XSEDE overview and Extended Collaborative Support Service (ECSS)

Nancy Wilkins-Diehr
Co-PI (XSEDE), Co-Director (ECSS)
wilkinsn@sdsc.edu
XSEDE: Umbrella organization of NSF HPC investments

• $110M program funded by National Science Foundation, Office of Advanced Cyberinfrastructure

• No funding for major hardware
  — coordination, support and creating a national/international cyberinfrastructure
  — coordinate allocations, support, training and documentation
Current XSEDE Compute Resources

• Stampede2 @ TACC
  – 18 PFLOPS (PF) Intel Knights
  Landing and Xeon Skylake nodes
• Comet @ SDSC
  – 2.1 PF cluster w/GPUs
• Bridges @ PSC
  – 1.3 PF w/ large memory (274 TB)
• XStream @ Stanford
  – 1.0 PF GPU Cray CS-Storm cluster
• SuperMIC @ LSU
  – 925 TF Dell Cluster w/ GPUs and Xeon PHIs
• Jetstream @ Indiana
  – 516 TF HPC Cloud
• Wrangler @ TACC
  – 62 TF data analytics system
• Open Science Grid
  – 60,000 CPU cores

https://www.xsede.org/web/xup/resource-monitor
Data, and Software Resources

• Storage
  – Ranch @ TACC
    • 61 PB tape
  – Pylon @ PSC
    • 10 PB disk
  – Wrangler @ TACC
    • 10 PB disk
  – Data Oasis @ SDSC
    • 4 PB tape

• Software: 100s of titles
  – domain software
    • chemistry, CFD, bioinformatics, physics, astronomy, biology, engineering, statistics,…
  – tools
    • middleware, visualization, scripting, performance analysis, data storage and management,…
  – compilers and libraries
    • most languages supported, math libraries, machine learning,…

https://www.xsede.org/web/xup/resource-monitor
https://portal.xsede.org/software/#
Searchable software catalog
www.xsede.org/software

Software

Find software available on XSEDE Service Provider sites. You can view by Resource, Site or Software type and then search for name, version, URL and more. To view details about a software package, click on the software name to see available versions. For more details click on the version to find out more about the software including how to access the software package.

Are you looking for software that is accessible via a science gateway? Visit the Science Gateways Application List

Help us gauge interest in potential future installations: if there's a software package you'd find useful, submit a ticket to let us know.

Search software

View by:

Resources
Science Category/Domain
Sites

Bridges
abaqus       Abinit       abyss
AIPS         allpaths-lg   anaconda
anaconda5    annovar      ansys
anvio        aragorn      aspera
Clouds, VMs, VCs, containers
New developments in NSF-allocated supercomputers!

https://jetstream-cloud.org/
Atmosphere, OpenStack

• Singularity
  – Containerized supercomputing
  – Reproducibility and archival software and environment stacks
  – Mobility of Compute (portable, sharable, distributable container images)
  – User defined and controlled environments (BYOE)
  – Integratable with existing shared infrastructures and scheduling subsystems
  – Properly make use of the existing high performance physical hardware
  – Make use of all of the work that has been done in Docker so far

Cloud-based and on-demand, the 24/7 system includes discipline-specific apps.

Computation, experimentation, teaching: Jetstream benefits researchers from a range of fields by focusing on usability and support. You can even create virtual machines that look and feel like your lab workstation or home machine, with thousands of times the computing power.

Tech overview
Extended Collaborative Support Services

• Extended
  – Complements Helpdesk, longer term engagement

• Collaborative
  – Need involvement from PI’s group

• Must have a definite goals
  – We’re not just another helping hand

• Five subareas
  – Extended Support for Research Teams
  – Novel and Innovative Projects
  – Extended Support for Community Codes
  – Extended Support for Science Gateways
  – Extended Support for Training, Education and Outreach
ECSS offers support on a variety of topics

• Optimize applications (e.g. increased parallelization, better memory management, choosing appropriate algorithms, better IO, better use of accelerators, data analytics)
  — Individual teams or community code developers

• Integrate Science Gateways and workflows

• Enhance visualization

• Engage members of underrepresented communities and domain areas and mentor them
ECSS is an XSEDE-allocated service

