Gateways to Discovery: Cyberinfrastructure for the Long Tail of Science

Richard L. Moore[a] rlm@sdsc.edu
Chaitan Baru[a] baru@sdsc.edu
Diane Baxter[a] dbaxter@sdsc.edu
Geoffrey C. Fox[b] gcf@indiana.edu
Amit Majumdar[a] majumdar@sdsc.edu
Phillip Papadopoulos[a] phil@sdsc.edu
Wayne Pfeiffer[a] pfeiffer@sdsc.edu
Robert S. Sinkovits[a] sinkovit@sdsc.edu
Shawn Strande[c] sstrande@ucar.edu
Mahidhar Tatineni[a] mahidhar@sdsc.edu
Richard P. Wagner[a] rpwagner@sdsc.edu
Nancy Wilkins-Diehr[a] wilkinsn@sdsc.edu
Michael L. Norman[a] mlnorman@sdsc.edu

[a] San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093
[b] School of Informatics and Computing, Indiana University, Bloomington, IN 47408
[c] National Center for Atmospheric Research, Boulder, CO 80305

ABSTRACT

NSF-funded computing centers have primarily focused on delivering high-performance computing resources to academic researchers with the most computationally demanding applications. But now that computational science is so pervasive, there is a need for infrastructure that can serve more researchers and disciplines than just those at the peak of the HPC pyramid. Here we describe SDSC's *Comet* system, which is scheduled for production in January 2015 and was designed to address the needs of a much larger and more expansive science community– the "long tail of science". *Comet* will have a peak performance of 2 petaflop/s, mostly delivered using Intel's next generation Xeon processor. It will include some large-memory and GPU-accelerated nodes, node-local flash memory, 7 PB of *Performance Storage*, and 6 PB of *Durable Storage*. These features, together with the availability of high performance virtualization, will enable users to run complex, heterogeneous workloads on a single integrated resource.

Categories and Subject Descriptors

C.5.1 Super (very large) computers, K.6 Management of computing and information systems

General Terms

Design, Management

Keywords

High performance computing, high throughput computing, science gateways, virtualization, GPU, solid-state drive, parallel file system, scientific applications, user support

XSEDE '14, July 13 - 18 2014, Atlanta, GA, USA

1. INTRODUCTION

For the past three decades, NSF-funded computing centers have designed and operated high-performance computing resources primarily for researchers with the most computationally demanding applications. But now that computational science is so pervasive, there is a need for infrastructure that can serve more researchers and disciplines than just those at the peak of the Branscomb Pyramid [1, 2]. As computational science evolves, it is critical that HPC resources be configured and operated to support additional users and new modalities of computing.

National initiatives and reports, including the *CIF21* [3, 4], and the ACCI task force reports [5] have gathered critical, cross-community requirements. Furthermore, a 2010 survey [6] of 5,000 NSF PIs and thus *potential users* of NSF's cyberinfrastructure, conducted by Indiana University and SDSC, identified tremendous unmet demand for new types of cyberinfrastructure, listing *small/modest-scale computing resources*, *data-centric resources*, *workflow support* and *virtual environments* as critical capabilities not yet available at national resource sites. Similarly a 2009 workshop identified the requirements of "wide" user communities and their emerging computing modalities that are not well served by current XSEDE resources [7]. These are the users and communities that we refer to as the "long tail of science".

SDSC's *Comet* system, scheduled for production in January 2015, will enable new modalities of computing and address the

