Scientific Computing at SLAC

Richard P. Mount
Director: Scientific Computing and Computing Services

DOE Review
June 15, 2005
Scientific Computing
The relationship between Science and the components of Scientific Computing

<table>
<thead>
<tr>
<th>Application Sciences</th>
<th>High-energy and Particle-Astro Physics, Accelerator Science, Photon Science …</th>
</tr>
</thead>
<tbody>
<tr>
<td>Issues addressable</td>
<td>Particle interactions with matter, Electromagnetic structures, Huge volumes of data, Image processing …</td>
</tr>
<tr>
<td>with “computing”</td>
<td></td>
</tr>
<tr>
<td>Computing techniques</td>
<td>PDE Solving, Algorithmic geometry, Visualization, Meshes, Object databases, Scalable file systems …</td>
</tr>
<tr>
<td>Computing architectures</td>
<td>Single system image, Low-latency clusters, Throughput-oriented clusters, Scalable storage …</td>
</tr>
<tr>
<td>Computing hardware</td>
<td>Processors, I/O devices, Mass-storage hardware, Random-access hardware, Networks and Interconnects …</td>
</tr>
</tbody>
</table>
Scientific Computing:
SLAC’s goals for leadership in Scientific Computing

Application Sciences

Issues addressable with “computing”

Computing techniques

Computing architectures

Computing hardware

SLAC + Stanford Science

The Science of Scientific Computing

Computing for Data-Intensive Science

Collaboration with Stanford and Industry
Scientific Computing:
Current SLAC leadership and recent achievements in Scientific Computing

PDE solving for complex electromagnetic structures
Huge-memory systems for data analysis
Scalable data management

GEANT4 photon/particle interaction in complex structures (in collaboration with CERN)

World’s largest database
Internet2 Land-Speed Record; SC2004 Bandwidth Challenge
SLAC Scientific Computing Drivers

• BaBar (data-taking ends December 2008)
  – The world’s most data-driven experiment
  – Data analysis challenges until the end of the decade

• KIPAC
  – From cosmological modeling to petabyte data analysis

• Photon Science at SSRL and LCLS
  – Ultrafast Science, modeling and data analysis

• Accelerator Science
  – Modeling electromagnetic structures (PDE solvers in a demanding application)

• The Broader US HEP Program (aka LHC)
  – Contributes to the orientation of SLAC Scientific Computing R&D
SLAC-BaBar Computing Fabric

- **Client**
- **Disk Server**
- **Tape Server**

**IP Network (Cisco)**

- 1700 dual CPU Linux
- 400 single CPU
- Sun/Solaris

- 120 dual/quad CPU
- Sun/Solaris
- ~400 TB Sun
- FibreChannel RAID arrays

- HPSS + SLAC enhancements to ROOT and Objectivity server code

- HEP-specific ROOT software (Xrootd) + Objectivity/DB object database

- 25 dual CPU
- Sun/Solaris
- 40 STK 9940B
- 6 STK 9840A
- 6 STK Powderhorn
- over 1 PB of data

Richard P. Mount, SLAC
June 2005
BaBar Computing at SLAC

- Farm Processors (5 generations, 3700 CPUs)
- Servers (the majority of the complexity)
- Disk storage (2+ generations, 400+ TB)
- Tape storage (40 Drives)
- Network “backplane” (~26 large switches)
- External network
Rackable Intel P4 Farm (bought in 2003/4)
384 machines, 2 per rack unit, dual 2.6 GHz CPU
Disks and Servers
1.6 TB usable per tray, ~160 trays bought 2003/4
Tape Drives
40 STK 9940B (200 GB) Drives
6 STK 9840 (20 GB) Drives
6 STK Silos (capacity 30,000 tapes)
BaBar Farm-Server Network
~26 Cisco 65xx Switches

Farm/Server Network
SLAC External Network (June 14, 2005)

622 Mbits/s to ESNet
1000 Mbits/s to Internet 2
~300 Mbits/s average traffic

Two 10 Gbits/s wavelengths to ESNET, UltraScience Net/NLR coming in July

[Graphs showing network utilization over time for SLAC ESNET Gateway and SLAC Stanford Gateway]
Research Areas (1)
(Funded by DOE-HEP and DOE SciDAC and DOE-MICS)