• Expert staff can be requested for collaborations lasting months to a year
  – Requests made through the XSEDE allocation system when requesting computational resources
• Typical collaborations require 20-25% staff time for one year
• Critical mass engenders success
  – ~28 FTEs (~70 individuals), 10 sites
  – Advanced degrees in a variety of science and technology fields
    • Some staff co-author publications or write proposals with PI team, a few are later supported by PI team
Diverse ECSS Expertise Available Because of Program’s Scale

• Fields of expertise: astrophysics, bioinformatics, CFD, chemistry, computer science, climate modeling, engineering, genomics, hydrology, humanities, machine learning, molecular dynamics, phylogenetics, physics, seismology, statistics.

• Technologies: clusters, large shared memory systems, MICs, GPUs, KNL, VMs

• Languages: C, C++, Fortran, MPI, OpenMP, Java, JavaScript, shell programming, CUDA, OpenACC, Python, R, MATLAB

• Techniques: benchmarking, cloud computing, data mining, databases, FFTs, finite element methods, grid generation, grid middleware, Lattice Boltzmann methods, libraries, linear algebra, Monte Carlo methods, parallel debugging, parallel I/O, petascale computing, scheduling, science gateways, visualization, workflows

Cross disciplinary talent is important to our success
Need staff with a variety of domain and CI expertise
Allows true intellectual collaborations

ECSS RY1 Projects Field of Science

- 110 - Mathematical Sciences
- 120 - Astronomical Sciences
- 130 - Physics
- 140 - Chemistry
- 150 - Materials Research
- 310 - Computer and Computation Research
- 340 - Advanced Scientific Computing
- 401 - Biological Sciences
- 410 - Molecular and Cellular Biosciences
- 440 - Behavioral and Cognitive Sciences
- 450 - Social and Economic Science
- 500 - Geosciences
- 510 - Atmospheric Sciences
- 520 - Earth Sciences
- 540 - Polar Programs
- 610 - Chemical and Thermal Systems
- 620 - Mechanical and Structural Systems
- 680 - Industrial Science and Technological Innovation
- 710 - Humanities
Steps to get started

- [http://portal.xsede.org](http://portal.xsede.org)
  - Get an account

- When logged in, submit a request

- Fill out the form
• Startup allocations up to 200,000 CPU hours are reviewed and approved continually

• Larger requests (up to tens of millions of hours) are reviewed quarterly

• Faculty and postdocs can be PIs
  – Students must convince their advisor or connect through the Boise Campus Champions allocation

• Plenty of advice available on how to write a successful request
  – Webinars, online materials
Or push the easy button

• Contact Boise State Campus Champions
  – Kyle Shannon and Jason Watt
• They can add you to their allocation where you have access to any resource right away
• Instant trial allocations on Comet and Jetstream
Plenty of training materials available as well
Whether or not you are using XSEDE resources
www.xsede.org/training

- Online training
- In person training
- Distributed training
  - Live at multiple sites simultaneously
- With demand, we can come to your site
ECSS Symposium
Staff and others can learn from project work

- [www.xsede.org/ecss-symposium](http://www.xsede.org/ecss-symposium)
- Third Tuesday of the month, 10am Pacific
  - Open to all
- Most recent talk, Oct 17
  - Geodynamo Simulation Code for Paleomagnetic Observations
    - Shiquan Su (NCAR) Chad Burdyshaw (NICS)
- Talks available on the XSEDE channel on YouTube
Science Successes as a result of ECSS collaborations
Tornado Genesis
Data Mining for Tornado Genesis

• Long term goal is to reduce the false alarm rate (3/4) and increase warning lead time (15 minutes)

• Use visualization and data mining to explore the factors that separate tornado formation from tornado failure in high resolution simulations of supercell storms
ECSS Contributions

• To date, Foss’ work has identified 6 weather features, that can potentially be used to distinguish storms that generate tornados.

