Comet – Tales from the Long Tail – Two Years in and 10,000 Users Later

Accelerated Data and Computing '17
Resource Management Session

July 17-18
San Jose, CA

Shawn Strande
and the Comet Team

SDSC SAN DIEGO SUPERCOMPUTER CENTER

UC San Diego

Presented on behalf of the project team at SDSC and Indiana University

Trevor Cooper	HPC Systems Manager
Haisong Cai	HPC Storage
Mike Dwyer	Documentation
Karen Flammer	EOT
Geoffrey Fox (IU)	Virtualization
Jerry Greenberg	User Services
Jim Hayes	HPC Systems Administration
Tom Hutton	Networking
Christopher Irving	HPC Systems Administration
Gregor von Laszewski (IU)	Virtualization
Amit Majumdar	Scientific Applications, Science Gateways
Dmitry Mishin	HPC Systems Programmer
Sonia Nayak	Financial reporting
Mike Norman	Principal Investigator
Phil Papadopoulos	Chief Architect, Co-PI
Wayne Pfeiffer	Scientific Applications
Susan Rathbun	EOT
Scott Sakai	Security
Jeff Sale	EOT
Bob Sinkovits	Scientific Applications, Co-PI
Shawn Strande	Project manager, Co-PI
Mahidhar Tatineni	User Services
Fugang Wang (IU)	Virtualization
Nancy Wilkins-Diehr	Science Gateways, Co-PI
Nicole Wolter	User Services & Accounting
Kenneth Yoshimoto	HPC Systems Programmer
Jan Zverina	Communications

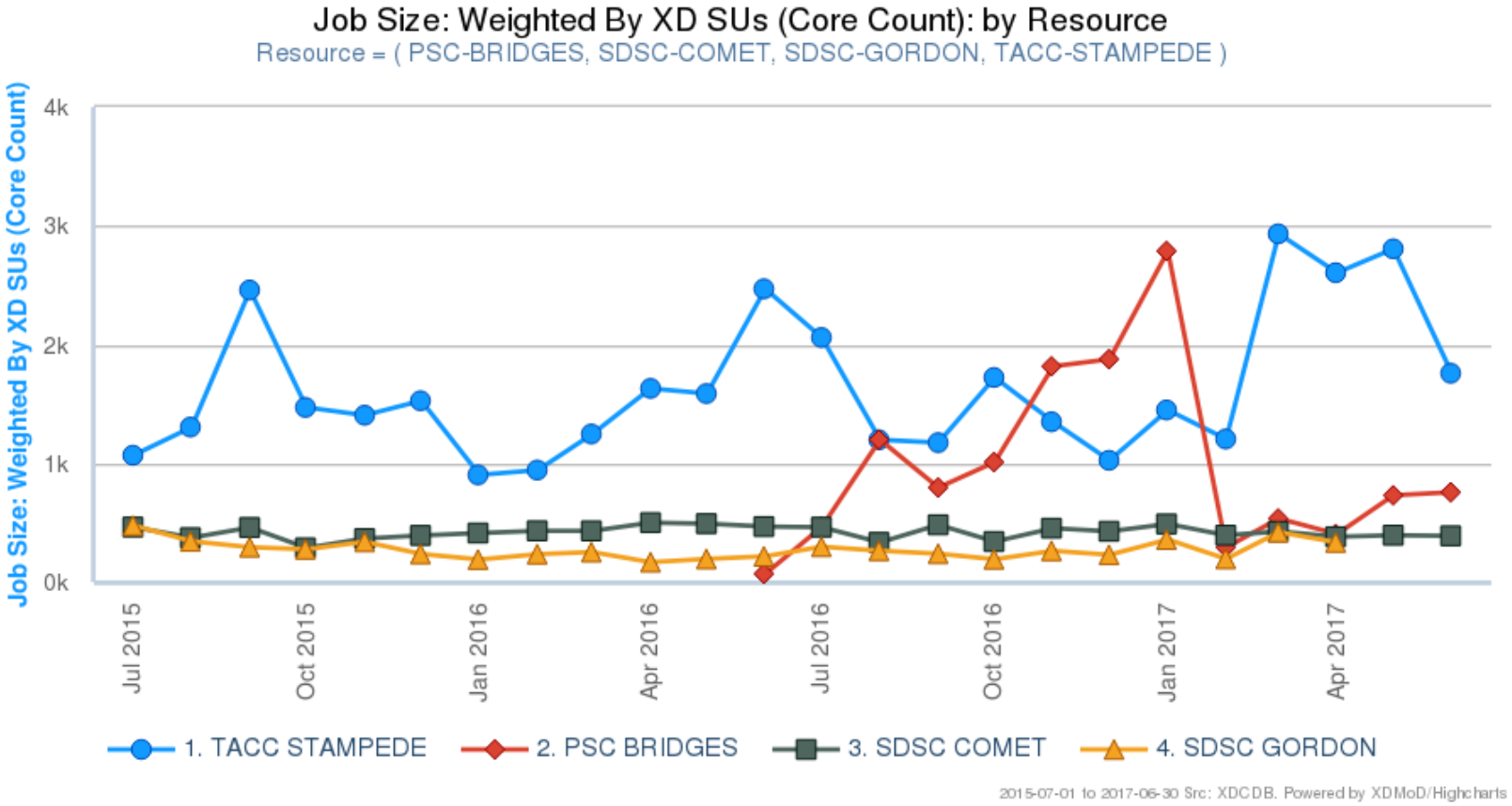
Acknowledgements: NSF OCA: 1341698; 1548562

