Selecting and Using TeraGrid/XD Resources for Maximum Productivity with Hands-On Examples

Kim Dillman
Research Programmer
Purdue Campus Champion
Rosen Center for Advanced Computing
Purdue University

July 18, 2011

Introduction and Overview

- A Few Things to Know…
- Becoming Familiar with Your Project's Job and Data Requirements
- Understanding What's Available and How to Choose
- Examples of User/Resource Matching Scenarios
- Getting on the Machine
- Understanding Queues and Policies
- Getting the Best Performance
- Let's try an example…

TeraGrid™

A Few Things to Know…

Assumptions for this Session

- You already know the basics of what TeraGrid/XSEDE is
- You have attended or understand the basic concepts presented in the "TeraGrid New User Training"
- You either have an existing allocation or have users or your own projects that may benefit from obtaining an allocation (if you want to try running an example you can either use your own allocation or a temporary training account)

TeraGrid™

A Few Comments Before We Start

- TeraGrid is in the process of transitioning to a new program called XSEDE
- This tutorial was created primarily using the resources available under the TeraGrid program and site
- A few comments will be included at various times during this tutorial with reference to the newer XSEDE program
- This tutorial and documentation will be upgraded to include the newer XSEDE web pages, links, and tools at a later date when they are made available on the XSEDE program site.

Becoming Familiar with Your Project's Job and Data Requirements

Ask Some Questions About…(1)

- **What software do you need?**
  - Is it 3rd party
  - Does it require a license?
  - Is it open source or "free"?
  - Is it already installed on a TeraGrid system?
    - https://www.teragrid.org/web/user-support/software_search
  - Can you install it yourself or will you need assistance?
  - Does your primary code have other software dependencies?
  - Do you have access to the source to recompile it?
  - Are there benchmarks already available for it on TeraGrid systems?
    - https://www.teragrid.org/web/user-support/benchmark
  - How much disk space does it take to install it (if not already available on a TeraGrid system)?

TeraGrid™

Ask Some Questions About…(2)

- How does your software/code function?
  - Do you need to run benchmarks to determine performance and best resource choice?
  - Is it serial or parallel?
  - How well does is scale (maximum number of CPUs/nodes before performance degrades)?
  - Does the code produce substantial network traffic (affects choice of node interconnect technology)?
  - Does it have specific hardware architecture requirements?(i.e. only Intel 64, Cray, SGI, etc.)?
  - Does it have specific OS requirements (i.e. RedHat or SuSE Linux, Windows, etc.)?
  - Does it require a "shared memory" architecture (i.e. SGI Altix, UltraViolet, etc.)?

TeraGrid™

Ask Some Questions About...(3)

- **What are some characteristics of your particular jobs?**
  - How long will your jobs run (min/max/average)?
  - How many total job runs will you need to make?
  - How much memory per processor or node does it need?
  - How much disk space does each job required for input and output (node local or system shared)?
  - How much total disk space do you need to complete your project?
  - Do you need to archive your data output to another system/storage area?
  - How long (time) does your job output need to reside on the system where it was created (i.e. for further processing, input to other runs, etc.)?

TeraGrid™

Ask Some Questions About…(4)

- Does the research require "special" systems?
  - Does it need any special hardware such as GPUs?
- Would it be a good candidate for the Purdue Condor Pool?
  - Serial jobs?
  - Large number of serial jobs?
  - Workflow management?
  - Short job duration (under 2 hours)
  - Access to the source code for recompile by "condor_submit" for longer job runs (note exceptions in the Condor Appendix of this document)
  - Other… (see Condor Appendix)

TeraGrid™

Ask Some Questions About…(5)

- **Would it be a good candidate for the Open Science Grid (OSG)?**
  - Some of the same requirements as the Purdue Condor Pool
    - Serial jobs / workflow / job submission script is almost identical
  - Additional possible requirements for OSG (HTPC – High Throughput Parallel Computing)
    - Ensembles of Single node (up to 16 core) parallel jobs using MPI, OpenMP, Linda, etc
  - Different methods of data access for the above job types
    - Scratch, Shared (Apps and Data), High Capacity Distributed File System (local and WAN),
    - Access through local filesystem (fuse) and Grid protocols (GridFTP, SRM)
    - Different Solutions: Xrootd, Hadoop, etc.
  - Other… (see OSG Appendix for more details)

TeraGrid™

Understanding What's Available and How to Choose

Becoming familiar with the resources

- What types of resources are available?
  - Resource categories
  - Pre-production / Production
  - Resource specifics
- Where do I find general info?
  - The resource catalog link:
  - https://www.teragrid.org/web/user-support/resources
- Where do I find specific info?
  - Individual resource "User Guides" (click on the icon next to the resource in the resource catalog)
  - XSEDE with also have User Guides in a common format in the new User Portal

TeraGrid™

TeraGrid Resource Categories

- Compute
- Special
- Visualization
- Future (New) Systems
- Other
- Data
- Advanced Support for TeraGrid Applications - ASTA (will have a new name in XSEDE)

TeraGrid™

Available "compute" Resources

- **Compute/Visualization Resource catalog:**
  - https://www.teragrid.org/web/user-support/compute_resources
- **There are 3 major types of "compute" resources currently available to TeraGrid users:**
  - SMP
    - Symmetric MultiProcessing) A multiprocessing architecture in which multiple CPUs, residing in one cabinet, share the same memory. SMP systems provide scalability.
  - MPP (Massively Parallel Processors)
    - A multiprocessing architecture that uses up to thousands of processors. MPP systems use a different programming paradigm than the more common symmetric multiprocessing (SMP) systems used.
  - Cluster

TeraGrid™

Available SMP Systems

- **NCSA – Ember:**
  - SGI UltraViolet system
  - 4 UV 1000 systems with 384 cores (Intel Nehalem EX 6 core processors) and 2 TB of memory each
  - Primarily for moderate to large shared memory applications
- **PSC – Blacklight:**
  - SGI UltraViolet system
  - UV 1000 system with 4094 cores (Intel Nehalem 8 core processors) and 32 TB of memory
  - Primarily for large shared memory applications

TeraGrid™

Available MPP Systems

- **NICS – Kraken:**
  - Cray XT5 – 112,896 processors
  - Intended for highly scalable applications
  - Minimum Startup allocation request is 100,000 SUs(?)
  - Find out if users are really going to use this system before adding them or requesting a startup allocation
- **TACC – Ranger:**
  - Sun Constellation – 62,976 processors
  - Intended for codes scalable to thousands of cores

TeraGrid™

Available Cluster Systems

- **Purdue – Steele:**
  - Dell PowerEdge 1950 – 7,144 processors (1560 are available in the longest running production queue)
  - Suited for a wide range of serial and small/medium parallel jobs
  - Longest Wall Time (720 hours)
  - Primarily 16 GB memory nodes and 1 GigE Ethernet interconnect
  - Access for up to 4 hours to some nodes with 32 GB memory and SDR Infiniband interconntect
- **TACC – Lonestar4:**
  - Dell PowerEdge Westmere – 22,656 cores (2 6-core processers per node)
  - 24 GB memory per node (2 GB per core)
  - Intended primarily for applications scalable up to thousands of cores
  - Can run serials jobs in a special queue

TeraGrid™

Special Resources

- **NCSA – Lincoln:**
  - Dell PowerEdge 1950 / NVIDIA Tesla S1070
  - Intended for applications that can make use of heterogeneous processors (CPU and GPU)
- **Purdue – Condor (high throughput)**
  - Pool of over 27,000+ processors
  - Various Architectures and OS
  - Designed for high-throughput computing and is excellent for parameter sweeps, Monte Carlo simulations, and other serial applications
  - See Condor Appendix A for more details.
- **IU – Quarry:**
  - Used for web services hosting and Science Gateways

TeraGrid™

Special Resources (cont.)

- **SDSC – Dash:**
  - Intel Nehalem – 512 processors
  - Primarily used for data intensive computing
  - vSMP (virtual shared memory) software from ScaleMP that aggregates memory across 16 nodes. This allows applications to address 768GB of memory.
  - 4 TB of Flash memory configurable as fast file I/O subsystem or extended fast virtual memory swap space.
- **SDSC – Trestles:**
  - 10,368 cores (4 8-core AMD Magny-Cours processors per node)
  - 20 TB DRAM (64 GB per node / 2 GB per core)
  - 38 TB SSD Flash Memory (120 GB per node)
  - QDR Infiniband interconnect
  - Intended for moderately scalable parallel applications and fast local I/O requirements

TeraGrid™

Visualization Resources

- **TACC – Longhorn:**
  - Dell/NVIDIA Visualization and Data Analysis Cluster
  - A hybrid CPU/GPU system designed for remote, interactive visualization and data analysis, but it also supports production, compute-intensive calculations on both the CPUs and GPUs via off-hour queues
- **TACC – Spur:**
  - Sun Visualization Cluster
  - 128 compute cores / 32 NVIDIA FX5600 GPUs
  - Spur is intended for serial and parallel visualization applications that take advantage of large per-node memory, multiple computing cores, and multiple graphics processors.

TeraGrid™

Visualization Resources (cont.)