• Huge-memory systems for data analysis
  (SCCS Systems group and BaBar)
  – Expected major growth area (more later)

• Scalable Data-Intensive Systems:
  (SCCS Systems and Physics Experiment Support groups)
  – “The world’s largest database” (OK not really a database any more)
  – How to maintain performance with data volumes growing like “Moore’s Law”? (intelligence plus brute force)
  – How to improve performance by factors of 10, 100, 1000, … ?
  – Robustness, load balancing, troubleshootability in 1000 – 10000-box systems
  – Astronomical data analysis on a petabyte scale (in collaboration with KIPAC)
Research Areas (2)
(Funded by DOE-HEP and DOE SciDAC and DOE MICS)

• Grids and Security:
  (SCCS Physics Experiment Support. Systems and Security groups)
  – **PPDG**: Building the US HEP Grid – OSG;
  – Security in an open scientific environment;
  – Accounting, monitoring, troubleshooting and robustness.

• Network Research and Stunts:
  (SCCS Network group – Les Cottrell et al.)
  – Land-speed record and other trophies

• Internet Monitoring and Prediction:
  (SCCS Network group)
  – **IEPM**: Internet End-to-End Performance Monitoring (~5 years)
    SLAC is the/a top user of ESNet and the/a top user of Internet2. (Fermilab
    doesn’t do so badly either)
  – **INCITE**: Edge-based Traffic Processing and Service Inference for High-
    Performance Networks
Research Areas (3)
(Funded by DOE-HEP and DOE SciDAC and DOE MICS)

• GEANT4: Simulation of particle interactions in million to billion-element geometries:
  (SCCS Physics Experiment Support Group – M. Asai, D. Wright, T. Koi, J. Perl …)
  – BaBar, GLAST, LCD …
  – LHC program
  – Space
  – Medical

• PDE Solving
  for complex electromagnetic structures:
  (Kwok ‘s advanced Computing Department + SCCS clusters)
Growing Competences

• Parallel Computing (MPI …)
  – Driven by KIPAC (Tom Abel) and ACD (Kwok Ko)
  – SCCS competence in parallel computing (= Alf Wachsmann currently)
  – MPI clusters and SGI SSI system

• Visualization
  – Driven by KIPAC and ACD
  – SCCS competence is currently experimental-HEP focused (WIRED, HEPREP …)
  – (A polite way of saying that growth is needed)
A Leadership-Class Facility for Data-Intensive Science

Richard P. Mount
Director, SLAC Computing Services
Assistant Director, SLAC Research Division

Washington DC, April 13, 2004
Technology Issues in Data Access

- Latency
- Speed/Bandwidth
- (Cost)
- (Reliability)
Latency and Speed – Random Access

Random-Access Storage Performance

Retrieval Rate Mbytes/s vs log10 (Object Size Bytes)

PC2100
WD200GB
STK9940B

Richard P. Mount,  SLAC  
June 2005
Latency and Speed – Random Access

Historical Trends in Storage Performance

![Graph showing historical trends in storage performance with different storage devices and object sizes. The x-axis represents log10 (Object Size Bytes), and the y-axis represents Retrieval Rate MBytes/s. The graph compares PC2100, WD200GB, STK9940B, RAM 10 years ago, Disk 10 years ago, and Tape 10 years ago.]
The Strategy

- There is significant commercial interest in an architecture including data-cache memory
- **But:** from interest to delivery will take 3-4 years
- **And:** applications will take time to adapt not just codes, but their whole approach to computing, to exploit the new architecture
- **Hence:** two phases
  1. **Development phase (years 1,2,3)**
     - Commodity hardware taken to its limits
     - BaBar as principal user, adapting existing data-access software to exploit the configuration
     - BaBar/SLAC contribution to hardware and manpower
     - Publicize results
     - Encourage other users
     - Begin collaboration with industry
  2. **Production-Class Facility (year 3 onwards)**
     - Optimized architecture
     - Strong industrial collaboration
     - Wide applicability
PetaCache
The Team

