Description of Facilities and Resources

Oak Ridge National Laboratory
and the
UT-ORNL Joint Institute for Computational Sciences

1. Oak Ridge National Laboratory

Computer Facilities. The Oak Ridge National Laboratory (ORNL) hosts three petascale computing facilities: the Oak Ridge Leadership Computing Facility (OLCF), managed for DOE; the National Institute for Computational Sciences (NICS) computing facility operated for the National Science Foundation (NSF); and the National Climate-Computing Research Center (NCRC), formed as collaboration between ORNL and the National Oceanographic and Atmospheric Administration (NOAA) to explore a variety of research topics in climate sciences. Each of these facilities has a professional, experienced operational and engineering staff comprising groups in high-performance computing (HPC) operations, technology integration, user services, scientific computing, and application performance tools. The ORNL computer facility staff provides continuous operation of the centers and immediate problem resolution. On evenings and weekends, operators provide first-line problem resolution for users with additional user support and system administrators on-call for more difficult problems.

Other Facilities. The Oak Ridge Science and Technology Park at ORNL is the nation's first technology park on the campus of a national laboratory. The technology park is available for private sector companies that are collaborating with research scientists. Laboratory officials anticipate that the new park will be used to help create new companies from technologies developed at ORNL.

1.1 Primary Systems

Titan is a Cray XK7 system consisting of 18,688 AMD sixteen-core Opteron™ processors providing a peak performance of more than 3.3 petaflops (PF) and 600 terabytes (TB) of memory. A total of 512 service input/output (I/O) nodes provide access to the 32 petabytes (PB) "Spider" Lustre parallel file system at more than 1 terabyte (TB/s). External login nodes (decoupled from the XK7 system) provide a powerful compilation and interactive environment using dual-socket, twelve-core AMD Opteron processors and 256 GB of memory. Each of the 18,688 Titan compute nodes is paired with an NVIDIA Kepler graphics processing unit (GPU) designed to accelerate calculations. With a peak performance per Kepler accelerator of more than 1TF, the aggregate performance of Titan exceeds 27PF. Titan is the Department of Energy's most powerful open science computer system and is available to the international science community through the INCITE program, jointly managed by DOE's Leadership Computing Facilities at Argonne and Oak Ridge National Laboratories and through the DOE Office of Science's ALCC program.

Gaea consists of a pair of Cray XE6 systems. The smaller partition contains 2,624 socket G34 AMD 16-core Opteron processors, providing 41,984 compute cores, 84 TB of double data rate 3 (DDR3) memory,

and a peak performance of 386 teraflops (TF). The larger partition contains 4,896 socket G34 AMD 16-core Interlagos Opteron processors, providing 78,336 compute cores, 156.7 TB of DDR3 memory, and a peak performance of 721 TF.

The aggregate system provides 1.106 PF of computing capability, and 248 TB of memory. The Gaea compute partitions are supported by a series of external login nodes and a single Lustre filesystem. The F1 file system is based on more than 2,000 Nearline-SAS drives and provides just under 6 PB (formatted) space to all compute partitions as well as the data transfer capability to the NOAA archive. Gaea is the NOAA climate community's most powerful computer system and is available to the climate research community through the Department of Commerce/NOAA.

The ORNL Institutional Cluster (OIC) consists of two phases. The original OIC consists of a bladed architecture from Ciara Technologies called VXRACK. Each VXRACK contains two login nodes, three storage nodes, and 80 compute nodes. Each compute node has dual Intel 3.4 GHz Xeon EM64T processors, 4 GB of memory, and dual gigabit Ethernet interconnects. Each VXRACK and its associated login and storage nodes are called a block. There are a total of nine blocks of this type. Phase 2 blocks were acquired and brought online in 2008. They are SGI Altix machines. There are two types of blocks in this family.

- Thin nodes (3 blocks). Each Altix contains 1 login node, 1 storage node, and 28 compute nodes within 14 chassis. Each node has eight cores and 16 GB of memory. The login and storage nodes are XE240 boxes from SGI. The compute nodes are XE310 boxes from SGI.
- Fat nodes (2 blocks). Each Altix contains 1 login node, 1 storage node, and 20 compute nodes within 20 separate chassis. Each node has eight cores and 16 GB of memory. These XE240 nodes from SGI contain larger node-local scratch space and a much higher I/O to this scratch space because the space is a volume from four disks.

EOS is a 744-node Cray XC30 cluster with a total of 47.6 TB of memory. The processor is the Intel® Xeon® E5-2670. Eos uses Cray's Aries interconnect in a network topology called Dragonfly. Aries provides a higher bandwidth and lower latency interconnect than Gemini. Support for I/O on Eos is provided by (16) I/O service nodes. The system has (2) external login nodes.