“Greg comes at the problem from a completely different perspective, and provides new ways of looking at the data that you wouldn't have thought of.” – Amy McGovern

“Greg's visualizations are of much higher quality than what meteorologists typically use. The four-dimensionality and high resolution provide a much fuller perspective on how storms and tornadoes evolve. I've been studying these storms for over a decade, and these visualizations have changed the way I think about them.” – Cory Potvin
Genomics
Improving the Quality and Efficiency of Assembling Transcriptomes

• A cell’s transcriptome is the set of genes activated by 'transcribing' them from DNA into RNA
• Researchers must first chemically cut the RNA into small pieces, sequence it, remove RNA not directing cell activity, and then match the overlapping fragments to reassemble the original RNA molecules
• Harder still, they must identify and correct sequencing mistakes, and deal with repetitive sequences that make the origin and number of repetitions of a given RNA sequence unclear
• While software tools exist to undertake all of these tasks, Ghaffari's report was the most comprehensive yet to examine a variety of factors that affect assembly speed and accuracy when these tools are combined in a start-to-finish workflow
ECSS Contributions

• The team helped select the most appropriate platform in XSEDE and then concentrated on optimizing the speed of data movement from storage to memory to the processors and back.
• They also incorporated new verification steps to avoid perplexing errors that arise when wrangling big data through complex pipelines.
• This work will provide tools to help other scientists improve our understanding of how living organisms respond to disease, environment and evolutionary changes.

“We are running some of the largest transcriptome assemblies ever done. Our effort focused on running all these big data sets many different ways to see what factors are important in getting the best quality.”
– Philip Blood
Artificial Intelligence
Poker—a Game of Incomplete Knowledge

• For the first time, an AI program has beaten the world experts at a game of incomplete knowledge, in this case No Limit Texas Hold ’em

• $10^{160}$ possibilities, so exhaustive search not possible

• Really searching for Nash equilibrium, a technique applicable to fields like business, military, cybersecurity and medicine
ECSS Contributions

• Parallelize the original algorithms to use MPI and OpenMP and then helped on optimization

• Facilitated the complex logistics of multi-PB data sets and real time competition usage (600 nodes of Bridges)
Let’s break for Q&A with ECSS staff

• Erin Hodgess
  – U Houston, Downtown
  – MS Economics, PhD Statistics
  – ECSS Affiliate
  – Campus Champion Fellow, 1 year collaboration with ECSS staff to
    • Develop tools for spatio-temporal functions in R
    • Extend skills in MPI and CUDA as they relate to R
    • Compile tools and produce an online class

• Derek Simmel
  – Pittsburgh Supercomputing Center
  – BS, MS Computer Science
  – ECSS Novel and Innovative Projects
  – XSEDE Cyberinfrastructure Integration - Requirements Analysis and Capability Delivery (XCI-RACD)

• Shiquan Su
  – National Center for Atmospheric Research
  – PhD Physics
  – ECSS staff, support for research teams, community codes, projects include
    • Surface-induced forcing and decadal variability and change of the East Asian climate, surface hydrology and agriculture
    • Improving Seasonal Prediction of the Indo-Pacific SST and Indian Monsoon with Multiple Ocean Analyses Ensemble
    • Tuning Geodynamo Simulation to Paleomagnetic Observations

• Kenneth Yoshimoto
  – San Diego Supercomputer Center
  – PhD Molecular Biology
  – Co-founder, Neuroscience Science Gateway
  – Scheduling expertise

– Grid computing, security expertise
Transportation
Autonomous and Connected Vehicles

- Autonomous vehicles sense their environment and navigate without human input
- Connected vehicles (CVs) can exchange messages containing location and other safety-related information with other vehicles, and with devices affixed to roadside infrastructure
- While the basic goal of CVs is safety, the data generated has enormous potential to support transportation planning and operations, which is her real interest
ECSS Contributions

• ECSS experts Weijia Xu and Amit Gupta (TACC) helped Ruiz-Juri and her colleagues figure out which TACC system to use, and settled on Rustler

• They also helped create a friendly user interface to remove some of the hurdles of using a command line

“The interface gave us the opportunity to look at this data now instead of, say, two years (later).”

“It would have been very hard without the help of Amit or Weijia to be able to have visibility and access to HPC from the interface of a preexisting code that we used for modeling.”

“For me the best part is not when we know what we want and they (ECSS) help us, but when they understand enough of what we're doing and can come up with new ideas on their own.”

— Natalia Ruiz-Juri
Humanities
Six Degrees of Francis Bacon

• Warren uses computational techniques such as machine learning, graph inferences, and web development to reconstruct and communicate the social networks (who knew whom) of early modern Britain from about 1500 to 1700.