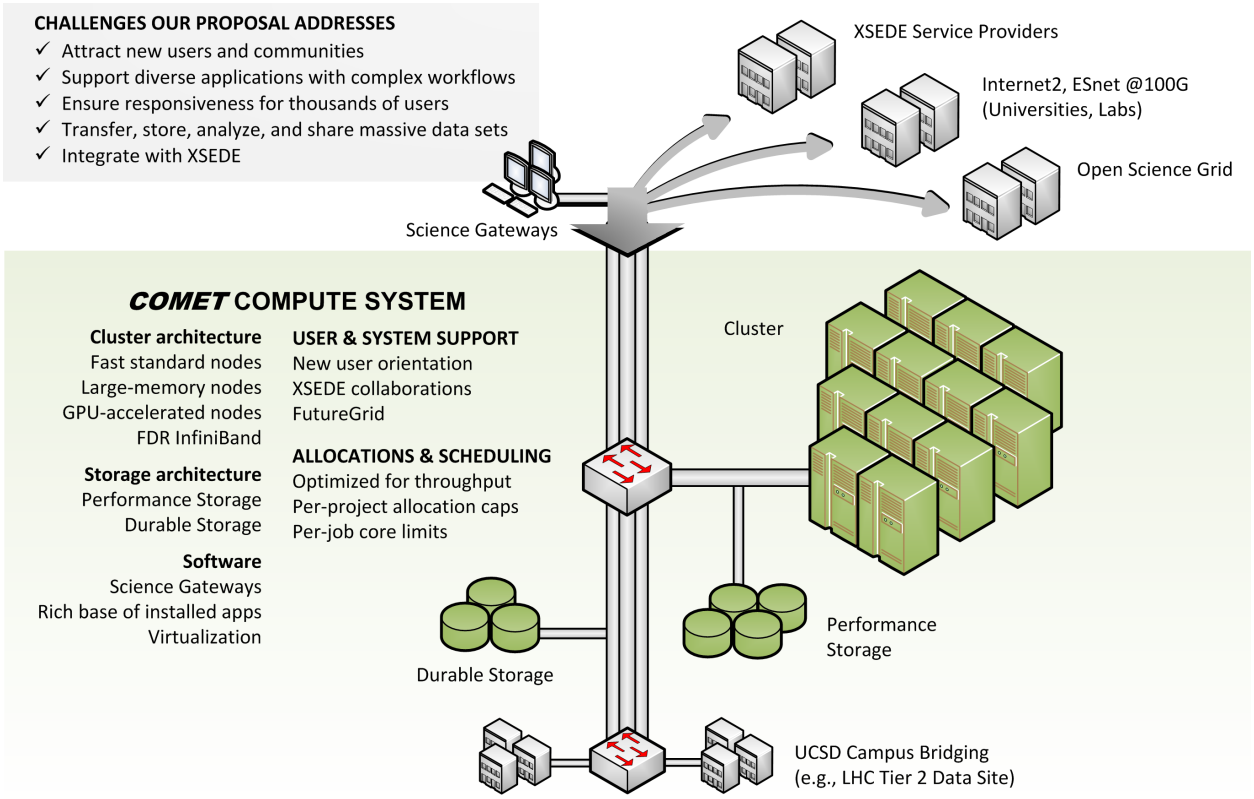

Figure 1. High-level view of *Comet* summarizing the architectural features, relationship between main components and how *Comet* will support a diverse community of users with a wide range of compute and data requirements.

needs of the long tail of science. *Comet* will support established HPC users requiring homogeneous systems with access to high-performance parallel file systems. To support targeted new user communities, the design pays special attention to requirements for heterogeneous computing and custom software stacks. Further scaling, to at least 10,000 users, is readily achievable via gateways.

Comet will:

- Deliver ~2 petaflop/s of capacity for the 98% of XSEDE jobs (50% of XSEDE core-hours) that use fewer than 1,000 cores. (See XDMoD [8] and many recent studies [9-12].) *Comet* will support larger jobs too, but its network is optimized for jobs of modest scale.
- Provide 7 raw PB of Lustre-based *Performance Storage* at 200 GB/s bandwidth for both scratch and allocated persistent storage, as well as 6 raw PB of *Durable Storage* for data reliability.
- Enable community-supported custom software stacks through fully virtualized HPC clusters. This is essential when the XSEDE software stack cannot meet community needs. Advances in hardware and software allow these virtual clusters to operate at near-native InfiniBand bandwidth/latency.
- Ensure high throughput and responsiveness to users by using allocation/scheduling policies proven effective on *Trestle*s. And have a "rapid-access" allocation to give users access within one day of their requests.
- Build upon and expand science gateway usage in XSEDE, via gateway-friendly software and operating policies.
- Provide high-speed internal and external communications to support data transfers to/from/between these systems and external research users/institutions.

The IU FutureGrid [13] team is contributing in several technical areas, including (i) easy "on-ramps" for new users, (ii) virtualization to allow application developers to create customized software environments, (iii) support for new software technologies, such as Hadoop, (iv) support for gateway development, and (v) bridging to its user base, which is dominated by computer scientists, many of whom are new to the HPC community.

Our focus on inclusivity and reaching new users and communities, particularly via gateways, will have broad impacts across scientific domains and underrepresented constituencies. Training modules related to the new technologies introduced into XSEDE by *Comet* will be contributed to XSEDE. The XSEDE Campus Champions, Novel and Innovative Projects, and MSI Outreach programs will be leveraged to further broaden participation.

2. RESOURCE SPECIFICATION

Comet will meet existing needs while laying new foundations to *expand* the range of computational science that can be

undertaken using emerging modalities of computing. We have designed these systems with our key vendor partners, Dell, Intel and Aeon, to support the typical mix of batch-based HPC while also enabling science gateways, high-performance virtualized clusters, and cloud-style computing. The integrated system architecture is shown in Figure 1 and summarized here.

2.1 Compute Nodes

Comet is designed to support modest-scale jobs that constitute the majority of current and projected computation workload in XSEDE [8]. It will be a heterogeneous system, consisting primarily of standard compute nodes with next-generation Intel Xeon processors, along with 36 NVIDIA GPU-accelerated nodes and four large-memory nodes. The system is expected to have a peak performance of ~2 petaflop/s, over 95% of which comes from Intel processors and the balance from GPUs.

(Note that performance parameters of the next-generation Intel Xeon processor remain proprietary and cannot be disclosed in this paper. Key information such as the clock speed, cores and flops per socket, and number of nodes will be disclosed as soon as possible.)