- David Leith, Richard Mount, PIs
- Randy Melen, Project Leader
- Bill Weeks, performance testing
- Andy Hanushevsky, xrootd
- Systems group members
- Network group members
- BaBar (Stephen Gowdy)
Development Machine Design Principles

- Attractive to scientists
  - Big enough data-cache capacity to promise revolutionary benefits
  - 1000 or more processors
- Processor to (any) data-cache memory latency < 100 $\mu$s
- Aggregate bandwidth to data-cache memory > 10 times that to a similar sized disk cache
- Data-cache memory should be 3% to 10% of the working set (approximately 10 to 30 terabytes for BaBar)
- Cost effective, but acceptably reliable
  - Constructed from carefully selected commodity components
- Cost no greater than (cost of commodity DRAM) + 50%
Development Machine
Design Choices

• Intel/AMD server mainboards with 4 or more ECC
dimm slots per processor
• 2 Gbyte dimms (4 Gbyte too expensive this year)
• 64-bit operating system and processor
  – Favors Solaris and AMD Opteron
• Large (500+ port) switch fabric
  – Large IP switches are most cost-effective
• Use of ($10M+) BaBar disk/tape infrastructure,
augmented for any non-BaBar use
Development Machine
Deployment – Proposed Year 1

Memory Interconnect Switch Fabric

650 Nodes, each
2 CPU, 16 GB memory

Storage Interconnect Switch Fabric

Cisco/Extreme/Foundry

> 100 Disk Servers
Provided by BaBar
Development Machine Deployment – Currently Funded

- Cisco Switch
  - Data-Servers: 64-128 Nodes, each Sun V20z, 2 Opteron CPU, 16 GB memory
  - Solaris
  - Up to 2TB total Memory

- Cisco Switches

- Clients: up to 2000 Nodes, each 2 CPU, 2 GB memory
  - Linux

- PetaCache
  - MICS Funding

- Existing HEP-Funded BaBar Systems
Latency (1)
Ideal

- Memory
- Client Application
Latency (2)
Current reality

Diagram:
- Data Server
  - OS
  - File System
    - Disk
- Client Application
  - OS
  - TCP Stack
    - NIC
- Network Switches
Latency (3)
Immediately Practical Goal

Diagram showing the components of a system with layered architecture:
- Memory
- Data Server
- OS
- File System
- Disk
- TCP Stack
- NIC
- Network Switches
- Client Application
- Data-Server-Client
- OS
- TCP Stack
- NIC
Latency Measurements
(Client and Server on the same switch)
Development Machine Deployment
 Likely “Low-Risk” Next Step

Data-Servers 80 Nodes, each 8 Opteron CPU, 128 GB memory
Up to 10TB total Memory
Solaris

Clients up to 2000 Nodes, each 2 CPU, 2 GB memory
Linux
Development Machine
Complementary “Higher Risk” Approach

• Add Flash-Memory based subsystems
  – Quarter to half the price of DRAM
  – Minimal power and heat
  – Persistent
  – 25 $\mu$s chip-level latency (but hundreds of $\mu$s latency in consumer devices)
  – Block-level access (~1kbyte)
  – Rated life of 10,000 writes for two-bit-per-cell devices
    (NB BaBar writes FibreChannel disks < 100 times in their entire service life)

• Exploring necessary hardware/firmware/software development with PantaSys Inc.
Object-Serving Software

- AMS and Xrootd (Andy Hanushevsky/SLAC)
  - Optimized for read-only access
  - Make 1000s of servers transparent to user code
  - Load balancing
  - Automatic staging from tape
  - Failure recovery

- Can allow BaBar to start getting benefit from a new data-access architecture within months without changes to user code

- Minimizes impact of hundreds of separate address spaces in the data-cache memory
Summary:
SLAC’s goals for leadership in Scientific Computing

Application Sciences

Issues addressable with “computing”

Computing techniques

Computing architectures

Computing hardware

SLAC + Stanford Science

The Science of Scientific Computing

for Data-Intensive Science

Collaboration with Stanford and Industry