epcc

An rough guide to accessing PRACE and XSEDE HPC resources

Gavin J. Pringle

Overview

- Three different opportunities for Accessing HPC resources

- PRACE (European)
  - Tier-1 platforms: DECI programme
  - Tier-0 Access

- XSEDE (United States)
  - Tier-0 Access

PRACE

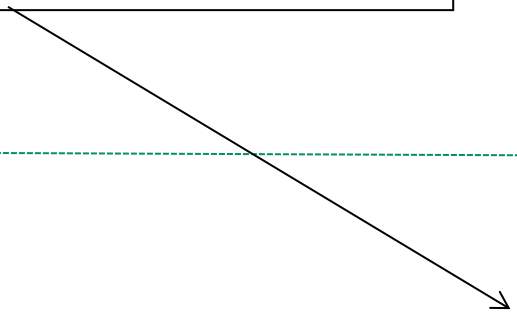
- DECI programme
  - Slides used by permission
  - Thanks to Chris Johnson
- Tier-0 access
  - Slides used by permission
  - Thanks to
- Acknowledgment
  - We acknowledge that the slides in this presentation have been provided using the PRACE Research Infrastructure. *The support of Marjolein Oorsprong from Finland and Chris Johnson from UK is gratefully acknowledged.*

- Tier-0 access
  - Slides used without permission
  - https://www.xsede.org/documents/384387/569747/Xsede+New+User+Tutorial.ppt
- Acknowledgment:
  - This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575.

DEISA (Distributed European Infrastructure for Supercomputing Applications)

- *http://www.deisa.eu*

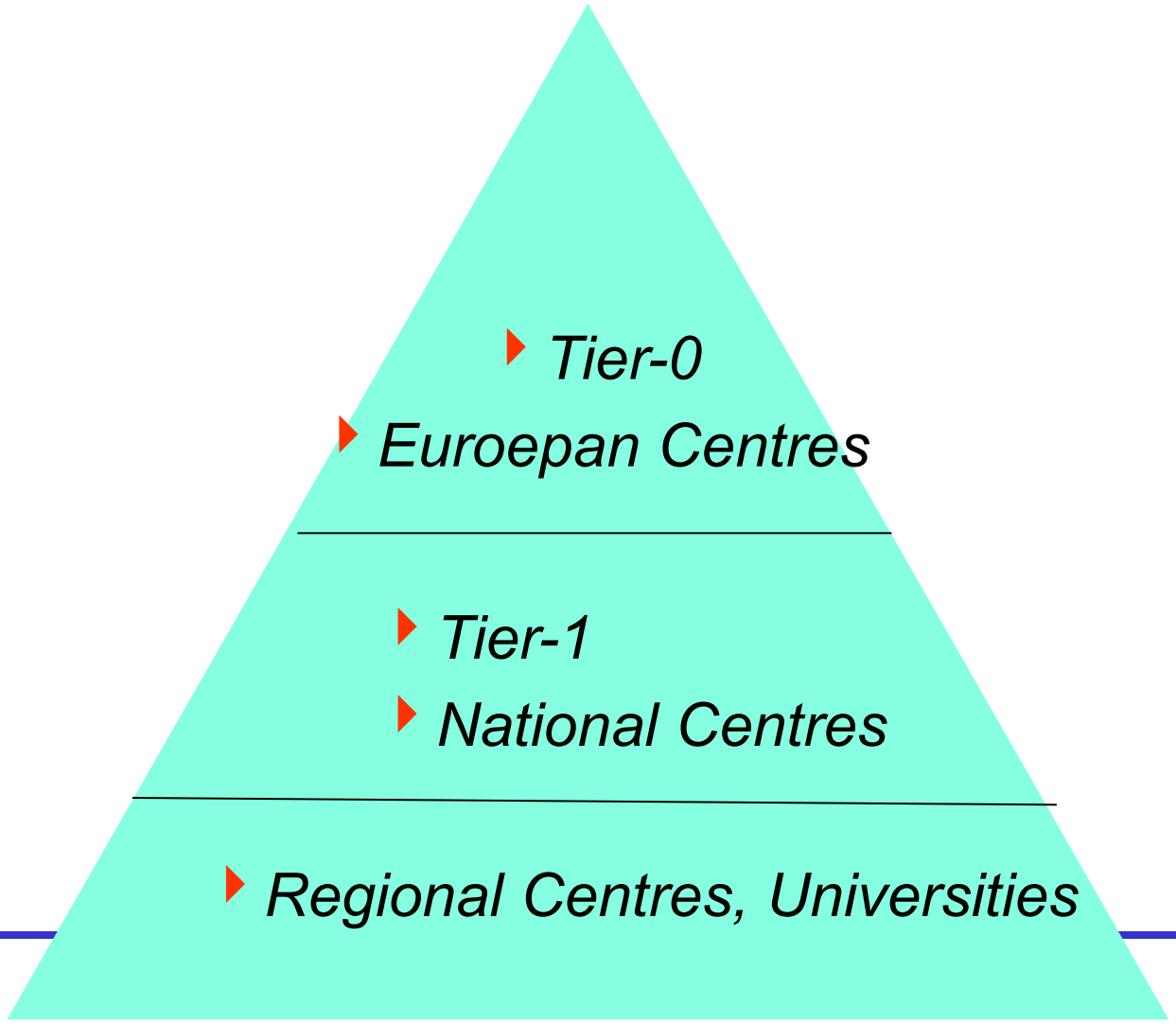
- DECI (DEISA Extreme Computing Initiative) — *DEISA 2005-2010*



- *PRACE 2010-2014* — DECI (Distributed European Computing Initiative)

- *PRACE (Partnership for Advanced Computing in Europe)*
  - *http://www.prace-project.eu/*

European resources



Tier-0

- *(7) JUQUEEN – FZJ*
- *IBM BlueGene/Q*

- *(9) SuperMUC – LRZ*
- *IBM System X iDataplex*

- *(12) FERMI – CINECA*
- *IBM BlueGene/Q*

- *(29) MareNostrum – BSC*
- *IBM System X iDataplex*

- *(15) CURIE – GENCI*
- *Bull x86 system*

- *(32) HERMIT – HLRS*
- *Cray XE6*

Tier-0 – how to apply

- There are currently no open Calls for Project Access
- The Call for Proposals for PRACE Preparatory Access is continuously open
  - http://www.prace-ri.eu/call-announcements/
  - http://www.prace-ri.eu/How-to-apply
- The general e-mail address for all questions related to Peer Review, Calls, etc.
  - peer-review@prace-ri.eu
- More later…

Tier-1 systems

- Around 20 systems for each DECI call
  - IBM Blue Gene/Q
  - Cray XE/XC
  - a range of large clusters including GPU resources
- Host of countries providing resources, e.g. for DECI-11:
  - Cyprus, Finland, Germany, Hungary, Ireland, Italy, Norway, Poland, Serbia, Spain, Switzerland, The Czech Republic, The Netherlands and United Kingdom

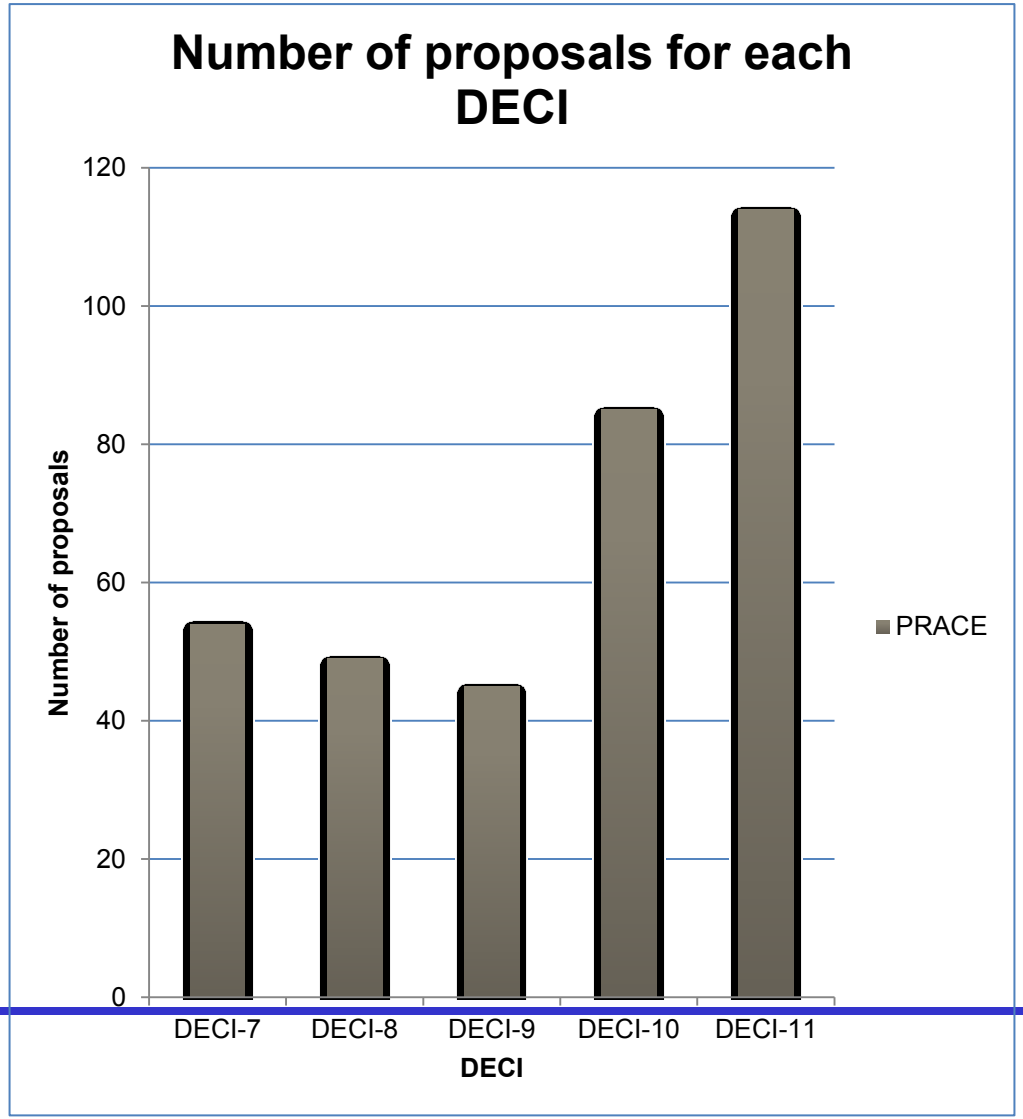
Resource exchange programme

- Participating countries contribute ~5% of national supercomputing resources annually…
  - …in exchange receive a similar amount back for projects from their country
- Non-participating countries
  - 15% of the overall allocation is given to non-contributing countries

