Production Grids

• Last Time
  » Administrative Info
  » Coursework
  » Signup for Topical Reports! (signup immediately if you haven’t)
  » Vision of Grids

• Today
  » Reality of High Performance Distributed Computing
  » Example Working Grids
  » Models of Use

• Reminders/Announcements
  » We will break 20 minutes early for “project team formation”

Outline

• TeraGrid
• Grid3/iVDGL
• A list of others…

• A perspective on important Grid research problems…
TeraGrid Vision: A Unified National HPC Infrastructure that is Persistent and Reliable

- Largest NSF compute resources
- Largest DOE instrument (SNS)
- Fastest network
- Massive storage
- Visualization instruments
- Science Gateways
- Community databases

E.g.: Geosciences: 4 data collections including high-res CT scans, global telemetry data, worldwide hydrology data, and regional LIDAR terrain data

TeraGrid Components

- **Compute hardware**
  - Intel/Linux Clusters, Alpha SMP clusters, POWER4 cluster, POWER3 cluster, SUN visualization system
- **Large-scale storage systems**
  - Hundreds of terabytes for secondary storage
- **Very high-speed network backbone**
  - Bandwidth for rich interaction and tight coupling
- **Grid middleware**
  - Globus, data management, ...
- **Next-generation applications**
Resources and Services
(40TF, 1.4PB disk, 12 PB tape)

<table>
<thead>
<tr>
<th>UC</th>
<th>Caltech</th>
<th>IU</th>
<th>NCSA</th>
<th>ORNL</th>
<th>PSC</th>
<th>Purdue</th>
<th>SDSC</th>
<th>TACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Online Storage</td>
<td>20 TB</td>
<td>170 TB</td>
<td>6 TB</td>
<td>600 TB</td>
<td>150 TB</td>
<td>540 TB</td>
<td>50 TB</td>
<td></td>
</tr>
<tr>
<td>Archival Storage</td>
<td>150 TB</td>
<td>1.5 PB</td>
<td>2.4 PB</td>
<td>150 TB</td>
<td>2 PB</td>
<td>2 PB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Networking (Gbps to hub)</td>
<td>30 Gbps CHI</td>
<td>30 Gbps LA</td>
<td>10 Gbps CHI</td>
<td>30 Gbps CHI</td>
<td>10 Gbps CHI</td>
<td>30 Gbps LA</td>
<td>10 Gbps CHI</td>
<td></td>
</tr>
<tr>
<td>Database &amp; Data Collections</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Instruments</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Visualization</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Dissemination</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

TeraGrid Compute Resources
Data Resources and Data Management Services

- **Approach**
  - Deploy core services
  - Drive the system with data intensive flagship applications
- **TG Data Services Plan**
  - High-speed cross-site data transfer capability
  - Parallel file systems
  - Mass storage systems
  - GridFTP and SRB-based access to data
  - Hosted data collections
  - Database capabilities
    - 2 dedicated 32-processor IBM p690 nodes

---

TeraGrid Runtime Environment

TeraGrid Services define what runtime environment a process can expect.

- **User**
- **User Proxy**
- **Globus Credential**
- **Site 1**
  - GRAM
  - GSI
  - Kerberos
  - Ticket
  - Process
- **Authenticated interprocess communication**
- **Site 2**
  - GRAM
  - GSI
  - Certificate
  - Public Key
  - Process

Single sign-on via "grid-id"  
Assignment of credentials to "user proxies"  
Mutual user-resource authentication  
Mapping to local ids
Common Authentication Service

- Standardized GSI authentication
  - A user can authenticate for SSH and Grid services to every TeraGrid system with a single user certificate
- Developed coordinated CA acceptance policy
  - TeraGrid-accepted Certificate Authorities include
    - NCSA, SDSC, PSC, Purdue, TACC CAs
    - DOEGrids CA, UK E-Science CA
- Procedures and tools to simplify the use and management of certificates
  - Certificate request, retrieval, distribution, installation

Grid Information Services

- Currently Leveraging Globus Grid Information Service
  - each service/resource is an information source
  - index servers at each of the TG sites
  - full mesh between index servers for fault tolerance
  - access control as needed
- Resource Information Management Strategies
  - TG GIS for systems level information
  - generic non-sensitive information
Grid Scheduling & Job Management: Condor-G, the User Interface

- Condor-G is the preferred job management interface
  - job scheduling, submission, tracking, etc.
  - allows for complex job relationships and data staging issues
  - interfaces to Globus layers transparently
  - allows you to use your workstation as your interface to the grid

- The ability to determine current system loads and queue status will come in the future.