The compute nodes are organized in blades. Each blade contains (4) nodes connected to a single Aries interconnect. Every node has (64) GB of DDR3 SDRAM and (2) sockets with (8) physical cores each. Intel's Hyper-threading (HT) technology allows each physical core to work as two logical cores so each node can functions as if it has (32) cores. Each of the two logical cores can store a program state, but they share most of their execution resources. Each application should be tested to see how HT impacts performance before HT is used. The best candidates for a performance boost with HT are codes that are heavily memory-bound. The default setting on Eos is to execute without HT, so users must invoke HT with the -j2 option to aprun.

In total, the Eos compute partition contains 11,904 traditional processor cores (23,808 logical cores with Intel Hyper-Threading enabled), and 47.6 TB of memory.

Rhea is a (196)-node commodity-type Linux cluster. The primary purpose of Rhea is to provide a conduit for large-scale scientific discovery via pre- and post-processing of simulation data generated on Titan. Users with accounts on INCITE- or ALCC-supported projects will automatically be given an account on Rhea. Director's Discretion (DD) projects may request access to Rhea.

Each of Rhea's nodes contains two 8-core 2.0 GHz Intel Xeon processors with Hyper-Threading and 64GB of main memory. Rhea is connected to the OLCF's 32PB high performance Lustre filesystem "Atlas".

1.2 The University of Tennessee

The University of Tennessee Knoxville (UTK) and Oak Ridge National Laboratory (ORNL) established the Joint Institute for Computational Sciences (JICS) in 1991 to encourage and facilitate the use of high-performance computing in the state of Tennessee. When UT joined Battelle Memorial Institute in April 2000, to manage ORNL for the Department of Energy (DOE), the vision for JICS expanded to encompass becoming a world-class center for research, education, and training in computational science and engineering. JICS advances scientific discovery and state-of-the-art engineering by enhancing knowledge of computational modeling and simulation through educating a new generation of scientists and engineers well versed in the application of computational modeling and simulation to solving the world's most challenging scientific and engineering problems.

The JICS facility, Figure 1, represents a large investment by the state of Tennessee and features a state-of-the- art interactive distance learning center with seating for 66 people, conference rooms, informal and open meeting space, executive offices for distinguished scientists and directors, and incubator suites for students and visiting staff.

Figure 1 Joint Institute for Computational Sciences building

The JICS facility is a hub of computational and engineering interactions. Joint faculty, postdocs, students, and research staff shares the building, which is designed specifically to provide intellectual and practical stimulation. The auditorium serves as the venue for invited lectures and seminars by representatives from academia, industry, and other laboratories, and the open lobby doubles as casual meeting space and the site for informal presentations and poster sessions, including an annual student poster session with over 200 presenters.

JICS employs professional research staff, joint faculty, postdoctoral fellows and students, and administrative staff. The joint faculty holds dual appointments as faculty members in departments at UT and as staff members in ORNL research groups.

One of JICS' main projects is the National Institute for Computational Sciences (NICS), originally founded in 2007. The mission of NICS is to enable the scientific discoveries of researchers nationwide by providing leading-edge computational resources and education, outreach, and training. NICS has a professional, experienced operational and engineering staff comprising groups in HPC operations, technology integration, user services, scientific computing, and application performance tools.

1.3 Computer Facilities

In June 2004, NICS moved into the then brand new 52,000 ft^2 building next door to the ORNL OLCF computer facility. The JICS building has a 1,500 ft^2 computer room, which is home to Beacon. The OLCF computer facility, located on the ORNL campus, is among the nation's most modern facilities for

scientific computing and currently is home to Darter, Nautilus, and Keeneland. The OLCF facility has 40,000 ft^2 divided equally into two rooms with 9 2.5MVA transformers, and another 27,000 ft^2 divided over two rooms in a recently added expansion building with 1 2.5 MVA transformer and the ability to expand; and finally 6,600 tons of chilled water – all of which is designed specifically for high-end computing systems. And finally, the UTK campus has a 2,116 ft^2 computer room as well.

JICS has a professional, experienced operational and engineering staff comprising groups in HPC operations, technology integration, user services, scientific computing, and application performance tools. JICS utilizes staff that provides continuous monitoring in all of these facilities and immediate problem resolution. On evenings and weekends, operators provide first-line problem resolution for users with additional user support and system administrators on-call for more difficult problems.

1.4. Resources

The following sections describe the resources the NICS currently operates.

Darter

NICS obtained a Cray XC30 system funded by UT in the first quarter of 2013. This system, named Darter, has 1,448 compute sockets each with an Intel Sandy Bridge processor (8 cores/socket). In total, the machine provides 240.9 TFlops of compute and 23.2 TB of memory. Darter is located in the OLCF machine room and is administered by NICS staff. This machine was purchased to provide NICS both early experience with the Cray XC30 architecture and provide computational cycles to UT and higher education institutions in the Tennessee system. There are 38 Tennessee projects allocated on Darter that have used tens of millions of cycles. From May 2014 to April 2015, 50% of the machine will be allocated to XSEDE users. Most of the users will be transfers from Kraken, which was decommissioned in April of 2014. Initially, there will be 18 transfers from Kraken totally approximately 15 million service units, with an expectation of more than 50 million service units.