- **NICS – Nautilus:**
  - Primarily used for three tasks
  - Visualizing data results from computer simulations with many complex variables
  - Analyzing large amounts of data coming from experimental facilities
  - Aggregating and interpreting input from a large number of sensors distributed over a wide geographic region
  - SGI UltraViolet
  - 1024 cores (Intel Nehalem)
  - 4 TB global shared memory

TeraGrid™

New Systems (Coming Soon)

- **NCSA – Forge**
  - Dell PowerEdge C6145 Quad Socket nodes (dual 8-core AMD Magny-Cour processors)
  - Each node supports 8 NVIDIA Fermi N2070 GPUs
  - Intended for applications that make use of GPGPUs
- **SDSC – Gordon**
  - Consists of a "compute cluster" and separate I/O nodes
  - 1024 dual socket compute nodes, 64 I/O nodes
  - Dual rail, QDR Infiniband, 3D torus interconnect
  - Designed for data intensive computer that spans domains such as genomics, graph problems, geophysics, and data mining.
  - Large memory super-nodes are ideal for serial or threaded applications that require a large amount of memory
  - Flash based I/O may provide significant performance increases for certain applications

TeraGrid™

Other Systems

- **FutureGrid: A grid testbed (IndianaU and many other partners)**
  - A collections of different systems
  - Not currently allocated via standard TeraGrid POPS allocation system(?)
  - More info:
    - **http://futuregrid.org/**
- **MATLAB on the TeraGrid (NO LONGER AVAILABLE?)**
  - Update: The request access link indicates that this system is no longer available for requests
  - Newer system from Cornell
  - Not currently allocated via standard TeraGrid POPS allocation system
  - More info:
    - **http://www.cac.cornell.edu/matlab/**
  - How to request access:
    - **http://www.cac.cornell.edu/matlab/status/interest.aspx**

TeraGrid™

OSG as an XSEDE Provider

and More …

OSG as an XSEDE Provider

- Currently a work in progress
  - Access to OSG resources via XSEDE Gateways
  - Will eventually be seamless
  - Including access from OSG interfaces to XSEDE
- OSG brings additional capabilities to your campus:
  - Create your own Campus Grid by federating local clusters
  - Single submit model to local campus resources and the OSG, and XSEDE
- First, lets talk about the unique capabilities of the OSG

The two familiar Models

- Capability Computing (HPC)
  - A few jobs parallelized over the whole system
  - Use whatever parallel s/w versions the sysadmin has installed
    - (MPI, OpenMP, Linda …)
- High Throughput Computing (HTC)
  - Run ensembles of single core jobs
  - The OSG is focused on serving HTC users

HTC is also what Campus Grids are about

- HTPC – High Throughput Parallel Computing (A hybrid model on the OSG)

Ensembles of small-way parallel jobs (single node, shared memory, up to 16 cores)

(10's – 1000's of jobs)

Use whatever parallel s/w (e.g. MPI, Linda) you want ☺

(It ships with the job)

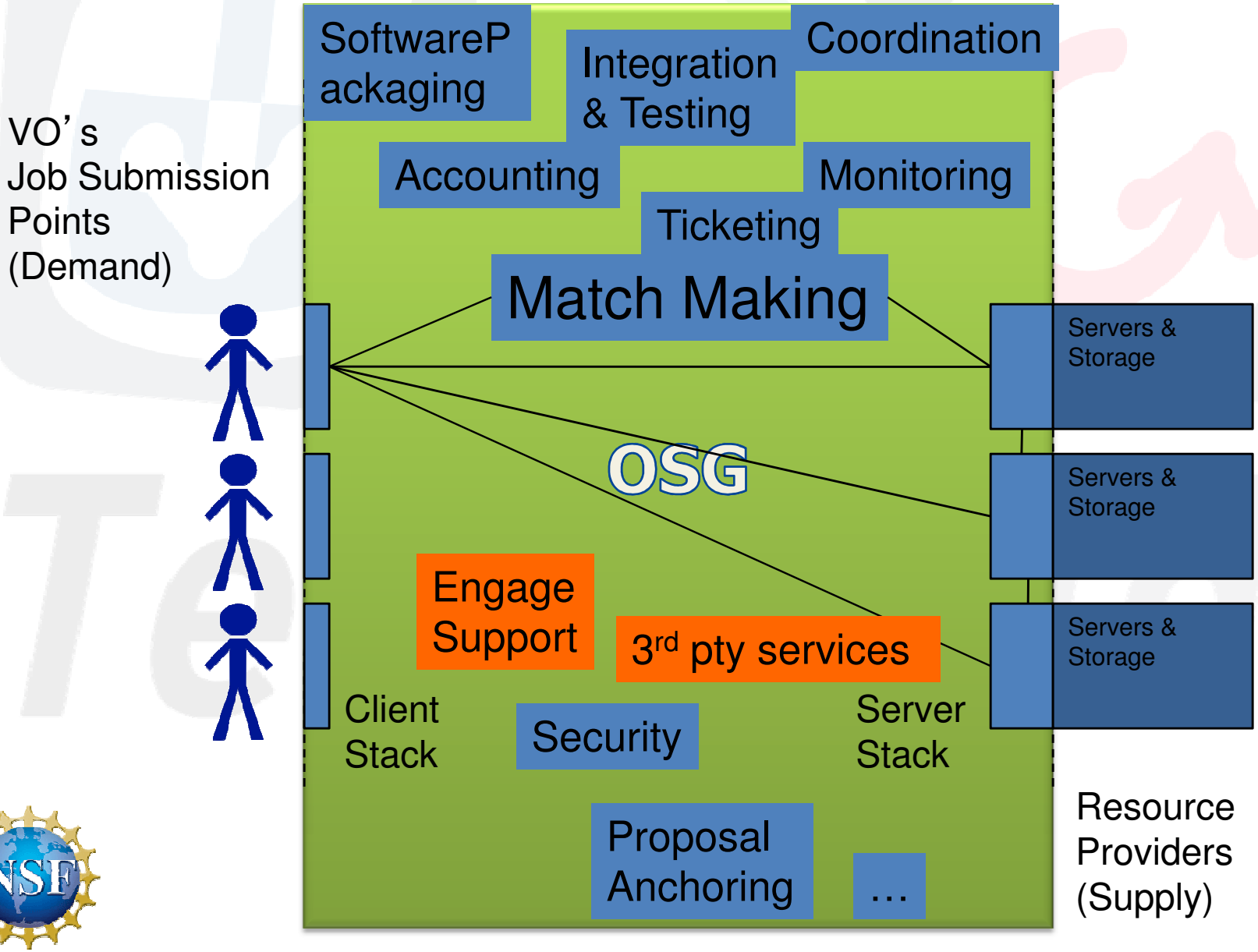
TeraGrid™

Making sense of the OSG

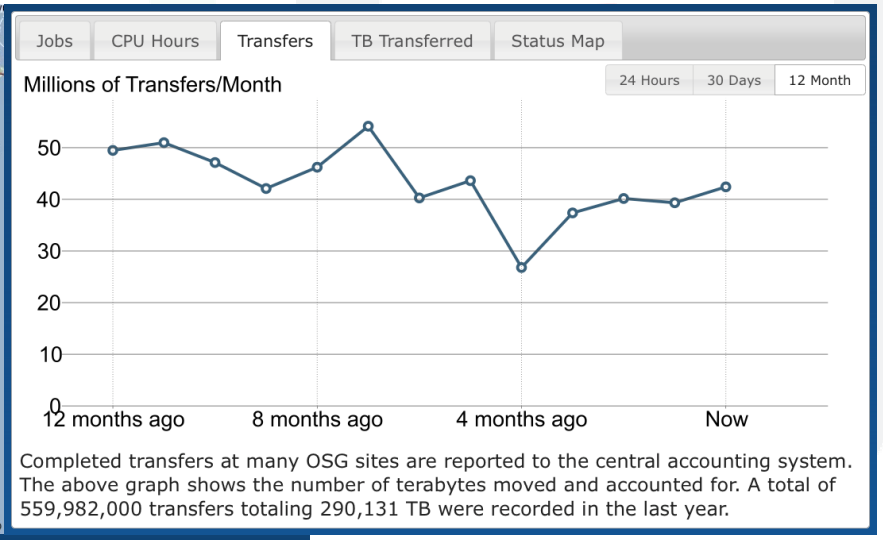
- OSG operates as a grid of distributed systems
  - Currently 70+ sites (120+ Resources)
  - Users access the "submit host" but do *not* login to any of the resources
  - Jobs are distributed across one or more resources
  - Heterogeneity of the resources is transparent to the user (by design)
  - There are no "allocations" on the OSG
    - Resources are optimistic
- OSG = Technology + Process + Sociology

TeraGrid™

The OSG is like eBay

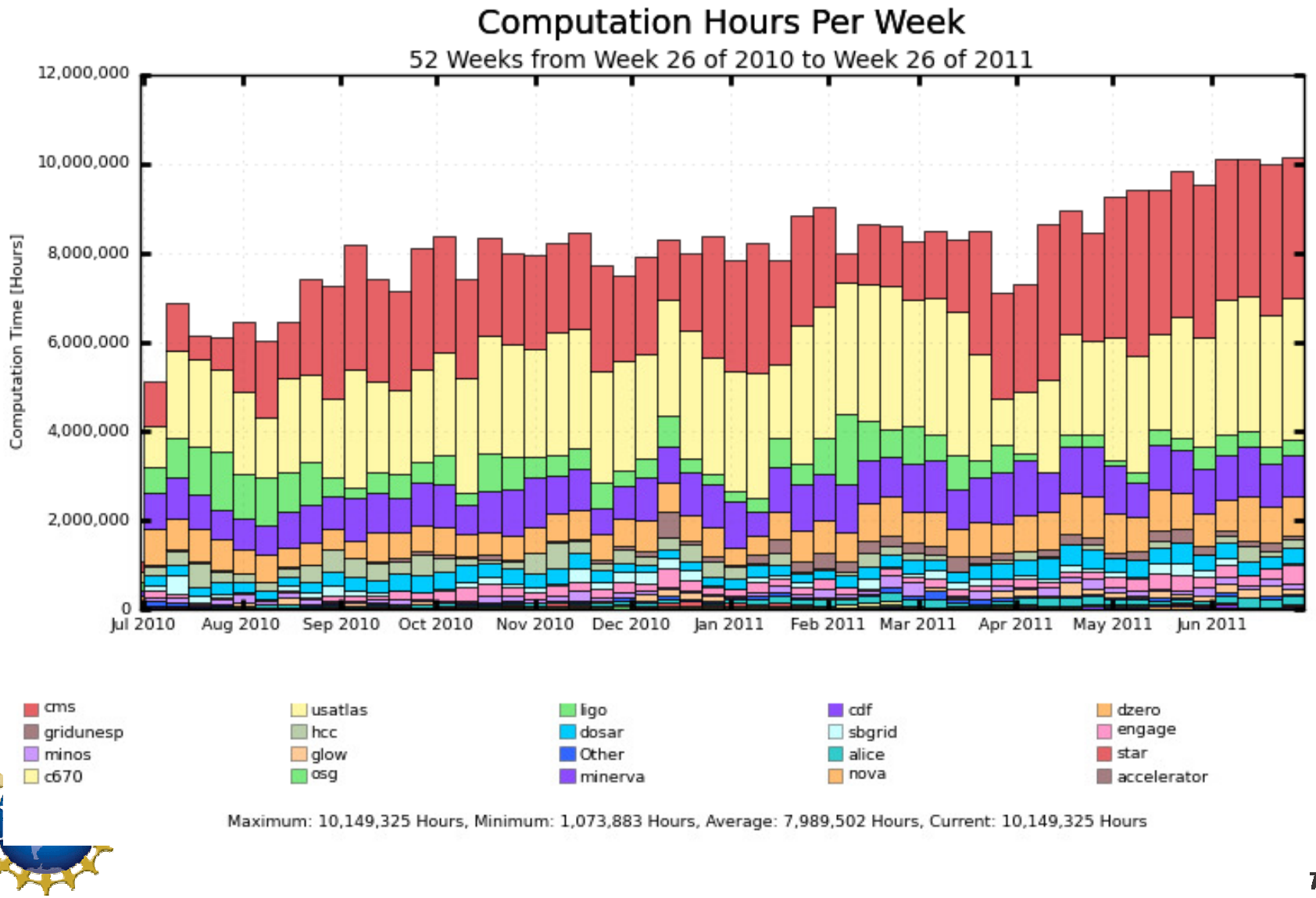


The OSG Now



http://display.grid.iu.edu/

Now > 10M hours / week



TeraGrid™

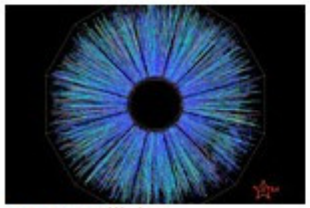
Science on the OSG Today

- Astrophysics
- Biochemistry
- Bioinformatics
- Earthquake Engineering
- Genetics
- Gravitational-wave physics
- Mathematics
- Nanotechnology
- Nuclear and particle physics
- Text mining
- And more…