Design and Operations

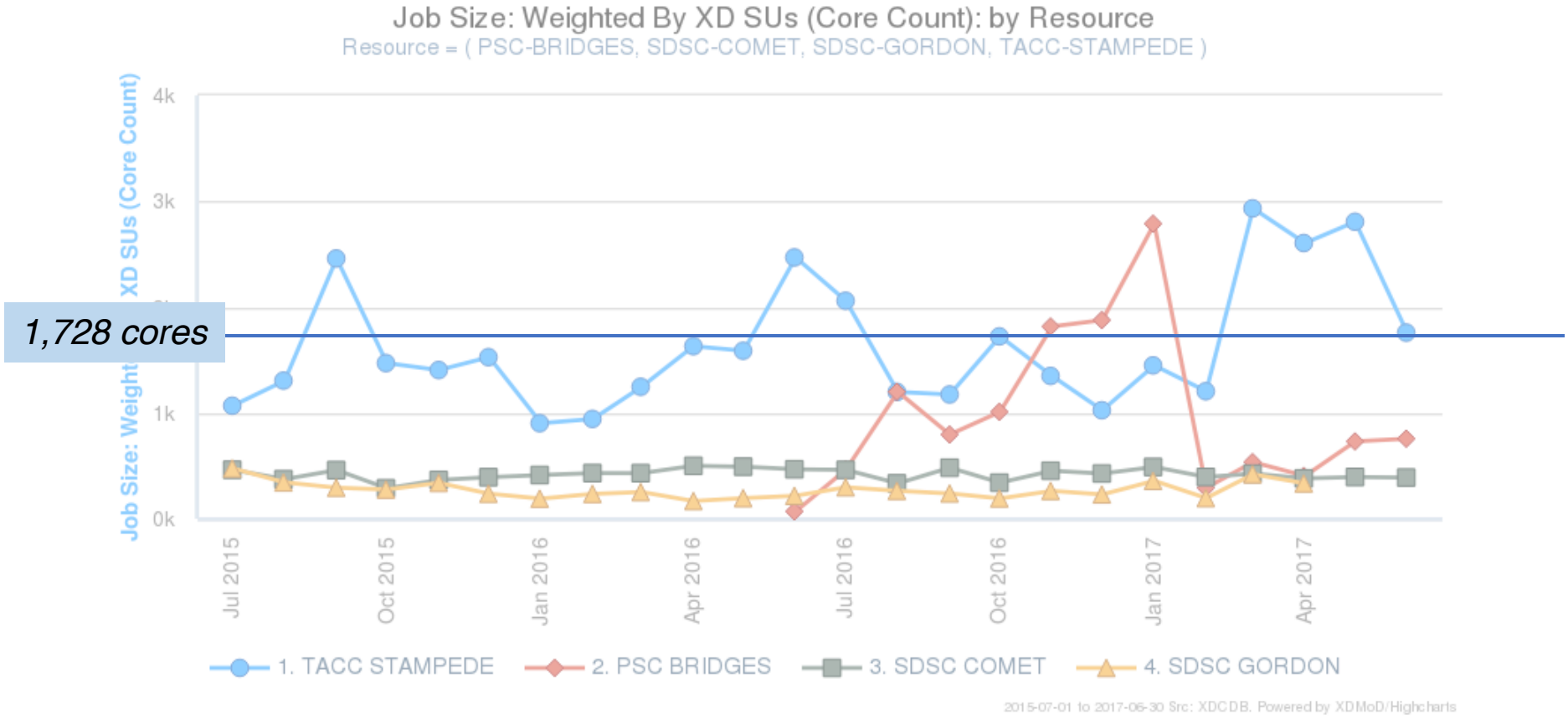
Comet: System Characteristics

- Total peak flops ~2.1 PF
- Dell primary integrator
  - *Intel Haswell processors w/ AVX2*
  - *Mellanox FDR InfiniBand*
- 1,944 standard compute nodes (46,656 cores)
  - *Dual CPUs, each 12-core, 2.5 GHz*
  - *128 GB DDR4 2133 MHz DRAM*
  - *2*160GB GB SSDs (local disk)*
- 72 GPU nodes (288 GPUs)
  - *36x with 2 NVIDIA K80 cards, each with dual Kepler3 GPUs*
  - *36x with 4 P100 cards*
- 4 large-memory nodes
  - *1.5 TB DDR4 1866 MHz DRAM*
  - *Four Haswell processors/node*
  - *64 cores/node*
- Hybrid fat-tree topology
  - *FDR (56 Gbps) InfiniBand*
  - *Rack-level (72 nodes, 1,728 cores) full bisection bandwidth*
  - *4:1 oversubscription cross-rack*
- Performance Storage (Aeon)
  - *7.6 PB, 200 GB/s; Lustre*
  - *Scratch & Persistent Storage segments*
- Durable Storage (Aeon)
  - *6 PB, 100 GB/s; Lustre*
  - *Automatic backups of critical data*
- Home directory storage
- Gateway hosting nodes
- Virtual image repository
- 100 Gbps external connectivity to Internet2 & ESNet

Average job size across XSEDE systems is below 2000 cores



One rack of Comet provides full bisection bandwidth up to 1,728 cores



Trestles and Gordon – A legacy of highly productive HPC systems for the long tail

University of Arkansas Acquires New Supercomputer for Research Support

May 04, 2015

FAYETTEVILLE, Ark. – A workhorse of a supercomputer is coming to the University of Arkansas.

The National Science Foundation and the San Diego Supercomputer Center at the University of California, San Diego, have agreed to transfer ownership of the computer cluster known as "Trestles" to the Arkansas High Performance Computing Center.

Once installed, the new supercomputer will more than double the center's computational capacity and allow it to run three times the

Courtesy of Ben Tolo, San Diego Supercomputer Center

Trestles was deployed at the San Diego

Flatiron Institute to Repurpose 'Gordon' Supercomputer at UC San Diego

The powerful resource from the San Diego Supercomputer Center will be used for astrophysics, biology and materials research.

March 15, 2017

The San Diego Supercomputer Center (SDSC) at the University of California San Diego and the Simons Foundation's Flatiron Institute in New York have reached an agreement under which the majority of SDSC's data-intensive *Gordon* supercomputer will be used by Simons for ongoing research following completion of the system's tenure as a National Science Foundation (NSF) resource on March 31.

Under the agreement, SDSC will provide high-performance computing (HPC) resources and services on *Gordon* for the

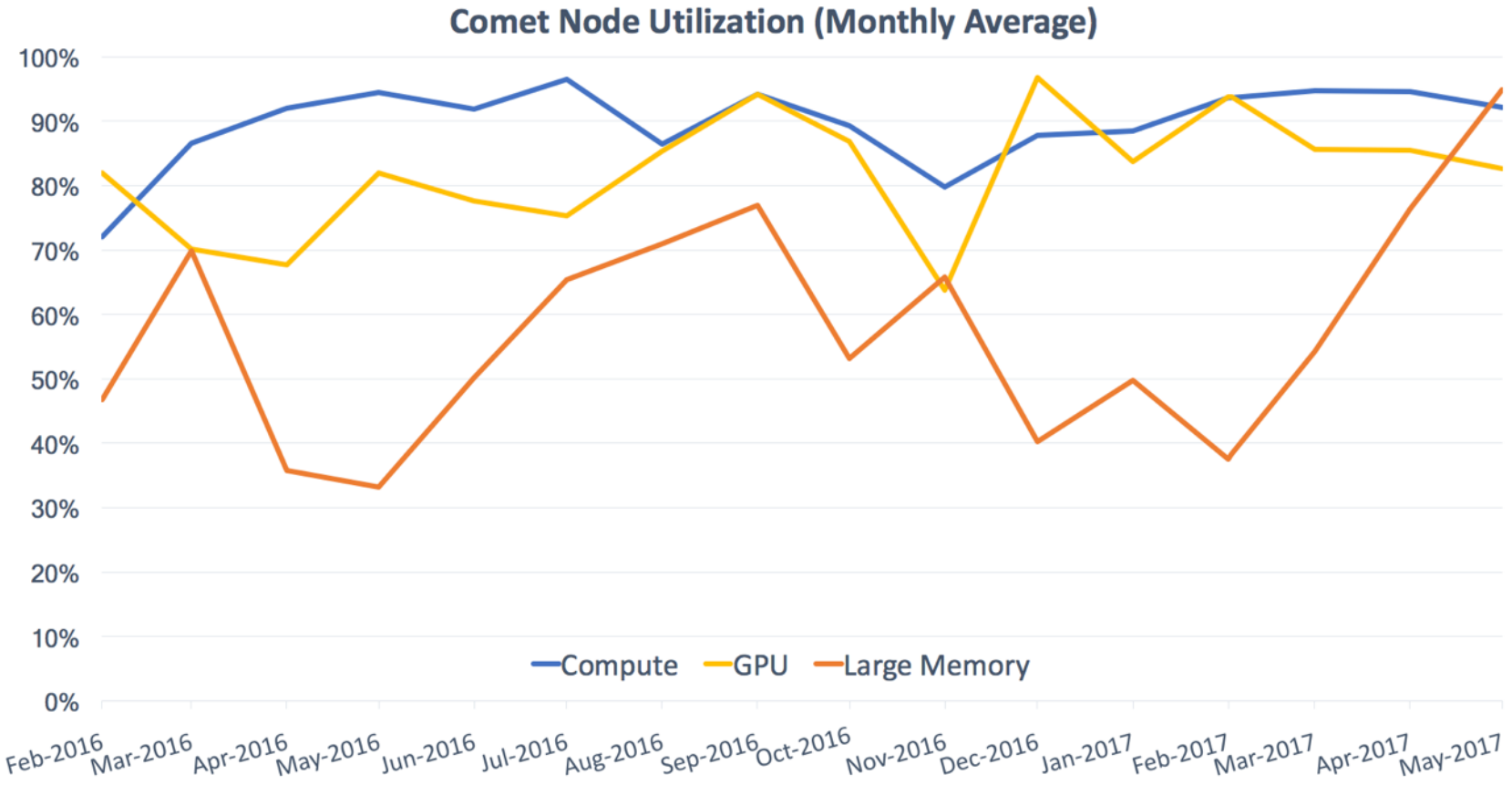
SDSC's data-intensive 'Gordon' supercomputer will be

Trestles and Gordon were both designed to serve a broader community than traditional HPC systems and endure well beyond their NSF-funded life.