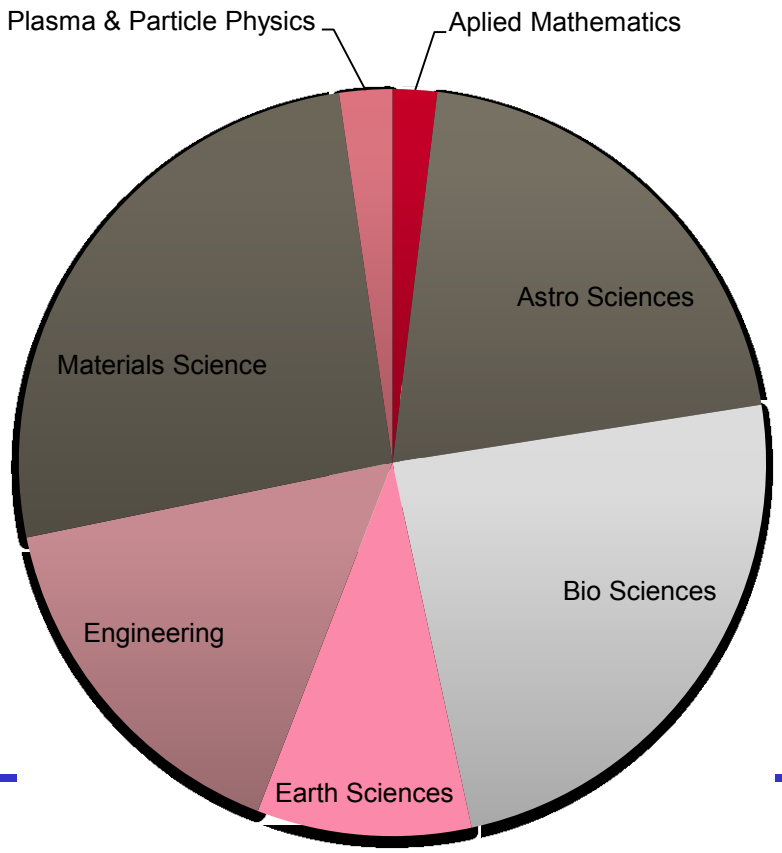
DECI calls

- Calls every 6 months
- Projects start 1st Nov/1st May
- Projects last 1 year
- 186 projects awarded time so far under PRACE
- Almost 700 million CPU hours awarded
- Projects received average of ~ 3.3 mi std hrs each
- Projects typically awarded 100,000-10,000,000 std hrs
- Enabling effort available

Proposals received



Time awarded for DECI-9/10 by Scientific Discipline



Additional features

- Single username
- Easy access to multiple resources
- Similar toolset/libraries across machines
- Ease-of-access/fast transfer
  - Gsissh
  - gridftp
- Uniform accounting tool (DART)

Future of DECI?

- PRACE-3IP project ended in August 2014
- Yet DECI-13 now open
- Probably ~18-month gap before new “PRACE”
- Will run DECI-12 and DECI-13 as “lightweight” DECI project to fill these gaps
- …keep an eye on mailings/PRACE website/etc.

Conversion factors

- To exchange resources we use standard hours
- Introduce concept conversion factors:
  - BlueGene/Q 1 core hour = 0.85 std hours
  - Cray XE 1 core hour = 1.25 std hours
  - Cray XC 1 core hour = 3.50 std hours
  - Intel Westmere EX@2.4 cluster

    1 core hours = 2.6 std hours
- For GPUs, multiply by further factor

Machines in DECI

- Fall broadly into 4 categories
  - Cray
  - BlueGene/Q
  - Cluster (without GPU)
  - Cluster (with GPU)

Cray

- SiSu, XC30, CSC, Finland

- ARCHER, XC30, EPCC, UK

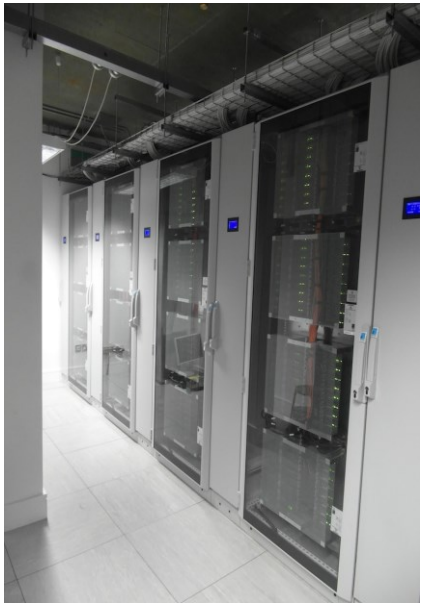
- Lindgren, XE6, PDC, Sweden

- Rosa, CSCS, XE6, Switzerland

IBM BlueGene

- BlueJoule, BG/Q, STFC, UK

Clusters



DECI-11 resources

Clusters		
BSC (Spain)	MinoTauro	2,430,000
CASTORC (Cyprus)	Cy-Tera	1,080,000
CYFRONET (Poland)	Zeus	3,153,600
CYFRONET (Poland)	Zeus BigMem	13,455,360
FZJ (Germany)	JuRoPA	6,000,000
ICHEC (Ireland)	Fionn-thin	11,437,056
ICHEC (Ireland)	Fionn-hybrid	578,160
ICHEC (Ireland)	Fionn-fat	404,712
ICM (Poland)	Boreasz	1,681,920
IPB (Serbia)	PARADOX	5,000,000
NIIF (Hungary)	NIIFI SC	2,000,001
PSNC (Poland)	Chimera	7,416,900
RZG (Germany)	Hydra	16,887,500
SURFSARA (The Netherlands)	Cartesius	12,342,000
UIO (Norway)	Abel	12,229,074
WCSS (Poland)	Supernova	4,086,990
VSB-TUO (Czech Republic)	Anselm	12,060,000
	Sub-total	**112,243,272**

Clusters - GPU		
CASTORC (Cyprus)	Cy-Tera-GPU	54,000
CINECA (Italy)	PLX	6,409,998
PSNC (Poland)	Cane	13,360,000
	Sub-total	**19,823,998**

Cray		
CSC (Finland)	Sisu	14,588,235
CSCS (Switzerland)	Rosa	11,250,000
EPCC (UK)	ARCHER	44,268,134
PDC (Sweden)	Lindgren	15,936,250
	Sub-total	**86,042,619**

BlueGene/Q		
STFC (UK)	Blue Joule	21,250,000
	Sub-total	**21,250,000**

	239,359,890

How to apply

- See call information here:
  - http://www.prace-project.eu/DECI-Projects
- Register here:
  - http://deci-peer-review.cines.fr/login

PRACE

PRACE Tier-0 Access

Criterion: Scientific Excellence

- Free of charge
- Obligation to publish results at the end of the award period
- PRACE calls are open to international projects

- Types of resource allocations for scientists
  - **Project Access (every 6 months)**
    - For a specific project, award period ~ 1 to 3 years
    - For individual researchers and research groups (no restriction of nationality for both researcher and centre)
    - Requires to demonstrate technical feasibility of project
  - **Programmatic Access**
    - purpose: to ensure a stable and reliable minimum access to the necessary computational resources for large-scale, long term projects of very high scientific quality and with a broad European scope, importance and relevance
    - maximum of 20% of the total resources available for programmatic access
  - **Preparatory Access**
    - Optionally with support from PRACE experts
    - Prepare proposals for Project Access

Peer Review Principles

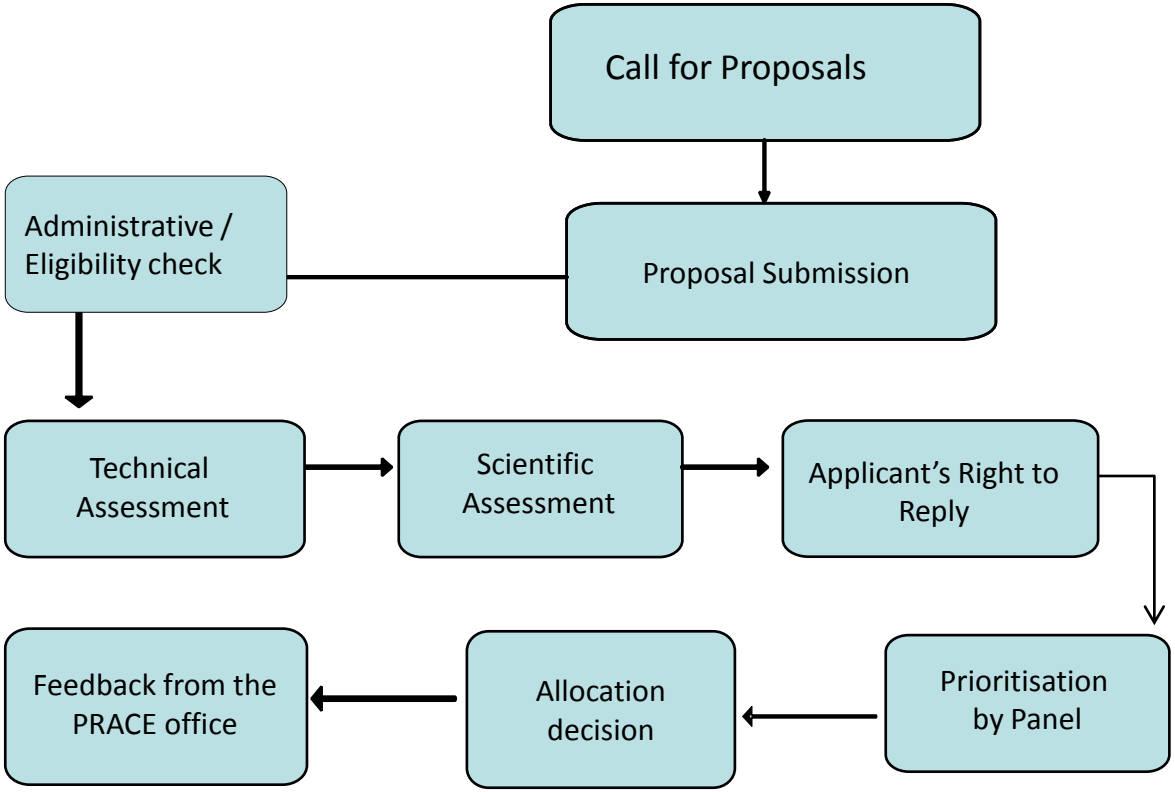
- Transparency
- Fairness
- Avoiding conflict of interests at all stages of the process
- Confidentiality
- Right to reply to technical and scientific evaluations
- Right to appeal the final decision

