TeraGrid Science Gateways

Nancy Wilkins-Diehr
TeraGrid Area Director for Science Gateways
wilkinsn@sdsc.edu

NBCR Summer Institute, August 4, 2009
Has anyone heard about the gateway program?

• Today I hope to answer
  – What is the TeraGrid?
  – What are gateways?
    • Why are gateways worth the effort?
    • What do they allow scientists to do that they couldn't do without gateways?
    • What are some specific examples of this?
    • Why are these examples important?
  – What about biology?
    • Should I use an existing gateway?
    • Should I develop my own?
    • Meeting the needs of next generation scientists
What is the TeraGrid?

A unique combination of fundamental CI components
What is the TeraGrid?

- One of the world’s largest grids for open scientific discovery
- Leadership class resources at eleven partner sites combined to create an integrated, persistent computational resource
  - High-performance networks
  - High-performance computers (>1 Pflops (~100,000 cores) -> 1.75 Pflops)
    - And a Condor pool (w/ ~25,000 CPUs)
  - Visualization systems
  - Data Collections (>30 PB, >100 discipline-specific databases)
  - Science Gateways
- User portal
- User services - Help desk, training, advanced application support
- Allocated to US researchers and their collaborators through national peer-review process
  - CPUs and advanced support available at no cost through peer review process
- Extremely user-driven
  - MPI jobs, ssh or grid (GRAM) access, etc.

Adapted from Dan Katz, TeraGrid

NBCR Summer Institute, August 4, 2009
• Very Powerful Tightly Coupled Distributed Memory
  – Ranger (TACC): Sun Constellation, 62,976 cores, 579 Tflop, 123 TB RAM
  – Kraken (NICS): Cray XT5, 66,048 cores, 608 Tflop, > 1 Pflop in 2009
• Shared Memory
  – Cobalt (NCSA): Altix, 8 Tflop, 3 TB shared memory
  – Pople (PSC): Altix, 5 Tflop, 1.5 TB shared memory
• Clusters with Infiniband
  – Abe (NCSA): 90 Tflops
  – Lonestar (TACC): 61 Tflops
  – QueenBee (LONI): 51 Tflops
• Condor Pool (Loosely Coupled)
  – Purdue- up to 22,000 cpus
• Gateway hosting
  – Quarry (IU): virtual machine support
• Visualization Resources
  – TeraDRE (Purdue): 48 node nVIDIA GPUs
  – Spur (TACC): 32 nVIDIA GPUs
• Storage Resources
  – GPFS-WAN (SDSC)
  – Lustre-WAN (IU)
  – Various archival resources
Gateways

A natural result of the impact of the internet on worldwide communication and information retrieval

Only 17 years since the release of Mosaic!

- **Implications on the conduct of science are still evolving**
  - 1980’s, Early gateways, National Center for Biotechnology Information BLAST server, search results sent by email, still a working portal today
  - 1992 Mosaic web browser developed
  - 1995 “International Protein Data Bank Enhanced by Computer Browser”
  - 2004 TeraGrid project director Rick Stevens recognized growth in scientific portal development and proposed the Science Gateway Program
  - Today, Web 3.0 and programmatic exchange of data between web pages

- **Simultaneous explosion of digital information**
  - Growing analysis needs in many, many scientific areas BIOLOGY!
  - Sensors, telescopes, satellites, digital images and video, genome sequencers!
  - #1 machine on Top500 today is 1000x more powerful than all combined entries on the first list in 1993

NBCR Summer Institute, August 4, 2009
Why are gateways worth the effort?

- Increasing range of expertise needed to tackle the most challenging scientific problems
  - How many details do you want each individual scientist to need to know?
- PBS, RSL, Condor
- Coupling multi-scale codes
- Assembling data from multiple sources
- Collaboration frameworks

#!/bin/bash

#PBS -q dque
#PBS -l nodes=1:ppn=2
#PBS -l walltime=00:02:00
#PBS -o pbs.out
#PBS -e pbs.err
#PBS -V
cd /users/wilkinsn/tutorial/exercise_3
../bin/mcell nmj_recon.main.mdl

# Full path to executable
executable=/users/wilkinsn/tutorial/bin/mcell

# Working directory, where Condor-G will write
initialdir=/users/wilkinsn/tutorial/exercise_3

# To set the working directory of the remote job, we specify it in this globus RSL, which will be appended
# to the RSL that Condor-G generates

globusrsl=(directory="/users/wilkinsn/tutorial/exercise_3")

