August 11, 2014

XSEDE at a Glance

Aaron Gardner (agardner@ufl.edu)
Campus Champion - University of Florida

XSEDE

Extreme Science and Engineering Discovery Environment

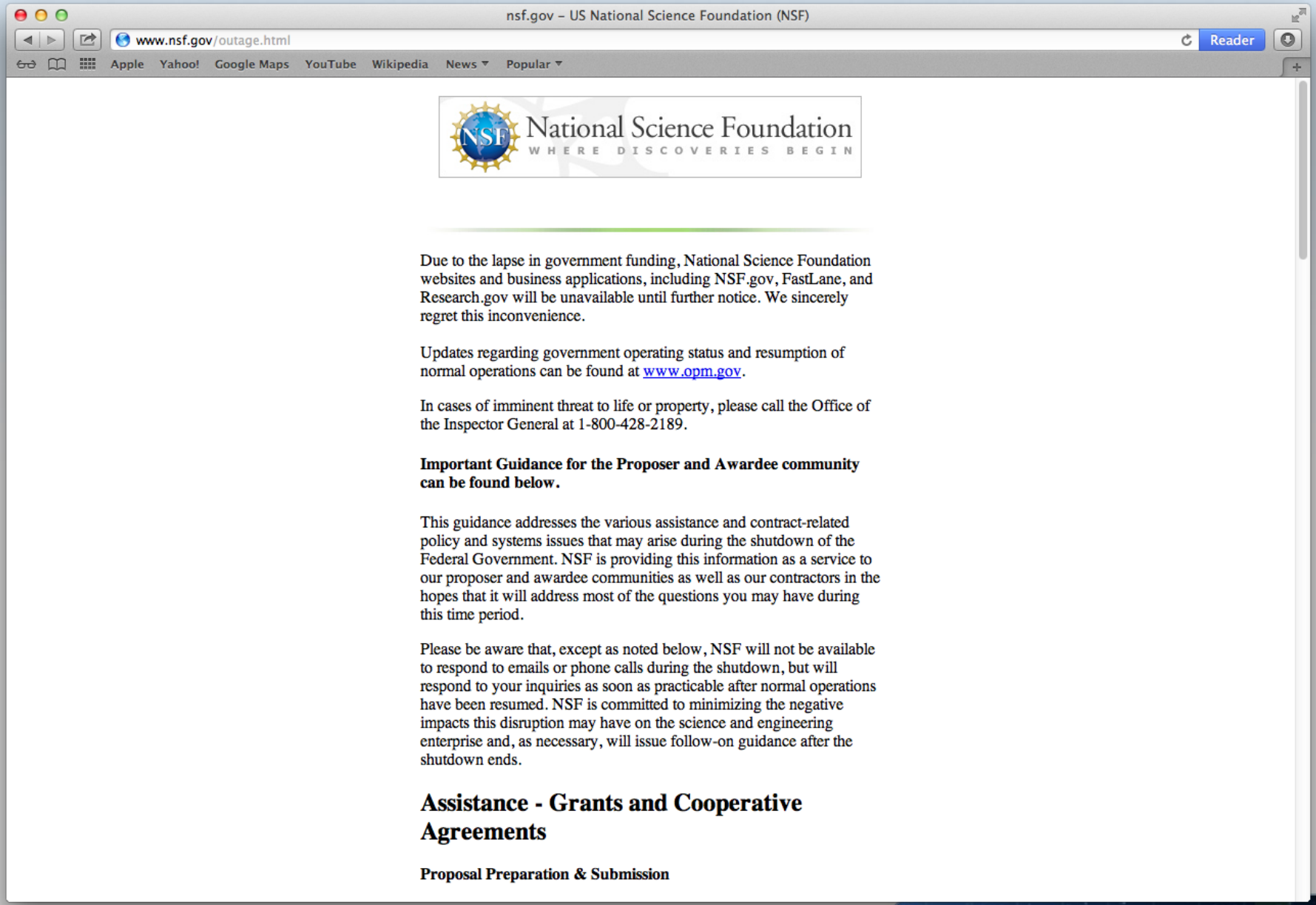
NSF

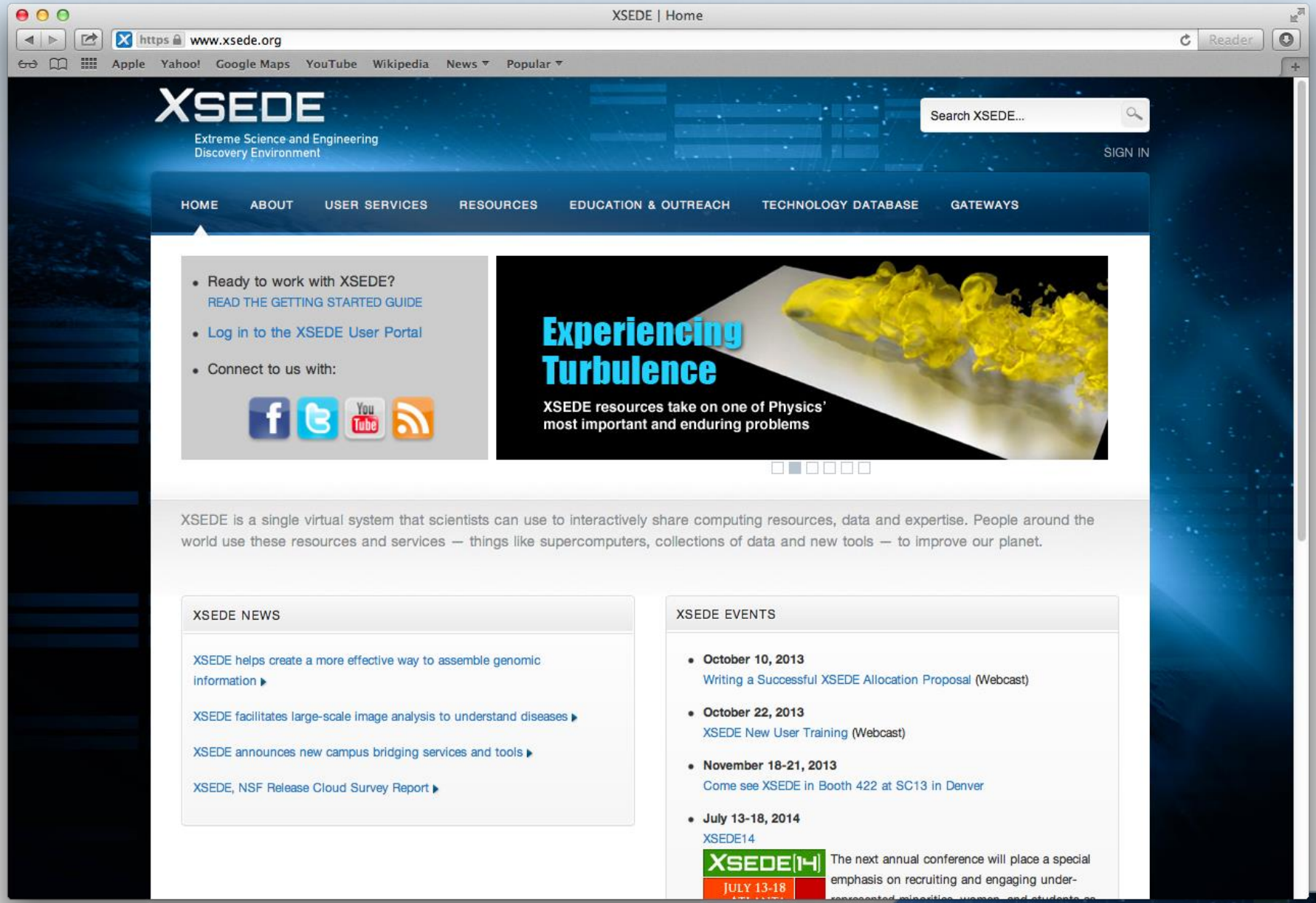
What is XSEDE?

The Extreme Science and Engineering Discovery Environment (XSEDE) is **the most advanced**, powerful, and **robust collection** of integrated **digital resources** and services in the world. It is a **single virtual system** that scientists can use to interactively share **computing resources**, **data**, and **expertise**.

What is XSEDE?

- World's largest infrastructure for open scientific discovery
- 5 year, $121 million project supported by the NSF
- Replaces and expands on the NSF TeraGrid project
  - More than 10,000 scientists used TeraGrid
- XSEDE continues same sort of work as TeraGrid
  - Expanded scope
  - Broader range of fields and disciplines
- Leadership class resources at partner sites combine to create an integrated, persistent computational resource
- Allocated through national peer-review process
- Free* (see next slide)


nsf.gov – US National Science Foundation (NSF)
www.nsf.gov/outage.html
Reader
Apple
Yahoo!
Google Maps
YouTube
Wikipedia
News
Popular
National Science Foundation
WHERE DISCOVERIES BEGIN
Due to the lapse in government funding, National Science Foundation websites and business applications, including NSF.gov, FastLane, and Research.gov will be unavailable until further notice. We sincerely regret this inconvenience.
Updates regarding government operating status and resumption of normal operations can be found at www.opm.gov.
In cases of imminent threat to life or property, please call the Office of the Inspector General at 1-800-428-2189.
Important Guidance for the Proposer and Awardee community can be found below.
This guidance addresses the various assistance and contract-related policy and systems issues that may arise during the shutdown of the Federal Government. NSF is providing this information as a service to our proposer and awardee communities as well as our contractors in the hopes that it will address most of the questions you may have during this time period.
Please be aware that, except as noted below, NSF will not be available to respond to emails or phone calls during the shutdown, but will respond to your inquiries as soon as practicable after normal operations have been resumed. NSF is committed to minimizing the negative impacts this disruption may have on the science and engineering enterprise and, as necessary, will issue follow-on guidance after the shutdown ends.
Assistance - Grants and Cooperative Agreements
Proposal Preparation & Submission