Each *Comet* standard node will have two Intel processors, 128 GB of DDR4 memory, and two 200-GB flash drives. The latter will support the operating system and provide fast scratch space.

Thirty-six of the standard nodes will each be augmented with four next-generation NVIDIA GPUs. Each of the four large-memory nodes will have four of the same next-generation Intel processors as in the standard nodes and 1.5 TB of DDR4 memory.

2.2 Interconnect and Networking

Each compute rack will have full-bisection InfiniBand FDR connectivity across 72 nodes in the rack. The overall system will have a 4:1 bisection FDR interconnect across the racks. Topology-aware scheduling will allow users with bandwidth-sensitive jobs that require 72 or fewer nodes to choose to run entirely withing a rack.

Each FDR link is rated at 7 GB/s. The MPI latency of the network is projected to be 2.4 µs. Each rack will use seven 36-port Mellanox FDR edge switches with a single FDR connection per node. The edge switches, in turn, will be connected to mid-tier switches via FDR links and then to the FDR core switch by additional FDR links. The FDR core switch will bridge to the Ethernet-based *Performance Storage* using Mellanox 36-port Virtual Protocol Interface (VPI) switches.

Comet will be integrated into the XSEDE network at a minimum of 10-Gbps connectivity and will reach a broad user base via a 100-Gbps link to Internet2 and ESnet. Bandwidth will be reservable on the XSEDE or 100-Gbps links for large-scale data transfers. On-demand Secure Circuits and Advance Reservation System (OSCARS) installations at UCSD, Internet2, ESnet, and other partners will facilitate end-to-end connections across the country. UCSD is one of the campuses chosen for the NSF Dynamic Network System (DYNES) installation [14], which provides OSCARS-configurable, reserved-circuit, guaranteed-bandwidth connections to support large, long-distance scientific data flows [15].

2.3 Major Storage Systems

Comet will include *Performance Storage*, a high-performance Lustre parallel file system based on an evolution of SDSC's *Data Oasis* design [16]. This system will provide 7 raw PB of scratch disk space and persistent allocated storage accessible at 200 GB/s. In addition, *Durable Storage* will provide another 6 raw PB of disk to store a second copy of critical user data, including persistent *Performance Storage* data.

SDSC pioneered high-performance storage over Ethernet fabrics [16]. Ethernet supports campus-area and wide-area networking, while InfiniBand does not without specialized equipment. Ethernet is thus preferred when making high-performance storage accessible beyond the data center. The rapid deployment of 100 GbE in Internet2 and at numerous research institutions around the country, including UCSD, indicates that higher-speed networking capability will become common over the next five years. The *Comet* networking infrastructure pays special attention to creating appropriately sized bridges between our Ethernet and InfiniBand fabrics. The design will allow 100 GbE wide-area connections to integrate easily into our storage, compute, and data systems.

Data Oasis, SDSC's high-performance storage architecture in production on *Trestles* and *Gordon* [16], motivated the design for the Lustre-based *Performance Storage* that will be used by *Comet*. This storage will be part of a 40-GbE fabric accessed via 4 Mellanox IB-Ethernet gateways to create a total of 72 40-GbE ports in the 360-GB/s IB-Ethernet bridge. The basic building block will be an Object Storage Server (OSS), consisting of a dual-socket, Intel server with 64 GB of memory, dual 40-GbE network cards, and 216 TB via 36 x 6-TB SAS drives. Each server will be capable of delivering more than 6.3 GB/s between network and disk. *Performance Storage* will have 32 OSSes for an aggregate capacity of 7 raw PB and a sustained performance of 200 GB/s.

To help protect users from data loss, we will re-deploy SDSC's current *Data Oasis* system (6 raw PB, 100 GB/s) as *Durable Storage* to provide a second copy of all data that is part of users' storage allocations (non-scratch *Performance Storage*). The storage will be accessed over the same storage fabric described above and offers a fast, cost-effective alternative to the deployment of a new tape-based archive.

2.4 Service Nodes and Ancillary Storage

The *Comet* design includes additional service nodes and storage systems to ensure a robust, usable environment for users. *Comet* will include: 4x 40 Gbps data mover nodes that host GridFTP and serve Globus Online; login nodes operated in a round-robin fashion; nodes for Rocks systems management; virtualized servers for hosting user/community gateway front ends and related application services; mirrored NFS servers for user home file systems; and an additional NFS storage repository specifically for virtual cluster images.

3. HIGH-PERFORMANCE VIRTUALIZATION

Virtualization provides a way to run systems and applications environments significantly different than those on a traditional batch system. The software stack, from operating system to application, is packaged within a single image file that can be instantiated on the production system. Our partner FutureGrid

will have the lead role in working with targeted user communities to develop such custom environments and prepare them for use on the production *Comet*.