Current TG Software Stack

- SuSE SLES
- X-cat
- MPICH, MPICH-G2, MPICH-VMI
- gm drivers
- VMI/CRM
- Globus
- Condor-G
- gsi-ssh
- GPT
- SoftEnv
- MyProxy
- Intel compilers
- GNU compilers
- HDF4/5
- SRB client
- Inca
- db2-client
- uberftp
- tg-policy
- tg_usage
- myprojects
- Kx509
- atlas
Summary

- Large number of processors
  - Possible to run across clusters (very parallel code)
- High bandwidth network between TeraGrid sites for fast data transfer
- Large parallel file system

Grid2003: an application grid laboratory

[Rob Gardner UC, 2003]
Grid3 at a Glance

- Grid environment built from core Globus and Condor middleware, as delivered through the Virtual Data Toolkit (VDT)
  - GRAM, GridFTP, MDS, RLS, VDS, VOMS, …
- …equipped with VO and multi-VO security, monitoring, and operations services
- …allowing federation with other Grids where possible, eg. CERN LHC Computing Grid (LCG)
  - USATLAS: GriPhyN VDS execution on LCG sites
  - USCMS: storage element interoperability (SRM/dCache)

Grid3 – a snapshot of sites

Facilities:
- Five national laboratories (Argonne, Brookhaven, Fermilab (LHC T1), Brookhaven (LHC T1), Berkeley Lab)
- Universities: LHC T2’s, campus computing facilities, department clusters

Types (all platforms Linux RH):
- Shared and dedicated
- Production clusters
- Development clusters
Application Overview

- 7 Scientific applications and 3 CS demonstrators
  - All iVDGL experiments participated in the Grid2003 project
  - A third HEP and two Bio-Chemical experiments also participated

- Over 100 users authorized to run on Grid3
  - Application execution performed by dedicated individuals
  - Typically 1, 2 or 3 users ran the applications from a particular experiment

Physics Workflow on Grid3

- Higgs $\rightarrow$ 4 lepton sample
  - Simulation and Reconstruction
  - 2000 jobs (X 6 subjobs); 100–200 events per job (~200K events)
  - 500 GB output data files

- Workflow
  - Jobs defined, stored and registered in RLS
  - Chimera, Pegasus, and Condor-G to submit to Grid3 sites
  - Data pushed to SE at Brookhaven, registered RLS
Grid3 Continued Operations

- CMS data challenge during first three months of 2004

Metrics Summary

<table>
<thead>
<tr>
<th>Metric</th>
<th>Target</th>
<th>Grid2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of CPUs</td>
<td>400</td>
<td>2762 (27 sites)</td>
</tr>
<tr>
<td>Number of users</td>
<td>&gt; 10</td>
<td>102 (16)</td>
</tr>
<tr>
<td>Number of Applications</td>
<td>&gt; 4</td>
<td>10</td>
</tr>
<tr>
<td>Number of site running concurrent applications</td>
<td>&gt; 10</td>
<td>17</td>
</tr>
<tr>
<td>Peak number of concurrent jobs</td>
<td>1000</td>
<td>1100</td>
</tr>
<tr>
<td>Data Transfer per day</td>
<td>&gt; 2-3 TB</td>
<td>4.4 TB (11.12.03)</td>
</tr>
</tbody>
</table>
Jobs by Month

**Grid2003 Metrics Results**

- **Hardware resources**
  - Total of 2762 CPUs
    - Maximum CPU count
    - Off project contribution > 60%
  - Total of 27 sites
    - 27 administrative domains with local policies in effect
    - All across US and Korea
- **Running jobs**
  - Peak number of jobs 1100
  - During SC2003 various applications were running simultaneously across various Grid3 sites
Data Consumed by Grid3

