Sun Workshop on Grid Computing

Eric Stahlberg
Senior Systems Manager
eas@osc.edu

What Is OSC?

Mission: To deliver high performance computing technology to higher education, government and industry in Ohio

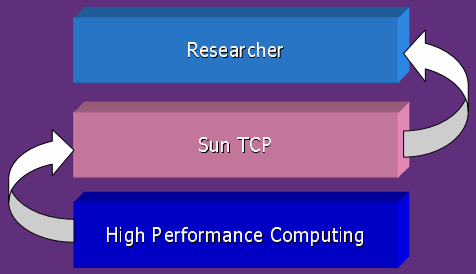
- Started in 1987
- Supported by Ohio Board of Regents
- Provides resources for all Ohio universities and colleges

Delivering High Performance Computing

- We have a variety of technology…
  - SGI IA32 Linux cluster
  - 146 processor Itanium cluster
  - 16 processor Cray SV1
  - 32 processor SGI Origin 2000
  - Sun Center of Excellence for HPC Environments
    - New 12- and 24- processor systems with *terabytes* of storage

Challenge:
To deliver hardware technology to the research desktop

Making the Connection…



Portal Challenges

- ✓ Integrate desktop applications
- ✓ Provide access to large variety of systems
- ✓ Provide access to large variety of applications
- ✓ Allow end-user customization (at some level)
- ✓ Keep it simple
  - ❑ Use
  - ❑ Administration
- ✓ Keep it effective

Life Sciences Resources Update

- Web-accessible resources
  - Entigen BioNavigator … no future
  - Sun TCP … in development and under review
- Desktop Research Environment
  - LabBook … available now
- High-performance bioinformatics hardware
  - Cray SV1 libraries … available now
  - Time Logic … available now
- Other life-science applications
  - Need input … presently
- Chemistry software
  - Amber … available now
  - Many more … available now

Cray SV1 Genomic Libraries

- Description
  - Source code libraries for customize applications
  - Extremely fast pattern recognition (100x-1000x)
  - Big Cray memory can store large data files
- Uses
  - Mapping SNPs, STRs or inverted repeats
  - Reference The Scientist, August 20, 2001 for applications of libraries at NCI
  - Very interested in Ohio opportunities to work with NCI in collaborative projects exploiting technology (See http://www-fbsc.ncifcrf.gov)

Time Logic Boards

- Genomic application acceleration
  - Circuit boards mounted in Sun systems
  - Command line and potential web access
  - > 100x speedup for many algorithms
- Applications available
  - NCBI BLASTALL, PSI-BLAST
  - HMMSearch, HMM-Framesearch
  - Smith-Waterman
  - ClustalW
  - Phrap

Quick and Dirty Integration

- Constraints
  - Make it quick
  - Don't alter production environments
- Objectives
  - Integrate multiple systems
  - Integrate multiple technologies
  - Create usable environment for feedback

Integration Strategy

- Use ssh for remote execution
  - Controllable
  - Preserves production environments
  - Downside: Setup overhead and lost accounting
- Use tiered indirection for flexibility
  - Tier 1: Map TCP application to installed location
  - Tier 2: Data mapping or transport to common area
  - Tier 3: Host selection
  - Tier 4: Local execution

Implementation

- Tier 1- Application mapping
  - Use soft links to map to more accessible areas
- Tier 2- Data transport
  - Create run-time directory with necessary permission and transfer uploaded project data
- Tier 3 – Host selection
  - Identify host and local executable for ssh if needed
- Tier 4 – Local execution
  - Run application and create output files

Time Logic DeCypher Example

- Characteristics
  - Runs on single non-execute host system
  - Uses separate job management system
  - Output must be retrieved
  - Tier1 to tier 3 are scripts, similar for all apps
- Tier4 unique
  - Use ssh to initiate search request, poll status, and retrieve text output
  - Post-process text output into html documents for use
- Tier 2 returns final results to project directory

Integrated Applications

- Cray SV1
  - Custom application using special libraries
- Time Logic DeCypher
  - Special job control on Sun Fire 6800
- Origin 2000
  - Multicoil bioinformatics application
- SunFire 6800
  - Predotar and Mitoprot
  - ClustalW parallel

Observations

- Application integration flexibility is important
- Ability to debug implementation outside of portal is advantageous
- Multiple technologies can be readily merged with minimal effort
- Relying on ssh is not managable for large number of users of common applications
- Data locality is an important issue