Performance within virtualized environments for applications that extend beyond a node has been a major barrier, thus far, to adoption of this technology in HPC [17-22]. Our testing last year confirmed that this was still the case for commercial services like Amazon EC2. However Single Root I/O Virtualization (SR-IOV) within InfiniBand adapters [23] will be available in *Comet*. This technology, which has now reached a level of maturity that makes it appropriate for production environments, reduces virtualization overhead to minimal levels and dramatically opens the landscape for both single-node and highly parallel applications to run efficiently in virtual environments on *Comet* [24].

We will use Rocks native functionality and OpenStack Compute (Nova) to manage virtual machines (VMs). Nova provides a graphical interface for users to manage images and VM instances. VMs will be provisioned as predefined compute appliances based on application needs. They will be incorporated into the batch system where users can request them like any other queue. Thus, users will have a choice whether or not to run in a virtual environment. Any user or gateway that prefers a custom software stack and environment can request a VM and run without special provisioning.

Virtualization can provide increased flexibility for science gateway developers and users, as gateways become more sophisticated and move beyond simple web-based job submission to HPC systems. For example, NSGPortal [25] allows users to submit C++ code to describe customized neuronal models, which are then compiled and run within the NEURON simulation environment. However, this can be a security risk and typically is not allowed on HPC systems. By virtualizing the environment, the "untrusted" code can be isolated from the rest of the cluster and be executed safely.

4. PERFORMANCE AND INNOVATION FOR SCIENCE AND ENGINEERING APPLICATIONS

Comet will support the vast majority of existing XSEDE users with more capacity on faster processors. It also will provide thousands of new users access to HPC via gateways and virtualized software environments.

4.1 Focus on modest-scale applications

During 2012, 98% of XSEDE jobs ran on 1,024 cores or less, and such jobs consumed ~50% of all service units across XSEDE resources [8]. *Comet* has been designed to target this applications sector and to enable responsive turnaround for this vast majority of modest-scale jobs, yet still will provide good performance for occasional jobs at higher core counts. Table 1 lists example analyses and codes that will excel on *Comet.*

4.1.1 Standard nodes for most XSEDE applications

Most applications and analyses that currently run on XSEDE resources will run on the standard nodes of *Comet* and achieve superior performance. These analyses encompass those in the physical science and engineering disciplines with years of experience using HPC as well as those in disciplines such as biology and social science that are relatively new to HPC. We had demonstrated the feasibility of creating an interface to the OSG job manager on *Gordon* for the analysis Large Hadron Collider data [26] and will continue to provide this functionality on *Comet*.

Biologists, chemists, and Earth scientists are doing their analyses on *Gordon* and *Trestles* through the convenience of compute-processing gateways. All of these analyses will run well on the standard nodes of *Comet*.

In all fields of science and engineering, parameter scans constitute a common workflow. They can involve thousands of jobs run in a pleasingly parallel fashion with individual jobs typically running on a few cores or even serially. The increased throughput provided by the *Comet* standard nodes will allow more extensive scans, and many will run efficiently in the shared-job queue.

Table 1. Example analyses and codes that will excel on Comet

Type of Analysis	Typical Codes	Nodes Used
Phylogenetic tree inference	BEAST, MrBayes, RAxML	Standard
Neural simulation	GENESIS, NEURON	Standard
High-frequency trading analysis	Custom	Standard + flash
Climate simulation	SAM-MMF, WRF	Standard
Visualizing simulation output	VisIt, ParaView	Standard
Analyses requiring virtualization	Custom	Standard
Quantum chemistry	Gaussian, Q-Chem	Standard + flash, large memory
Structural analysis	Abaqus	Standard + flash, large memory
De novo assembly of genomes	SOAPdenovo2, Velvet	Large memory
Molecular dynamics	Amber, CHARMM, Gromacs, NAMD	GPU-accelerated

4.1.2 Flash drives for applications with large scratch needs or large file counts

Most quantum chemistry codes, e.g., Gaussian [27] and Q-Chem [28], create temporary scratch files to store one- and two-electron integrals. Storing these integrals on locally attached flash memory instead of disk can greatly improve performance, reduce I/O traffic on the interconnect, and avoid overwhelming the Lustre metadata servers. Structural mechanics codes, such as Abaqus [29], also benefit from storing the stiffness matrix on flash memory. Such performance benefits have attracted many XSEDE users to run on *Gordon* and *Trestles*,and similar benefits will be provided by the flash memory on every *Comet* node.

A noteworthy example is the work of Alan Aspuru-Guzik's group at Harvard. As part of the Clean Energy Project [30], his team is screening millions of organic semiconductors to find better photovoltaic materials for solar cells. Each semiconductor is analyzed in a single Q-Chem job on either *Gordon* or *Trestles*. Hundreds of serial jobs run at a time, all using flash memory to improve performance and minimize the impact on other users that disk I/O would entail. This is a massive undertaking that will take years to complete, and the availability of *Comet* will accelerate the search.