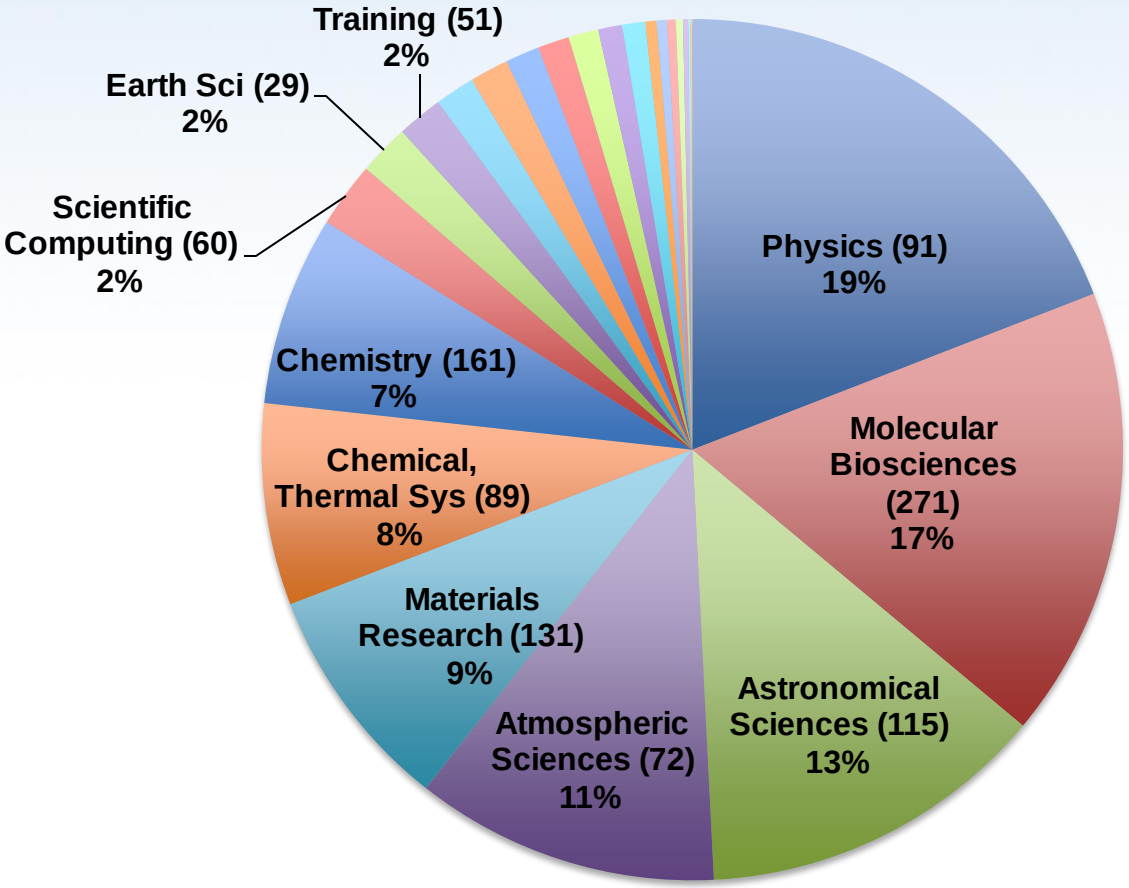

XSEDE | Home
https www.xsede.org
Reader
Apple Yahoo! Google Maps YouTube Wikipedia News Popular
XSEDE
Extreme Science and Engineering Discovery Environment
Search XSEDE...
SIGN IN
HOME ABOUT USER SERVICES RESOURCES EDUCATION & OUTREACH TECHNOLOGY DATABASE GATEWAYS
Ready to work with XSEDE?
READ THE GETTING STARTED GUIDE
Log in to the XSEDE User Portal
Connect to us with:
Experiencing Turbulence
XSEDE resources take on one of Physics' most important and enduring problems
XSEDE is a single virtual system that scientists can use to interactively share computing resources, data and expertise. People around the world use these resources and services — things like supercomputers, collections of data and new tools — to improve our planet.
XSEDE NEWS
XSEDE helps create a more effective way to assemble genomic information
XSEDE facilitates large-scale image analysis to understand diseases
XSEDE announces new campus bridging services and tools
XSEDE, NSF Release Cloud Survey Report
XSEDE EVENTS
October 10, 2013
Writing a Successful XSEDE Allocation Proposal (Webcast)
October 22, 2013
XSEDE New User Training (Webcast)
November 18-21, 2013
Come see XSEDE in Booth 422 at SC13 in Denver
July 13-18, 2014
XSEDE14
XSEDE[H]
JULY 13-18
The next annual conference will place a special emphasis on recruiting and engaging under-

What is Cyberinfrastructure?