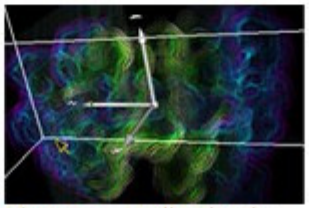
ATLAS Detector
Copyright CERN
Permission Information

SDSS Telescope
Image Credit Fermilab
Permission Information

CDMS photo
Image Credit Fermilab
Permission Information



STAR Collision
Image Credit Brookhaven National Laboratory/STAR Collaboration
Permission Information



BioMOCA Application in nanoHUB
Image Credit Shawn Rice, Purdue University
Permission Information

CMS Detector
Copyright CERN
Permission Information

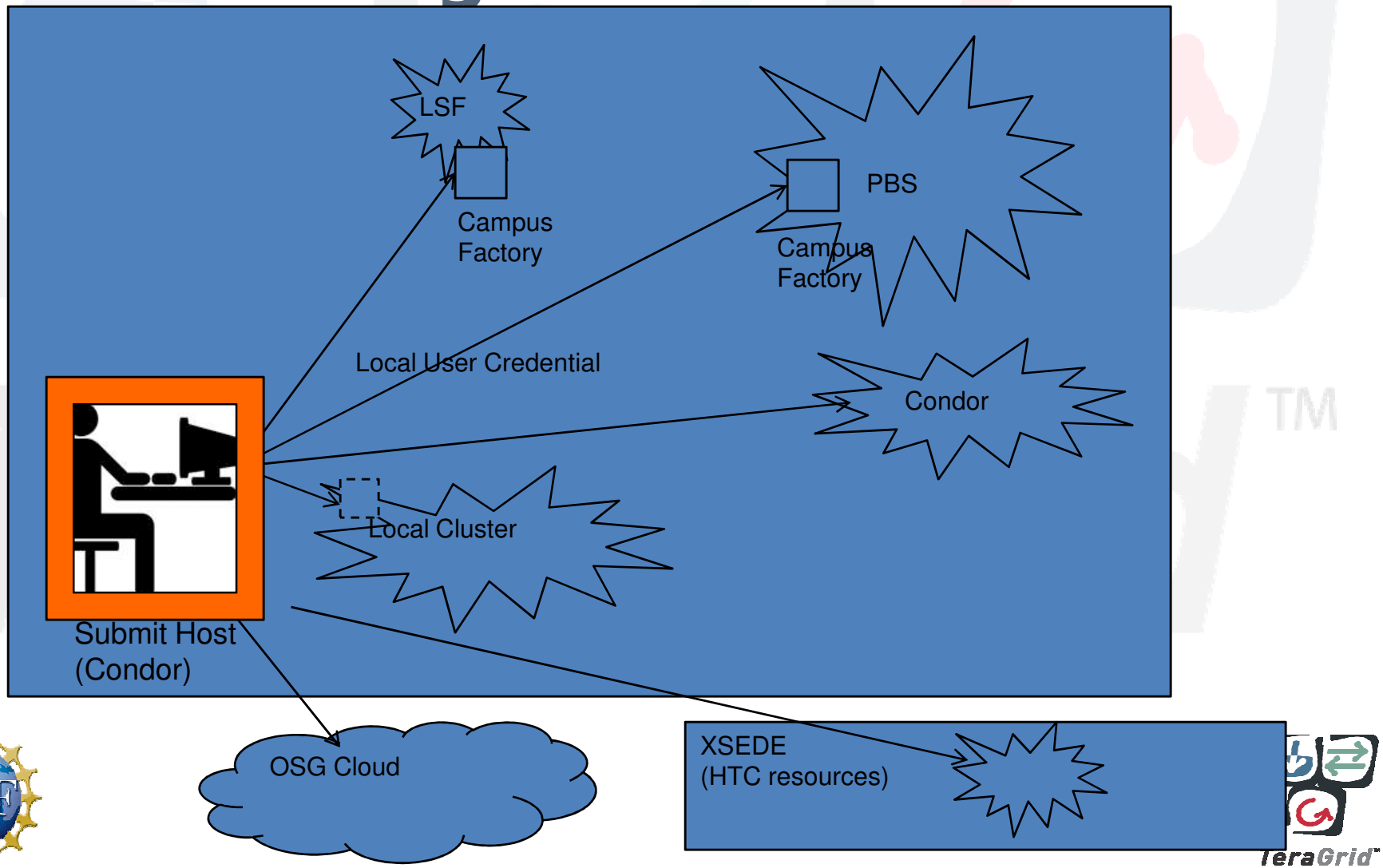
Auger photo
Image Credit Pierre Auger Observatory
Permission Information

MiniBooNE photo
Image Credit Fermilab
Permission Information

DZero Detector
Image Credit Fermilab
Permission Information

TeraGrid

OSG Enabling a Campus Grid: All Resources Can be Accessed Via a Single Interface



Campus Grids

- Can be easily set up on your campus
- Gives access to local resources (federated) through a single interface
- Also can submit through the same interface to OSG and XSEDE (HTC) resources

TeraGrid™

Submit File for OSG (Campus Grid)

```
Universe = vanilla
#notify_user = <user email address>
Executable = serial64ff
#transfer_executable = false
# Files (and directory) in the submit host: log, stdout, stderr, stdin, other files
InitialDir = run_1
Log = serial64ff.log
Output = serial64ff.stdout
Error = serial64ff.stderr
#input = serial64.in
fetch_files=serial64ff.in
#Arguments = <arg1><arg2><argn>
should_transfer_files = IF_NEEDED
when_to_transfer_output = ON_EXIT

# Required for access to XSEDE resources
+TGProject = "TG-XXXNNNNNNN"
# Required for access to OSG resources via (GlideinWMS)
requirements = IS_GLIDEIN == True
x509userproxy=/tmp/x509up_u20003
Queue
```

TeraGrid™

Available "data" resources in OSG (Data is the tricky part)

- Local file system
  - Well defined "scratch" space
  - Available during the execution of the job
- Shared file system
  - Opportunistic use
  - Well defined areas for applications and data (OSG_APP, OSG_DATA)
  - Performance varies (GPFS, Lustre, NFS)
- High capacity Distributed File System (local and WAN)
  - Access through local file system (fuse) and Grid protocols (GridFTP, SRM)
  - Different solutions: Xrootd, Hadoop, …
- TeraGrid Data Replication Service (Access to
  - Access through iRODS command line and other compatible clients

Submit to Run

- No difference from regular Condor job submission
- All Condor data movement mechanisms work as well
- condor_submit <submit_file>

OSG Summary

- Make sure your application is suited to the "high throughput computing" paradigm
- OSG usage is "opportunistic"
  - No allocations necessary
- OSG also brings the ability to federate resources on your own campus
- XSEDE Gateway access to the OSG is a work in progress.
  - Should eventually be transparent
- Data is tricky – work with the experts

For questions of more information, contact:

Marco Mambelli
marco@hep.uchicago.edu

TeraGrid™

Available "data" Resources

- **Data Resource Catalog:**
  - https://www.teragrid.org/web/user-support/data_resources
- **Archive systems**
  - Use tapes for long-term storage
  - Very high capacity (petabytes or tens of petabytes)
- **Wide-Area and Global file systems**
  - Extension of parallel file systems over wide-area networks
  - One file system, available on multiple sites/resources
- **Data Replication Service**
  - A distributed replication and data management service, utilizing archive and disk systems at mulitple sites (see Access). All data is replicated across two separate storage systems.

TeraGrid™

Check out the available resource specifications (1)

- **Compute and Visualization List:**
  - https://www.teragrid.org/web/user-support/compute_resources

User Guide

Resource Name Platform	Description	Specifications
IU		
Big Red IBM e1350 This resource will be available to allocated users of TeraGrid through 2011-03-31.	Big Red has 768 IBM JS21 compute nodes, each with two dual-core 2.5 GHz PowerPC 970MP CPUs, 8 GB memory, 72 GB of local scratch disk, and a PCI-X Myrinet-2000 interconnect for high-bandwidth low-latency MPI applications. It has access to 266 TB of local GPFS scratch space, the TeraGrid-wide GPFS-WAN file system, and to the 535 TB Lustre file system provided by the Data Capacitor. NOTE: 6.5 TFlops is available for TeraGrid usage. **Recommended Use** Big Red is a distributed shared-memory cluster, intended to run parallel as well as serial applications. **Status** In production and accepting allocation requests **Startup Allocation Limit** 30000 SUs	**Machine Type** MPP **Operating System** SuSE Linux Enterprise Server 9 **Teraflops** 30.7 **Disk Size** 266 TB
Quarry Dell AMD	Quarry Gateway Web Services Hosting consists of multiple Dell AMD systems geographically distributed for failover. Each system has at least 8 cores and 32 GB of memory. Persistent storage is available by utilizing IU NFS home directories with a 10GB default quota or the 335TB Data Capacitor WAN (Lustre) file system. The system utilizes OpenVZ to provide virtual hosting of RPM-based Linux distributions. The host operating system is Red Hat Enterprise Linux. **Recommended Use** This machine is used for hosting Scientific Gateway and Web Service allocations. The Quarry resource is restricted to members of approved XRAC grants which have a web service component. **Status** In production and accepting allocation requests **Startup Allocation Limit**	**Machine Type** SMP **Operating System** RedHat Enterprise Linux Server **Teraflops** 0 **Disk Size** 335 TB

that PIs request some time on resources that will be continuing beyond March 2010.

- **New TeraGrid users with experience with comparable architectures elsewhere:** please describe your previous experience in your resource request.
- **Users new to and without prior experience on these systems and architectures:** we strongly encourage you to select resources that have an end date at least 12 months from the expected allocation start. If not, the TeraGrid and/or TRAC may recommend alternate systems.
- **Users wishing to use Roaming:** Due to the retirement in March of resources that are in the TG-Wide Roaming Access pool, the Roaming Access resource has been discontinued. Please select the set of individual resources you may need, with an appropriate allocation request for each. You can use a POPS Transfer Request to rebalance your allocation distribution later on in the allocation period.

