6}$
- $10^{-2}$
- 1

About 25 orders of magnitude…

Stockpile Stewardship Program needs exceed Peta-scale, even Exo-scale…

March 14, 2002
Materials science has four scales of grand challenges for predictive simulations of materials properties

Successful implementation requires closely coupled experiments and models

(Yield surface)

Materials properties and performance under extreme conditions

Single crystal constitutive relation

Rules for dislocation motion and multiplication

(Adapted from a presentation by de la Rubia)
ASCI Computing Systems Roadmap

- working with industry to reach 100 TeraOP/s performance level in 2004/5 -

ASCI PathForward

BlueGene/L (2004/5) ➔ ASCI Advanced Architectures

RFP Responses due 4/29

IN BAFO stage ➔ Sandia Red Storm ~20 TeraOP

Q 33 TeraOP / 33 TB

White 12.3 TeraOP / 8 TB

Blue 3+ TeraOP / 1.5-2.6 TB

Red 1.8 ➔ 3.0 TeraOP / 0.5 TB

Purple 60-100 TeraOP / 30-50TB

Plan

Develop

Use

Time (Calendar Year)

'97 '98 '99 '00 '01 '02 '03 '04 '05 '06

March 14, 2002
Computational landscape through 2004 as seen by Japanese Earth Simulator Project

Earth Simulator Project in the World

The diagram illustrates the computational landscape through 2004, highlighting the performance capabilities of various supercomputers, including the ASCI Purple, which is a significant project managed by the Japanese Earth Simulator.
High performance platforms are based on large clusters of commodity SMP’s

- Commodity portion differs
  - CMOS manufacturing process and ASIC design techniques
  - Microprocessor
  - Chip Sets
  - Memory Components

- Interconnect differs
  - Commodity interconnects (Ethernet)
  - Semi-custom interconnects (Myrinet, QsNet)
  - Custom interconnects (IBM Colony, SGI ccNUMA, Sun)

- Size of SMP differs
  - 2-128-way defines the 99.999% of market
  - Sweetspot of marketplace is actually 2-4 → 8 CPUs
By the end of White, the Livermore model (clustered SMPs with incrementally improving functionality) will have provided about half the programming model longevity (19 years) that the Vector era did.
Livermore model provides consistent programming model across scalable platforms generations and implementations.

Idea: Provide a consistent programming model for multiple platform generations and across multiple vendors!

Idea: Incrementally increase functionality over time!
The Architecture of the Earth Simulator is a cluster of (vector) SMP's:

- Peak performance/AP: 8Gflops
- Peak performance/PN: 64Gflops
- Shared memory/PN: 16GB
- Total peak performance: 40Tflops
- Total number of PNs: 640
- Total main memory: 10TB
- Total system cost: $350M USD
- Total number of APs: 5120

Commodity component is CMOS technology ASIC design and manufacturing.
With over 22.5 teraOP/s, Livermore Computing is the world's (second) largest scientific simulation site.

Secure computing facility: 19.0 TF peak, 15,506 CPUs, 12.4 TB memory

Open computing facility: 3.6 TF peak, 3,116 CPUs, 2.2 TB memory

Unprecedented Capability Resources

Capacity resources for “routine” calculations
ASCI White was used this year to meet demanding ASCI Mileposts at all three labs

ASCI White delivery began in June and was complete in September, 2000

Serial number 1 system at 512 nodes requires intensive 1 year integration process

IBM systems in customer service – we made a major effort to serve University Alliances
- White – NH2 (12.3 TF)
- Ice – NH2 (.6 TF)
- SKY – Silver (3.9 TF)
- Frost – NH2 (1.6 TF)
- Blue – Silver (.74 TF)

White has delivered to the ASCI Tri-Laboratory Program
- Demonstrated 7.99 teraFLOP/s on Linpack and 3.9 on sPPM
- Completed FY00 and FY01 Applications, DisCom² and PSE milestones

12 TF, 8 TB memory, 100+ TB disk

Siting White in B451 Required Major Facilities Upgrade
ASCI White System Architecture