- "**Cyberinfrastructure** is a technological solution to the problem of efficiently connecting data, computers, and people with the goal of enabling derivation of novel scientific theories and knowledge."[1]
- Term was used by the NSF Blue Ribbon committee in 2003 in response to the question: "How can NSF… remove existing barriers to the rapid evolution of high performance computing, making it truly usable by all the nation's scientists, engineers, scholars, and citizens?"
- The **TeraGrid**[2] is the NSF's response to this question.
- **Cyberinfrastructure** is also called **e-Science**[3]

[1]Source: Wikipedia
[2]More properly, the TeraGrid in it's current form: the "Extensible Terascale Facility"
[3]Source: NSF

Who Uses XSEDE?



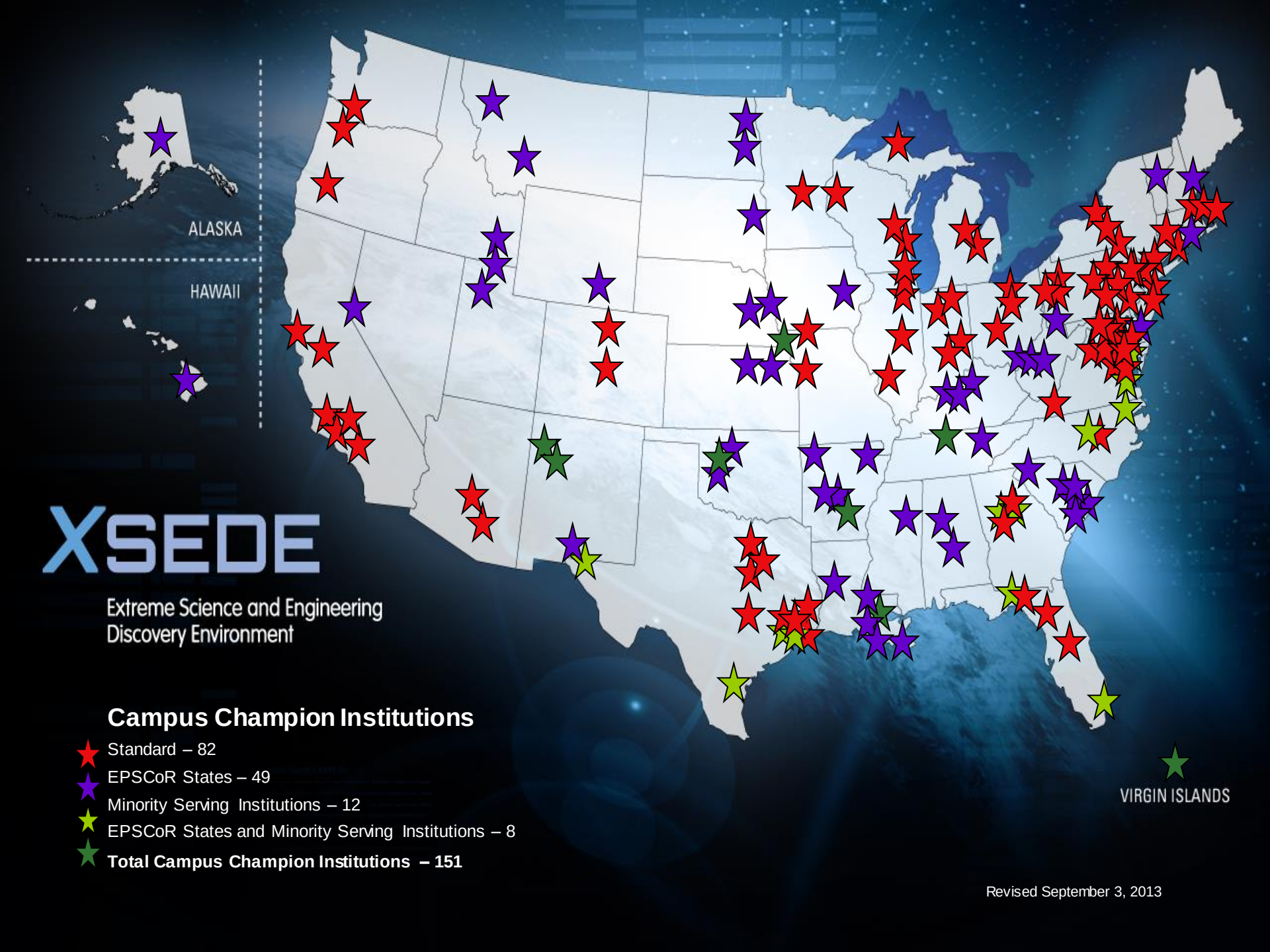
- >2 billion cpu-hours allocated
- 1400 allocations
- 350 institutions
- 32 research domains