Comet's operational polices and software are designed to support long tail users

- Allocations
  - Individual PIs limited to 10M SU
  - Gateways can request more than 10M SUs
  - Gateways exempt from "reconciliation" cuts
- Optimized for throughput
  - Job limits are set at jobs of 1,728 cores or less (a single rack)
  - Support for shared node jobs is a boon for high throughput computing and utilization
  - Comet "Trial Accounts" provide 1000 SU accounts within one day
- Science gateways reach large communities
  - There 13 gateways on Comet, reaching thousands of users through easy to use web portals
- Virtual Clusters (VC) support well-formed communities
  - Near native IB performance
  - Project-controlled resources and software environments
  - Requires the allocation team possess systems administration expertise

Smaller job sizes, shared node jobs and gateways lead to very high utilization



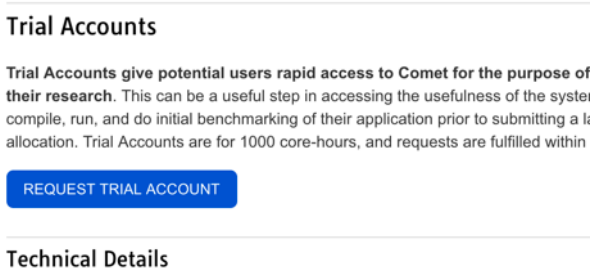
User Experience

When asked "What do you like about Comet?" users commented on the quick turnaround time, usability, and excellent user support

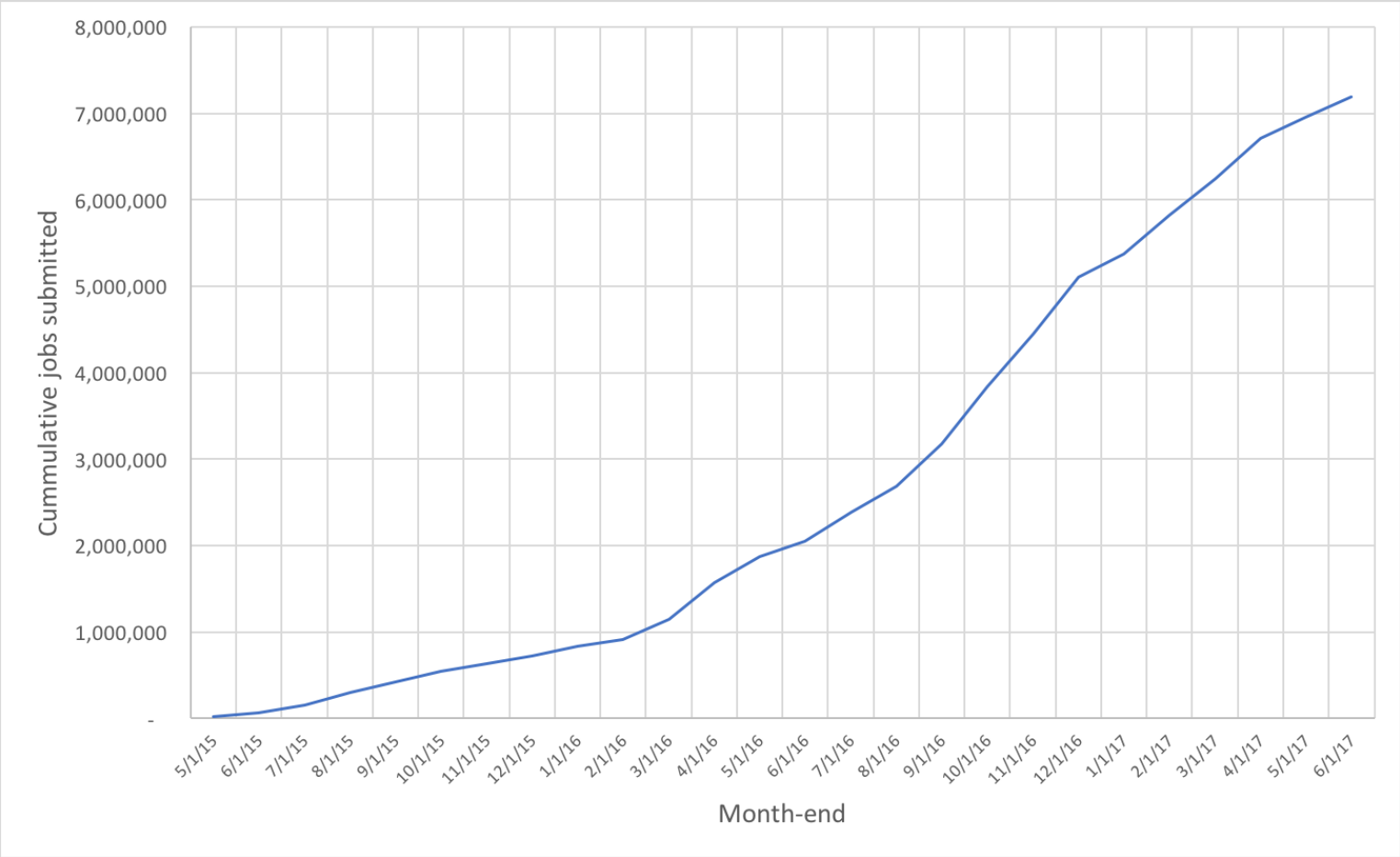
- *"Short queue times and large processor/GPU-to-node ratio, in comparison to other HPC resources I have used."*
- *"Many standard compute nodes with 12 cores per socket. Good scheduling policy resulting in lesser queue times, large memory, large flash memory enabling faster data read/write."*
- *"Smaller jobs run quickly; the system is quite stable."*
- *"It is easily accessed, it has very thorough documentation from both SDSC and XSEDE, and it has a diverse set of nodes, and it is a very large cluster."*
- *"Low pend time. Low compute time. Relatively large number of cores per node works well for my code.*
- *"Excellent help desk. Worked with me personally, immediately to solve all my problems."*

Trial Accounts, an SDSC innovation to provide rapid access for evaluation of Comet

- *Lowers barrier to entry for new users/communities.*
- *Integrated process with XSEDE.*
- *1,000 core-hours awarded in 1 day.*
- *Users request trial account by clicking on link on the XSEDE portal or SDSC Comet page.* Users need a XSEDE portal account (which they can self create)
- *Designated User Services staff at SDSC serve as PIs* on an allocation that serves as the pool for the rapid user accounts.
- ***566*** rapid user accounts created since inception ***236*** of these have been converted to full allocations.