Best science supported at the highest level through a fair Peer Review Process

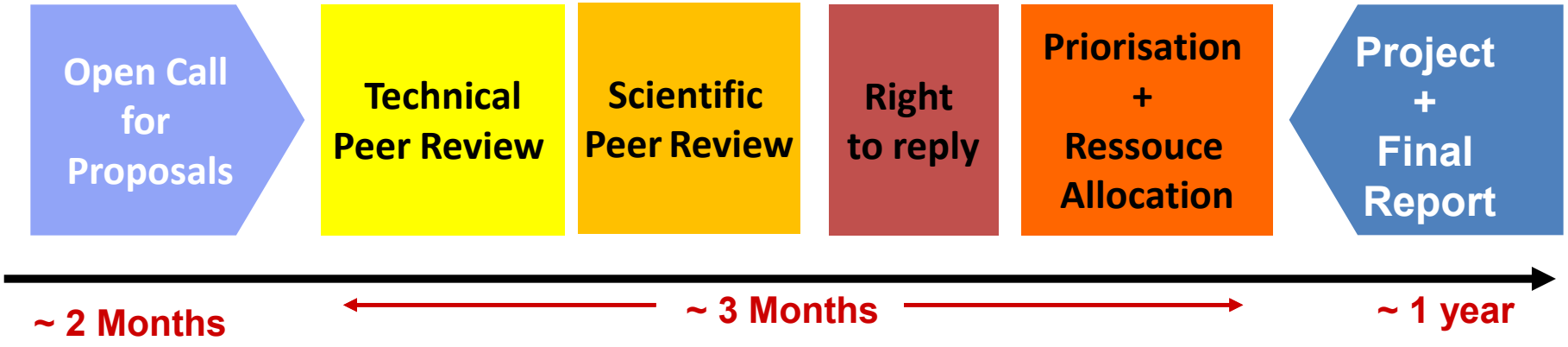
Criteria for scientific assessment

- **Scientific excellence**. Demonstrate scientific excellence and a potential for high European and international impact
- **Novelty and transformative qualities**.
  - Proposals should be novel, develop an important scientific topic of major relevance to European research
  - Describe possible transformative aspects, and expected advances
- **Relevance** to the call if a specific scope is stated in the call
- **Methodology**
- **Dissemination**
  - The planned channels and resources for dissemination and knowledge exchange
  - List of recent publications relevant to the proposed project.
- **Management**. Solid management structure in the project
- These criteria should be **fully addressed** in the application.

PRACE Peer Review Process



Project Access



Preparatory Access

- Permanently open with quarterly cut-off dates (03/06/09/12)
- Intended to prepare proposals for Project Access
- *Not for production runs nor research activities*
  - Testing scalability: **Type A**, allocation for 2 months
  - Code development or optimisation:
    - **Type B**, allocation for 6 months
    - **Type C**, allocation for 6 months, including support
- Fixed amount of resources, depending on the system
- Technical review only
- On-line application: https://prace-peer-review.cines.fr/

November 6, 2014

XSEDE New User Tutorial

Ken Hackworth,
James Marsteller,
Marcela Madrid,
Steve Tuecke

XSEDE New User Tutorial

- Today's session is a general overview of XSEDE for prospective and new XSEDE users, including the allocation process and safety. It is not going to teach you computational science or programming, but rather, how to use XSEDE resources.

- The lecture will consist of four parts:
- Writing and Submitting a Successful Allocation (Ken Hackworth)
- Information Security Training for XSEDE Researchers (Jim Masteller)
- Using XSEDE Resources (Marcela Madrid)
- Using Globus Online (Steve Tuecke)

After attending this tutorial…

You will know how to:

- Submit an allocation for XSEDE resources.
- How to protect XSEDE resources and your own work and how to choose a secure password.
- Login in to XSEDE resources, transfer files, submit jobs.

What is XSEDE?

- The most powerful integrated advanced digital resources and services in the world. Funded by NSF.

- Scientists and Engineers around the world use XSEDE resources and services: supercomputers, collections of data, help services.

- XSEDE lowers technological barriers to the access and use of computing resources. Using XSEDE, researchers can establish private, secure environments that have all the resources, services, and collaboration support they need to be productive.

- Consists of supercomputers, high-end visualization, data analysis and storage around the country.

November 6, 2014

Using XSEDE Resources

Extreme Science and Engineering Discovery Environment

Marcela Madrid, PSC
mmadrid@psc.edu

OUTLINE

- XSEDE Resources
- XSEDE User Portal
- Accessing Resources: Single Sign On
- File Transfers
- Running: file systems, batch jobs
- Why has my job not run?
- Batch schedulers. Improving your job turnaround
- Managing Your Software Environment
- Getting Help
- Exercise: Actually running on a supercomputer

Stampede
User Guide
Debuting Intel's new MIC technology on a massive scale, TACC's Stampede system is currently in early user mode and goes into production in January, 2013.

Blacklight
User Guide
Blacklight is a SGI shared memory system intended for applications that require a large shared memory for computational tasks.

Gordon
User Guide
Gordon is a unique, a flash-based supercomputer designed for data-intensive applications.

Keeneland
User Guide
Keeneland is a balanced hybrid CPU/GPGPU system for use with codes that can take advantage of accelerator performance. Keeneland has over 200 nodes with 3 GPUs per node making it an excellent resource for scaling highly parallel codes. Applications from scientific domains such as materials, combustion, molecular dynamics, and astrophysics have seen substantial performance increases on Keeneland.

Kraken
User Guide
The Kraken system is a Cray XT5 system with compute nodes interconnected with SeaStar, a 3D torus. Kraken is intended for highly scalable parallel applications.

Lonestar
User Guide
Lonestar, a Dell Linux Cluster, is a powerful, multi-use cyberinfrastructure HPC and remote visualization resource. Lonestar is intended primarily for parallel applications scalable to thousands of cores. Normal batch queues will enable users to run simulations up to 24 hours. Lonestar also provides access to large memory nodes, and nodes containing NVIDIA GPU's, giving users access to high-throughput computing and remote visualization capabilities respectively.

Ranger
User Guide
The Ranger Sun Constellation Cluster intended for users with codes scalable to thousands of cores (1024 and above). A batch queue is available to assist users develop, test, and scale codes up to 1024 compute cores. Four separate login nodes will provide interactive connectivity to the system for compiling and interfacing with the batch queuing system.

Steele
User Guide
Steele compute cluster consists of five logical sub-clusters, each with a different combination of memory and interconnect. Steele is well suited for a wide range of both serial and parallel jobs.

Trestles
User Guide
Trestles employs flash-based memory and is designed for modest-scale research providing very fast turnaround time. It is intended for moderately scalable parallel applications with an emphasis on improving productivity for a broad spectrum of users. Trestles is ideal for applications with fast local I/O requirements that can benefit from the flash memory available on each compute node.

xsede.org

XSEDE

xsede.org

- User Services:
- Getting started guide

Supporting Non-traditional Disciplines: How Can XSEDE Help Meet Challenges?

- Massively parallel clusters
- Large shared-memory nodes
- Parallel file systems
- Fast networking
- Very large databases
- Efficient data movement tools
- Consulting services provided by experts (expertise on OpenMP, workflow, science gateway, scientific database, visualization and more)

XSEDE Services: Resources

NIP Reach out and support areas or communities that have not traditionally used high performance computing.

- Economics: finance trading data
- Census data
- Word mining.
- Social sciences
- Humanities
- Art

XSEDE RESOURCES: Services:
1. Extended Collaborative Support (ECS)

- Can solicit ECS support for help:
- Porting applications to new resources
- Providing help for portal and gateway development
- Implementing algorithmic enhancements
- Implementing parallel math libraries
- Improving scalability of codes to higher processor counts
- Optimizing codes to efficiently utilize specific resources
- Assisting with visualization, workflow, data analysis, and data transfer
- Inquire at help@xsede.org

Campus Champions: campus representatives. Easiest way to get started is to contact your local campus champion:

https://www.xsede.org/campus-champions

How do I get started using XSEDE resources?

To get started using XSEDE a researcher needs to:

- apply for an XSEDE allocation
- request to be added to an existing one.

You do either of these through the Portal.

Or

- Contact your local campus champion.

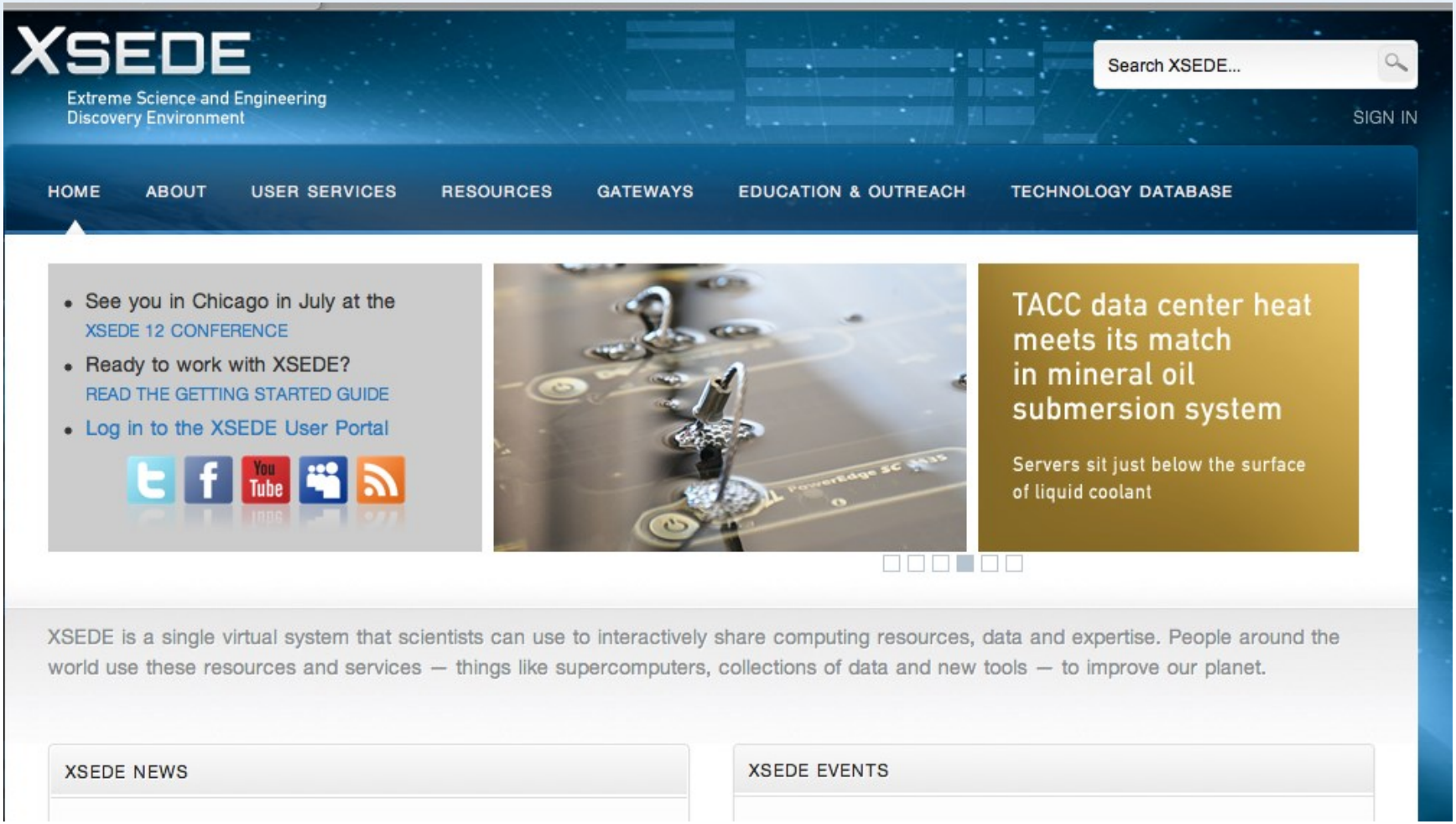
XSEDE User Portal (XUP)