Within academic finance, Mao Ye of UIUC is analyzing high frequency trading on *Gordon* to identify abnormal trading activity and suggest regulatory solutions. After performance tuning by SDSC staff, Ye's software now runs more than 5,000 times faster, so he can analyze the combined NYSE, NASDAQ and BATS exchanges (~ 8,000 stocks) for the heaviest days of trading in two hours using a single node of *Gordon*. The new approach requires the generation of nearly 100,000 intermediate files and will make very heavy use of *Comet's* flash drives.

4.1.3 Large memory applications can use Comet's 1.5 TB nodes

The standard nodes of *Comet* will have 128 GB of memory each, twice that on *Gordon* and *Trestles*. Most shared-memory codes that require access to specialized large-memory XSEDE resources will run on *Comet's* standard nodes but do so faster. *Comet* will also have four nodes with even more memory: 1.5 TB each. These large-memory nodes will enable analyses with exceptionally large shared-memory requirements.

Determining the genome sequence of a species for the first time requires *de novo* assembly of the so-called short reads produced by current DNA sequencers. For vertebrate genomes, the assemblers require large amounts of shared memory: hundreds of GB for SOAPdenovo2 [31] and more than a TB for Velvet [32]. As new sequencers producing longer reads become available, these memory requirements will grow. The 1.5-TB nodes on *Comet* will be capable of performing assemblies using both applications. Moreover, all but the first of the four steps in a SOAPdenovo2 assembly will run on the 128-GB nodes of *Comet.*.

Researchers from Cornell used Abaqus on *Gordon* to model the response of a rat vertebra to stress and hope to model human vertebrae. An attempted high-resolution analysis using 750 GB of flash memory was still not complete after 100 hours. However, the same analysis using 750 GB of DRAM aggregated across 16 *Gordon* nodes with the vSMP software completed in 10 hours. This analysis will run even faster on a 1.5-TB node of *Comet*.

4.1.4 Vectorizable applications will use Comet's GPU-accelerated nodes

Applications that can be vectorized are good candidates for running on GPUs rather than on x86 CPUs. Many molecular dynamics codes have vectorizable kernels. Four codes that are heavily used within XSEDE – Amber [33, 34], CHARMM [35], Gromacs [36], and NAMD [37] – have been implemented on NVIDIA GPUs and run faster on them than on x86 CPUs. To support such applications, *Comet* will include 36 GPU-accelerated nodes, each with four GPUs.

Amber PMEMD, in particular, has been highly optimized and achieves outstanding performance running only on the GPU, not the CPU. For a standard benchmark with 23k atoms, Amber PMEMD runs about four times faster on a single NVIDIA K20 GPU as on a single node of *Gordon* [38]. We expect similar benefits for each GPU in *Comet*. Moreover, four independent simulations can be run on the GPUs in a *Comet* node with no loss of speed. This cost-effective increase in throughput will allow better statistics in studies that use enhanced sampling methods.

4.2 Gateways

Gateways provide browser-based access to remote resources for compute processing, data processing, or data serving. Compute- and data-processing gateways have proven ideal in helping new users run community codes on XSEDE resources. Such gateways are particularly effective for researchers who only occasionally compute. They do not need to get individual allocations, and the browser interface relieves them from having to learn about HPC details, such as parallelization and job scheduling. Users simply select code options and upload their data files via the interface without logging onto an HPC system.

SDSC is a leader in deploying and supporting all of these types of gateways within XSEDE. SDSC pioneered gateway-friendly operational policies on *Trestles*, extended them to *Gordon*, and will adopt similar policies for *Comet*. For compute- and data-processing gateways, these policies include flexible allocations in response to growing demand, quick turnaround for short jobs, a long-job queue for codes without restart capability, and a shared-job queue to allow cost-effective execution of codes that do not scale to a full node.

During 2012, 66% of all XSEDE gateway usage in unnormalized core hours was on *Gordon* and *Trestles*. Currently SDSC provides access to computationally demanding community codes running on *Gordon* and *Trestles* via five NSF-funded gateways:

- CIPRES [39] runs BEAST [40], MrBayes [41], and RAxML [42] to infer phylogenetic trees;
- CyberGIS [43] runs pRasterBlaster [44] to project maps and TauDEM [45] to determine water flow from topography;
- GridChem [46] runs Gaussian [27] to calculate electronic structures;
- The Neuroscience Gateway [25] runs GENESIS [47] and NEURON [48] to model individual neurons and networks of neurons; and
- UltraScan [49] runs its eponymous software to analyze ultracentrifugation experiments.

The much larger throughput of *Comet* will allow these existing compute-processing gateways as well as emerging ones to serve many more users and jobs.

5. USER SUPPORT

The new users, communities, and modalities of computing that we will promote through *Comet* require distinct competencies and new models of support. Two of these are particularly relevant – new users and gateway communities.