ASCI White System

- 4 Login/Network nodes w/16 GB SDRAM
- 32 Visualization nodes w/16 GB SDRAM
- 460 PBATCH nodes w/16GB SDRAM
- Colony Double-Single interconnect
- 12.8 GB/s delivered global I/O performance
- 5.12 GB/s delivered local I/O performance
- 24 Gb Ethernet External Network

Programming/Usage Model

- Mohonk PSSP Software release
- 4,095 MPI/US, 2,048 MPI/IP tasks
- Application launch over ~492 NH-2 nodes
- 16-way MuSPPA, Shared Memory, 32b and 64b MPI
- Likely usage is 4 MPI tasks/node with 4 threads/MPI task
- Single STDIO interface

IBM ASCI White has already achieved 5 key ASCI Milestones
NightHawk-2 Node Architecture

IBM RS6000 SP NH-2 Node
- 24 GF /Node, many with 16 GB
  - 375MHz Power3-2
  - 8MB L2 cache
  - 512 Power2 ID = 5.75 NH2
- 16-Way UMA SMP
- 137 clocks to memory
  - Four memory banks
  - 16 GB/s memory bandwidth
- New Colony Switch/Adapters
  - Delivered with Colony SS = 1.0 GB/s peak.
  - Future SW Colony DS = 2.0 GB/s
  - GPFS performance improved by Colony DS (KLAPI) comms
- Each node has local disk with independent OS (AIX 4.3.3)

Memory B/F = 16.0/24.0 = 0.666
Interconnect B/F = 2.0/24.0 = 0.084 Colony DS
Big systems are only one element in a highly integrated simulation environment

- Visualization of results is as important as simulation
- High speed networking required to move data
- Archival of simulation and experimental or observational data is essential

Operationally “full-system” calculations very similar to high-energy accelerators usage model

- Large team does upfront development and planning
- Access slot to resource politically scheduled
- Large team does run over weeks to months
- Post run data analysis takes months
Blue-Pacific & White are part of an integrated and balanced simulation environment.

FY01 Platform Plan
ASCI White at 12.3TF
ASCI Blue-Pacific at 3.9TF

FY01 I/O Plan
I/O speeds determined by ASCI Apps Milestone requirements in FY01

FY01 DVC Plan
Data Visualization Corridor
2 TB vis Disk cache

We have delivered over 160 MB/s parallel FTP to the archive and stored over 7.2 TB in one day!
ASCI's Data Visualization Theater allows scientists to “see and understand” the results of their calculations.
Distributed computing essential

- Grid computing utilizes multiple sites
- Remote usage of tera-scale computers requires different data transport models
  - Data analysis (visualization) must be distributed
- Scientific collaboration usually implies physically distributed teams
  - Large scale simulations can’t be done by individuals any more.
DisCom² networking delivers tera-scale data sets between the tri-laboratory campuses

Data migration, manipulation and storage

Key user requirements

- Effective movement of large data sets
- Secure (for classified) data transmission

OC-3c = 155 Mb/s
OC-12c = 622 Mb/s
OC-48c = 2.5 Gb/s
Target Purple hardware architecture includes scalable I/O, login, visualization and compute resources

Finally, “the network is the computer”

Purple System
- 60-100 teraOP/s, 30-60 TB memory, 1.2-2.0 PB disk
- Parallel batch/interactive, visualization nodes
- 2-8 Login/network nodes
- Clustered I/O services for global I/O
- Login/network nodes for login/NFS
- Infiniband™ for parallel FTP
- All external networking is 1-10Gb/s Ethernet

Programming/Usage Model
- Application launch over all compute nodes
- 1 MPI task/CPU and Shared Memory, full 64b support
- Scalable MPI (MPI_allreduce, buffer space)
- Expect multiple MPI tasks/node with 4-16 OpenMP/MPI task
- Single STDIO interface
- Parallel I/O to single file, multiple serial I/O (1 file/MPI task)

SAN Model
- Compute, Visualization and HPSS scalable clusters
- Multiple clusters for multiple purposes
- Linux and other UNIX OS’s
- Single SAN wide global file system
- High-end I/O subsystem drives implementation
- Have Purple I/O subsystem outlive compute
- Studio quality and office quality tera-scale displays