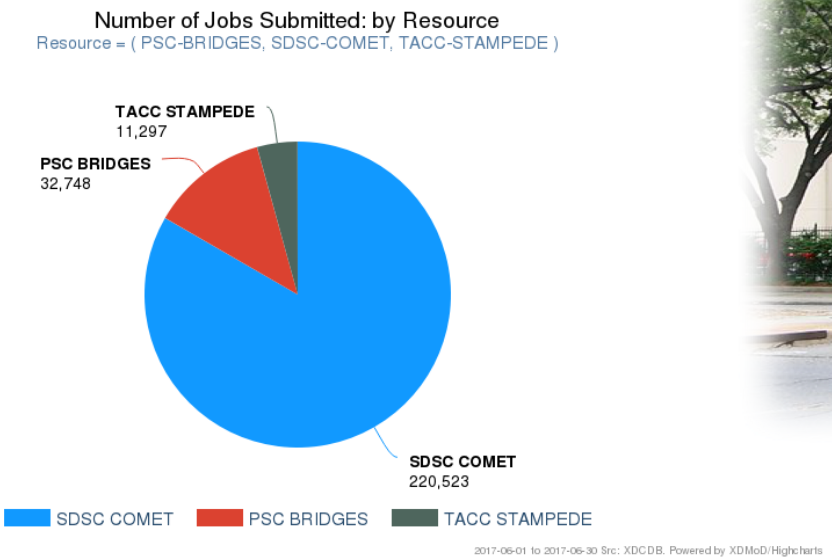


Users have submitted over 7M jobs since Comet entered production. Job success rate is > 99.8%



Keeping the Slurm train running on time

- Problem: High job submission rates of > 50K jobs/day from gateways and HTC workloads like OSG pushed Slurm pre-run database accounting checks to timeout causing all job submissions to stop
- Solution: Implement memcached database to hold accounting information and alleviate Slurm database calls
- Result: Comet can support > 50K jobs/day



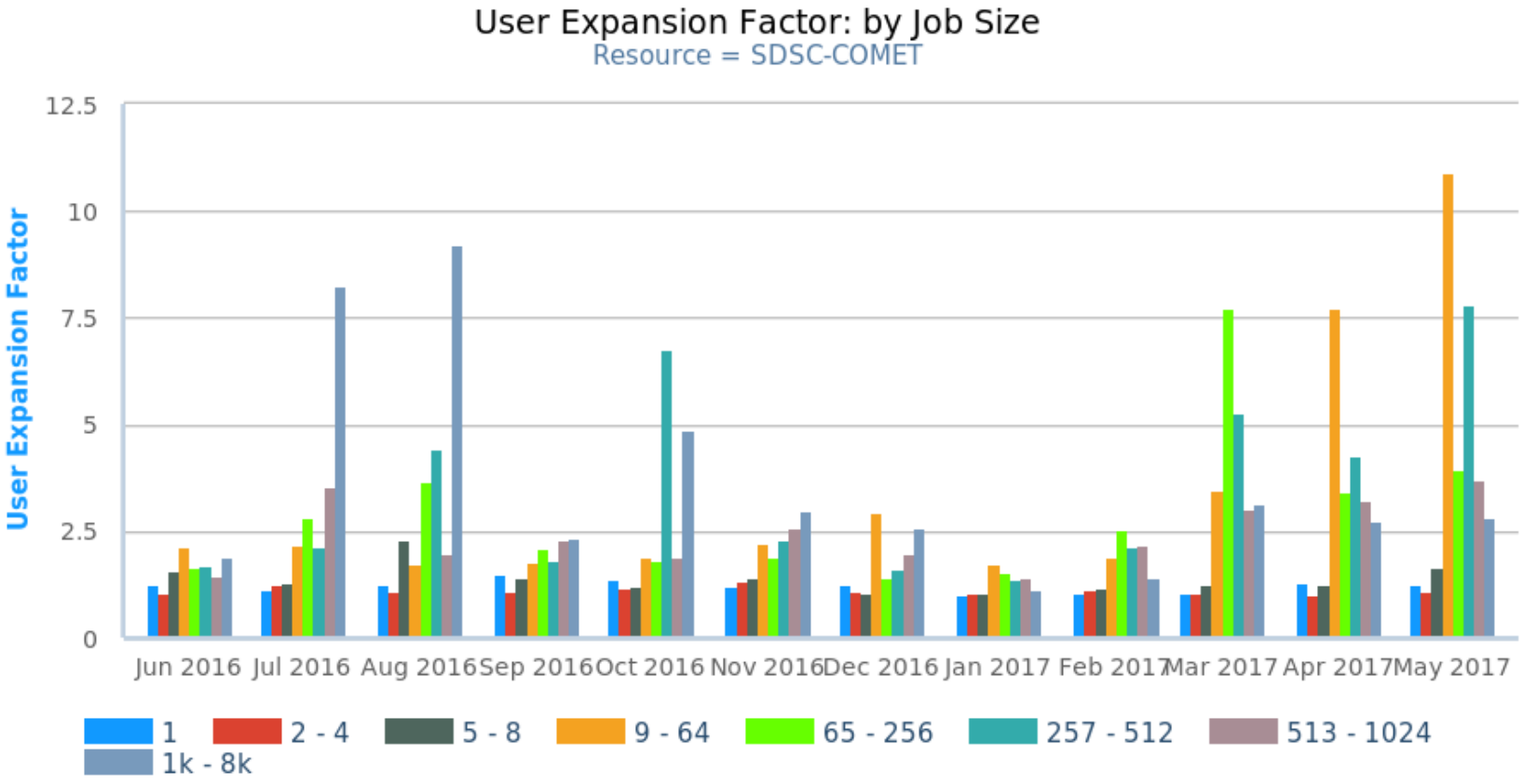
By Diego Delso, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=17764259

Gateways on Comet serve 100x the number of users per SU than traditional users

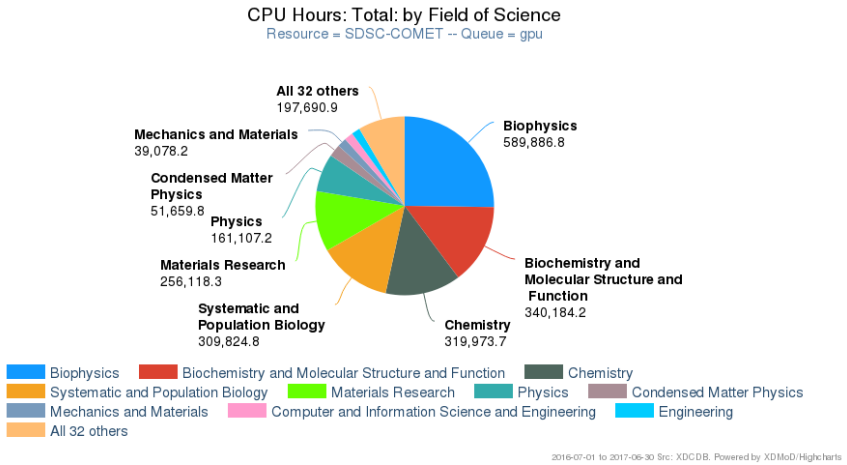
User type	Users served	SUs consumed	Ratio of SUs consumed to # users
Gateway	15,276	30M	500 Users/1M SU
Traditional	2,604	428M	6 Projects/1M SU

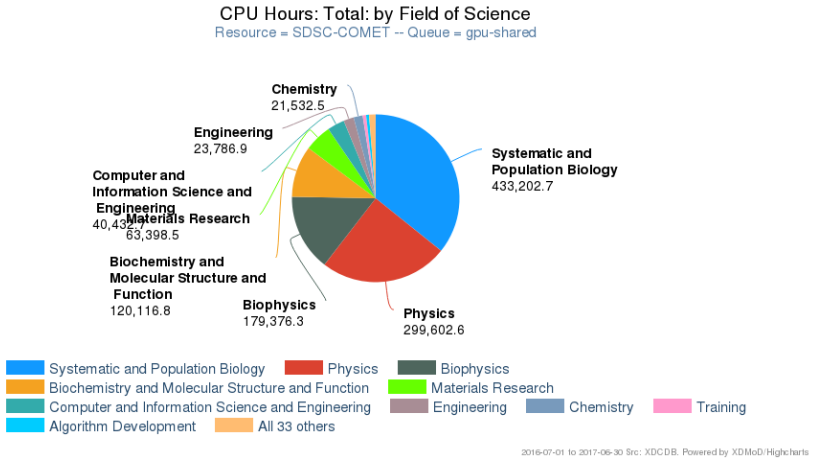
-> Gateways lower the barrier to making productive use of high performance computing. *No allocation request needed for gateway users!*