# Arguments to pass to executable.
arguments=nmj_recon.main.mdl

+( &
   (resourceManagerContact="tg-login1.sdsc.teragrid.org/jobmanager-pbs")
   (stdout="/users/birnbaum/tutorial/exercise_3/globus.out")
   (stderr="/users/birnbaum/tutorial/exercise_3/globus.err")
   (executable=/users/wilkinsn/tutorial/bin/mcell)
   (args=nmj_recon.main.mdl)
   (login1.sd=login1.sdsc.teragrid.org)
   (jobqueue=condor)
   (hostCc=login1.sdsc.teragrid.org)
   (ma=condor)
   (dir=.)
   (std=condor)
   (universe=globus)
)

# Files to receive sdout and stderr.
output=condor.out
error=condor.err

# Specify the number of copies of the job to submit to the condor queue.
queue 1
Gateways democratize access to high end resources

• Almost anyone can investigate scientific questions using high end resources
  – Not just those in the research groups of those who request allocations
  – Gateways allow anyone with a web browser to explore
    • Opportunities can be uncovered via google
      – My 11-year-old son discovered nanoHUB.org himself while his class was studying Bucky Balls

• Foster new ideas, cross-disciplinary approaches
  – Encourage students to experiment

• But used in production too
  – Significant number of papers resulting from gateways including GridChem, nanoHUB
  – Scientists can focus on challenging science problems rather than challenging infrastructure problems

NBCR Summer Institute, August 4, 2009
Today, there are approximately 35 gateways using the TeraGrid
Not just ease of use

What can scientists do that they couldn’t do previously?

• Linked Environments for Atmospheric Discovery (LEAD) - access to radar data
• National Virtual Observatory (NVO) – access to sky surveys
• Ocean Observing Initiative (OOI) – access to sensor data
• PolarGrid – access to polar ice sheet data
• SIDGrid – expensive datasets, analysis tools
• GridChem – coupling multiscale codes

• How would this have been done before gateways?
Gateways for Biology

- www.teragrid.org/gateways
  - List of all TeraGrid gateways

- Biodrugscore
- RENCI Science Portal
- Open Life Sciences Gateway
- Robetta
- 454 pyrosequencing data
Derive and validate scoring functions

- Create training sets using structural and binding data from multiple databases including PDBbind and PDBcal
- Define the components of scoring functions by picking from among a list of pre-computed terms
  - Partial least-squares regression analysis

Validate scoring functions

Apply custom scoring functions for the ranking of chemical libraries that are pre-docked against a large set of binding cavities from the human proteome

- If the receptor of interest is not available, biodrugscore makes it possible for users to dock libraries against their target on the TeraGrid using their own account.

NBCR Summer Institute, August 4, 2009
• Compute resources

• Service projects
  – Quantum to Continuum Mechanics Tools
  – Data Analysis Tools for Molecular Sequences
  – Heart Modeling
  – Visualization and multi-scale modeling
  – Grid services and Telescience

• Tools and downloads
  – 40+ packages, databases, services
• 125 biology applications
  – From Antigenic to WordMatch and everything in between
  – RENCI Science Desktop
  – BlastMaster desktop

• Talk at NBCR Summer Institute, Aug 6, 1pm
Open Life Sciences Gateway
http://lsgw.uc.teragrid.org

- Bioinformatics applications and data collections
- Portal access, direct Web services calls, workflows with Taverna
  - And now google gadgets!

NBCR Summer Institute, August 4, 2009
Robetta
http://www.robetta.org

• Protein structure prediction server
  – Rosetta code from the David Baker laboratory

• Also available
  – RosettaAntibody Server
  – RosettaDesign Server
  – RosettaDock Server
  – Rosetta Commons
  – FoldIt
  – Rosetta@home
  – Human Proteome Folding Project

NBCR Summer Institute, August 4, 2009
New Pyrosequencing Technology
The future of individualized medicine

• Allegheny General Hospital
  – Early adopter of DNA sequencing platform from 454 Life Science Inc./Roche

• Significant upgrade to this sequencing platform
  – Massively parallel, clone-free DNA pyrosequencing technology well suited for a variety of applications
  – Improved sequencing inline with NIH goal of making individual DNA sequencing available for $1k

• TeraGrid to be used for data processing

NBCR Summer Institute, August 4, 2009

http://www.roche-applied-science.com/publications/multimedia/genome_sequencer/flx_presentation/wbt.htm
What if you want to build your own gateway?

