Applications Challenges in the XSEDE Environment

John Towns
PI and Project Director, XSEDE
Director, Collaborative eScience Programs, NCSA
jtowns@ncsa.illinois.edu

XSEDE

Extreme Science and Engineering Discovery Environment

XSEDE – *accelerating scientific discovery*

*XSEDE aspires to be **the** place to go to access digital research services.*

Accelerate scientific discovery by enhancing the productivity of researchers, engineers, and scholars through the use of advanced digital services and infrastructure.

XSEDE's Strategic Goals

- *Deepen* and *extend* the use of the XSEDE ecosystem
  - *deepen* use of XSEDE by existing researchers
  - *extend* use of XSEDE to new communities
  - prepare the current and next generation via education, training, and outreach
  - raise the general awareness of the value of advanced digital services
- *Advance* the XSEDE infrastructure
  - create an open and evolving infrastructure
  - enhance the array of technical expertise and support services offered
- *Sustain* the XSEDE infrastructure
  - sustain a reliable and secure infrastructure
  - provide excellent user support services
  - operate an effective and innovative virtual organization

XSEDE is a large and complex project

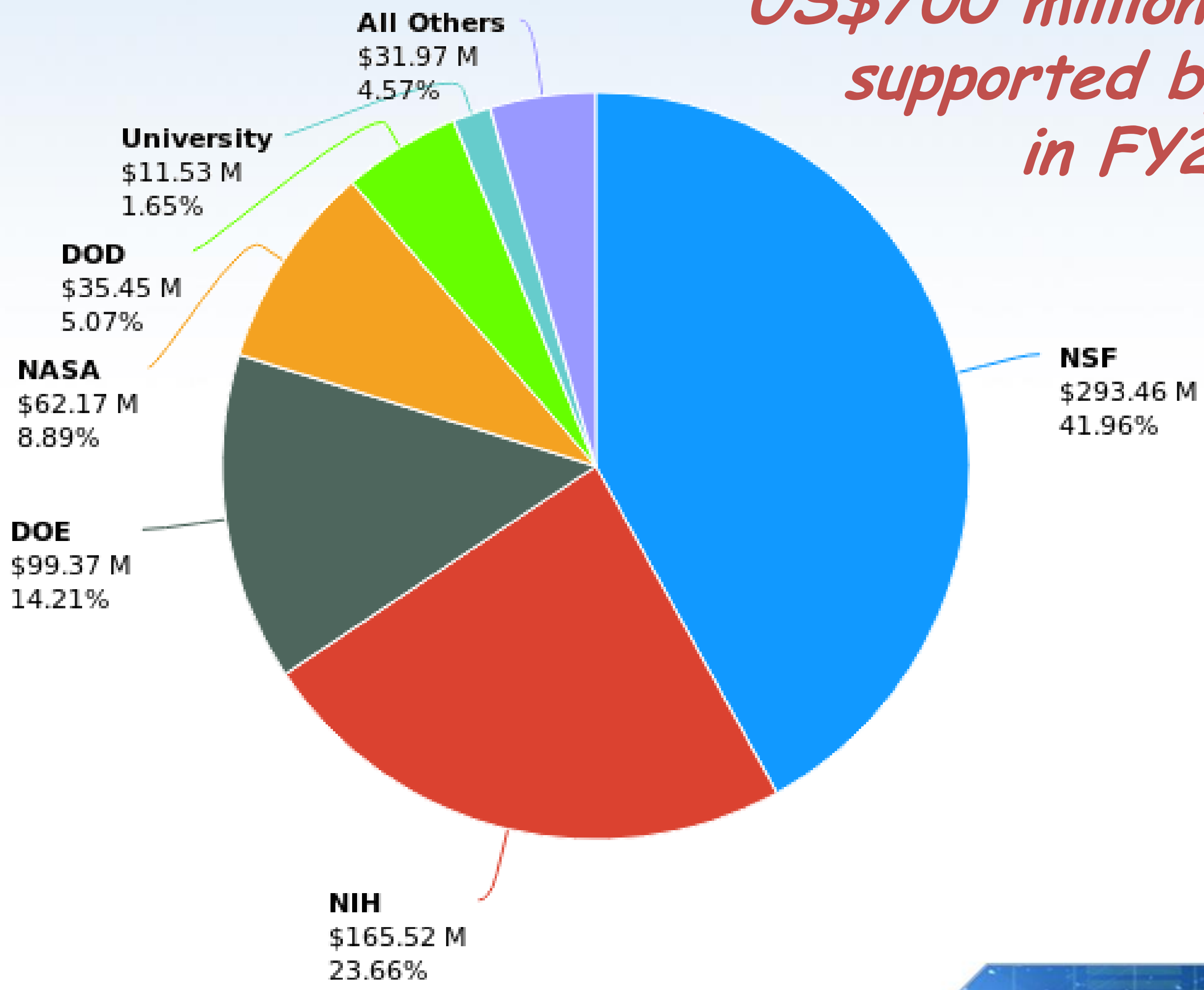
- 5 year, $130M project
  - includes $9M, 5 year Technology Investigation Service
    - separate award from NSF
  - option for additional 5 years of funding upon major review after PY3

- No funding for major hardware
  - coordination, support and creating a national/international eScience infrastructure
  - coordinate allocations, training and documentation for >$100M of concurrent project awards from NSF

- ~140 FTE (~250 individuals) across 20 partner institutions

Total Research Funding Supported by XSEDE in FY2013

US$700 million in research supported by XSEDE in FY2013



What is XSEDE?