Expansion factors for small jobs are excellent, but things are getting busy

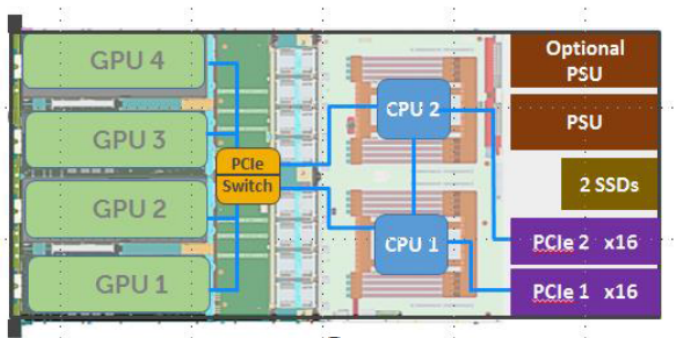


GPU nodes support researcher from over 70 institutions across many fields of science





- *36 nodes with144 K80*
- *36 nodes with144 P100*
- *Support for dedicated and shared-node jobs*
- *SDSC is the largest source of GPUs in XSEDE*



P100 nodes use a PCIe switch for high GPU-GPU bandwidth, low latency

Science Gateways & Virtual Clusters

Science Gateways
Revolutionizing and Democratizing Science

science gateway /sī′ əns gāt′ wā′/ *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.

- **Research is digital and increasingly complex**
- **Gateways provide *broad access* to advanced resources and allow *all* to tackle today's challenging science questions**

SDSC leading on several fronts to adapt to increase in gateway users

- Allocations
  - Allocations limited to decrease wait times
  - Comet to award gateways XRAC-recommended CPU
    - No further reductions in “reconciliation” process
- Scheduling
  - More jobs per user for gateway allocations
- Documentation
  - SDSC helping to drive gateway-friendly machines as an XSEDE resource type
  - Contributing to cross-XSEDE VM library
    - Led by Jetstream and Bridges teams
- Science Gateway Community Institute
- XDMoD per user gateway queries
  - Talking with XDMoD team to move this up in priority for easier generation of gateway user counts

Gateways using Comet

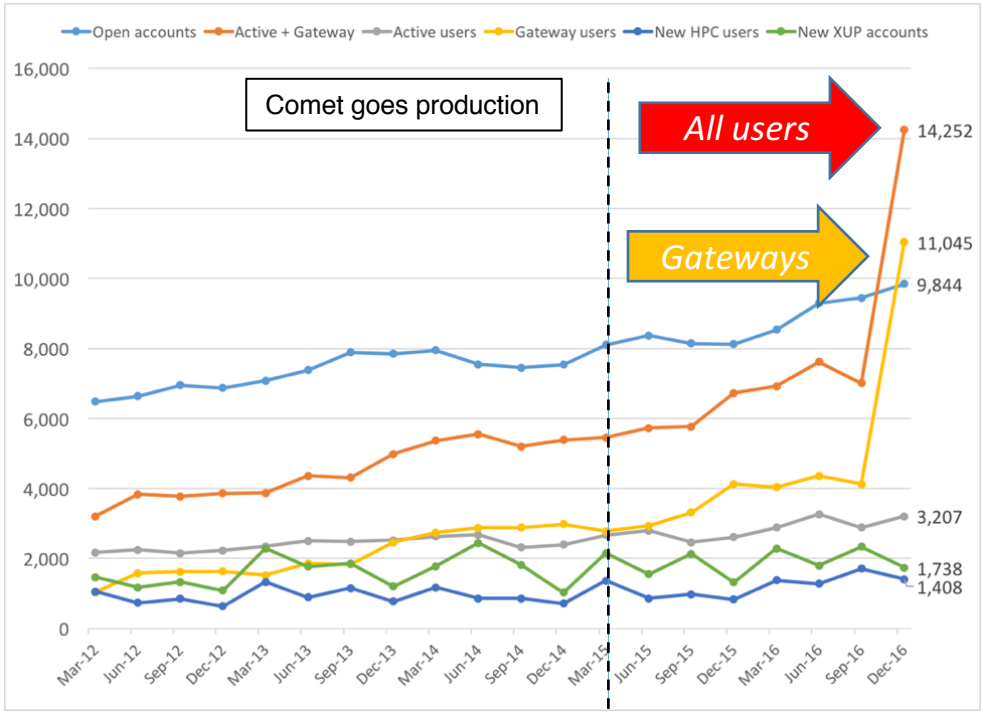
Gateway	Domain	CPU
CIPRES; UC San Diego	Systematic and population biology	11,088,219
Neuroscience Gateway; UC San Diego	Neuroscience	3,115,557
SEAgrid; IU	Chemistry, engineering	2516016
I-TASSER; U Michigan	Biochemistry, molecular structure and function	905929
Ultrascan3; UT Hlth Sci Ctr Houston	Biophysics	89921
TAS; SUNY Buffalo	XDMoD jobs	49522
waterhub; Purdue U	Earth sciences	14684
SciGaP; Indiana U	Gateway development	1503
ChemCompute; Sonoma State U	Chemistry	1216
COSMIC2; UC San Diego	Biochemistry, molecular structure and function	355
dREG; Cornell U	Veterinary medicine, gene expressions	323
UCI Social Science Gateway; UC Irvine	Anthropology	278
vdjserver; UT Southwestern Med Ctr	Molecular Biosciences	24

In two years, Comet has served nearly 18,000 unique users, easily surpassing the project goal of 10,000

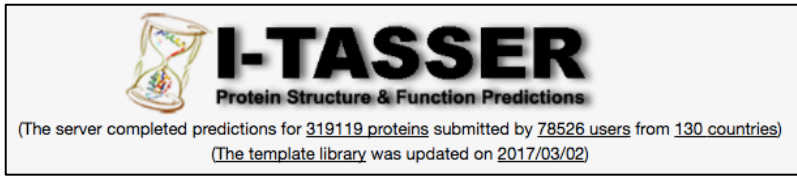
User type	Counts
Traditional login	2,604
CIPRES	6,310
I-TASSER	8,015
All other gateways	951
Total unique user counts	17,880

This was driven in large part by two gateways, I-TASSER and CIPRES, but we are seeing growth in other gateways as well.

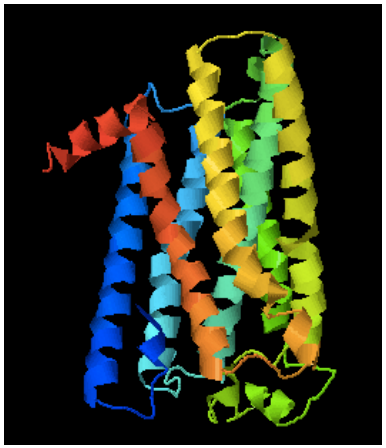
In Q4 2016 gateway users are 77% of active XSEDE users



XSEDE users



- 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



This is largely due to the CIPRES and I-TASSER gateways, but others are gaining

Science Gateways Community Institute

Designed to help the community build gateways more effectively

Press Release 16-088

NSF commits $35 million to improve scientific software

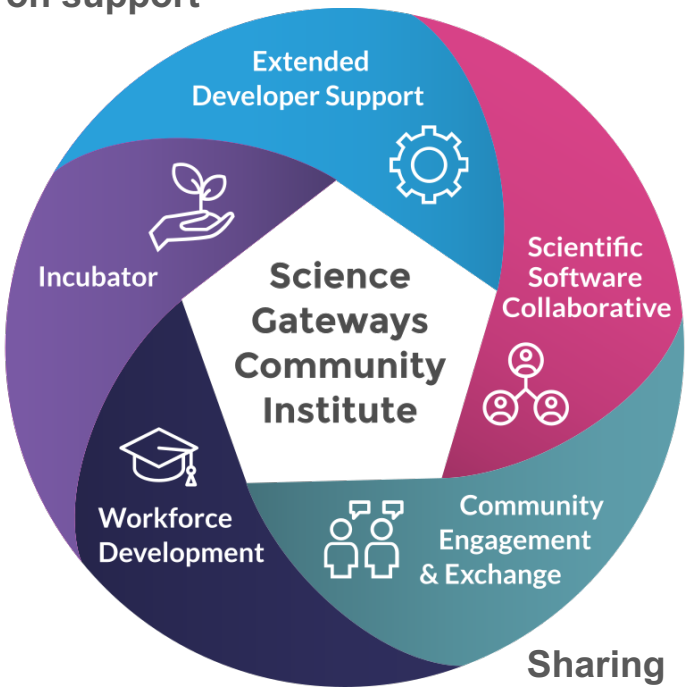
Science Gateways Community Institute

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."