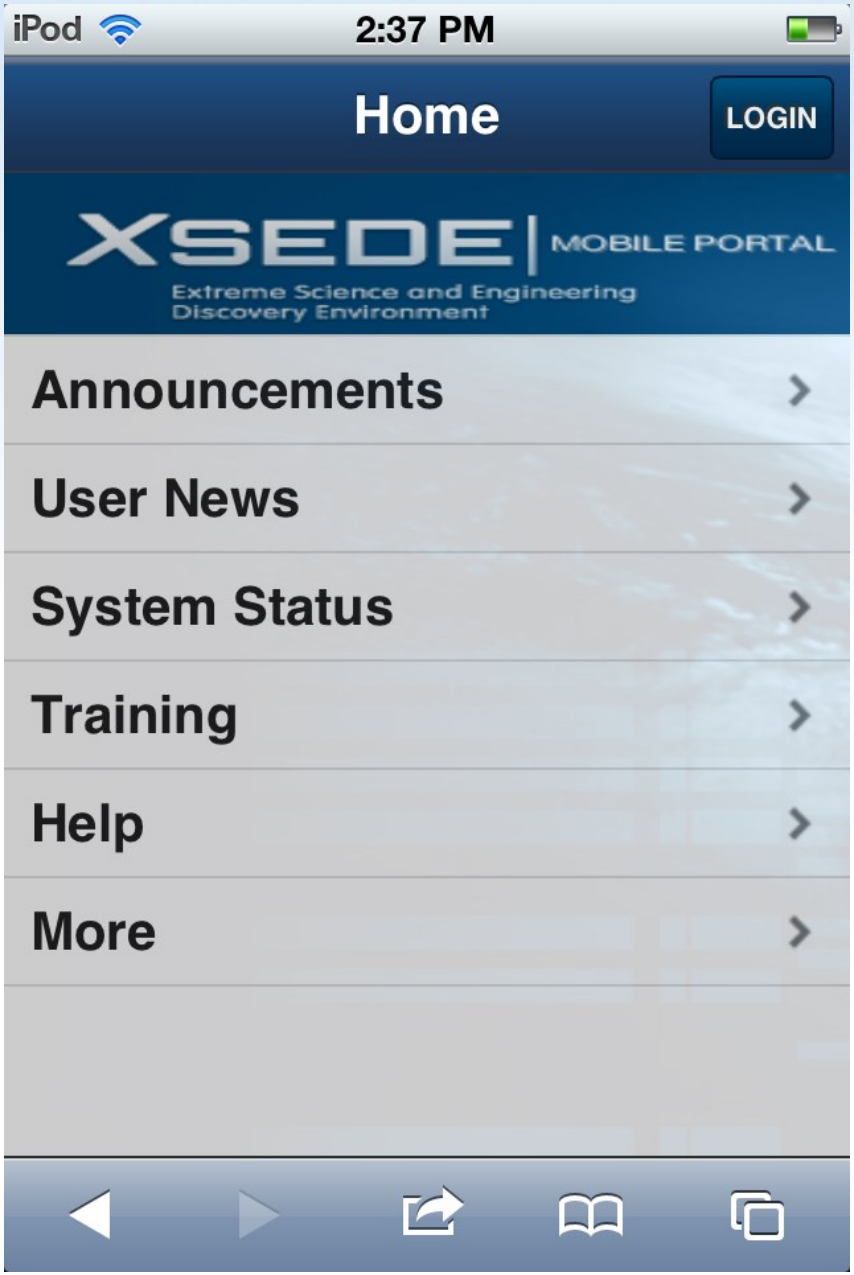
Web-based single point of contact that provides:

- Continually updated information about your accounts.
- Access to your XSEDE accounts and allocated resources: *The Portal provides a single location from which to access XSEDE resources. One can access all accounts on various machines from the Portal or with your Portal Username and Password.*
- Interfaces for data management, data collections, and other user tasks and resources.
- Monitor your allocation
- Access to the Help Desk.

XSEDE HOME PAGE: xsede.org

Can create a Portal account at any time





Also available:

Mobile User Portal
https://mobile.xsede.org

portal.xsede.org
User Responsibilities & Security

- When you login to the Portal at the beginning of each allocation term, you will be asked to accept the User Responsibilities form:
- Explains acceptable use to protect shared resources and intellectual property. You are responsible for your account and for protecting your passwords:
- Choose a strong password: Not a word that can be found in a dictionary or that can be easily guessed.
- Includes not sharing passwords, not writing passwords down where they can be easily found, and not using tools which expose passwords on the network (such as telnet). This includes private keys: make sure they are password-protected.
- Do not share accounts. One account per user. Can add members of your group to your grant through the Portal.

User Responsibilities form

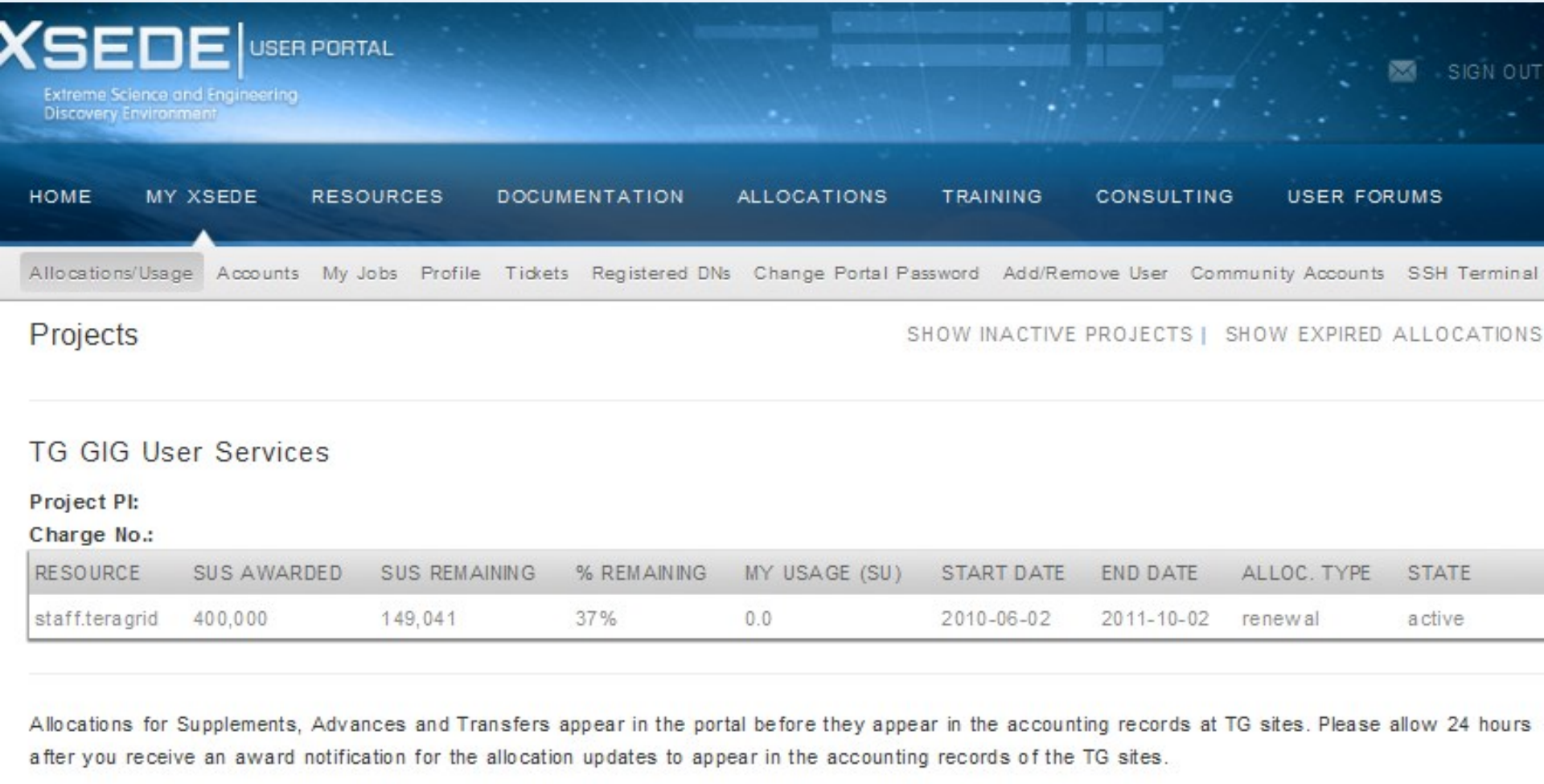
- Appropriate Behavior
  - Protecting computing, closing SSH terminals when done, logging out of the User Portal when done, etc.

- Acknowledgment in publications.