See the XD Transition page for more information and links to help.

LEGEND

Availability Status Indicators

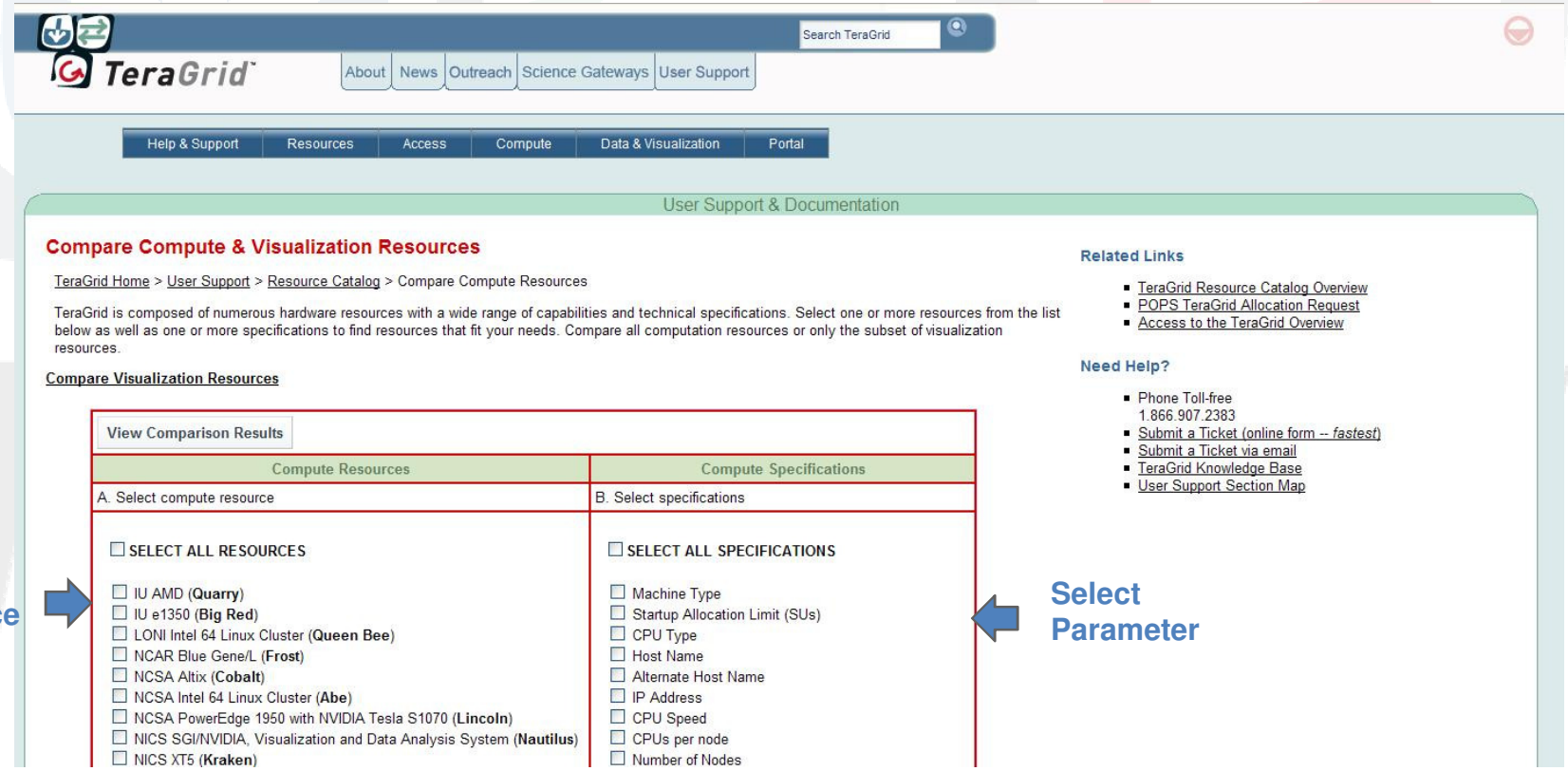
- Future Resource
- Pre-production: Available for Allocations
- In Production + Allocable
- In Production / No Future Allocations
- Decommissioned

Resource Details -- click on icons in resource display to link to resource-specific documentation

User Guide icon User Guide
User News icon Resource User News
Applications icon Applications

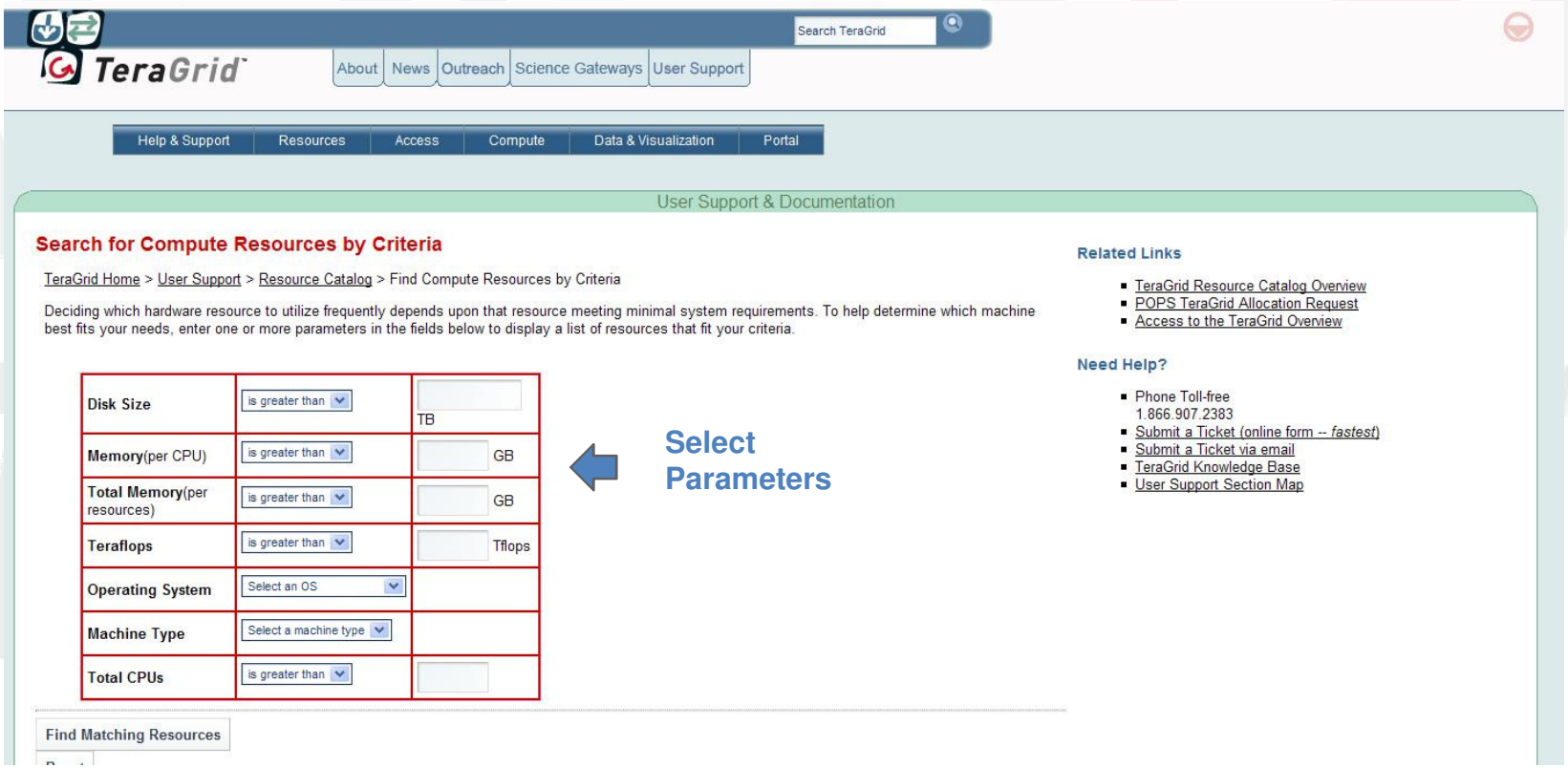
Check out the available resource specifications (2)

- **Compute and Visualization Compare:**
  - https://www.teragrid.org/web/user-support/compare_compute



Check out the available resource specifications (3)

- **Compute and Visualization Search:**
  - https://www.teragrid.org/web/user-support/resource_search



Check out the available resource specifications (4)