Longer-term, hands-on support

Software & visibility for gateways



Diverse expertise on demand

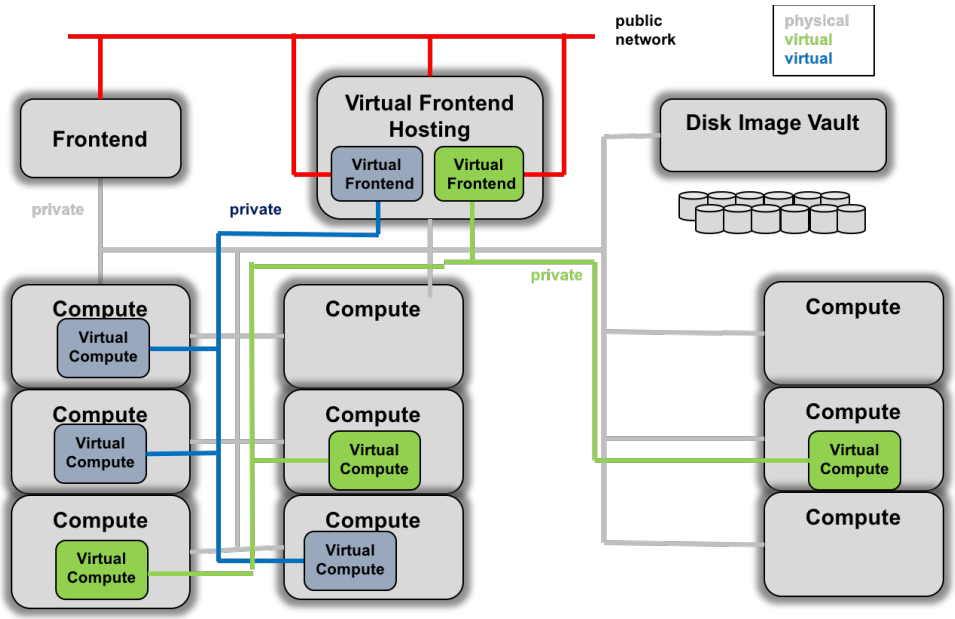
Sharing experiences & knowledge as a community

Student opportunities & educator resources

SGCI

Comet's Virtual Cluster Feature Provides Near-bare metal performance will full software customization to research teams

On-ramping via Indiana University team



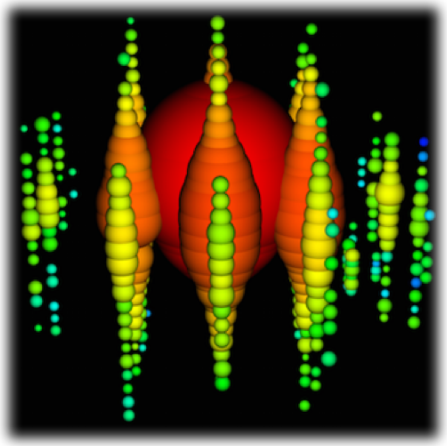
Technology	Capability
KVM	Virtual Machine Creation
SR-IOV	Near native IB performance in Virtual Machines
Rocks	Systems Managment
ZFS + IMG-STORAGE	Virtual Machine Disk Image Management
VLANs	Virtual Cluster Private Management Network Isolation
PKEYs	Virtual Cluster Infiniband Network Isolation
Nucleus API	Virtual Machine Management and Provisioning, Status and Console Access
Cloudmesh Client	Convenient Access and Automation Support

Recently, Singularity has become popular for its ability to on-ramp users from their campus/lab software environment to Comet

Virtual Cluster Case Study: Open Science Grid (OSG)

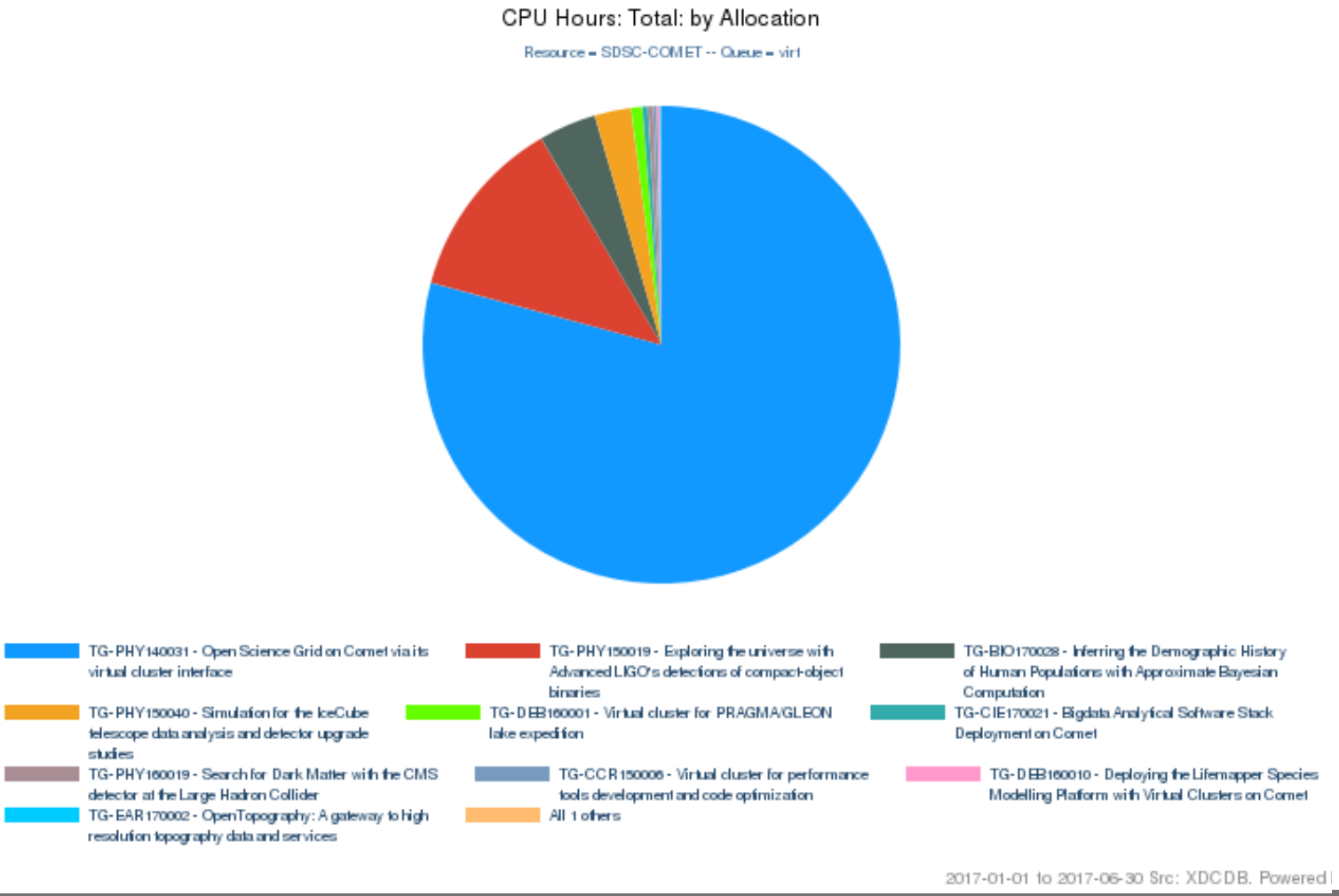
- Supporting small scale multi-core, (future: large memory & GPU) , large IO jobs, i.e., functions as a **portal** to Comet via the virtual cluster interface in order to provide capabilities that are otherwise rare on OSG.
  - One admin to manage a VC with many nodes, and to submit jobs on behalf of ***multiple allocations*** (LIGO-gravitational wave, CMS-dark matter, IceCube, etc.)
- Feedback/Justification from OSG
  - The VC interface is a very powerful tool for experts.
  - Was **very easy** for OSG to use the VC interface.
    - a day or two and we were running close to the full diversity of science on the VC onramp environment.
  - Having access to a virtual cluster on Comet that can be **customized** opens the door for **a whole new set of applications we were never able to support before.**