Conclusions and Outlook

- Grid2003 taught us many lessons about how to deploy and run a production grid
- Breakthrough in demonstrated use of “opportunistic” resources enabled by grid technologies
- Grid3+ will be a critical resource for continued data challenges, and environment to learn how to operate and upgrade large scale production grids
- Challenge will be to maintain vital R&D effort while doing sustained production operations
Grid3 Update, 2/2005

Grid3: An Operational National Grid
- 35 sites, 3500 CPUs: Universities + 4 national labs
- Part of LHC Grid
- Running since October 2003
- Applications in HEP, LIGO, SDSS, Genomics, CS

http://www.ivdgl.org/grid3

Grid3 Applications
- High energy physics
  - US-ATLAS analysis (DIAL),
  - US-ATLAS GEANT3 simulation (GCE)
  - US-CMS GEANT4 simulation (MOP)
  - BTeV simulation
- Gravity waves
  - LIGO: blind search for continuous sources
- Digital astronomy
  - SDSS: cluster finding (maxBcg)
- Bioinformatics
  - Bio-molecular analysis (SnB)
  - Genome analysis (GADU/Gnare)
- CS Demonstrators
  - Job Exerciser, GridFTP, NetLogger-grid2003
Grid3 Shared Use Over 6 months

Usage: CPUs

time: 0/10/2004

day: time

atlas dc2

cms dc04

Sep 10

Open Science Grid

Applications

BarBar, STAR, PHENIX, etc.
Biology
Computer Science
Astrophysics
Run 2 CDF, D0
LHC, Atlas, CMS, Alice

Persistent Grid Infrastructure

General Facility for any Community e.g. Tier-Grid

Laboratory Serving Multiple Communities e.g. Fermilab, BNL, NERSC
Community Facility e.g. US ATLAS or CMS Tier-1/Tier-2
University Facility e.g. UF, Ithaca, Buffalo
University Community Facility e.g. GLOW

Facilities

User Support Center
Middleware Providers
Grid Operations Center
Certificate Authorities
Database Operators

http://www.opensciencegrid.org
Other large Grids…

- *PlanetLab*
- UK E-science Grid
- European Union Data Grid
- Naregi Grid (Japan)
- China Grid
- UniGrid
- ...

- A list of hundreds at this point…
And Thousands of…

- Desktop Grids
  - Condor, Seti@home, Folding@home, GIMPS, ComputeAgainstCancer, ...
- Database Grids (Oracle 10g)
- Condor
- Platform Computing LSF
- Portable Batch System
- enFuzion
- ...

Critical Grid Research Issues: Perspective and Lessons from Large-Scale Grids

- HPDC-13 Panel -- June 6, 2004
- Grid2003 Rob Gardner, U Chicago
- Planetlab Jeff Chase, Duke
- Condor Miron Livny, U Wisconsin
- Globus Ian Foster, U Chicago
- Andrew Chien, UCSD (Moderator)

- Panel Charge
  - Top 5 Things Learned
    - What ARE major problems (and need extensive research)
    - What are NOT major problems
    - Two "takeaways" for every HPDC researcher
Not Major Problems (but often studied extensively)

- Performance
- Meta scheduling
- Grid economy
- Communication overhead
- Reservations
- Predictions
- Scalability to millions of devices
- Heterogeneity (hide differences behind familiar low-level APIs)
- Basic resource access, monitoring, etc.
- Grid “plumbing” shift to service frameworks and standardization efforts.
- Identifying interesting Grid applications -- There are many
- Compilers and programming languages

Major Problems

- Troubleshooting & problem determination and Debugging (Root causes)
- Autonomic management (problem scope, provide QoS)
- Accountability, Trust, Security
- Resource allocation (load control)
- Complex Resource Management Policies
- Application models
  - Integrating on-demand resources
- Heterogeneous schema
  - Integrating data, services, etc.
- Application Development: integrated, dynamic distributed systems
- Workflows and interfaces, data exchange with other grids
Summary

- TeraGrid – a tightly-coupled production grid
- Grid3/iVDGL – a loosely-coupled federation of heterogeneous resources
- … and many others… Grids are large and in large-scale use!

- Expert perspective on important Grid research problems…