XSEDE Supports a Breadth of Research

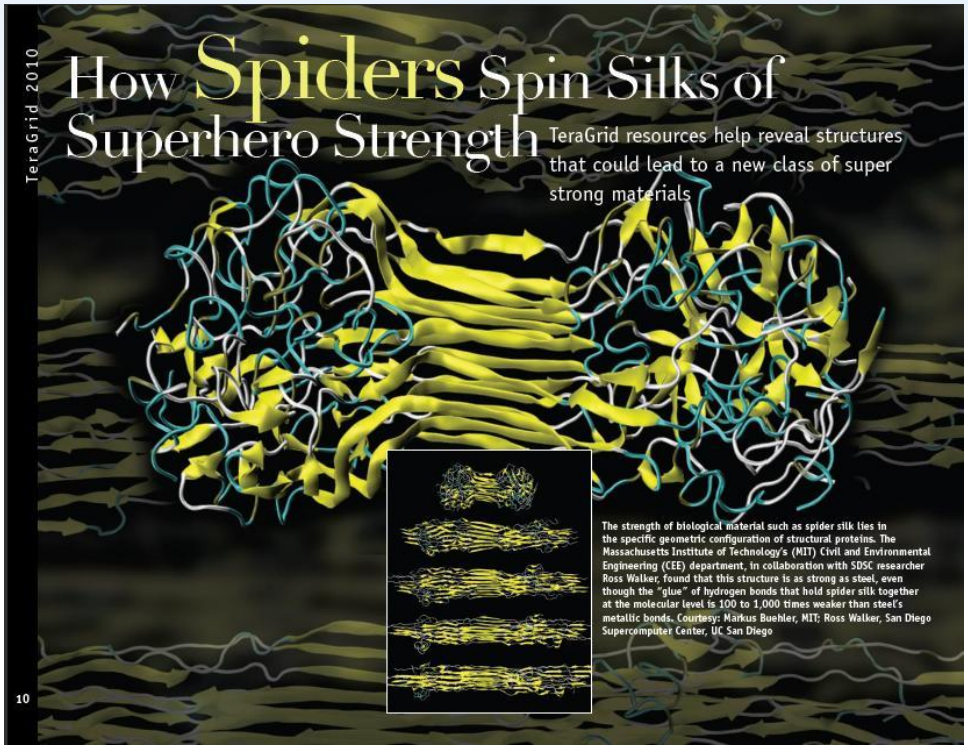
From direct contact with user community as part of requirements collections:

- Earthquake Science and Civil Engineering
- Molecular Dynamics
- Nanotechnology
- Plant Science
- Storm modeling
- Epidemiology
- Particle Physics
- Economic analysis of phone network patterns
- Brain science
- Analysis of large cosmological simulations
- DNA sequencing
- Computational Molecular Sciences
- Neutron Science
- International Collaboration in Cosmology and Plasma Physics

Sampling of much larger set. Many examples are new to use of advanced digital services. Range from petascale to disjoint HTC, many are data driven. XSEDE will support thousands of such projects.


ALASKA
HAWAII
XSEDE
Extreme Science and Engineering Discovery Environment
Campus Champion Institutions
Standard – 82
EPSCoR States – 49
Minority Serving Institutions – 12
EPSCoR States and Minority Serving Institutions – 8
Total Campus Champion Institutions – 151
VIRGIN ISLANDS
Revised September 3, 2013

Who Uses XSEDE? – Spider Silk



PI: Markus Buehler

Institution: MIT

"We found that the structure of spider silk at the nanoscale can explain why this material is as strong as steel, even though the "glue" of the hydrogen bonds holding spider silk together at the molecular level is 100 to 1,000 times weaker than steel's metallic bonds." says Buehler.

Excerpts from "TeraGrid Science Highlights 2010"

Data Mining and Text Analysis

PI: Sorin Matei, David Braun
Institution: Purdue University

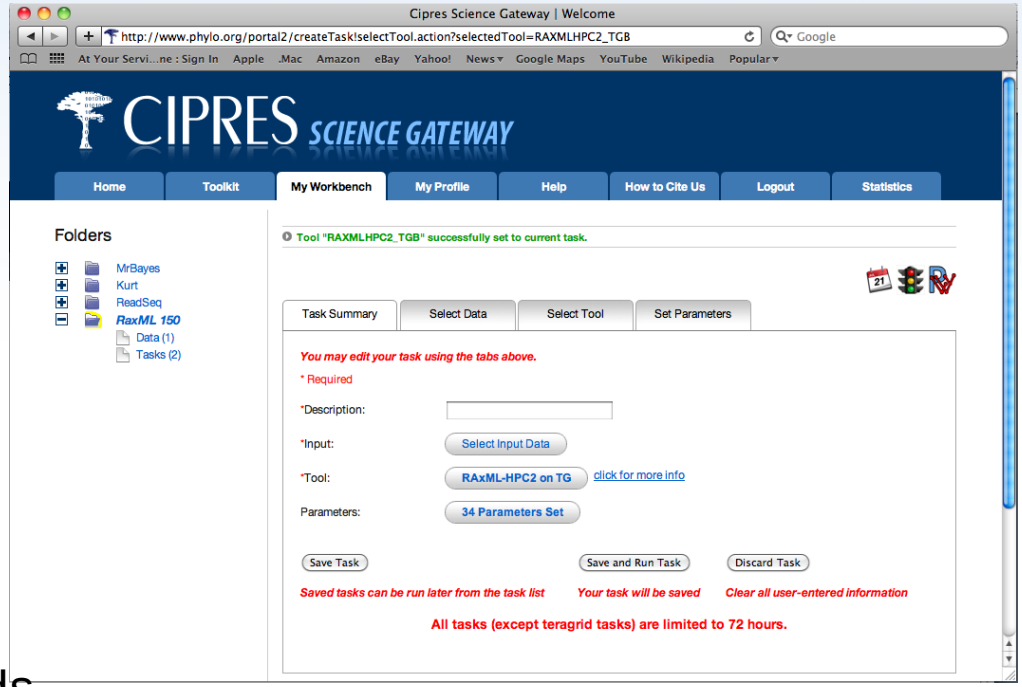
Purdue researchers led by Sorin Adam Matei are analyzing the entire collection of articles produced in Wikipedia from 2001-2008, and all their revisions – a computationally demanding task made possible by TeraGrid resources.