IceCube Neutrino Observatory
https://icecube.wisc.edu/science/highlights



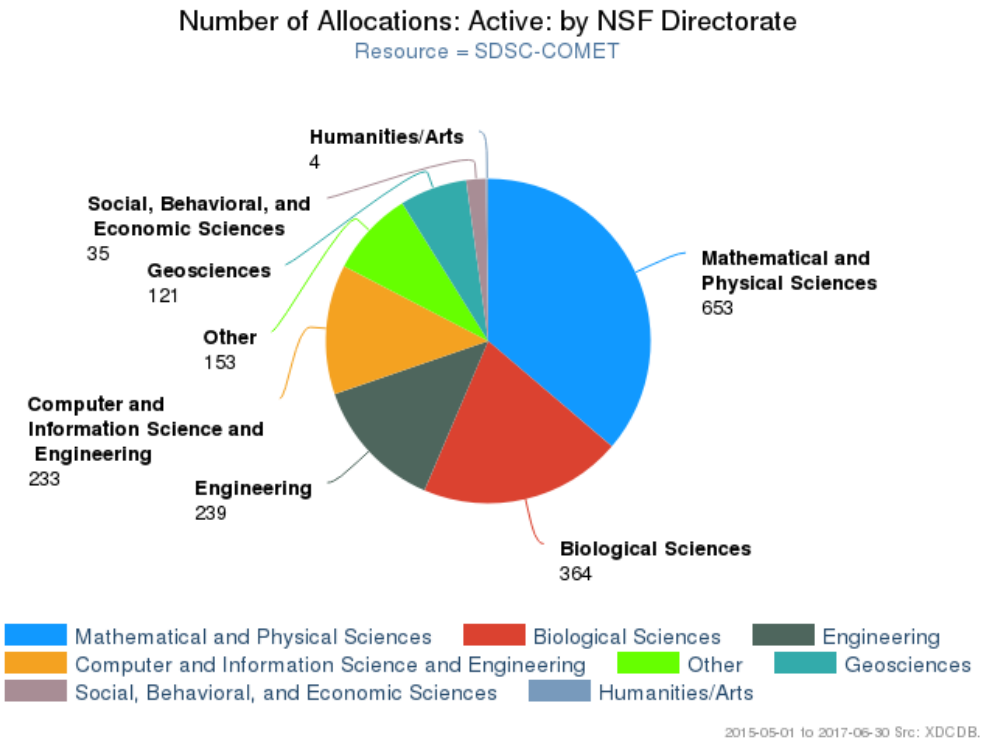
CMS Detector/LHC
http://cms.web.cern.ch/news/what-cms

Any XSEDE Allocated Project Can Run in a Comet OSG Virtual Cluster



Science Impact

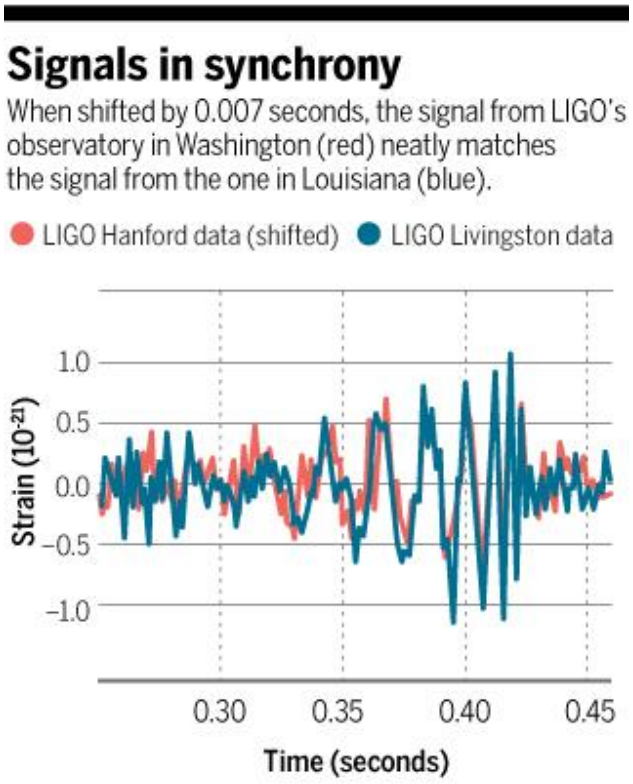
Comet has supported over 1,700 allocations and led to 700 publications