5.1 New users

Relevant and up-to-date documentation, targeted training, and a well-staffed help desk, are essential to assist new users in getting started in HPC. Our documentation and training will highlight the unique characteristics of *Comet* with example-based materials and will be integrated into the XSEDE user portal to provide ease of access. Widely used applications, such as Matlab, Hadoop, and R, will be documented with step-by-step instructions, sample scripts, and screenshots wherever applicable. The emphasis will be on running applications in parallel (e.g., MATLAB Distributed Computing Server and the R multicore package) and other HPC related topics. Much of this material has already been developed for *Gordon* and *Trestles*, and we will be updated for *Comet*.

A large set of open-source software packages will be provided on both resources, as is the case on *Gordon* and *Trestles* today. These will be made available to users via the XSEDE-standard Modules environment. Since commercial packages can be expensive for large-scale HPC systems, we will carefully select only those that complement the capabilities of open-source software and are not generally available on other national HPC systems. A robust and relevant software offering is critical to broadening the user base and bridging from campus to national-scale resources.

New user training on *Comet* will cover programming/debugging tools (DDT), compilers, code parallelization (MPI, OpenMP, OpenACC), use of the batch systems, as well as software like MATLAB, which is a familiar development and run-time environment for many new user communities. We will also leverage XSEDE's excellent archived courses and tutorials covering basic HPC topics.

The XSEDE start-up allocation process can be a barrier to some new user communities, especially when all they want is to test-drive the HPC systems to see if they will work for them. We intend to work with XSEDE on a new "rapid-access" allocations approach that will cut the wait time for an introductory test account from 1-2 weeks to just one day. Similar to what is done now for training accounts, we will pre-populate a fixed number of trial accounts to avoid having to create one for each new request. The allocation amounts will be limited (much less than start-up accounts), but sufficient for evaluating the software environment and running simple test jobs.

XSEDE NIPS (Novel and Innovative Project Support) and Campus Champions have proven to be effective ways to bring in new users and communities from participating campuses,. We will work with both programs to identify and support new users and communities most likely to benefit from *Comet's* powerful architecture.

5.2 Gateway communities

By providing web front-ends to domain-specific applications and data on XSEDE, science gateways are enabling large, self-identified research communities to access national resources through a common interface configured for ease of use [50]. While leveraging gateways will be crucial, especially to achieve NSF's objectives with respect to inclusivity and supporting large numbers of users and new communities, we will not have the resources to develop gateways per se, which would be best done by the disciplinary communities themselves. Rather, we will train XSEDE gateway developers in the use of *Comet* as a gateway resource, and seek their guidance in ensuring that *Comet* is configured as a gateway-friendly resource.

Over the last several years, SDSC has led development of innovative allocations policies, software, security protocols, and user access/accounting to enable creation of gateways such as CIPRES and GridChem. We will continue this work by reaching out to new communities and assisting them in building such gateways and porting their current web portal environments to *Comet*. Specifically, we will: (1) deploy common software elements needed to support gateway functionality on *Comet*; (2) provide dedicated nodes for virtual machines, integrated in the system, for gateway developers to host gateways/portals; (3) work with XSEDE to establish gateway-friendly allocations policies that will accommodate the rapid growth in gateway usage; (4) configure scheduling policies to support large numbers of jobs from gateways; (5) support shared-node runs for multiple small jobs that can be packed into a single node; (6) support long-running jobs (current practice on *Trestles/Gordon*); and (7) augment limited resources with targeted support/training for gateway developers and community trainers through venues like the Gateway Science Institute [51] XSEDE ECSS, and communities that express interest in migrating to the gateway model.

Gateways are becoming increasingly sophisticated making it possible, for example, for users to compile user code or specify complex workflows through the gateway [25, 52]. This increased functionality also implies a more flexible software environment—different gateways may require different software stacks. The FutureGrid team will assist in providing support for an "on-ramp" environment on *Comet* where gateway developers can use virtualization to develop these specialized software stacks/images. Once developed and tested, they can be migrated to production where they can execute within virtualized containers. We will dedicate a set of *Comet* nodes for this on-ramp effort using the XSEDE start-up and educational allocation mechanisms.

6. ALLOCATION POLICIES

Comet will be allocated via the XRAC allocation process, with 90% of its available cycles allocated to XSEDE users. We project at least 10,000 users on *Comet* (although this number could be substantially higher as gateways continue to gain adoption).

Comet will be operated to ensure high throughput for a large number of users with modest-scale jobs. SDSC proposed the *Trestles* system in 2010 with these objectives and pioneered a number of allocations and scheduling policies to accomplish them [9, 53, 54]. We will: (1) limit the amount of time any single project can be allocated to ~2% of available resources (gateways will be excepted as they represent many users); (2) limit jobs sizes to <2,000 cores, with exceptions allowed by special request, (3) maintain utilization at slightly lower levels than most HPC systems to improve throughput/productivity [9]; (4) maintain a modest allocation contingency to allow generous start-up allocations and elasticity for gateway allocations (with hard-to-predict usage); (5) proactively monitor queue wait times and expansion factors; (6) implement topology-aware scheduling algorithms to optimize internode performance; (7) enable short-pool, user-settable and on-demand reservations [53]; (8) define a pool of nodes for development/debugging, available on short notice via the scheduler; and (9) establish

metrics for user access and throughput to assess the effectiveness of our procedures.