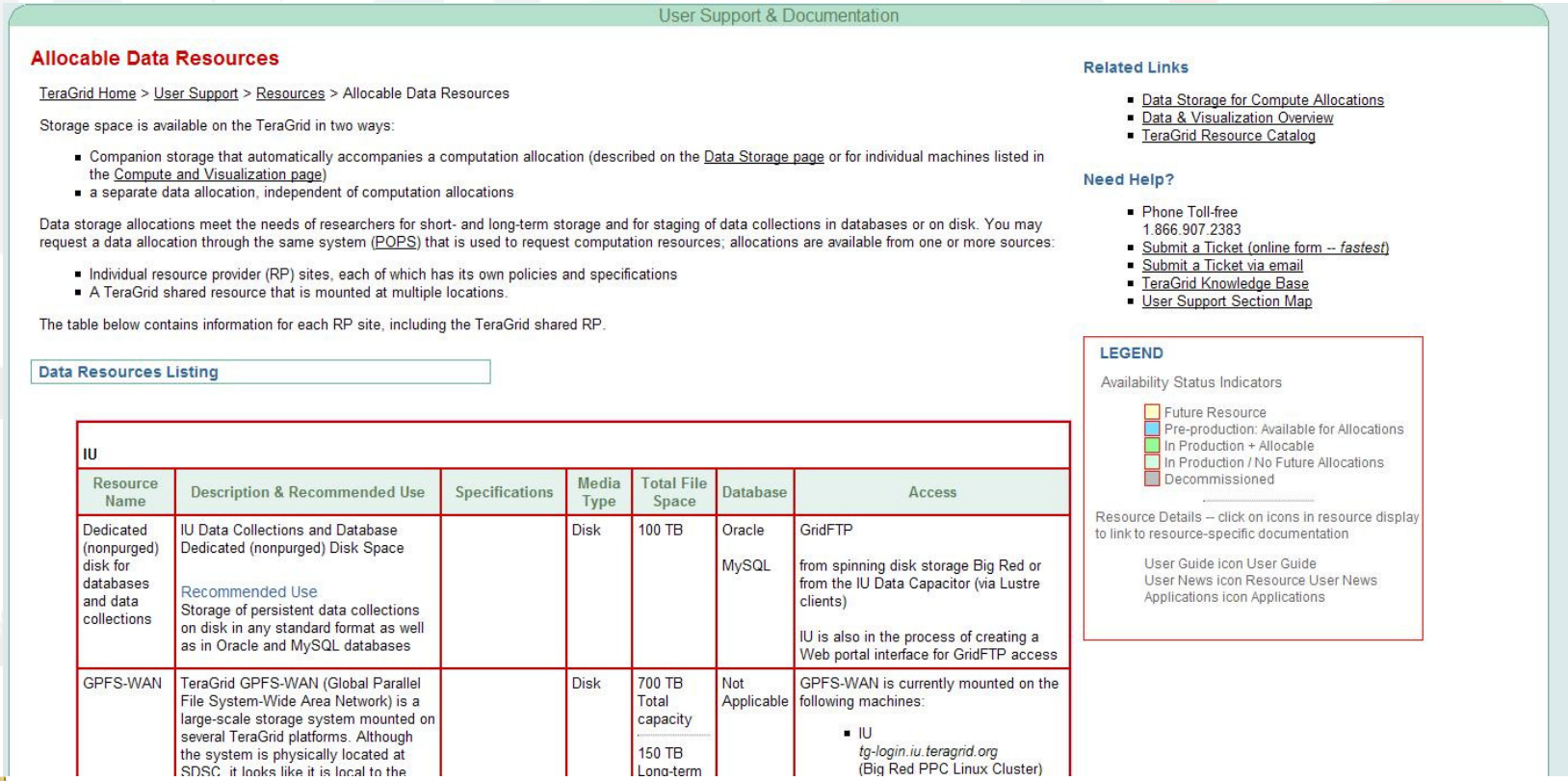
- **Local Data Storage List:**
  - https://www.teragrid.org/web/user-support/storage

Site	File System	Type	Quota	Purge	Backup
IU	**Big Red**				
	HOME (/N/u/<username>/BigRed)	NFS	10 GB	none	nightly
	TG_CLUSTER_PFS	GPFS		60 days*	None; user can get account on IU HPSS for backup
	TG_CLUSTER_GPFS	GPFS			
	TG_CLUSTER_SCRATCH (/N/gpfsbr/<username>)	GPFS			
	TG_NODE_SCRATCH (/scratch)	Local scratch		60 days*	None; user can get account on IU HPSS for backup
	TG_COMMUNITY (/N/comm)	NFS	As requested		
	HPSS	archival	5 TB	none	dual copy
	IA-64 Linux Cluster				
	TG_CLUSTER_HOME		5 GB	never	nightly
	TG_CLUSTER_PFS (scratch)	GPFS	none	Files not accessed in 5 days	none
	TG_NODE_SCRATCH (Node local scratch)		none	files are automatically deleted after batch job is completed	none
	TG_CLUSTER_SCRATCH (NFS Scratch)	NFS	none	Time since last access: File Size: Removed after <= 25 MB: 14 days > 25 GB: 5 days	none
	SGI Altix				

TeraGrid™

Check out the available resource specifications (5)

- **Data Resource List:**
  - https://www.teragrid.org/web/user-support/data_resources



User Support & Documentation

Allocable Data Resources

TeraGrid Home > User Support > Resources > Allocable Data Resources

Storage space is available on the TeraGrid in two ways:

- Companion storage that automatically accompanies a computation allocation (described on the Data Storage page or for individual machines listed in the Compute and Visualization page)
- a separate data allocation, independent of computation allocations

Data storage allocations meet the needs of researchers for short- and long-term storage and for staging of data collections in databases or on disk. You may request a data allocation through the same system (POPS) that is used to request computation resources; allocations are available from one or more sources:

- Individual resource provider (RP) sites, each of which has its own policies and specifications
- A TeraGrid shared resource that is mounted at multiple locations.

The table below contains information for each RP site, including the TeraGrid shared RP.

Data Resources Listing

IU

Resource Name	Description & Recommended Use	Specifications	Media Type	Total File Space	Database	Access
Dedicated (nonpurged) disk for databases and data collections	IU Data Collections and Database Dedicated (nonpurged) Disk Space Recommended Use Storage of persistent data collections on disk in any standard format as well as in Oracle and MySQL databases		Disk	100 TB	Oracle MySQL	GridFTP from spinning disk storage Big Red or from the IU Data Capacitor (via Lustre clients) IU is also in the process of creating a Web portal interface for GridFTP access
GPFS-WAN	TeraGrid GPFS-WAN (Global Parallel File System-Wide Area Network) is a large-scale storage system mounted on several TeraGrid platforms. Although the system is physically located at SDSC, it looks like it is local to the		Disk	700 TB Total capacity 150 TB Long-term	Not Applicable	GPFS-WAN is currently mounted on the following machines: ▪ IU *tg-login.iu.teragrid.org* (Big Red PPC Linux Cluster)

Related Links

- Data Storage for Compute Allocations
- Data & Visualization Overview
- TeraGrid Resource Catalog

Need Help?

- Phone Toll-free
  1.866.907.2383
- Submit a Ticket (online form -- *fastest*)
- Submit a Ticket via email
- TeraGrid Knowledge Base
- User Support Section Map

LEGEND

Availability Status Indicators

- Future Resource
- Pre-production: Available for Allocations
- In Production + Allocable
- In Production / No Future Allocations
- Decommissioned

Resource Details -- click on icons in resource display to link to resource-specific documentation

User Guide icon User Guide
User News icon Resource User News
Applications icon Applications

Campus Champions Resource Selection Tool (coming soon)

- **Resource Selection Spreadsheet - Summary:**
  - Currently under revision and will move to new location soon

	A	B	C	D	E	F	G	H	I	J	K	L	
1	Startup SUs	Resource	CoA?	Status	Type	Institution	Machine	Available End Date	OS	Tflops	Disk	Comments	User Guide URL
2	30000	Big Red		Prod	MPP	IU	IB e1350	3/31/2011	SuSE LES 9	30.7	266 TB		http://racinfo.indiana.e
3	1	Quarry		Prod	Web	IU	Dell AMD		RH REL	0	336 TB		http://rtinfo.uits.iu.edu
4	50000	Queen Bee	AQSL	Prod	Cluster	LONI	Dell Intel 64	7/31/2011	RH REL 4	50.7	192 TB		http://www.loni.org/te
5	50000	Frost		Prod	MPP	NCAR	IBM Blue Gene/L	4/1/2011	SuSE LES 9	22.936	110 TB		http://www2.cisl.ucar.e
6	50000	Abe	AQSL	Prod	Cluster	NCSA	Dell Intel 64		RH REL 4	89.47	400 TB		http://www.ncsa.illinoi
7	30000	Cobalt		Prod	SMP	NCSA	SGI Altix		SGI ProPack 5	8.2	100 TB	will be replace by Ember	http://www.ncsa.illinoi
8	30000	Lincoln		Prod	Special	NCSA	Dell/Intel PowerEdge 1950 with NVIDIA Tesla S1070		RH REL 4	47.5	400 TB		http://www.ncsa.illinoi
9	200000	Kraken		Prod	MPP	NICS	Cray XT5		CLE 2.2 UP01	1030	2400 TB	minimum startup is 100000	http://www.nics.tenne
10	3000	Nautilus		Pre-Prod	SMP	NICS	SGI UV/NVIDIA		SuSE v11	0	960 TB	Available July 2010?	N/A
11	30000	NSTG		Prod	Cluster	ORNL	IBM IA-32	3/31/2011	Sci Linux 5.2	0.34	2.14 TB		http://www.ornl.teragr
12	30000	Pople		NA	SMP	PSC	SGI Altix 4700		SuSE Linux	5	150 TB		http://www.psc.edu/m
13	200000	Condor Pool		Prod	Special	Purdue	Various		Various	60	170 TB		http://www.rcac.purdu http://www.rcac.purdu
14	50000	Steele	AQSL	Prod	Cluster	Purdue	Dell 1950		RH REL 5	66.59	130 TB		http://www.rcac.purdu http://www.rcac.purdu
15	200000	TeraDRE		Prod	Special	Purdue	Nvidia GeForce 6600 GT		Various	60	170 TB		http://teradre.rcac.pur
16	30000	Dash		Prod	Special	SDSC	Appro		CentOS	4.9	75 TB		http://www.sdsc.edu/u
17	50000	Lonestar	AQSL	Prod	Cluster	TACC	Dell PowerEdge 1955		CentOS	62.16	106.5 TB		http://services.tacc.ute
18	3000	Longhorn		Prod	Special	TACC	Dell/Nvidia		RH REL 5	20.7	18.7 TB		http://services.tacc.ute
19	200000	Ranger		Prod	MPP	TACC	Sun Constellation		CentOS	579	1700 TB		http://services.tacc.ute
20	1000	Spur		Prod	Special	TACC	Vis Cluster		RH REL 4	1.13	1730 TB		http://services.tacc.ute

TeraGrid™

Campus Champions Resource Selection Tool (coming soon)

- **Resource Selection Spreadsheet – All Queues:**
  - Currently under revision and will move to new location soon