• Their website allows scholars, students, and citizen humanists to improve the network — add relationships to validate or reject some of the inferences that have been made.

• He has been working with the Oxford Dictionary of National Biography (ODNB), the gold standard of British lives from the Roman Empire to the present.
“I'm not sure we would have become involved in XSEDE if it were not for ECSS.... The collaborative support model was attractive because someone with my background and training was intimidated by the prospect of using supercomputers. Knowing that there was a process to get our team up to speed was incredibly influential in bringing us on board.... ECSS allowed us to extend our early work and move forward with it... I can't say enough about the impact that the ECSS program has had for the project.” – Christopher Warren
American History
Throughout the 1930s and 1940s, Roy Stryker, head of the Information Division for the Farm Security Administration, sent photographers to capture images depicting rural life. Stryker used photography to change public and even Congressional perception of what was needed to alleviate the dire situation of farmers post-Dust Bowl and post-Great Depression. This study analyzes those images. It serves many disciplines in the arts, humanities, American history, etc. 175,000 photographs preserved, now in the Library of Congress.
Significance of this Project

• With the HPC infrastructures, they are empowering researchers to ask big questions like
  – How did the images in this database depict rural life?
  – What visual narratives appear across the corpus?
  – Are there significant differences in approach from photographer to photographer?
  – What can we learn from the images that Stryker killed?

• Most photographs are available in three or four different sizes. They noticed that using lower and higher resolution images does not present the same results. They are pursuing implications.
• This project will benefit not just this corpus of images, but all other image collections, as well as photographers and data miners

• For this research, Wuerffel and her team were awarded the Best Accelerating Discovery Paper at XSEDE16

“Without the ECSS, we wouldn't have been able to do this project. They have been great as far as providing support implementing the algorithms. Their familiarity with the hardware and what can be done has really enabled us to do this project.” – Jeffrey Will
Significant success in anthropology
PI Curtis Marean, Arizona State University

• South African regional climate model ported to XSEDE systems
• Model used to hind-cast ancient glacial climates
• Allows research team to investigate several questions in human origins studies
NIP projects are complex and can require significant support
But they pave the way for others

Our project involved taking the South African regional climate model, moving it to XSEDE systems, and then using that model to hind-cast ancient glacial climates so as to address several questions in human origins studies. This was a very complex task made more difficult by the fact that the lead climate modeler (Dr. Francois Engelbrecht) is located in South Africa. We have now successfully run these simulations, overcoming many hurdles, and none of that would have been possible without David O’Neal. Simply put, the project would have failed without him. On a personal level, I have found him a joy to work with. He is always available and helpful and is an extraordinary problem solver.

Dr. Curtis Marean, Arizona State University
Typical needs from underrepresented domains

• Interactivity
• Familiar environment and tools
  – Matlab, R, Python, others
  – Virtualization can help with some of this
• Science gateways
  – Text analytics, video analytics
• Data focus
  – Transfer, hosting, preparation
  – Advanced analysis including leading technologies and methods like graph analytics and deep learning
Number of gateway users explodes in 2016
Gateways account for 74% of all users

This is largely due to the CIPRES and I-TASSER gateways, but others are gaining
• Consistently ranked as one of the best methods for automated protein structure prediction in the community-wide CASP (Critical Assessment of protein Structure Prediction) experiments
• One of the most widely used systems in the field for online, full-length protein structure and function prediction
• Now available to all, on supercomputers
I-TASSER up and going quickly
No ECSS support needed because support for gateways is well-defined

“The generous grant of computing resources from XSEDE is very helpful in improving the capacity of the I-TASSER system to serve the broader biomedical community by providing faster and higher quality simulations of protein models.”