7. ACKNOWLEDGMENTS

Comet is supported by NSF grant: ACI #1341698 Gateways to Discovery: Cyberinfrastructure for the Long Tail of Science. The authors thank Stephanie Sides and Ben Tolo for their important contributions.

8. REFERENCES

[1] *NSB 93-205 -- NSF Blue Ribbon Panel on High Performance Computing*. NSF, Arlington, VA, 1993.

[2] *NSF Advisory Committee for Cyberinfrastructure Task Force on Campus Bridging. Final Report*. NSF, Arlington, VA, March 2011.

[3] *Cyberinfrastructure Framework for 21st Century Science and Engineering: Vision*. NSF, Arlington, VA, May 2012.

[4] *Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21)*, http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=504730

[5] *ACCI Task Force reports, available at* http://www.nsf.gov/od/oci/taskforces/, NSF, Arlington, VA.

[6] Stewart, C. A., Katz, D. S., Hart, D. L., Lantrip, D., McCaulay, D. S. and Moore, R. L. *Survey of Cyberinfrastructure Needs and Interests of NSF-funded Principal Investigators*. Indiana University, Bloomington, IN, January 2011.

[7] Katz, D. S., Keahey, K. and Jul, S. *TeraGrid eXtreme Digital 'Wide Users' Requirements Elicitation Meeting, Computation Institute Technical Report CI-TR-10-0811*. University of Chicago and Argonne National Laboratory, 2011.

[8] *XDMoD - XSEDE Metrics on Demand, NSF award OCI-1025159*.

[9] Moore, R. L., Jundt, A., Carson, L. K., Yoshimoto, K., Ghadersohi, A. and Young, W. S. *Analyzing throughput and utilization on Trestles, Proceedings of XSEDE12*. ACM, (Chicago, IL, 2012).

[10] Furlani, T. R., Schneider, B. I., Jones, M. D., Towns, J., Hart, D. L., Patra, A. K., DeLeon, R. L., Gallo, S. M., Lu, C.-D. and Ghadersohi, A. *Data analytics driven cyberinfrastructure operations, planning and analysis using XDMoD, SC12 Conference, Salt Lake City, UT*, 2012.

[11] Hart, D. *Deep and wide metrics for HPC resource capability and project usage, Supercomputing '11, November 2011, Seattle, WA, USA*. ACM.

[12] Schneider, B. *A Data History of TeraGrid/XSEDE Usage: Defining a Strategy for Advanced CyberInfrastructure (ACI)*, April 2012.

[13] *FutureGrid, https://portal.futuregrid.org/*.

[14] Boyd, E., Newman, H., McKee, S. and Sheldon, P. *MRI-R2 Consortium: Development of Dynamic Network System (DYNES), NSF ACI award 0958998*, 2010.

[15] Cortese, J. *New Dynamic Circuit Provisioning Available on Pacific Wave,* http://pacificwave.net/p=433/, November 26, 2012.

[16] *2012 Annual HPCwire Readers' Choice Awards, November 2012,*http://www.hpcwire.com/specialfeatures/2012_Annual_HPCwire_Readers_Choice_Awards.html.

[17] Jorissen, K., Vila, F. D. and Rehr, J. J. A high performance scientific cloud computing environment for materials simulations. *Computer Physics Communications*, 183, (9) 2012, 1911-1919.

[18] Rehr, J. *SI2-SSE: Cloud-Computing-Clusters for Scientific Research, NSF ACI award 1048052*. NSF, 2010.

[19] Rehr, J. J., Vila, F. D., Gardner, J. P., Svec, L. and Prange, M. Scientific computing in the cloud. *Computing in Science & Engineering*, 12, (3) 2010, 34-43.

[20] Yelick, K., Coghlan, S., Draney, B. and Cannon, R. S. *The Magellan Report on Cloud Computing for Science, U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research*. December 2011.

[21] Jackson, K. R., Ramakrishnan, L., Muriki, K., Canon, S., Cholia, S., Shalf, J., Wasserman, H. J. and Wright, N. J. *Performance analysis of high performance computing applications on the amazon web services cloud*. IEEE 2nd International Conference on Cloud Computing Technology and Science (CloudCom), 2010.

[22] Mehrotra, P., Djomehri, J., Heistand, S., Hood, R., Jin, H., Lazanoff, A., Saini, S. and Biswas, R. *Performance evaluation of Amazon EC2 for NASA HPC applications. Proceedings of the 3rd workshop on Scientific Cloud Computing,* ACM, (Delft, The Netherlands, June 2012).

[23] *Overview of Single Root I/O Virtualization (SR-IOV),* http://msdn.microsoft.com/enus/library/windows/hardware/hh440148%28v=vs.85%29.aspx.