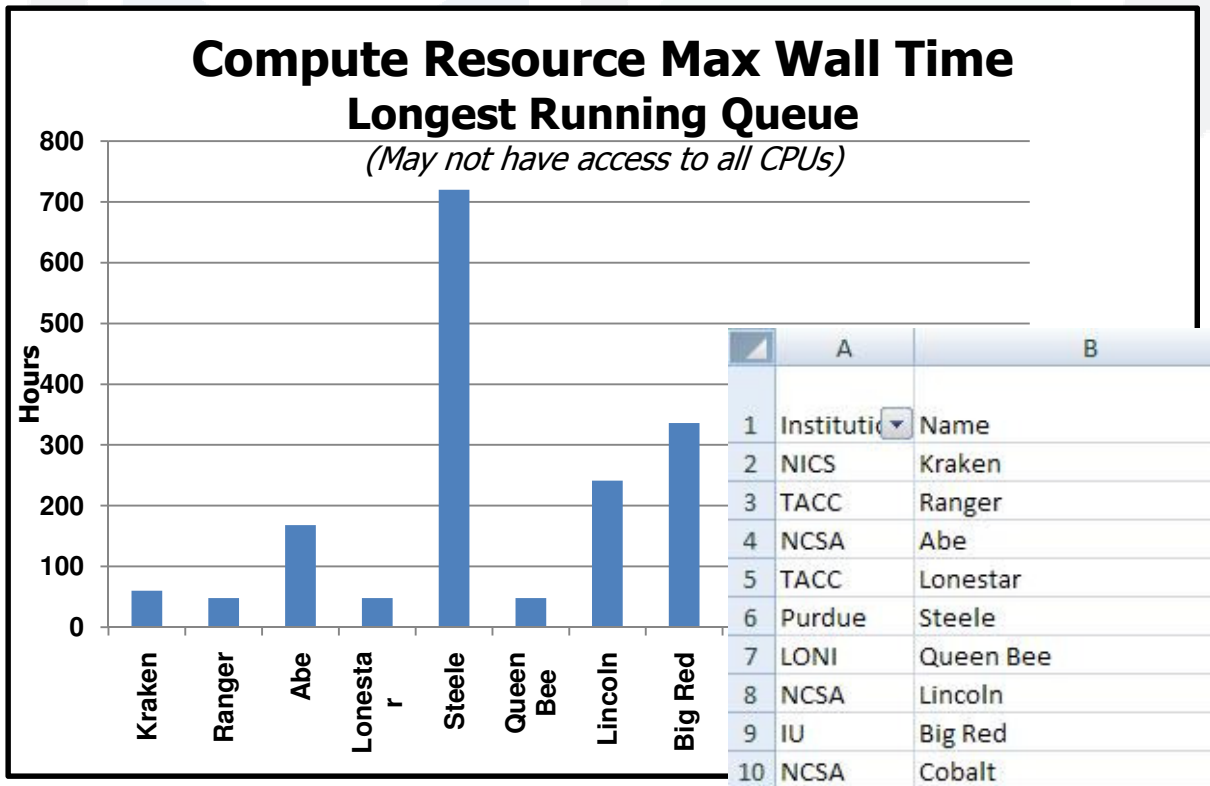
	A	B	C	D	E	F	G	H
1	Name	Institutio	System	Roamin	Queue	Wall Ti	CPUs	URL
13	Abe	NCSA	Dell Intel 64 Linux Cluster	Y	debug	30	128	http://www.ncsa.uiuc.edu/UserInfo/Resources/Hardware/Intel64Cluster/Doc/Jobs.html#sched
14	Abe	NCSA	Dell Intel 64 Linux Cluster	Y	normal	48	4800	http://www.ncsa.uiuc.edu/UserInfo/Resources/Hardware/Intel64Cluster/Doc/Jobs.html#sched
15	Abe	NCSA	Dell Intel 64 Linux Cluster	Y	wide	48	9568	http://www.ncsa.uiuc.edu/UserInfo/Resources/Hardware/Intel64Cluster/Doc/Jobs.html#sched
16	Abe	NCSA	Dell Intel 64 Linux Cluster	Y	long	168	4800	http://www.ncsa.uiuc.edu/UserInfo/Resources/Hardware/Intel64Cluster/Doc/Jobs.html#sched
17	Lonestar	TACC	Dell PowerEdge Linux Cluster	Y	serial	12	1	http://services.tacc.utexas.edu/index.php/lonestar-user-guide
18	Lonestar	TACC	Dell PowerEdge Linux Cluster	Y	normal	48	512	http://services.tacc.utexas.edu/index.php/lonestar-user-guide
19	Lonestar	TACC	Dell PowerEdge Linux Cluster	Y	high	48	512	http://services.tacc.utexas.edu/index.php/lonestar-user-guide
21	Lonestar	TACC	Dell PowerEdge Linux Cluster	Y	development	0.5	16	http://services.tacc.utexas.edu/index.php/lonestar-user-guide
22	Steele	Purdue	Dell Intel 64 Linux Cluster	Y	standby	4	7144	http://www.rcac.purdue.edu/userinfo/resources/steele/userguide.cfm#run_pbs_queues
23	Steele	Purdue	Dell Intel 64 Linux Cluster	Y	standby-8	8	400	http://www.rcac.purdue.edu/userinfo/resources/steele/userguide.cfm#run_pbs_queues
24	Steele	Purdue	Dell Intel 64 Linux Cluster	Y	tg_workq	720	1600	http://www.rcac.purdue.edu/userinfo/resources/steele/userguide.cfm#run_pbs_queues
25	Queen Bee	LONI	Dell Intel 64 Linux Cluster	Y	workq	48	2048	http://www.loni.org/teragrid/users_guide.php#queue_limits
51								
52								
53								
54								
55								

MaxWall | Policies

TeraGrid™

Campus Champions Resource Selection Tool (coming soon)

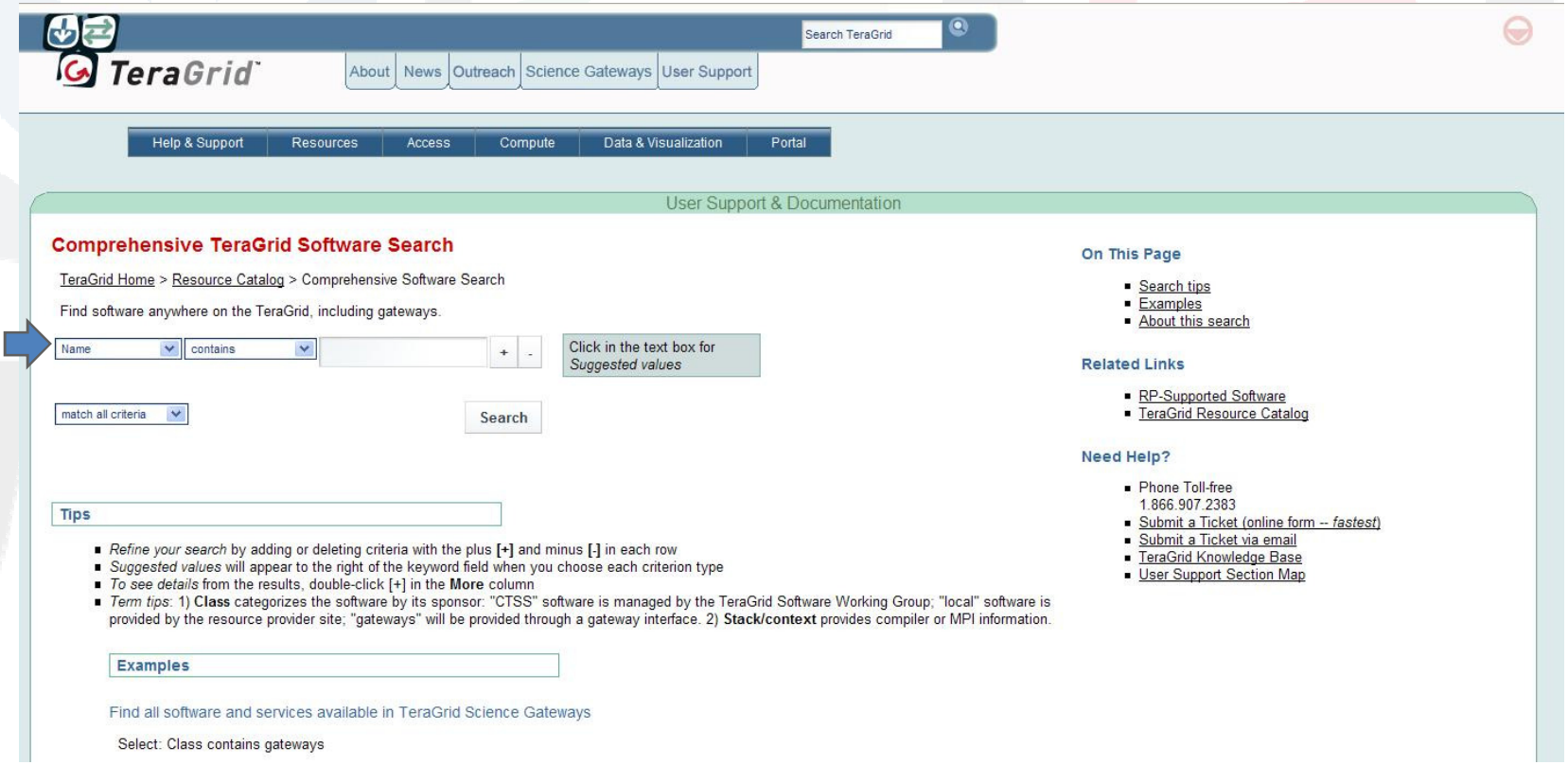
- **Resource Selection Spreadsheet - MaxWall:**
  - Currently under revision and will move to new location soon



	A	B	C	D	E
1	Institution	Name	Max Wall Time	Max CP	
2	NICS	Kraken	60	256	1
3	TACC	Ranger	48	1024	1
4	NCSA	Abe	168	4800	1
5	TACC	Lonestar	48	512	1
6	Purdue	Steele	720	1600	1
7	LONI	Queen Bee	48	2048	1
8	NCSA	Lincoln	241	1536	1
9	IU	Big Red	336	128	1
10	NCSA	Cobalt	200	128	1
11	NCAR	Frost	24	1024	1
12	PSC	Pople	168	36	1
13	ORNL	NSTG	168	28	1

Check out the available software

- **Software Search:**
  - https://www.teragrid.org/web/user-support/software_search



Examples of User/Resource Matching Scenarios

User "A"

- **Problem Description**
  - **Code :**
    - **Is open source**
    - **Is parallel**
    - **Is not currently installed anywhere on TeraGrid systems**
    - **Does not scale too well across nodes**
    - **Has several other code dependencies**
  - **Job requirements are:**
    - **Many runs (small parameter study)**
    - **Somewhat memory intensive**
    - **May run much longer than 48 hours**
- **Solution**
  - **Steele and QueenBee for benchmarking (newer choice would be Lonestar or Trestles instead of QueenBee)**

TeraGrid™

User "B"

- **Problem Description**
  - Code :
    - Is open source
    - Is parallel
    - Is currently installed on at least a few TeraGrid systems
    - Scales fairly well if the node interconnect is fast
    - Has benchmarks available for some of the TeraGrid systems (although not for specific module and usage scenario)
  - Job requirements are:
    - A few runs
    - Rather large number of interations/time steps per job
  - Benchmarks on specific problem were run with Campus Champions assistance
    - Steele was too slow due to the interconnect technology
    - Kraken was much better
- **Solution**
  - Kraken was chosen for startup allocation request for production runs

TeraGrid™

User "C"

- **Problem Description**
  - Code :
    - Is custom
    - Is serial
    - Can be statically compiled into a single executable
    - Can be compiled for different OS/Architecture combinations
  - Job requirements are:
    - A VERY large number of runs (over 500,000)
    - Each job runs in about an hour or two
    - Jobs are similar except for the input file and some parameters
- **Solution**
  - Condor was chosen due to the serial nature of the code, the large number of jobs to be run, the ability to compile the code into a single executable on multiple OS/Architecture combinations, and the relatively short execution time (This would also be suited for OSG)