"We looked at how article production is distributed across users contributions relative to each other over time. The work includes visualizations of patterns to make them easier to discern," says Matei.

Excerpts from "TeraGrid Science Highlights 2010"

XSEDE "Science Gateways" for Bioinformatics

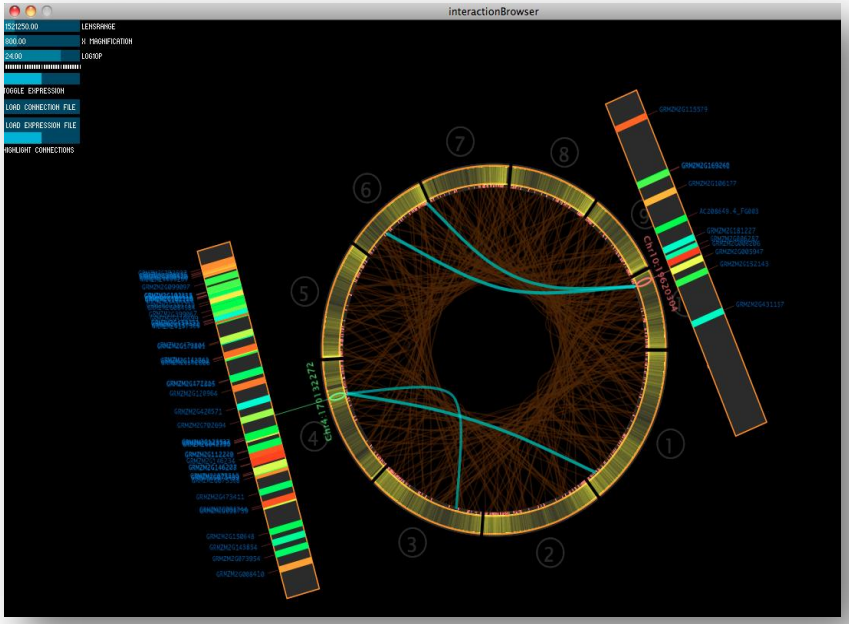
- Web-based Science Gateways provide access to XSEDE
- CIPRES provides:
  - BEAST
  - GARLI
  - MrBayes
  - RAxML
  - MAFFT
- High performance, parallel applications run at SDSC
- 4000 users of CIPRES and hundreds of journal citations



**Adapted from information provided by Wayne Pfeiffer, SDSC*

Who Uses XSEDE? – iPlant

Science goals by 2015: Major emerging computational problem is deducing Phenotype from Genotype, e.g. QTL (Quantitative Trait Locus) mapping - accurate prediction of traits (e.g. drought tolerance for maize) based on genetic sequence. Based on data collected in hundreds of labs around the world and stored in dozens of online distributed databases.

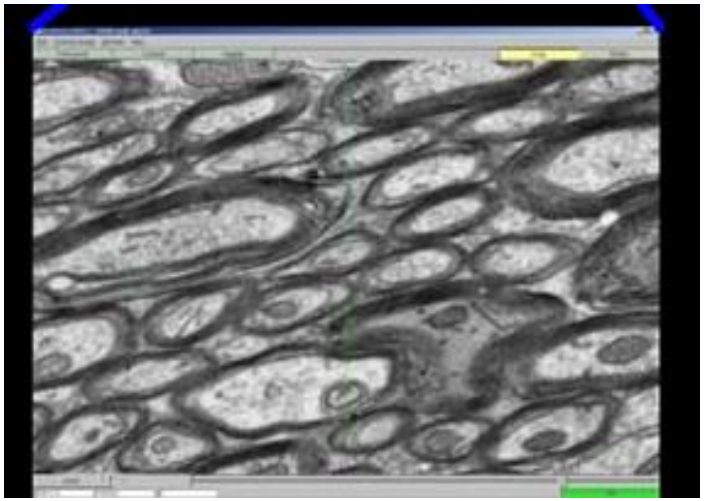


Infrastructure needs: This data-driven petascale combinatoric problem requires high speed access to both genotypic and phenotypic databases (distributed at several sites).