- An ecosystem of advanced digital services
  - support a growing portfolio of resources and services
    - advanced computing, high-end visualization, data analysis, and other resources and services
    - interoperability with other infrastructures
- A virtual organization providing
  - dynamic distributed infrastructure
  - support services, and technical expertise to enable researchers engineers and scholars
    - addressing the most important and challenging problems facing the nation and world
- A project funded by the National Science Foundation

XSEDE offers access to a variety of resources

- Leading-edge distributed memory systems
- Very large shared memory systems
- High throughput systems, including Open Science Grid (OSG)
- Visualization engines
- Accelerators like GPUs and Xeon PHIs

Many scientific problems have components that call for use of more than one architecture.

Current XSEDE Compute Resources

- Stampede @ TACC
  - 9.5 PFLOPS (PF) Dell Cluster w/ GPUs and Xeon PHIs
- Kraken @ NICS
  - 1.2 PF Cray XT5
- Keeneland @ GaTech/NICS
  - 615 TF HP GPU cluster
- Gordon @ SDSC
  - 341 TF Appro Distributed SMP cluster
- Lonestar (4) @ TACC
  - 302 TF Dell Cluster
- Trestles @ SDSC
  - 100 TF Appro Cluster
- Blacklight @ PSC
  - 37 TF SGI UV (2 x 16TB shared memory SMP)
- Mason
  - 3.8 TF HP Cluster with large memory nodes (2TB/node)

https://www.xsede.org/web/xup/resource-monitor

Current XSEDE Visualization and Data Resources

- Visualization
  - Longhorn @ TACC
    - 20.7 TF Dell/NVIDIA cluster
    - 18.7 TB disk

https://www.xsede.org/web/xup/resource-monitor#advanced_vis_systems

- Storage
  - Ranch @ TACC
    - 40 PB tape
  - HPSS @ NICS
    - 12 PB tape
  - Data Supercell @ PSC
    - 4 PB disk
  - Data Oasis @ SDSC
    - 4 PB tape

https://www.xsede.org/web/xup/resource-monitor#storage_systems

Challenges/Hinderances (1)

- Memory bandwidth on MIC
  - needed to implement OMP threads on MIC to obtain sufficient memory bandwidth
    - 240 threads per MIC
  - Stride one memory access were critical to good performance
    - induces significant code restructuring in many cases
- Vectorization by compiler was poor
  - compiler was confused by data structures and did not recognized opportunities for vectorization
  - needed to restructure data layout
  - loops with branches also noted as a challenge

Challenges/Hinderances (2)

- Thread affinity
  - by default, threads were poorly located with respect to communications patterns
  - needed to use directives to assign thread distribution
    - best distribution varied by application
- Alignment issues
  - non-aligned vector access have very high overhead
  - compiler did not recognize these and hand directives needed to be inserted

Challenges/Hinderances (3)

- Allocated arrays on MIC are not persistent
  - by default data assigned to offloaded array are not persistent between kernel calls
  - needed to implement conditional array allocation and free-ing functions to avoid overhead of unnecessary data movement
  - this represented significant coding effort
- Splitting computation between CPUs and MICs required to fully utilize system
  - Represents significant effort in balancing the workload and communication requirements

Challenges/Hinderances (4)

- Long expressions difficult to optimize
  - Frequently noted that very long expression do not perform well
  - Need to split these into multiple statements
- Lack of tools!
  - most work guided by manually instrumenting code
  - current tools provide some support, but limited in capability

Challenges/Hinderances (5)

- I/O subsystems inadequately support disparate needs
  - interactive use, e.g. `ls -l` on a large number of files
  - metadata heavy use, e.g. many file creates
  - I/O server heavy use, e.g. many I/O operations
- Filesystem focus on scaling across nodes and not within a node
  - effective use of filesystem by a single node requires multiple threads but sill limited by node's connection
- Random I/O very painful
  - often inherent to algorithms used
  - libraries sometime help; more is needed here

Questions?

Our reach will forever

exceed our grasp, but,

in stretching our horizon,

we forever improve our world.

XSEDE

Extreme Science and Engineering
Discovery Environment