- Report Suspicious Activity: ***If you have any suspicion that your account or personal computer has been compromised send email to help@xsede.org or call 24/7: 1-866-907-2383 immediately***

My XSEDE: Allocations/Usage

When the allocation is active it will show in the Portal, MyXSEDE, Allocations/Usage tab:



portal.xsede.org→MY XSEDE→Profile

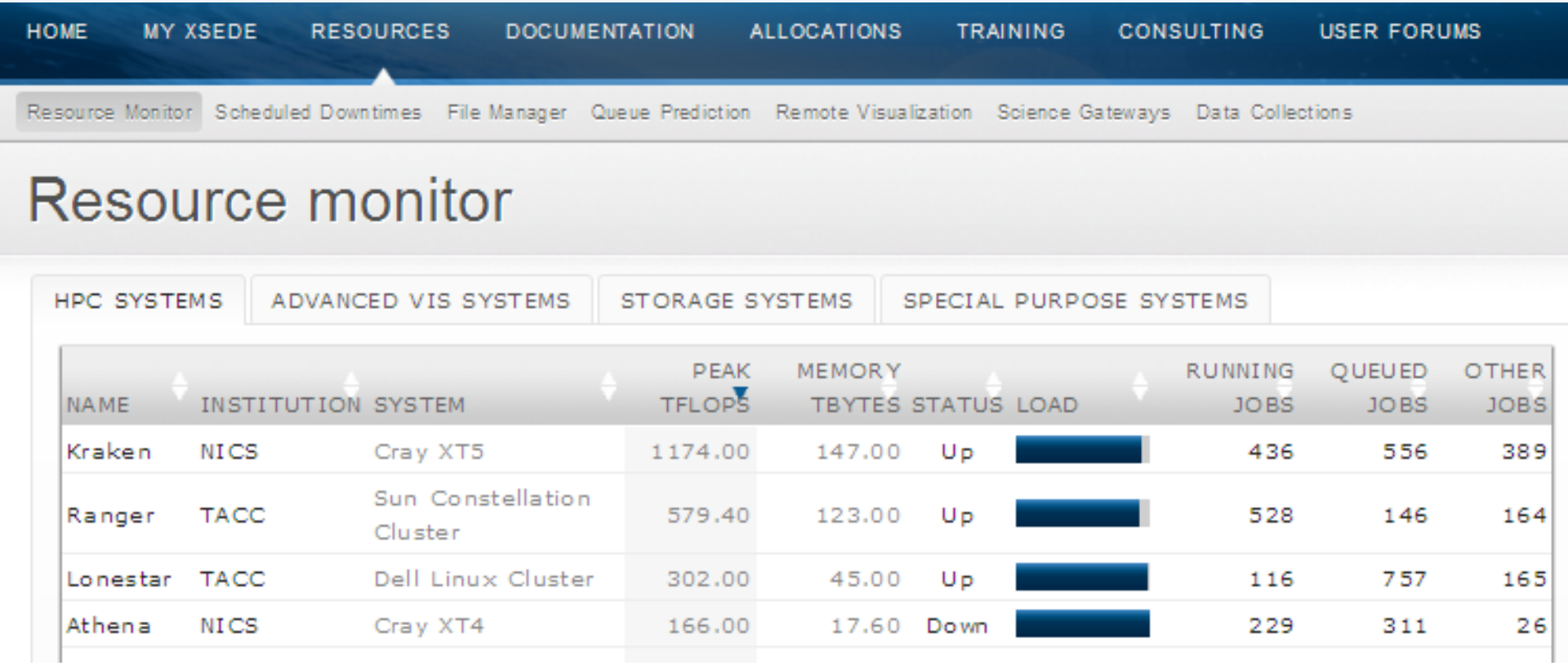
Profile

- View and or change your user information (organization, address).
- Important because the email that is here is where XSEDE will communicate with you regarding your allocation.
- Can choose what mailing lists to subscribe to.

Add Publication: Anyone who is on an XSEDE project can add a publication and associate it with their project.

portal.xsede.org->Resources-> Systems Monitor

View system information: TFLOPS, memory, today's load, jobs running in queue. **Status: up or down**: takes you to the news announcements that tells you when the machine is expected to come back up.



NAME	INSTITUTION	SYSTEM	PEAK TFLOPS	MEMORY TBYTES	STATUS	LOAD	RUNNING JOBS	QUEUED JOBS	OTHER JOBS
Kraken	NICS	Cray XT5	1174.00	147.00	Up		436	556	389
Ranger	TACC	Sun Constellation Cluster	579.40	123.00	Up		528	146	164
Lonestar	TACC	Dell Linux Cluster	302.00	45.00	Up		116	757	165
Athena	NICS	Cray XT4	166.00	17.60	Down		229	311	26

Accessing XSEDE Resources

- Several methods are possible:
- Single Sign On (SSO):

  ssh login.xsede.org

  Through portal
- Direct login access (ssh) (Need to request direct passwds)

Login to XSEDE Resources

- Open up a terminal on your laptop or workstation. Windows users may download one of many ssh clients, free ones include "putty". Type:

ssh login.xsede.org

(the Hub) : Enter your **Portal username and password**.

- After this you can connect to those machines on which you have an allocation, using the command

gsissh machine-name

gsissh ranger.tacc.xsede.org

gsissh lonestar.tacc.xsede.org

gsissh blacklight.psc.teragrid.org

No need to enter another passwd.

Login from the Hub: **ssh login.xsede.org**

How many seconds your credential is valid for:
[user@login ~]$ grid-proxy-info -timeleft
42466 (seconds)

Once your Credential expires, you won't be able to gsissh. To refresh the credential, you can logout of the SSO hub and log back in.

You can type '**man myproxy-logon**' to learn more about options to this command.

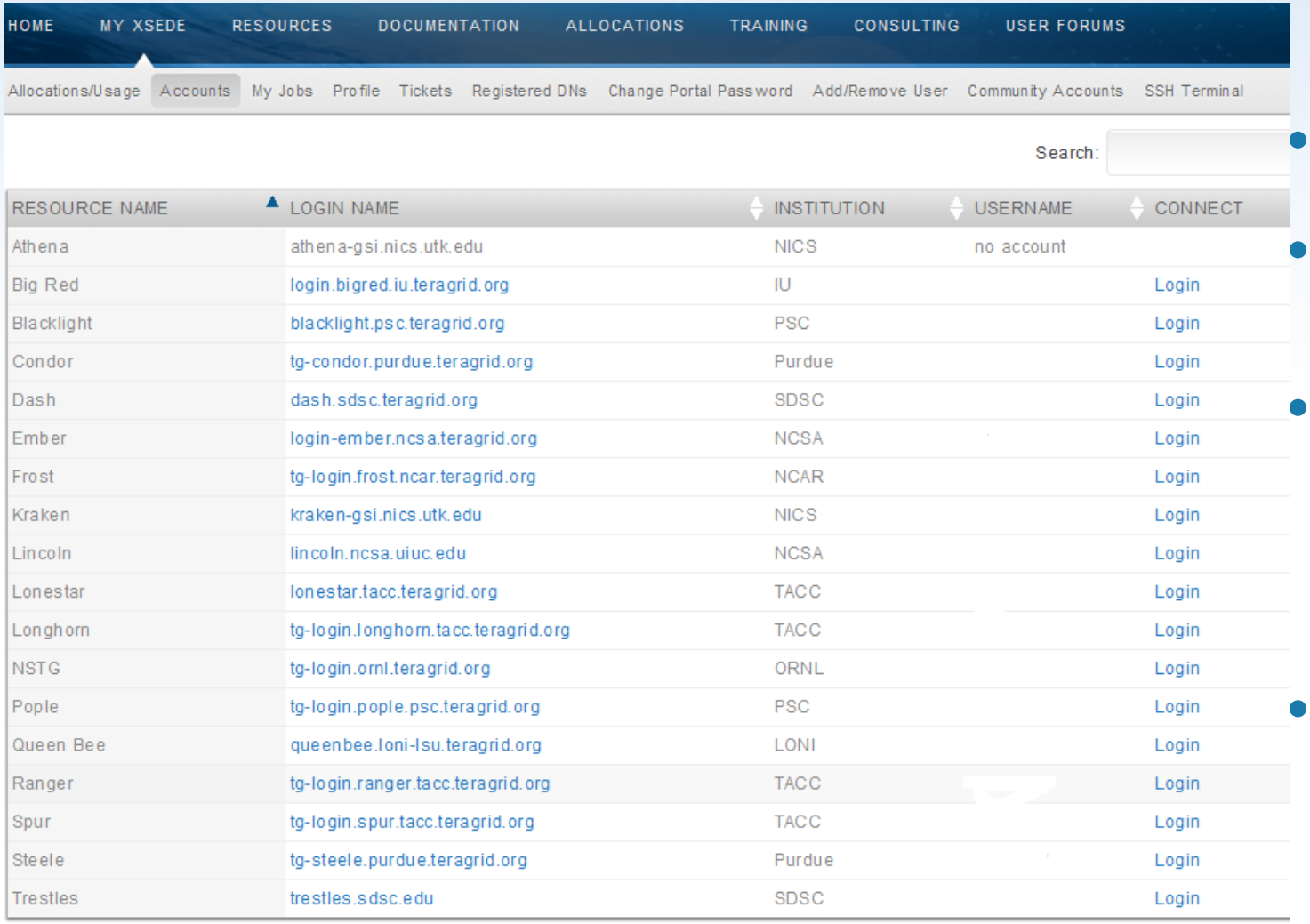
Helpful Tip: Set up host aliases in a config file to make it easy to logon to an abbreviated alias rather than using the full XSEDE login host.