XSEDE will provide the coordinated networking and workflow tools needed for this type of work.

Brain Science-Connectome

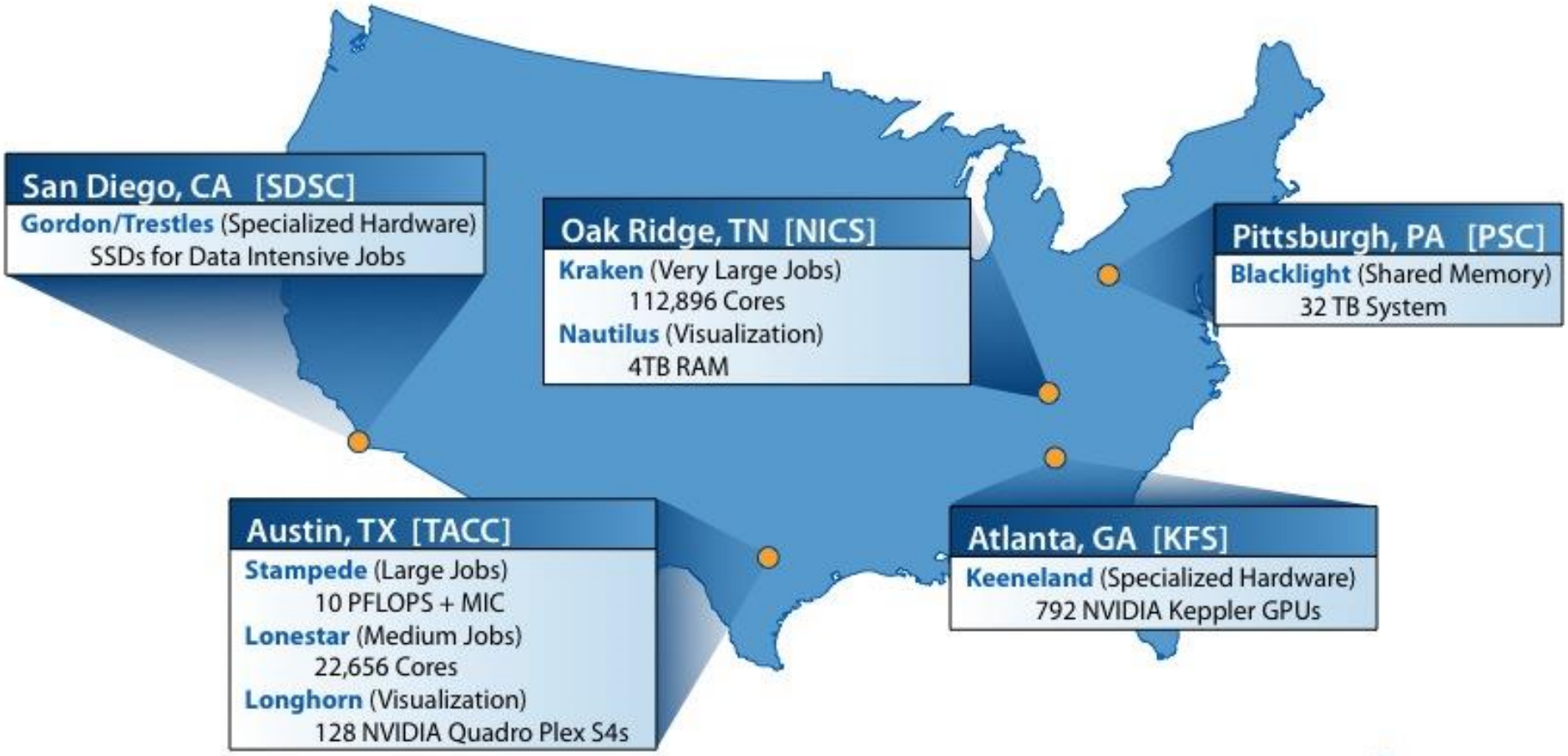
Science goals by 2015: Capture, process, and analyze ~1 mm^3 of brain tissue, reconstructing complete neural wiring diagram at full synaptic resolution; present resulting image data repository to national community for analysis and visualization



Infrastructure Needs: High-throughput transmission electron microscopy (TEM) high-resolution images of sections must be processed, registered (taking warping into account), and assembled for viewing; Raw data (>6 PB), must be archived; TEM data must be streamed in near real time at sustained ~1 GB/s. Results in ~3 PB of co-registered data.

As with all large datasets that researchers throughout the country will want to access, XSEDE's data motion, network tuning, and campus bridging capabilities will be invaluable.

What Resources does XSEDE Offer?