“It was a pleasure to me when I found the system is very easy to use and many of the settings are identical to our local computer system. In particular, it allows uploading of our library, which relieved my major worry that I had before. So it is actually more useful than what we thought”

- Dr. Yang Zhang, U Michigan School of Medicine

Yang Zhang, PI
Chengxin Zhang, PhD student
Cyberinfrastructure for Phylogenetic Research (CIPRES)
PI Mark Miller, SDSC, www.phylo.org

- 210 US research universities
  - Harvard, Yale, UC Berkeley, Stanford, etc.
  - Non-PhD granting colleges (including one all-women's college, community colleges, and Hispanic-serving institutions)
- 3 K-12 school systems
- 43 non-governmental organizations
  - Museums including the Smithsonian Institution, the American Museum of Natural History, and the Field Museum,
  - Botanical gardens, (e.g. Chicago, Rancho Santa Ana, and New York)
  - Institutes (e.g. JCVI and Broad)
- 10 US governmental agencies
  - Including NIH, USDA, NOAA, US Forest Service
- Curriculum delivery (76)
- 2,000+ publications since 2010
11 codes are supported by CIPRES; most have modest scalability; some run for days

<table>
<thead>
<tr>
<th>Code</th>
<th>Version</th>
<th>Language</th>
<th>Cores charged</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEAST</td>
<td>1.8.2</td>
<td>Java + C++</td>
<td>2, 4, or GPUs</td>
</tr>
<tr>
<td>BEAST2</td>
<td>2.1.3</td>
<td>Java + C++</td>
<td>1, 2, or 3</td>
</tr>
<tr>
<td>DPPDiv</td>
<td>1.0</td>
<td>C++</td>
<td>16</td>
</tr>
<tr>
<td>FastTree</td>
<td>2.1.8</td>
<td>C</td>
<td>3</td>
</tr>
<tr>
<td>GARLI</td>
<td>2.0.1</td>
<td>C++</td>
<td>1 to 24</td>
</tr>
<tr>
<td>jModelTest2</td>
<td>2.1.6</td>
<td>Java + C</td>
<td>8</td>
</tr>
<tr>
<td>MAFFT</td>
<td>7.187</td>
<td>C</td>
<td>8</td>
</tr>
<tr>
<td>MrBayes</td>
<td>3.2.6</td>
<td>C + C++</td>
<td>8 or 16</td>
</tr>
<tr>
<td>Migrate</td>
<td>3.6.11</td>
<td>C</td>
<td>1 to 72</td>
</tr>
<tr>
<td>Phylobayes</td>
<td>1.5a</td>
<td>C++</td>
<td>64</td>
</tr>
<tr>
<td>RAxML</td>
<td>8.1.24</td>
<td>C</td>
<td>12, 24, or 48</td>
</tr>
</tbody>
</table>
CIPRES’ reach is deep and broad
Nature article, Feb 2016
Mass. state science fair, July 2012
Connecting people and resources to accelerate discovery by empowering the science gateway community

Science Gateways Community Institute

Nancy Wilkins-Diehr, SDSC
Michael Zentner, Purdue
Marlon Pierce, Indiana U

Maytal Dahan, TACC
Katherine Lawrence, U of Michigan
Linda Hayden, ECSU

Award Number
ACI-1547611
What is a science gateway?

**science gateway** /sɪˈəns ɡætˈwæ/ n.

1. an online community space for science and engineering research and education.

2. a Web-based resource for accessing data, software, computing services, and equipment specific to the needs of a science or engineering discipline.
Gateways are changing the face of science

Predicting protein structures with a multiplayer online game

Seth Cooper¹, Firas Khatib², Andrew Leaver-Fay¹, David C. B. L. Aziz¹, Karina S. G. W. L. Melo³, Parameswaran Iyer⁴, Andrew T. S. D. Brown⁴, Robert J. M. Z. A. M. H. Turner⁵

Brunel University London research expertise in science gateways and e-Infrastructures has had a huge impact on building a better connected and open research community across African countries and between Europe and Africa.
NSF commits $35 million to improve scientific software

The second award, led by the University of California, San Diego, establishes the Science Gateways Community Institute, a multi-institutional consortium that will increase the capabilities, number and sustainability of science gateways. Gateways are mobile or web-based applications that provide broad access to the nation's shared cyberinfrastructure to scientists and citizens alike.