[user@login ~]$ gsissh blacklight
instead of: [user@login ~]$ gsissh blacklight.psc.teragrid.org

Create a config file under ~/.ssh directory.
Sample contents of the config file:

```
## SSH client Host aliases for XSEDE
## - - - - - - - - - - - - - - - - - - - - - - - - ##
Host kraken
Hostname gsissh.kraken.nics.xsede.org
Host nautilus
Hostname gsissh.nautilus.nics.xsede.org
Host blacklight
Hostname tg-login1.blacklight.psc.teragrid.org
Host lonestar
Hostname lonestar.tacc.utexas.edu
Host longhorn
Hostname longhorn.tacc.utexas.edu
```

SSO via User Portal



RESOURCE NAME	LOGIN NAME	INSTITUTION	USERNAME	CONNECT
Athena	athena-gsi.nics.utk.edu	NICS	no account	
Big Red	login.bigred.iu.teragrid.org	IU		Login
Blacklight	blacklight.psc.teragrid.org	PSC		Login
Condor	tg-condor.purdue.teragrid.org	Purdue		Login
Dash	dash.sdsc.teragrid.org	SDSC		Login
Ember	login-ember.ncsa.teragrid.org	NCSA		Login
Frost	tg-login.frost.ncar.teragrid.org	NCAR		Login
Kraken	kraken-gsi.nics.utk.edu	NICS		Login
Lincoln	lincoln.ncsa.uiuc.edu	NCSA		Login
Lonestar	lonestar.tacc.teragrid.org	TACC		Login
Longhorn	tg-login.longhorn.tacc.teragrid.org	TACC		Login
NSTG	tg-login.ornl.teragrid.org	ORNL		Login
Pople	tg-login.pople.psc.teragrid.org	PSC		Login
Queen Bee	queenbee.loni-lsu.teragrid.org	LONI		Login
Ranger	tg-login.ranger.tacc.teragrid.org	TACC		Login
Spur	tg-login.spur.tacc.teragrid.org	TACC		Login
Steele	tg-steele.purdue.teragrid.org	Purdue		Login
Trestles	trestles.sdsc.edu	SDSC		Login

- Make sure you are logged into the XSEDE User Portal
- Go to ‘My XSEDE’ tab
- Go to the ‘Accounts’ link
- Resources you have access to will be indicated by a ‘login’ link
- Click on the ‘login’ link of the resource you would like to login to

Accessing Resources: Direct Access

- SSO is the default method.

- However, if you choose direct access:

  you'll need to request the direct access password for the machine (by submitting a ticket).
  - Use a secure shell (ssh) client.
  - From Linux or Mac terminal window:
    - ssh –l <username> <machinename>
    - E.g.: ssh –l username ranger.tacc.utexas.edu
  - From Windows:
    - Download one of many ssh clients
    - Free ones include "putty"
    - Most campuses have a site license for a fancier one.

File transfers between XSEDE resources

or

Between your laptop and an XSEDE resource

USAGE MODE	TRANSFER METHOD	PROS	CONS
GRAPHICAL USER INTERFACE	PORTAL	Easy to use, single sign-on via user portal, desktop download available	Basic/beginner usage
	Globus Online	Easy to use web interface, can use Portal (SSO), desktop download available	Basic/beginner usage
COMMAND LINE INTERFACE	Globus-url-copy uberftp	Advanced syntax available for scripting, can use XSEDE single sign on, handles large and small files	Advanced knowledge required for authentication and scripting capabilities
	scp sftp	Easy to use	Must use local username and password, meant for small files < 2GB

File Transfers: Small (< 2 GB) Files

- To transfer ***small files*** between XSEDE Resources and/or your own workstation you can:

- *scp* or *sftp* from the Unix command line. Easy to use. Provides poor performance for very large files. ***Requires local username and password***.

- From Linux or Mac, you can run these commands directly from the terminal.

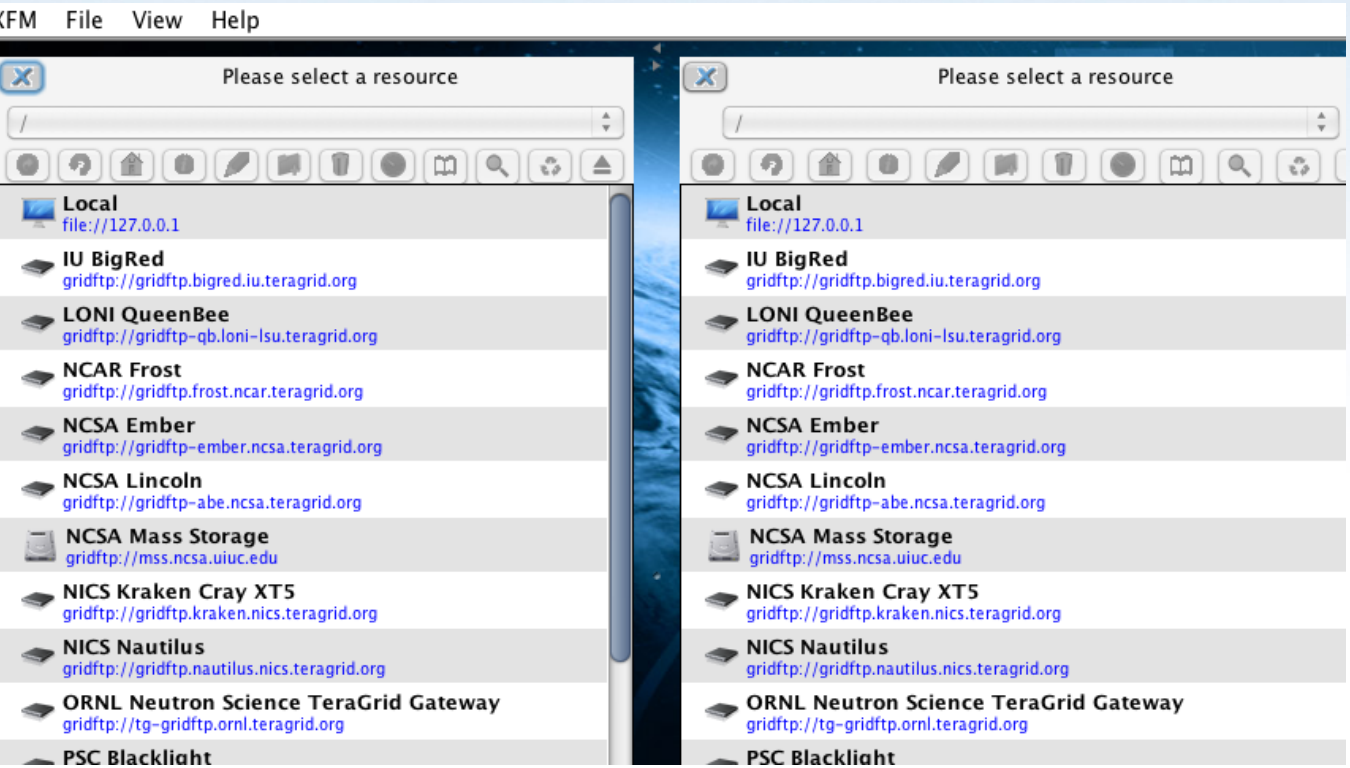
- From Windows, use your ssh client to do this (putty has free downloads for these tools, too! – just Google “putty sftp”).

These are easy to use and secure, but provide poor performance for large files.

File Transfer: User Portal File Manager

- Log into the XSEDE User Portal
- Select “Resources” tab
- Select “File Manager” tab
- (now wait for Java Applet to load)

portal.xsede.org→Resources → File Manager

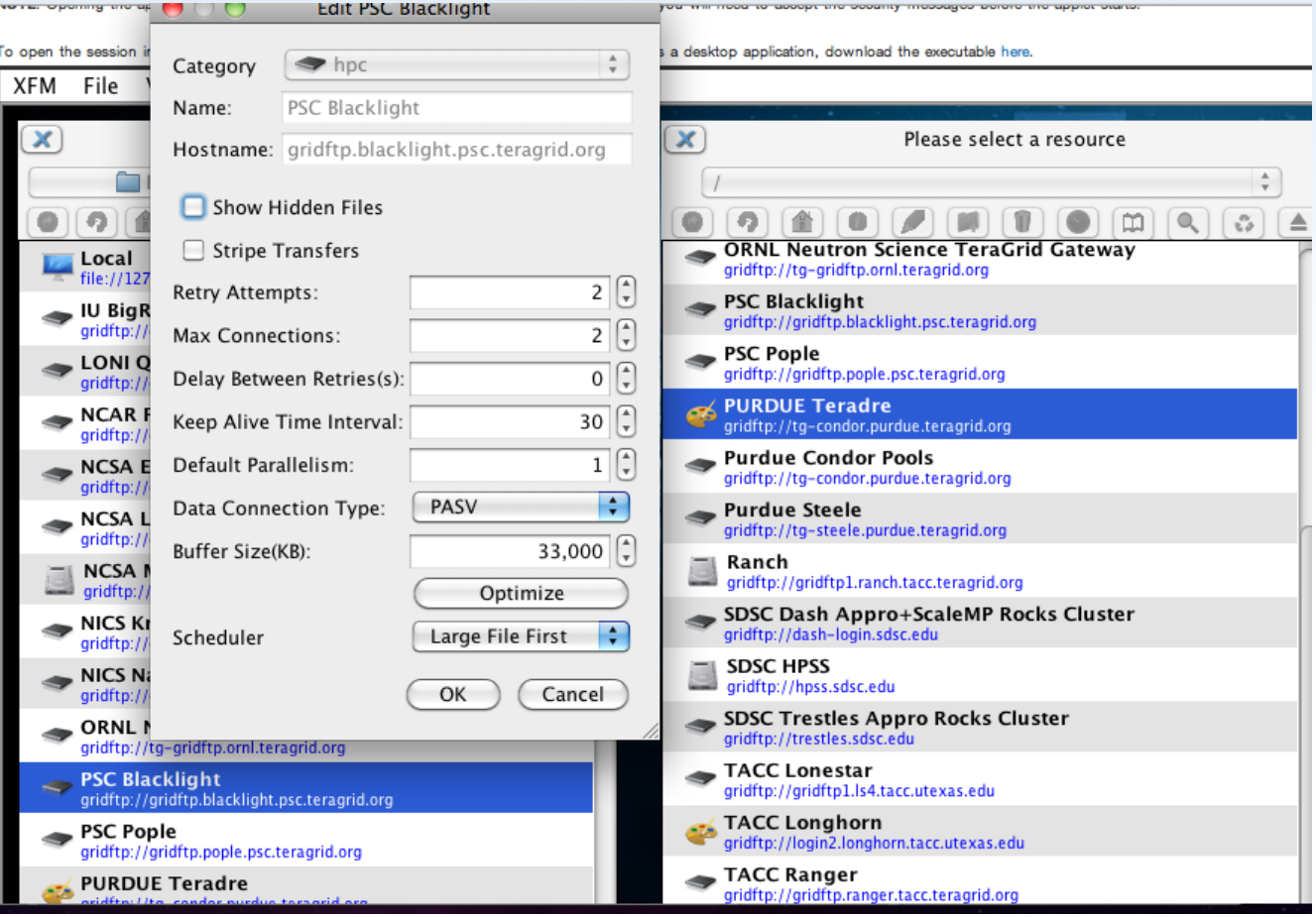


Click on the X (XSEDE logo) to list all machines. *This includes:*

- your local machine.
- XSEDE$Share: 2 GB of space to collaborate. Allows you to share files with your collaborators.

- Click on the 'View' tab and select "Show Transfers" to get information on past and ongoing file transfers and speed.
- Drag files between resources.

portal.xsede.org → Resources → File Manager
Large (> 2 GB) File Transfers



- On the left panel of the file browser, click the XSEDE logo to display your list of available resources.