Programming/Usage Model
- ~100GB/s delivered on Purple
- Auto (parallel) migration to HPSS tape
- Uniform name space
- Need to know security
- Quality of service guarantees
- Transparent use of Visualization and other clusters
- /home (1:1), /var/tmp (n:1), and parallel file system (n:m)
Artist's Representation of the Japanese Earth Simulator

- **Earth Simulator**
- **Processor Node Cabinets**
- **Interconnection Network Cabinets**
- **Disks**
- **Cartridge Library System**
- **Air Conditioning System**
- **Power Supply System**
- **Double Floor for PN-IN Cables**
- **Seismic Isolation System**

**Measurements:**
- 65m (71yd)
- 50m (55yd)

**Building Information:**
- Building is 35,145 ft² and 7.5 MW for computer and 12.5 MW total.
The Purple system requires a new generation of physical plant: Terascale Simulation Facility (TSF)

$92M TSF line item in place …..
Groundbreaking in April 4, 2002
Two computer rooms for multiple system integration
22 MW power
48,000 square feet of computer room floor space
First computer room complete in June 2004
Ready for Purple and BlueGene/L
Complex complete 2006
- Open Source software development model is a disruptive technology

- Commodity (IA-32 and possibly IA-64) Microprocessors are a disruptive technology trend

- Commodity (or semi-commodity) networking is a disruptive technology trend
Why Open Source?

*HPTC customers require*
- Source for quick bug fix ability
- Hedge against “change in support” status
- Each site has unique and demanding requirements that vendors can’t make money supporting

*Joint development model has been proven effective with HPSS*
- IP model needs to change to Open Source
- Developers and early adopters are the target market
- Multiple sites find more bugs, provide incredible marketing potential
- LEVERAGE GOVERNMENT DEVELOPMENT EFFORTS!

*Next wave of software development*
- Disruptive SW technologies has important backers
- Community development will ultimately improve ROI = REVENUE/INVESTMENT
Disruptive Technology Trends

- Disruptive technology tends to start out small with much faster growth rate
  - Innovators dilemma is when to switch...
- Next trend in software development will cause tremendous shift in business plans
  - Open source coupled with the commoditization of hardware
- This transition will require rethinking of what is “value add” and what is “intellectual property”
  - Business models will be based on service
    - Competitive differentiation in
      - Quality of integration
      - End-to-end costs
      - Time to market
  - HPTC market different from general server and cluster market
    - Joint development
    - Expertise
    - Early adopters

For technology companies, the time to switch to open source is NOW!
Open Source Business Model

- Don’t sell software
- Don’t sell hardware
- Sell end to end solutions
  - Price based on customers perceived value (demand) and supply
    - Not strongly related to actual hardware, software and support costs
  - Requires different marketing approach
  - Requires tight integration of service organization (the folks who build web sites, provide e-business solutions) with HW/SW development and Marketing

If the French can make money selling water, IBM should be able to figure out how to make money selling free software - Irving Ledowskiburger
The GFMD code is the first to return a timing/scaling study for PCR (P4 Linux Cluster) – results are exciting!

Problem size per processor fixed at 400 atoms – near perfect scaling to 250 processors

**Diagram:***

- **PCR** – 128-2P nodes (870 GF/s)
  - Linux NetworX
  - 1.7-GHz Intel Pentium-4
- **TC2K** – 128-4P nodes (681 GF/s)
  - Compaq AlphaServer ES40
  - 667-MHz Alpha EV67
- **Frost** – 68-16P nodes (1632 GF/s)
  - IBM SP3 Night-Hawk 2
  - 375-MHz Power3+
- **Blue** – 336-4P nodes (892 GF/s)
  - IBM SP2 Silver
  - 332-MHz PowerPC 604e

**Better:***
Recent MDCASK timing/scaling study for PCR also confirm effectiveness of Linux-Cluster strategy

Problem size fixed at 2,048,000 atoms – problem distributed over variable number of processors