- Request an allocation
  - Only a 1 paragraph abstract required for up to 200k CPU hours

- Register your gateway
  - Visibility on public TeraGrid page

- Request a community account
  - Run jobs for others via your portal

- Staff support is available!

www.teragrid.org/gateways
Tremendous Opportunities Using the Largest Shared Resources - Challenges too!

What’s different when the resource doesn’t belong just to me?
- Resource discovery
- Accounting
- Security
- Proposal-based requests for resources (peer-reviewed access)
  - Code scaling and performance numbers
  - Justification of resources
  - Gateway citations

Tremendous benefits at the high end, but even more work for the developers

Potential impact on science is huge
- Small number of developers can impact thousands of scientists

But need a way to train and fund those developers and provide them with appropriate tools
Gateways can further investments in other projects

• Increase access
  – To instruments, expensive data collections, we’ll see an example today

• Increase capabilities
  – To analyze data, we’ll see an example today

• Improve workforce development
  – Can prepare students to function in today’s cross-disciplinary world, we’ll see an example today

• Increase outreach

• Increase public awareness
  – Public sees value in investments in large facilities
  – Pew 2006 study indicates that half of all internet users have been to a site specializing in science
  – Those who seek out science information on the internet are more likely to believe that scientific pursuits have a positive impact on society

NBCR Summer Institute, August 4, 2009
Now for a look at a few non-bio gateways
Gateways in the marketplace
Kids control telescopes and share images

• “In seconds my computer screen was transformed into a live telescopic view”
  – “Slooh's users include newbies and professional astronomers in 70 countries”

• Observatories in the Canary Islands and Chile, Australia coming soon

• 5000 images/month since 2003

• Increases public support for investment in these facilities
Linked Environments for Atmospheric Discovery (LEAD)

- Providing tools that are needed to make accurate predictions of tornados and hurricanes
- Data exploration and Grid workflow

NBCR Summer Institute, August 4, 2009
Social Informatics Data Grid
Collaborative access to large, complex datasets

- SIDGrid is unique among social science data archive projects
  - Streaming data which change over time
    - Voice, video, images (e.g. fMRI), text, numerical (e.g. heart rate, eye movement)
  - Investigate multiple datasets, collected at different time scales, simultaneously
    - Large data requirements
    - Sophisticated analysis tools

http://www.ci.uchicago.edu/research/files/sidgrid.mov
Viewing multimodal data like a symphony conductor

• “Music-score” display and synchronized playback of video and audio files
  – Pitch tracks
  – Text
  – Head nods, pause, gesture references

• Central archive of multi-modal data, annotations, and analyses
  – Distributed annotation efforts by multiple researchers working on a common data set
    • History of updates

• Computational tools
  – Distributed acoustic analysis using Praat
  – Statistical analysis using R
  – Matrix computations using Matlab and Octave

Source: Studying Discourse and Dialog with SIDGrid, Levow, 2008
Future Work

• Web technologies change fast
  – Must be able to adapt quickly

• Gateways and gadgets
  – Gateway components incorporated into any social networking page
  – 75% of 18 to 24 year-olds have social networking websites

• iPhone apps?

• Web 3.0
  – Beyond social networking and sharing content
  – Standards and querying interfaces to programmatically share data across sites
    • Resource Description Framework (RDF), SPARQL

NBCR Summer Institute, August 4, 2009
Tremendous Potential for Gateways

• In only 17 years, the Web has fundamentally changed human communication
• Science Gateways can leverage this amazingly powerful tool to:
  – Transform the way scientists collaborate
  – Streamline conduct of science
  – Influence the public’s perception of science
• Reliability, trust, continuity are fundamental to truly change the conduct of science through the use of gateways
  – High end resources can have a profound impact
• The future is very exciting!

NBCR Summer Institute, August 4, 2009
Thank you for your attention!

Questions?

Nancy Wilkins-Diehr, wilkinsn@sdsc.edu
www.teragrid.org