[24] Lockwood, G. K., Tatineni, M. and Wagner, R. P. *SR-IOV: Performance Benefits for Virtualized Interconnects, Submitted to XSEDE14, (Atlanta GA, July 13-18 2014)*.

[25] *Neuroscience Gateway Portal,* http://www.nsgportal.org.

[26] Wagner, R., Tatineni, M., Hocks, E., Yoshimoto, K., Sakai, S., Norman, M. L., Bockelman, B., Sfiligoi, I., Tadel, M. and Letts, J. *Using Gordon to accelerate LHC science*. Proceedings of XSEDE13, ACM , (San Diego, CA, July, 2013).

[27] *Gaussian,* http://www.gaussian.com. City.

[28] Kong, J., White, C. A., Krylov, A. I., Sherrill, D., Adamson, R. D., Furlani, T. R., Lee, M. S., Lee, A. M., Gwaltney, S. R. and Adams, T. R. Q-Chem 2.0: a high-performance ab initio electronic structure program package. *Journal of Computational Chemistry*, 21, (16) 2000, 1532-1548.

[29] Hibbitt, Karlsson and Sorensen *ABAQUS/Standard user's manual*. Hibbitt, Karlsson & Sorensen, 2001.

[30] *Clean Energy Project, https://cleanenergy.harvard.edu*.

[31] Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q. and Liu, Y. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1, (1) 2012, 18.

[32] Zerbino, D. R. and Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, 18, (5) 2008, 821-829.

[33] Götz, A. W., Williamson, M. J., Xu, D., Poole, D., Le Grand, S. and Walker, R. C. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized Born. *Journal of chemical theory and computation*, 8, (5) 2012, 1542-1555.

[34] Salomon-Ferrer, R., Götz, A. W., Poole, D., Le Grand, S. and Walker, R. C. Routine microsecond molecular dynamics simulations with Amber on GPUs. 2. Explicit solvent particle mesh Ewald. *Journal of Chemical Theory and Computation*, 9, (9) 2013, 3878-3888.

[35] Brooks, B. R., Brooks, C. L., MacKerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C. and Boresch, S. CHARMM: the biomolecular simulation program. *Journal of computational chemistry*, 30, (10) 2009, 1545-1614.

[36] Hess, B., Kutzner, C., Van Der Spoel, D. and Lindahl, E. GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation*, 4, (3) 2008, 435-447.

[37] Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., Chipot, C., Skeel, R. D., Kale, L. and Schulten, K. Scalable molecular dynamics with NAMD. *Journal of computational chemistry*, 26, (16) 2005, 1781-1802.

[38] *AMBER (PMEMD) Benchmarks,* http://ambermd.org/gpus/benchmarks.htm.

[39] *CIPRES,* http://www.phylo.org.

[40] Drummond, A. J. and Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, 7, (1) 2007, 214.

[41] Huelsenbeck, J. P. and Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17, (8) 2001, 754-755.

[42] Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22, (21) 2006, 2688-2690.

[43] *CyberGIS,* http://cybergis.cigi.uiuc.edu.

[44] Behzad, B., Liu, Y., Shook, E., Finn, M. P., Mattli, D. M. and Wang, S. A Performance Profiling Strategy for High-Performance Map Re-Projection of Coarse-Scale Spatial Raster Data, In *Auto-Carto 2012, a cartography and geographic information society research symposium*, Columbus, OH, 2012.

[45] Tarboton, D. G. Terrain analysis using digital elevation models (TauDEM). *Utah State University, Logan*2005).

[46] *Computational Chemistry Grid,* http://www.gridchem.org.

[47] Bower, J. M. and Beeman, D. *The book of GENESIS: exploring realistic neural models with the GEneral NEural SImulation System*. The Electronic Library of Science, 1995.

[48] Carnevale, N. T. and Hines, M. L. *The NEURON book*. Cambridge University Press, 2006.

[49] *UltraScan Analysis Software,* http://ultrascan.uthscsa.edu.

[50] Wilkins-Diehr, N., Gannon, D., Klimeck, G., Oster, S. and Pamidighantam, S. TeraGrid science gateways and their impact on science. *Computer*, 41, (11) 2008, 32-41.

[51] *Science Gateway Institute,* http://sciencegateways.org.

[52] Miller, M. A., Pfeiffer, W. and Schwartz, T. *The CIPRES science gateway: enabling high-impact science for phylogenetics researchers with limited resources*, *Proceedings of XSEDE12*, ACM, (Chicago, IL, 2012).

[53] Moore, R. L., Hart, D. L., Pfeiffer, W., Tatineni, M., Yoshimoto, K. and Young, W. S. *Trestles: a high-productivity HPC system targeted to modest-scale and gateway users*. *Proceedings of TeraGrid 11*, ACM, (Salt Lake City, UT, 2011).

[54] Yoshimoto, K., Choi, D., Moore, R., Majumdar, A. and Hocks, E. Implementations of Urgent Computing on Production HPC Systems. *Procedia Computer Science*, (9) 2012, 1687-1693.