TeraGrid™

User "D"

- **Problem Description**
  - **Code :**
    - **Is custom**
    - **Is serial**
    - **Can be statically compiled into a single executable**
  - **Job requirements are:**
    - **Very few runs**
    - **Runs between 24-48 hours**
    - **Requires a very large amount of memory (>80 GB)**
- **Solution**
  - **Cobalt and Pople were chosen since the code is serial but requires a large amount of "shared" memory (Newer choices would be Ember and Blacklight)**

TeraGrid™

Getting on the Machine

Common methods to ALL TeraGrid systems

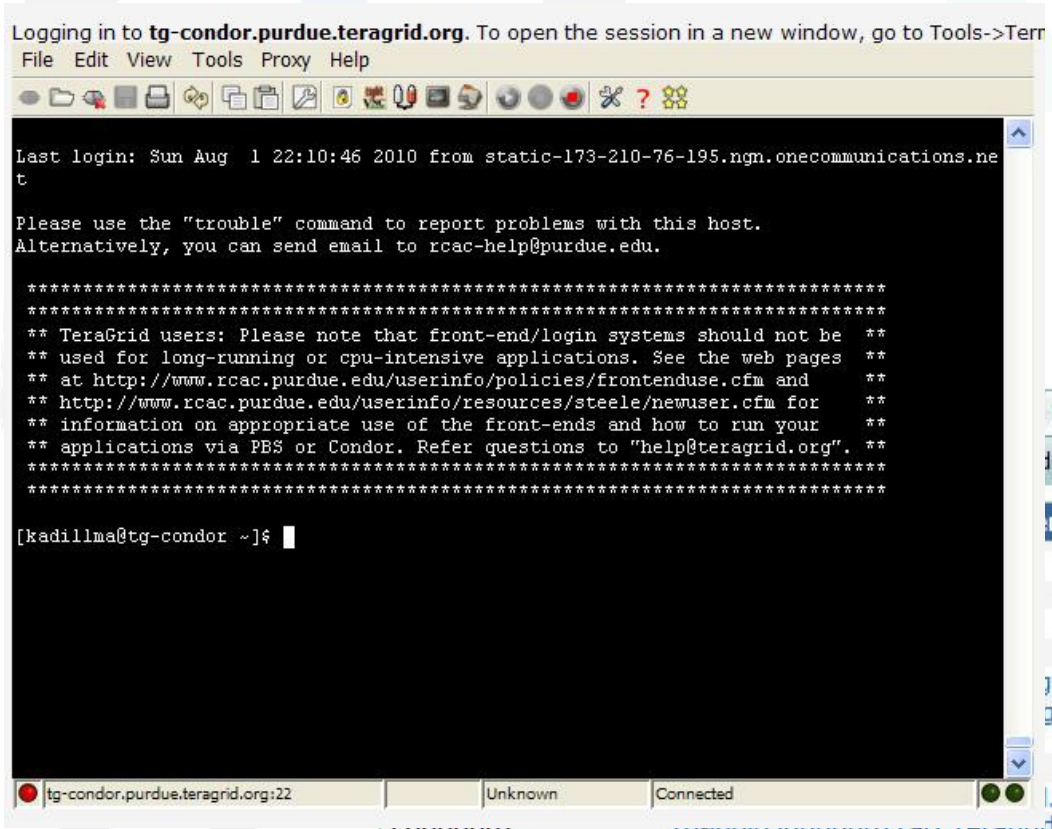
- Login to the TeraGrid User Portal
  - Go to the User Portal
  - Select the "Accounts" tab
  - Use the "login" link under the "Connect" column
- Download, setup and use the GSI-SSHTerm Java Application
  - Download:

  https://security.ncsa.illinois.edu/gsi-sshterm
  - Step by Step Tutorial:

  http://www.rcac.purdue.edu/teragrid/userinfo/tutorials/HowToUseGSISSHApplet.pdf
- Globus Toolkit/MyProxy (UNIX/Mac)

TeraGrid™

TeraGrid User Portal Method

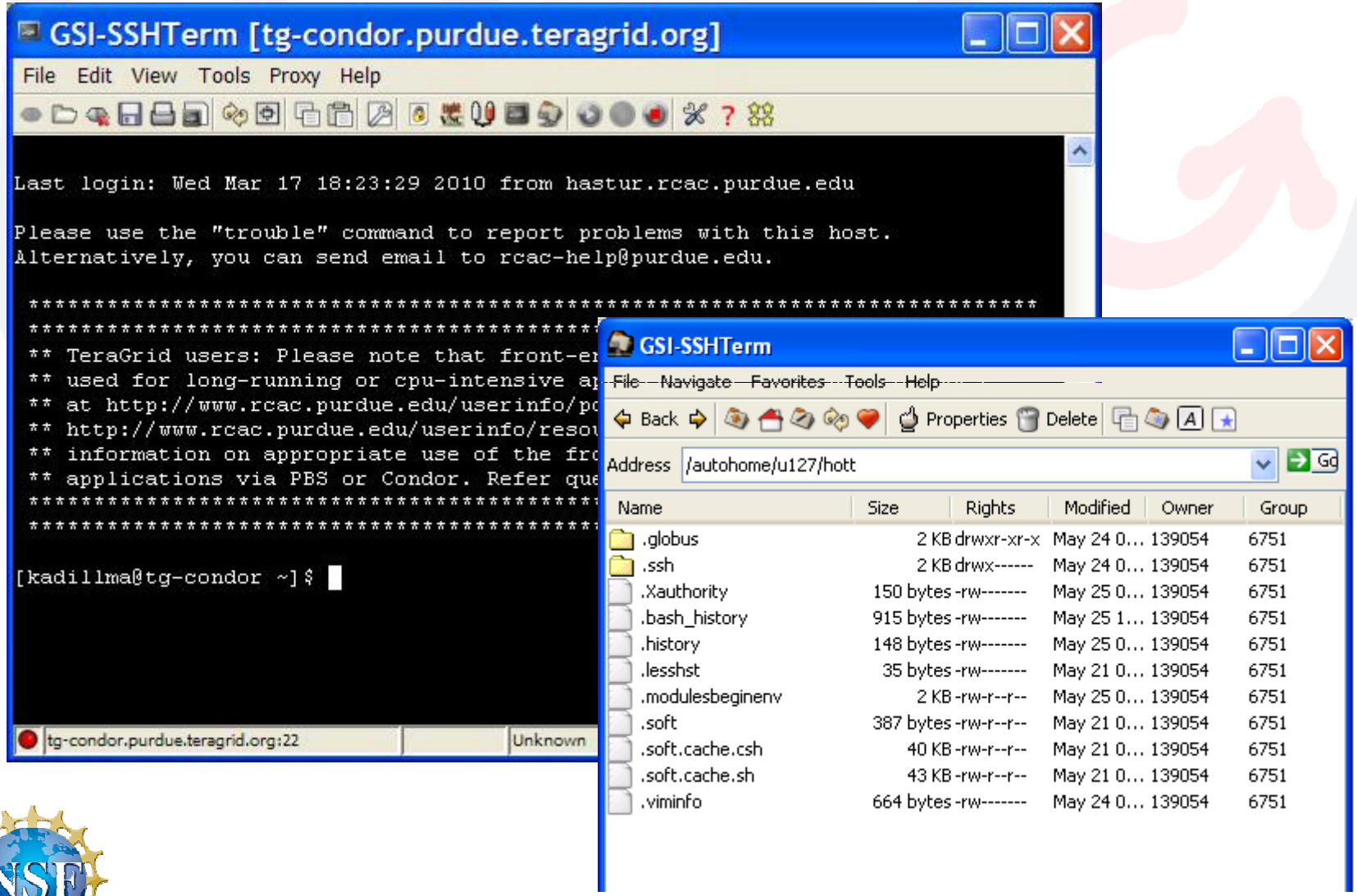


m Accounts

Institution	Username	Connect
NCSA	kadillma	Login
IU	tg-kdillma	Login
NCSA	kadillma	Login
Purdue	kadillma	Login
NCAR	kdillman	Login
NICS	kadillma	Login
NCSA	kadillma	Login
TACC	tg803828	Login
TACC	tg803828	Login
ORNL	kadillma	Login
PSC	kadillma	Login
LONI	kdillman	Login
TACC	tg803828	Login
TACC	tg803828	Login
Purdue	kadillma	Login

NSTG	tg-login.ornl.teragrid.org
Pople	tg-login.pople.psc.teragrid.org
Queen Bee	queenbee.loni-lsu.teragrid.org
Ranger	tg-login.ranger.tacc.teragrid.org
Spur	tg-login.spur.tacc.teragrid.org
Steele	tg-steele.purdue.teragrid.org

TeraGrid GSI-SSHTerm Application Method



TeraGrid

Other methods (optional per resource)

- Local site password (where supported)
  - NCSA (Lincoln, Ember)
  - PSC (Blacklight)
  - SDSC (Dash, Trestles)
  - TACC (Ranger, Lonestar4, Longhorn, Spur)
- SSH Key Pairs (where supported)
  - Purdue (Steele, Condor)
- SecureID Card (FOB) and password
  - NICS (Kraken, Nautilus)
- Common TeraGrid authentication only
  - ?

TeraGrid™

Understanding Queues and Policies

What you should know about queues in general

- There are many different "job manager" systems in use on the TeraGrid resources
  - PBS Pro, Torque, SGE, LoadLeveler, etc.
- Each resource can define the name and queue limits based on their own requirements
  - No "standard" names to reflect queue usage models or policies
- Not all queues on a resource are necessarily available to TeraGrid users
  - Ex. Steele currently has 54 queues, but only 4 are available to TeraGrid users – the rest are for local users or staff only
- Each Resource may required you to submit jobs to a queue in a different way
  - Submit directly to the correct execution queue that meets your job's specific requirements
  - Submit to a "general" queue which routes your job to the correct execution queue that matches your job's requirements

What you should know about queue commands

- There are commands to list what queues are defined
  - Ex: PBS Pro
    - qstat –Q
- There are commands to list the details of a specific queue:
  - Ex: PBS Pro
    - qstat –Qf <queue_name>
- There are commands to list what jobs are in the queue
  - Ex: PBS Pro
    - qstat –a <queue_name>

TeraGrid™

What you should know about queue commands

- There are commands to list the status of all of your jobs currently in the system
  - Ex: PBS Pro
    - qstat –u <userid>
- There are commands to list the details of a specific job in the system
  - Ex: PBS Pro
    - qstat –f <job_id>
- Some systems have commands to show you when your job might start
  - Ex: Moab
    - showstart

TeraGrid™

What you should know about policies

- Job Manager software can implement "policies" that either LIMIT and/or PRIORITIZE jobs in the queue
- These "policies" can vary GREATLY between resources
- Limits may be applied to jobs in a queue or across all queues, for example:
  - Number of jobs queued or running per user
  - Number of jobs queued or running per project allocation number
  - Number of total nodes per user
- Priorities can be based on a variety of attributes including but not limited to:
  - Wall time requested
  - Number of CPUs requested
  - Time already spent waiting in the queue
  - Number of jobs previously run by the user

TeraGrid™

Getting the Best Performance

Benchmarking

- Check out the AUS section to see if your code has been benchmarked on any of the current TeraGrid systems
  - This does not guarantee that your jobs will perform in the same manner but it should help you determine what systems might be best to start running sample test jobs
  - Set up your own test runs to provide benchmark results on these systems for you particular job setup
  - https://www.teragrid.org/web/user-support/aus_projects
- Set up your own benchmarks by planning a "matrix" of job runs varying the number of processors for either:
  - A fixed wall time per job
  - A specific number of "iterations" or time steps.