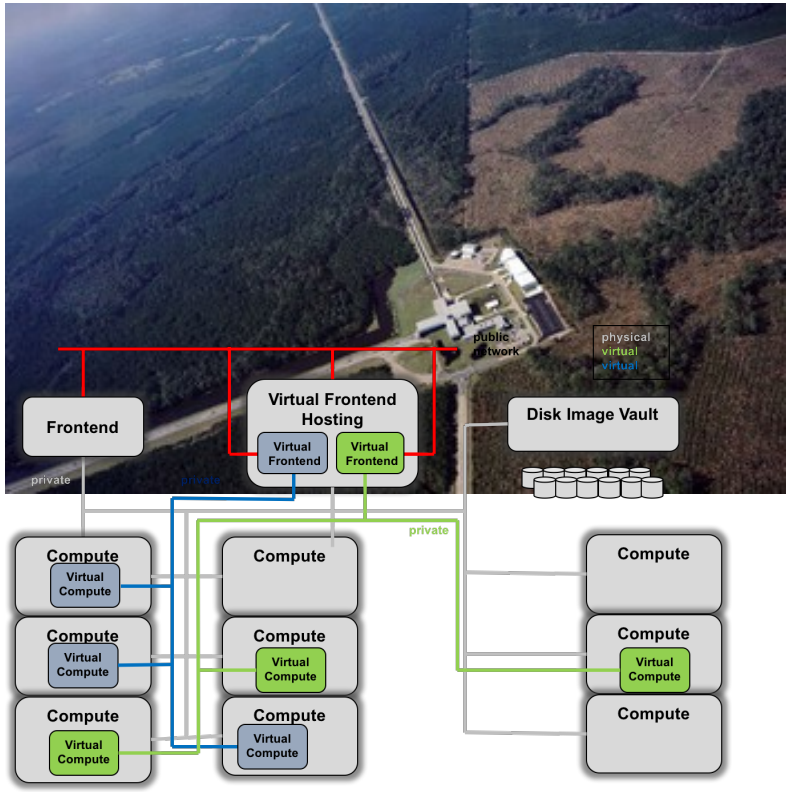


380 institutions, with increasing diversity. E.g.,

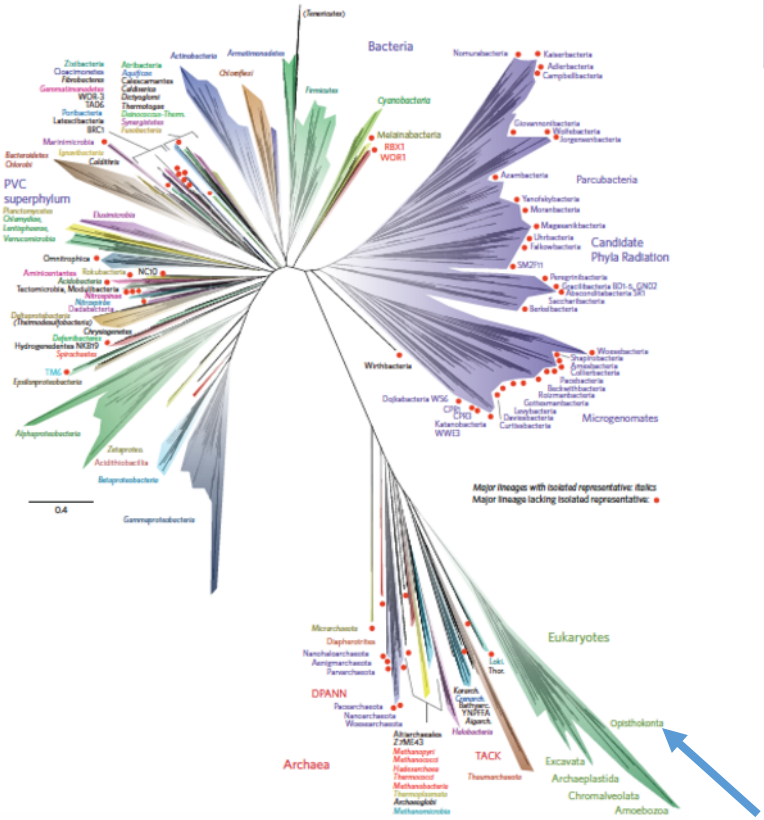
- Non-profit research institutes, hospitals, zoos and museums
  - La Jolla Institute for Allergy and Immunology
  - American Museum of Natural History
  - City of Hope
  - San Diego Zoo Institute for Conservation Research
  - Beth Israel Deaconess Medical Center
  - Burnham Institute
- FFRDCs, state and federal government
  - California Department of Water Resources
  - Federal Reserve Bank of New York
  - Fermi National Accelerator Laboratory
  - National Institute of Standards and Technology

OSG Virtual Cluster used to help confirm LIGO discovery





Phylogenetic tree inference using CIPRES gateway



nature microbiology

LETTERS

PUBLISHED: 11 APRIL 2016 | ARTICLE NUMBER: 16048 | DOI: 10.1038/NMICROBIOL.2016.48

OPEN

A new view of the tree of life

Laura A. Hug[1†], Brett J. Baker[2], Karthik Anantharaman[1], Christopher T. Brown[3], Alexander J. Probst[1], Cindy J. Castelle[1], Cristina N. Butterfield[1], Alex W. Hernsdorf[3], Yuki Amano[4], Kotaro Ise[4], Yohey Suzuki[5], Natasha Dudek[6], David A. Relman[7,8], Kari M. Finstad[9], Ronald Amundson[9], Brian C. Thomas[1] and Jillian F. Banfield[1,9*]

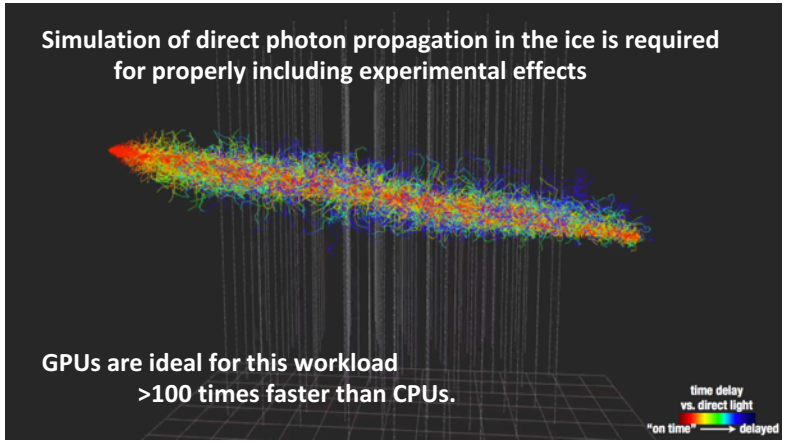
Tree was generated with RAxML on 48 cores of Comet in a 3-day run via CIPRES. Phyla with red dots are based upon metagenomic analyses without isolated representatives

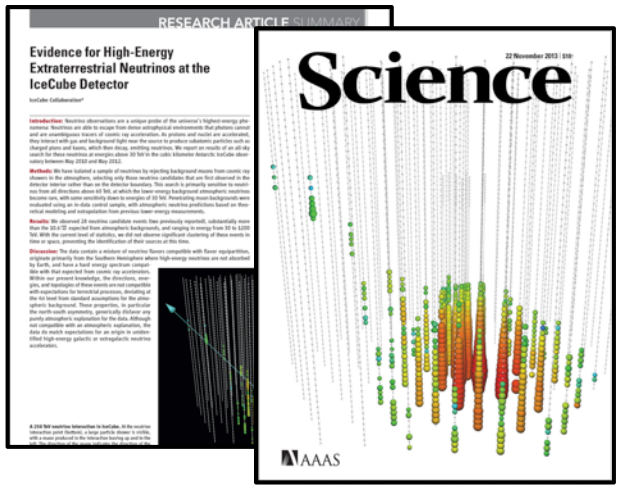
You are here in Opisthokonta, which includes animals & fungi

Science Gateway

IceCube Neutrino Observatory

IceCube found the first evidence for astrophysical neutrinos in 2013 and is extending the search to lower energy neutrinos. The main challenge is to keep the background low and a precise simulation of signal and background is crucial to the analysis.





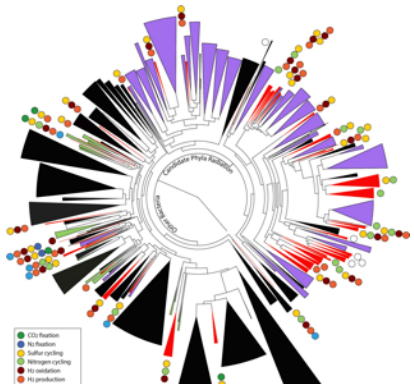
Comet's GPU nodes are a valuable resource for IceCube and integration with the experiment workload management system was very smooth thanks to previous OSG work on Comet

GPU

Summary Metrics for the Long Tail

Rue Comet

Comet Quarter



Rue Comet

18,000 unique users

Rue Comet

678 publications

Rue Comet

380 institutions

Rue Comet

505 research allocations

Rue Comet

7,000,000 jobs

Rue Comet

745 startups

Rue Comet

Science Gateways

Rue Comet

68 education allocations

Rue Comet

Virtual Clusters

Design and Resource Management Lessons from the Long Tail

- Comet's design has proven to be a versatile and highly usable resource for a large community of users. Important to pay careful attention to interconnect, accelerators, SSDs, scheduler, software stack.
- Allocations and operations policies, like capping allocations, and supporting shared node jobs, are important tools for the long tail and drive high utilization.
- Science gateway allocations hundreds more users per core-hour than standard allocations.
- Comet's approach to high performance virtualization has been effective, through on-ramping users can be time consuming. The emergence of Singularity and related technologies is also opening new possibilities for bridging new communities.
- A tremendous about of important science can be supported by a modest allocation.

Acknowledgments

NSF ACI: 1341698; 1548562