CONTACT: HELP@XSEDE.ORG

Data Storage and Transfer

- SDSC Gordon
  - SSD system with fast storage
- NCSA Mass Storage System
  - http://www.ncsa.illinois.edu/UserInfo/Data/MSS
- NICS HPSS
  - http://www.nics.utk.edu/computing-resources/hpss/
- Easy data transfer
  - In-browser SFTP or SCP clients through Portal SSH
- Standard data transfer
  - SCP to move data in/out of XSEDE systems
    - Requires SSH key setup
  - Rsync to move data in
- High performance data transfer
  - Globus Online: https://www.globusonline.org/

Support Resources

- Local Campus Champion
  - That's me!
- Centralized XSEDE help
  - help@xsede.org
- Extended one-on-one help (ECSS):
  - https://www.xsede.org/ecss
- Training
  - http://www.xsede.org/training

Other Resources

- Science Gateways
- Extended Support
- Open Science Grid
- FutureGrid
- Blue Waters (NCSA)
- Titan (ORNL/NICS)
- ALCF (Argonne)
- Hopper (NERSC)

“Why Should I Care About XSEDE?”

- Tap into community knowledge (see next slide)
- Extended Collaborative Support Service (ECSS)
- Resources with complementary characteristics to those found locally
- Extending network of collaborators
- The XSEDE community (noticing a theme yet?)

“Why Should I Care About XSEDE?”

San Diego Supercomputer Center: Scientists Help Tame Tidal Wave of Genomic Data Using SDSC's Trestles

www.sdsc.edu/News%20Items/PR091913_genome.html

Apple Yahoo! Google Maps YouTube Wikipedia News Popular

Contact Us Site Map Staff Directory

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

Search SDSC GO

RESEARCH & COLLABORATION | RESOURCES & SERVICES | USER SUPPORT | DISCOVERIES | NEWS CENTER | ABOUT SDSC

Home > News Center > SDSC Headlines

News Center

- SDSC Headlines
- News Nuggets
- In The News
- Events
- Profiles
- Publications
- Multimedia
- Subscribe
- For Media

09/19/2013

Scientists Help Tame Tidal Wave of Genomic Data Using SDSC's Trestles

Xifeng Yan, UC Santa Barbara

Sequencing the DNA of an organism, whether human, plant, or jellyfish, has become a straightforward task, but assembling the information gathered into something coherent remains a massive data challenge. Researchers using computational resources at the San Diego Supercomputer Center (SDSC) at the University of California, San Diego, have created a faster and more effective way to assemble genomic information, while increasing performance.

In a paper presented the past month at the 39th International Conference on Very Large Databases (VLDB2013) in Riva del Garda, Italy, Xifeng Yan, the Venkatesh Narayanamurti Chair of Computer Science at the University of California, Santa Barbara, explains how he used SDSC's *Trestles* compute cluster to help develop a new algorithm called MSP (minimum substring partitioning) that helps to assemble genomes with extreme efficiency. MSP is a critical part of a pipeline, or a group of software that assembles entire genomes, with each piece of the software doing one part of the job. Yan and his colleagues were able to optimize one of two steps to use a mere 10 gigabytes of memory without runtime slowdown.

"High-quality genome sequencing is foundational to many critical biological and medical problems," said Yan. "With the advent of massively parallel DNA sequencing technologies, how to manage and process the big sequence data has become an important issue. Experimental results showed that MSP can not only successfully complete the tasks on very large datasets within a small amount of memory, but also achieve better performance than existing state-

"OK I Care, How Do I Get Started?"

- **Campus Champion**
  - Get your feet wet with XSEDE
  - < 10k cpu-hours
  - 2 day lead time
- **Start-Up**
  - Benchmark and gain experience with resources
  - 200k cpu-hours
  - 2 week lead time
- **Education**
  - Class and workshop support
  - Short term (1 week to 6 months)
- **XSEDE Research Allocation (XRAC)**
  - Up to 10M cpu-hours
  - 10 page request, 4 month lead time

https://www.xsede.org/how-to-get-an-allocation

Steps to Getting Your Allocation

FREE

- **Step One** – Campus Champion Allocation
  - Log onto the Portal and get an account
  - Send Campus Champion (me!) your portal account ID
- **Step Two** – Start-Up/Educational Allocation
  - Sign up for a startup account
  - Do benchmarking
- **Step Three** – XRAC
  - Requires written proposal and CVs

Campus Champion Role Summary

- What I will do for you:
  - Help setup your XSEDE portal account
  - Get you acquainted with accessing XSEDE systems
  - Walk you through the allocations process
  - Answer XSEDE related questions that I have the answers to
  - Get you plugged in with the right people from XSEDE when I don't know the answer
  - Pass issues and feedback to the community with high visibility
  - Help evangelize XSEDE at events you host or directly with your colleagues

- What I won't do for you:
  - Fix code
  - Babysit production jobs
  - Plug the government back in

Acknowledgements & Contact

Presentation Content Thank You
Jeff Gardner (University of Washington)
Kim Dillman (Purdue University)
Other XSEDE Campus Champions
The XSEDE Community at Large

Aaron Gardner
agardner@ufl.edu


Our reach will forever
exceed our grasp, but,
in stretching our horizon,
we forever improve our world.
XSEDE
Extreme Science and Engineering
Discovery Environment