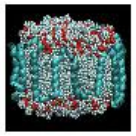
TeraGrid™

Benchmarking (AUS sample)

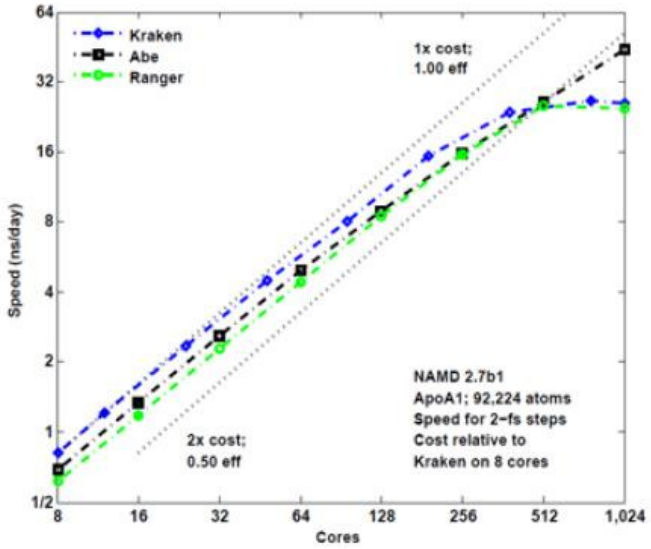
Apolipoprotein A1 (ApoA1)

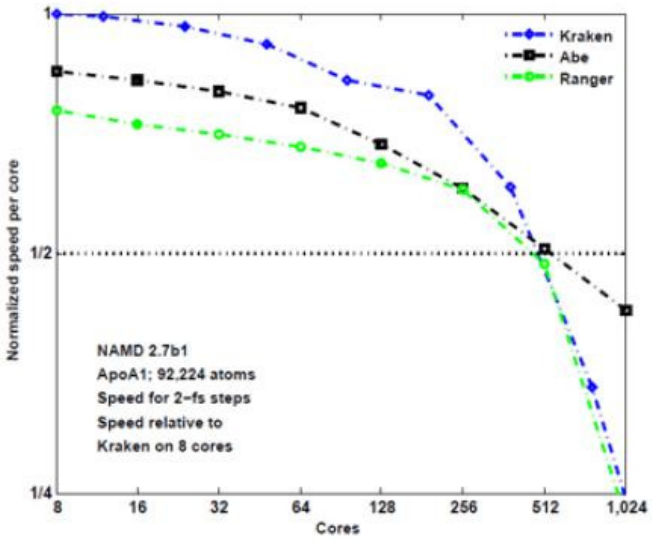
This benchmark consists of 92,224 atoms and uses the following parameters that directly affect performance:

- real space cutoff of 12 angstroms
- PME electrostatics every other timestep
- constant volume and energy (NVE)

[View image]  [View hi-res image]

[Download benchmark suite] [View data page]





TeraGrid™

AUS/ASTA Support

- Request AUS ASTA support to help optimize your code
  - Additional "resource" on startup or research allocation
  - Can also be requested as a "supplemental" allocation request on an existing allocation
- Submit a ticket to the help desk for assistance with performance problems on a specific resource or to ask for recommendations on which resource(s) to use for best performance

TeraGrid™

Let's Try an Example

Example Details

- Basic C program to Integrate the function X**2
- There are serial and parallel versions
- There are 3 different code versions:
  - Input from "stdin" and output to "stdout"
  - Input/Output to files whose names are "hard-coded" in the program
  - Input/Output to files whose names are passed to the program on the command line

Example Downloads

- Condor
  - https://springboard.hubzero.org/resources/32
- Steele
  - https://springboard.hubzero.org/resources/30

- GSI-SSHTerm Tutorials
  - http://www.rcac.purdue.edu/teragrid/userinfo/tutorials/HowToUseGSISSHApplet.pdf
  - http://www.rcac.purdue.edu/teragrid/userinfo/tutorials/HowToCustomizeandSaveGSISSHApplet.pdf

TeraGrid™

Resource Quick Summary Chart

Allocation Number: TG-TRA900337N					
Resource	Login Node	Queue Name	Example Path	Serial	Parallel
Steele	tg-steele.purdue.teragrid.org	R1776954	~kadillma/public/examples/Steele.zip	gcc	mpicc
Condor	tg-condor.purdue.teragrid.org	N/A	~kadillma/public/examples/Condor.zip	gcc	mpicc -static
Blacklight	blacklight.psc.teragrid.org	training	~kadillma/public/examples/Blacklight.zip	icc	icc -lmpi
Ranger	tg-login.ranger.tacc.teragrid.org	development	~tg803828/public/examples/Ranger.zip		mpicc -o <program name> -fast -tp barcelona-64 <program name.c>

TeraGrid™

Resources and Host Names

Resource Name	Login Name	Institution
Athena	athena-gsi.nics.utk.edu	NICS
Big Red	login.bigred.iu.teragrid.org	IU
Blacklight	blacklight.psc.teragrid.org	PSC
Condor	tg-condor.purdue.teragrid.org	Purdue
Dash	dash.sdsc.teragrid.org	SDSC
Ember	login-ember.ncsa.teragrid.org	NCSA
Frost	tg-login.frost.ncar.teragrid.org	NCAR
Kraken	kraken-gsi.nics.utk.edu	NICS
Lincoln	lincoln.ncsa.uiuc.edu	NCSA
Lonestar	lonestar.tacc.teragrid.org	TACC
Longhorn	tg-login.longhorn.tacc.teragrid.org	TACC
NSTG	tg-login.ornl.teragrid.org	ORNL
Pople	tg-login.pople.psc.teragrid.org	PSC
Queen Bee	queenbee.loni-lsu.teragrid.org	LONI
Ranger	tg-login.ranger.tacc.teragrid.org	TACC
Spur	tg-login.spur.tacc.teragrid.org	TACC
Steele	tg-steele.purdue.teragrid.org	Purdue
Trestles	trestles.sdsc.edu	SDSC

Appendix A: Condor Pool Resource Specifics

Potential Indicators for Condor Pool Job Match

- Serial jobs
- Batch "non-interactive" mode only
- Short job duration (under 2 hours)
- Access to source code for "condor_compile"
  - if longer job durations are required
  - See restriction sections for when you can't use this
- Statically compiled executable
- Parametric studies or "monte carlo" type runs that could benefit from a workflow manager (aka DAGMAN)
- Jobs runs with dependencies (again, could benefit from a workflow manager)
- Does not have "large" data input or output files.
- Need access to multiple/different hardware/OS architectures (i.e. Linux/Windows, 32 bit/64 bit, intel/AMD/PowerPC)

Condor Compile Restrictions Standard Universe (1)

- **Taken from:**
  - http://www.cs.wisc.edu/condor/manual/v7.3/2_4Road_map_Running.html
- **Limitations:**
  - **Multi-process jobs are not allowed. This includes system calls such as fork(), exec(), and system().**
  - Interprocess communication is not allowed. This includes pipes, semaphores, and shared memory.
  - Network communication must be brief. A job ***may*** make network connections using system calls such as socket(), but a network connection left open for long periods will delay checkpointing and migration.
  - Sending or receiving the SIGUSR2 or SIGTSTP signals is not allowed. Condor reserves these signals for its own use. Sending or receiving all other signals *is* allowed.
  - Alarms, timers, and sleeping are not allowed. This includes system calls such as alarm(), getitimer(), and sleep().
  - Multiple kernel-level threads are not allowed. However, multiple user-level threads *are* allowed.

Condor Compile Restrictions Standard Universe (2)

- **Limitations (cont):**
  - Memory mapped files are not allowed. This includes system calls such as mmap() and munmap().
  - File locks are allowed, but not retained between checkpoints.
  - **All files must be opened read-only or write-only. A file opened for both reading and writing will cause trouble if a job must be rolled back to an old checkpoint image. For compatibility reasons, a file opened for both reading and writing will result in a warning but not an error.**
  - A fair amount of disk space must be available on the submitting machine for storing a job's checkpoint images. A checkpoint image is approximately equal to the virtual memory consumed by a job while it runs. If disk space is short, a special *checkpoint server* can be designated for storing all the checkpoint images for a pool.
  - On Linux, your job must be statically linked. *condor_compile* does this by default. Dynamic linking is allowed on Solaris.
  - **Reading to or writing from files larger than 2 GB is not supported.**
  - Compiler choices limited by "condor_compile".

TeraGrid™

Condor Compile Restrictions Vanilla Universe

- **Recommendations per Phil Cheeseman (Purdue Condor Guru)**
- **Limitations:**
  - Multi-process jobs are not allowed. This includes system calls such as fork(), exec(), and system().
  - Static linking recommended to widen range of potential execution platforms.
  - Execution very similar to other batch systems in the case of *vanilla* codes.
  - Not well suited for jobs running more than a few hours without self-restart to recover from preemption *(eviction).*
  - Must be used for Windows applications.

TeraGrid™

Condor Compile Restrictions General

- **Recommendations per Phil Cheeseman (Purdue Condor Guru)**
- **Limitations:**
  - Neither universe should be expected to enable completion of work that cannot finish within PBS limits. Standard universe jobs requiring 2+ CPU weeks (without restart capability) to run are a risky proposition for Condor just as they are for PBS. Multi-hour vanilla universe jobs w/o self-restart capability are better suited for PBS (or a rewrite of the time consuming application(s)).
  - **Preemption is a certainty for 'long' jobs** although much less probable for jobs running less than an hour. Preemption is *always* a possibility.

TeraGrid™