  Before clicking on the resource, Right click on the resource you're going to transfer data from and select Edit. This will bring up the file transfer parameters:
  - Click on the "Optimize" button
  - Click the checkbox next to "Stripe Transfers"
  - Click OK
  - Repeat for the other panel using the destination resource
  - View → show transfer
  - Repeat this every time you change Resources
- Drag and drop the file from source to destination to transfer

Globus On-line

- Fast, reliable service for high performance file transfer
- High performance: Move terabytes of data in thousands of files
- Automatic fault recovery - Across multiple security domains
- Designed for researchers - Easy "fire and forget" file transfers
- No client software installation - New features automatically available –
- Consolidated support and troubleshooting - Works with existing GridFTP servers
- Ability to move files to any machine (even your laptop) with ease
- Create a free account at www.globusonline.org ; For help: send email to support@globusonline.org

File Systems on XSEDE Resources

Where your data resides on XSEDE and the appropriate storage is your responsibility. In general, all resources provide:

- $HOME: Permanent space, but small. A good choice for building software and working file collections of small to medium sized files, where a medium sized file is less than 50 MB.
- $SCRATCH: More space, but TEMPORARY. Recommended for running jobs. **Backup your files left here! They get deleted (purged).**
- Archival (mass) storage: Long-term storage of large amounts of data, is accessible from all sites, slower access time. Tar files before storing.

Login nodes

- When you login, you are in the login node.
- Login nodes should only be used for basic tasks such as file editing, code compilation, data backup, and job submission.
- Login nodes should not be used to run production simulations. Production work should be performed on the system's compute resources.

Batch Jobs

- Compute jobs *cannot* be run on the login nodes.
- All XSEDE compute resources use some form of batch scheduler.
- There are several batch systems in use, but all work basically the same way. Create a job script specifying:
  - Number/type of nodes you need.
  - How long you need to run.
  - Where your output files should be written to.

Create a script

```
#!/bin/csh
#PBS -l ncpus=16
#ncpus must be a multiple of 16
#PBS -l walltime=5:00
#PBS -j oe
#PBS -q batch

set echo

ja

#move to my $SCRATCH directory
cd $SCRATCH

#copy executable to $SCRATCH
cp $HOME/mympi .

#run my executable
mpirun -np $PBS_NCPUS ./mympi

ja -chlst
```

- Example script for running an MPI job on Blacklight at PSC.
- Actual commands are site and machine specific, but they follow general principles.
- Needs to be modified to run on other XSEDE machines.

Submitting/Manipulating Batch jobs

- Batch system should be used to run your job.
- Do not run on the login nodes.
- Submit the script that you have created:

Actual commands are machine specific, but they follow general principles.

qsub jobname

qstat –a

qstat -u username

qdel jobid

man qsub

Batch command examples

- **qsub amber.job**
- **qstat –a**

```
Job ID  Username Queue  Jobname  SessID NDS  Tasks Memory Time  S Time
------- -------- ------ -------- ------ ---- ----- ------ ----- - -----
29668   user1   batch   job2    21909   1    256   --   08:00  R    02:28
29894   user2   batch   run128   --     1    128   --   02:30    Q   --
29895   user3   batch   STDIN   15921   1     1    --   01:00  R 00:10
29896   user2   batch   jobL    21988   1   2048   --   01:00 R 00:09
29897   user4   batch   STDIN   22367   1     2    --   00:30   R 00:06
29898   user1   batch   amber    25188  1     1    --   01:10  R    00:00
```

- **qdel 29668**
- **After job 29898 runs: user1 should get file amber.job.o29898 with output/errors (log file)**

Why has my job not run?

- Never made it to the queue:

- Job not accepted by the queue:

Core requests on Kraken must be a multiple of twelve. You have requested an invalid number of cores (8). Please resubmit the job requesting an appropriate number of cores.

- Solution: Change the job script to request correct number of cores or memory for the resource.

My job did not complete

- Check the log files created job.e.89890
  job.o.89890
- One common problem: job run out of CPU time.
- Check the job script: time and memory requested, directory where you are writing files to.
- Do the input files exist in the directory where you specified?
- Do you have permission to use software?
- Waiting a very long time in the queue…
- If all fails… submit a ticket

Queue structure: job priority

- Job priority in the batch queues is based on the number of cores and wall clock time requested. Differs by site. Examples:

- Blacklight: *approx*. FIFO system. (Mechanisms in place to prevent a single user from dominating the batch queue and to prevent idle time on the machine).

- *Flexible time request* can improve your turnaround. *Packing small jobs*.

- Kraken: Priority to jobs that request large number of cores (over 32K processors) (except capability and dedicated jobs). Jobs with smaller core counts run on other systems (Trestles). However, they can run effectively on Kraken as *backfill*.

Queue structure: Backfill, Flexible time

- *Backfill*: While the scheduler is collecting nodes for larger jobs, those with short wall clock limits and small core counts may use those nodes without delaying the start time of the larger job.

- The system will not start a job that will not finish before the system maintenance time begins. Ex. Will run a 512 core jobs next. Waiting to finish a 256 core job that will take 4 more hours. Can run jobs that add up to 256 cores and will finish in 4 hours.

- To take advantage of this, request flexible walltime in your job script. A flexible walltime request can improve your job's turnaround in several circumstances.

Improving job turnaround

- Try to be as accurate as possible in estimating the walltime request for your job. Asking for more time than your job will actually need will almost certainly result in poorer turnaround for your job: Asking for the maximum walltime you can ask for a job will almost always result in poorer turnaround.

- Use flexible walltime

Improving job turnoaround: Flexible time

-l walltime_min=HH:MM:SS

-l walltime_max=HH:MM:SS

- Using flexible walltime limits increases the opportunity for your job to run on backfill blades.
- **Example:** if your job requests 64 cores and a range of walltime between 2 and 4 hours and a 64-core slot is available for 3 hours, your job could run in this slot with a walltime request of 3 hours. If your job had asked for a fixed walltime request of 4 hours it would not have been started.

Flexible time

- If the system starts one of your jobs with a flexible walltime request, it selects a walltime within the two specified limits. This walltime will not change during your job's execution. Can determine the walltime your job was assigned by

qstat –f $PBS_JOBID | grep Resource_List.walltime

- Your program should begin writing checkpoint files sufficiently in advance of the walltime so that the file writing is completed when the limit is reached. ***Save time to allow your job to transfer files after your program ends but before your job ends.***

timeout --timeout=$PROGRAM_TIME -- mpirun -np 32 ./mympi

Packing your jobs

- Running many small jobs places a great burden on the scheduler and is probably inconvenient for you.

- Pack many executions into a single job, which you then submit to PBS with a single qsub command.

Packing your jobs

Run each program execution in the background and place a wait command after each execution. Sample job to pack serial executions:

```
#!/bin/csh
#PBS -l ncpus=96
#PBS -l walltime=5:00
#PBS -q batch
dplace -c 0 ./myserial1 < serial1.dat &
dplace -c 32 ./myserial2 < serial2.dat &
dplace -c 64 ./myserial3 < serial3.dat &
wait
```

Packing your jobs, serial or MPI:

```
dplace -c 0 ./myserial1 < serial1.dat &
dplace -c 32 ./myserial2 < serial2.dat &
dplace -c 64 ./myserial3 < serial3.dat &
wait
```

- The dplace command insures that each execution will run on its own set of 32 cores. The executions will run concurrently.
- Same approach using the dplace command can be used to pack MPI executables.
-

Packing your jobs: OpenMP

- To pack OpenMP executables, replace the dplace command with the omplace command. Sample job to pack OpenMP executables:

- omplace -nt 32 -c 0 ./myopenmp1 < myopenmp1.dat &
- omplace -nt 32 -c 32 ./myopenmp2 < myopenmp2.dat &
- omplace -nt 32 -c 64 ./myopenmp3 < myopenmp3.dat &
- omplace -nt 32 -c 96 ./myopenmp4 < myopenmp4.dat &
- wait

Managing Your Environment: **Modules**

- Allows you to manipulate your environment.
- 'module list' shows currently loaded modules.
- 'module avail' shows available modules.
- 'module show' <name> describes module.

http://modules.sourceforge.net/

```
% module load gcc/3.1.1
% which gcc
/usr/local/gcc/3.1.1/linux/bin/gcc

% module switch gcc/3.1.1 gcc/3.2.0
% which gcc
/usr/local/gcc/3.2.0/linux/bin/gcc

% module unload gcc
% which gcc
gcc not found
```

Using modules in your job script

- First,

source /usr/share/modules/init/*shell-name*

- module load module-name

- Check on-line documentation

Exercises

- Portal account
- ssh –l username login.xsede.org
- And from there:
- gsissh blacklight.psc.teragrid.org
- gsissh lonestar.tacc.xsede.org
- or
- gsissh trestles.sdsc.xsede.org

Exercise

- Run the Molecular Dynamics Code Amber
- Amber needs as input:
- a structure
- a topology file
- An input file with the details of what we want to run.

- ssh –l *username* login.xsede.org
- And from there:
- gsissh blacklight.psc.teragrid.org
- Copy input files from
- cp ~*username*01/Blacklight/Input/* .

- ssh –l *username* login.xsede.org
- And from there:
- gsissh trestles.sdsc.xsede.org
- Copy input files from
- cp ~*user1*/Input/* .