<table>
<thead>
<tr>
<th>System</th>
<th># Processors</th>
<th>Performance (GF/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PCR</strong></td>
<td>128-2P nodes</td>
<td>870</td>
</tr>
<tr>
<td></td>
<td>Linux NetwoRX</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.7-GHz Intel Pentium-4</td>
<td></td>
</tr>
<tr>
<td><strong>TC2K</strong></td>
<td>128-4P nodes</td>
<td>681</td>
</tr>
<tr>
<td></td>
<td>Compaq AlphaServer ES40</td>
<td></td>
</tr>
<tr>
<td></td>
<td>667-MHz Alpha EV67</td>
<td></td>
</tr>
<tr>
<td><strong>Frost</strong></td>
<td>68-16P nodes</td>
<td>1632</td>
</tr>
<tr>
<td></td>
<td>IBM SP3 Night-Hawk 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>375-MHz Power3+</td>
<td></td>
</tr>
<tr>
<td><strong>Blue</strong></td>
<td>336-4P nodes</td>
<td>892</td>
</tr>
<tr>
<td></td>
<td>IBM SP2 Silver</td>
<td></td>
</tr>
<tr>
<td></td>
<td>332-MHz PowerPC 604e</td>
<td></td>
</tr>
</tbody>
</table>

Better

March 14, 2002
UMT2000 Purple Marquee Benchmark gives great performance on Pentium4
Proposed Dual P4 Linux cluster 11.1 teraFLOP/s system architecture†

System Specifications

- 11.1 teraFLOP/s, 23 TB of memory, 138 TB local, 144 TB global disk, 8.2 GB/s global I/O
- <5 μs, 300 (360) MB/s MPI latency and Bandwidth over QsNet
- Support 120 MB/s transfers to Archive over dual Jumbo Frame Gb-Enet and QSW links from each Login node
- Open Source OS, File System, Management, Tools

†Cluster wide file system leverages ASCI PathForward Lustre development

Estimated cost $15-20M (compare to White at $120M)
QsNet Elan3 modified fat-tree interconnect for a 1,152 node, 11+ TF/s cluster

F = 1,154x9.6 GF/s = 11.08 TF/s
Node B:F = 680 MB/s / 9.6 GF/s = 0.071
System B:F = 130.56 GB/s / 11.06 TF/s = 0.0118
Node : System = 6.0
Requires 16 QsNet Elan3 128-way switches

1,152 Elan3 ports
Commodity node scalable cluster has modest power, cooling and floor space requirements

11.1 TF/s peak (4 TF/s Linpack)  
23 TB memory, 138 TB local disk  
144 TB of global disk and 8.1 GB/s I/O

This System is 34’ wide by 25’ tall or 850 ft²  
Requires about 160 kW of power
Follow the money

Volume segment migrating from desk top to game machines
- These machines do greate fixed point arithmetic and visualization

Volume segment also migrating to Embedded
- CY02 Mercedes Benz has 32 microprocessors in it (that’s not a car, it’s a mobile cluster)
- System on a chip design technology has interesting possibilities
- This segment focuses on MIPS/Watt, MIPS/$
ASCI’s advanced-architecture effort is targeted at long range solutions to critical issues.

BlueGene/L is an R&D project with IBM
- Designed to deliver exceptional price/performance on a small and growing set of science applications
- Utilizes over 65K processors and multiple networks
- Based on system on a chip ASIC technology

BlueGene/L is addressing the FIVE KEY CHALLENGES facing ASCI Platforms today
- Power and cooling
- Floor space
- Cost of systems
- Interconnect
- Distance to memory (single node performance)

BlueGene/L is targeted for the 2004 time frame, at a price/performance unobtainable with conventional architectures and design approaches.
Summary

- Terascale scientific simulation has dramatically changed the way simulation and science is done.
- Requirements for predictive scientific simulations with large economic benefit are changing the state of the art.
  - ASCI Program as example
  - HPTC market driver
  - Japanese ESP second example
- Disruptive technologies including commodity clusters and open source software put HPTC on a different cost performance curve.

First ever 1B atom MD simulation shows supersonic crack formation and leads theory and experiment to show this phenomena exists.

First ever ab initio deuterium shock simulation confirms experimental evidence (that was previously ignored) that deuterium shocks are 50% more dense than theory predicts.