"Gateways foster collaborations and the exchange of ideas among researchers and can democratize access, providing broad access to resources sometimes unavailable to those who are not at leading research institutions," said Nancy Wilkins-Diehr, associate director of the San Diego Supercomputer Center and principal investigator for the project. "Sharing expertise about basic infrastructure allows developers to concentrate on the novel, the challenging, and the cutting-edge development needed by their specific user community."
Science Gateways Community Institute
*Designed to help the community build gateways more effectively*

- Diverse expertise on demand
- Longer-term, hands-on support
- Student opportunities & educator resources
- Extended Developer Support
- Software & visibility for gateways
- Scientific Software Collaborative
- Community Engagement & Exchange
- Sharing experiences & knowledge as a community

Science Gateways Community Institute
Incubator: Consulting expertise and in depth training

- Building a gateway can take many types of expertise
  - But projects cannot afford to hire this in
- Consulting services provide cost effective solution
  - Get expertise you need when you need it, release it when you’re finished
- In depth cohorts for training, group interactions, mentoring
  - Customized structure, content and goals
- ACTION: Request services at https://sciencegateways.org/request-services
Incubator Bootcamp
Held twice per year

- Full 5 days
- Knowledge dissemination
- Interactivity
- Community formation
- Putting away the normal daily routine
- Homework
Extended Developer Support

• Help building new gateways
  • Or portions of new gateways
• Dedicated support for months to a year
• Similar in structure to XSEDE ECSS program
• Different in that gateways can be built from the ground up
  • And can use any type of resource
• ACTION: Request services at https://sciencegateways.org/request-services
<table>
<thead>
<tr>
<th>Consultation</th>
<th>Description</th>
<th>Discipline</th>
<th>Gateway Type/ Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>COSMIC^2</td>
<td>Helping extend the SDSC Workbench framework to support large Cryo-EM data sets.</td>
<td>Cellular &amp; Molecular Medicine</td>
<td>Data Analysis</td>
</tr>
<tr>
<td>CitSci.org</td>
<td>Working with gateway team to improve efficiency of diverse queries required by broad range of citizen science user communities.</td>
<td>Computer Science/Multiple Disciplines</td>
<td>Platform for Creating Citizen Science Projects</td>
</tr>
<tr>
<td>USD Cybergateway</td>
<td>Developing a campus cluster-centric gateway with University of South Dakota.</td>
<td>Computer Science</td>
<td>Science Gateway Platform as a Service (SciGaP)</td>
</tr>
<tr>
<td>Computational Systems Biology Gateway</td>
<td>Collection of tools to computationally support drug discovery</td>
<td>Systems Biology</td>
<td>Computational tools</td>
</tr>
<tr>
<td>Aquavit</td>
<td>Using modeling and sharing for water management.</td>
<td>Freshwater Systems</td>
<td>Collaboration Portal</td>
</tr>
<tr>
<td>ENIGMA</td>
<td>Metadata management and sharing for bipolar disorder experiments.</td>
<td>Neuroscience, Clinical Psychiatry, and Genetics</td>
<td>Data sharing</td>
</tr>
<tr>
<td>Ecological Society of America</td>
<td>Nonprofit organization of scientists who conduct research, teach, and use ecological science to address environmental issues</td>
<td>Ecology</td>
<td>Knowledge hub</td>
</tr>
<tr>
<td>Indiana Watershed Initiative</td>
<td>Monitor the impacts of cover crops and two-stage drainage ditches on soil health and water quality</td>
<td>Biology</td>
<td></td>
</tr>
</tbody>
</table>
Scientific Software Collaborative

• Find software used to build gateways
  • List your software
  • Become an SGCI Affiliate and help others build gateways using your software

• Find existing gateways
  • Just use them
  • Learn what they’re built with
  • Converse with developers
  • Find gateways used in curricula

• https://catalog.sciencegateways.org
Community Engagement and Workforce Development

• Annual Gateways conference (since 2005)
  • Interact with other developers
  • Tutorials, presentations, posters, reception, open space discussions
  • Publish in special journal issue, joint with international developers
  • Gateways 2018
    • September 25-27, Austin

• Student fellowships, internships, travel support

• ACTION: Follow us via mailing list or Twitter (@sciencegateways)
• https://sciencegateways.org/community/newsletter
Thank you!

• Mailing list for news and updates: https://sciencegateways.org/community/newsletter
• Ready to request our services? https://sciencegateways.org/request-services
• Have a question? help@sciencegateways.org
• @sciencegateways on Twitter
• http://sciencegateways.org/linkedin