Amber input file: md.in

```
Test case
4ps MD of continuation dynamics from restart file
 &cntrl
  imin = 0, irest = 1, ntx = 7,
  ntb = 2, pres0 = 1.0, ntp = 1,
  taup = 2.0, ig= -1,
  cut = 10.0, ntr = 0,
  ntc = 2, ntf = 2,
  tempi = 300.0, temp0 = 300.0,
  ntt = 3, gamma_ln = 1.0,
  nstlim = 10, dt = 0.002,
  ntwr = 10, ntwx=2
/
```

Job script

```
#!/bin/csh
#PBS -l ncpus=16
#PBS -l walltime=10:00
#PBS -j oe
#PBS -q batch
set echo
cd $PBS_O_WORKDIR
source /usr/share/modules/init/csh
module load amber
mpirun -np 16 pmemd.MPI  -p prmtop -c restrt -r restrt_new -x
mdcrd_new -o mdout_new -O -i md.in
```

Trestles:

- #!/bin/bash
- #PBS -l nodes=1:ppn=4
- #PBS -l walltime=10:00
- #PBS -j oe
- #PBS -q normal
- #PBS -A uic157
- #PBS -N MPI_JOB
- #PBS -V
- set echo
- cd $PBS_O_WORKDIR
- . /etc/profile.d/modules.sh
- #run my executable
- module load amber
- echo $PBS_NODEFILE
- export EXE=`which pmemd.MPI`
- mpirun_rsh -np 4 -hostfile $PBS_NODEFILE $EXE -p prmtop -c restrt -r restrt_new -x mdcrd_new -o mdout_new -O -i md.in

Exercise:

- Submit the job (qsub amber.job)
- Monitor the job (qstat –a)
- Make sure you got the output files

Output files: amber.job.o270137

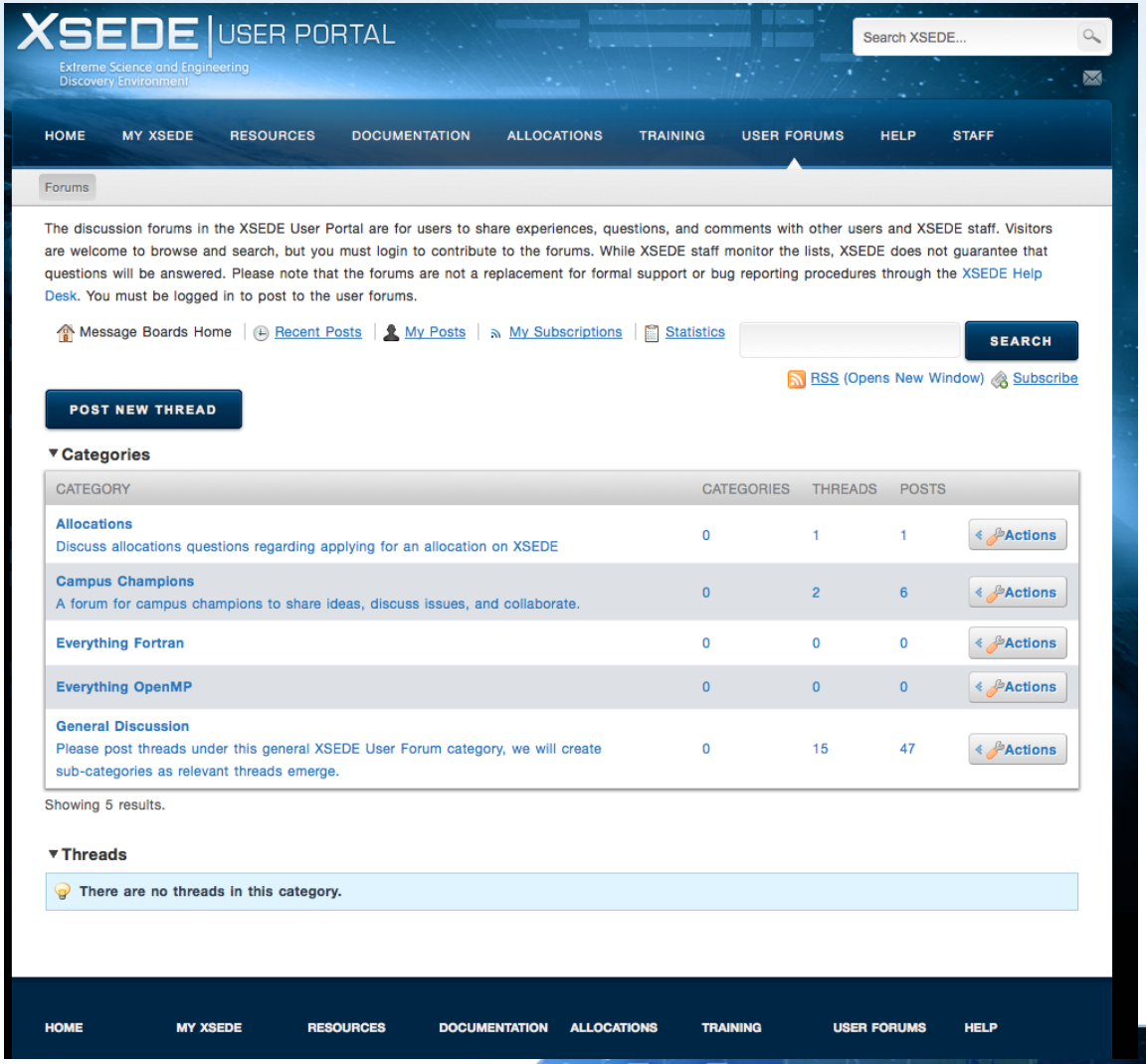
- setenv _LMFILES_ /usr/local/opt/modulefiles/torque/2.3.13_psc:/usr/share/modules/modulefiles/mpt/2.04:/usr/local/opt/modulefiles/ifort/11.1:/usr
- /local/opt/modulefiles/icc/11.1:/usr/local/opt/modulefiles/imkl/10.3.3:/usr/local/opt/modulefiles/psc_path/1.0:/usr/local/opt/modulefiles/globus
- /5.2.2:/usr/local/opt/modulefiles/xdusage/1.0-r7:/usr/local/opt/modulefiles/amber/12
- setenv LOADEDMODULES torque/2.3.13_psc:mpt/2.04:ifort/11.1:icc/11.1:imkl/10.3.3:psc_path/1.0:globus/5.2.2:xdusage/1.0-r7:amber/12
- setenv AMBER12 /usr/local/packages/AMBER12
- mpirun -np 16 pmemd.MPI -p prmtop -c restrt -r restrt_new -x mdcrd_new -o mdout_new -O -i md.in

Trestles:

- more MPI_JOB.o1519321
- /opt/torque/aux//1519321.trestles-fe1.sdsc.edu
- FORTRAN STOP
- FORTRAN STOP
- FORTRAN STOP
- FORTRAN STOP
- Nodes: trestles-1-5

Discussing your problems…
User Portal: User Forums

- The User Forums are a great place to ask questions, get help, or discuss ideas about XSEDE.



- We are here to help you…

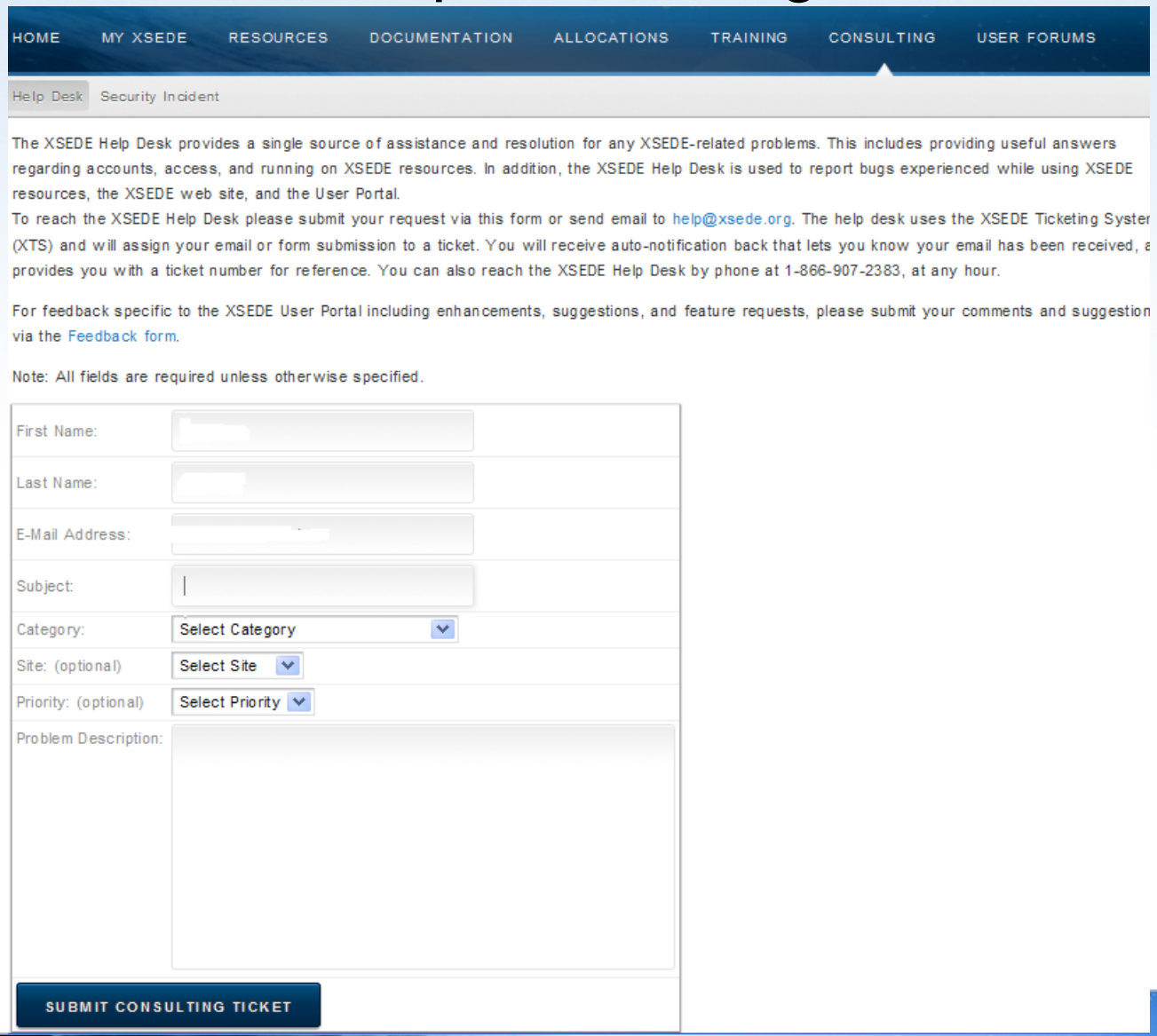
Need help? Reporting and Tracking Issues

- portal.xsede.org → Help

  Submit ticket

- portal.xsede.org → My XSEDE → Tickets
  - Submit ticket
  - View past tickets (both open and closed)
- Can also email help@xsede.org or call 1-866-907-2383, at any hour (24/7)

Help: Submitting a ticket



Preferred method:

Helps ensure all relevant details are provided for quick routing

XSEDE

More "helpful" resources

xsede.org→User Services

- Resources available at each Service Provider
  - User Guides describing memory, number of CPUs, file systems, etc.
  - Storage facilities
  - Software (Comprehensive Search)
- Training: portal.xsede.org → Training
  - Course Calendar
  - On-line training
- Get face-to-face help from XSEDE experts at your institution; contact your local Campus Champions.
- Extended Collaborative Support (formerly known as Advanced User Support (AUSS))

November 6, 2014

Thanks for listening and welcome to XSEDE!