High-Performance Storage

NICS currently supports two Lustre file systems. A direct attached scratch file system is available as a Lustre file system comprised of two couplets of Cray Sonexion 1600 storage controllers and back end disk, accessed through an FDR InfiniBand storage area network (SAN). The Sonexion scratch file system provides approximately 350 TB of short-term, high-performance storage to users, with a peak I/O rate of 11 GB/s. The Medusa file system is a multi cluster file system implemented as a site-wide file system at NICS. Medusa is currently running off of three couplets of DDN 10K controllers and their back end disk, and accessed through a QDR InfiniBand storage area network (SAN). The Medusa file system provides approximately a 1.3 PB of capacity with a peak I/O rate of 30 GB/s.

High-Performance Data Transfer Nodes

NICS currently supports multiple GridFTP data transfer nodes for high-performance data transfer between NICS and external computer systems. Currently, NICS has two sets of four data transfer nodes each with 10 Gb network connectivity and InfiniBand access to Lustre storage resources. These two sets of four data transfer nodes can support file transfer capability over four parallel streams simultaneously provided by GridFTP services. For the period October 2012 through September 2013 over 75 million file transfer requests and over 1 PB of data has been transferred in and out of NICS using GridFTP services.

Beacon

Beacon is an experimental cluster within the UT Application Acceleration Center of Excellence (AACE) funded through a NSF Strategic Technologies for Cyberinfrastructre (STCI) grant. The NSF-funded Beacon consists of sixteen compute nodes, a login node, and a management node. Each compute node is based on a dual socket Intel Xeon E5-2670 system with two Intel Knight's Ferry Cards. This cluster is being used to prepare NSF application teams and their applications for future systems based on the Intel Many Integrated Core (MIC) architecture.

In late 2012, UTK invested in Beacon and expanded the cluster with 48 new compute nodes, six IO nodes, and two more management nodes. Each of these compute nodes is a dual socket Intel Xeon E5-2670 system with four Intel Xeon Phi 5110P coprocessors. Using 36 of these new compute nodes; a Green500 run was done that demonstrated a new world record for power efficiency delivering just under 2.5 GFlops per Watt in November of 2013. Delivering over 100 TFlops of performance, the expanded Beacon will be used to prepare applications for the high levels of parallelism needed to exploit future system architectures. The machine is located in the JICS building machine room and is administered by NICS staff.

Thunderhead

NICS deployed a cloud testbed system in the Spring of 2014. This system, named Thunderhead, has 20 compute nodes, 5 storage nodes, and 4 service nodes. In total, the system has 40 Intel Sandy Bridge processors, 1.28 TB of memory, 20 TB of node-local storage, and 40 TB of shared storage. The system will include both 10 Gigabit Ethernet and quad-data-rate (QDR) InfiniBand networks; the 10 Gigabit Ethernet network will be available for traditional cloud applications, while the InfiniBand network will be available for more experimental cloud applications.

Nautilus

The UT Center for Remote Data Analysis and Visualization (RDAV) was a NICS activity focusing on data analysis, visualization, and data-intensive computing for the NSF community. RDAV was sponsored by NSF through a 4-year, TeraGrid eXtreme Digital award. The centerpiece hardware resource at RDAV is Nautilus, an SGI Altix UV1000 shared-memory platform featuring 1,024 Intel Nehalem processor cores and 4 terabytes of memory within a single system image. During the span of the RDAV NSF award, Nautilus was a TeraGrid/XSEDE resource, providing allocations to users through the XRAC. During its lifetime as a TeraGrid/XSEDE resource, Nautilus provided 10.7 million CPU hours to the NSF user community and provided service to over 1,500 users. After the end of the RDAV NSF award in September of 2013, NICS has continued to operate Nautilus as a UT resource. The machine is located in the OLCF machine room and is administered by NICS staff.

Also included in the Nautilus ecosystem are the Harpoon nodes, four SGI UV10 systems with 32 cores and 128 gigabytes of memory each. These nodes were acquired to be an area where inexperienced users could prototype their shared-memory visualization and analysis applications before moving them to the UV1000. As with Nautilus, the Harpoon nodes are located in the OLCF machine room and are administered by NICS staff.

Mars

NICS obtained a Cray hybrid XE6m/XK6m machine from a trade-in of the Cray XT4 (first instantiation of Kraken). This system, called Mars, has 16 XK compute nodes, each with 16 GB of memory, a 16-core AMD 2.2 GHz Opteron processor, and an NVIDIA X2090 GPGPU with 6 GB of memory. It also has 20 XE nodes with 32 GB of memory and two 16 core AMD 2.2 GHz Opteron processors; all nodes are connected via a Gemini interconnect. This machine is being used to test experimental GPGPU codes and

the Gemini interconnect. The machine is located in the JICS building machine room and is administered by NICS staff.

Keeneland

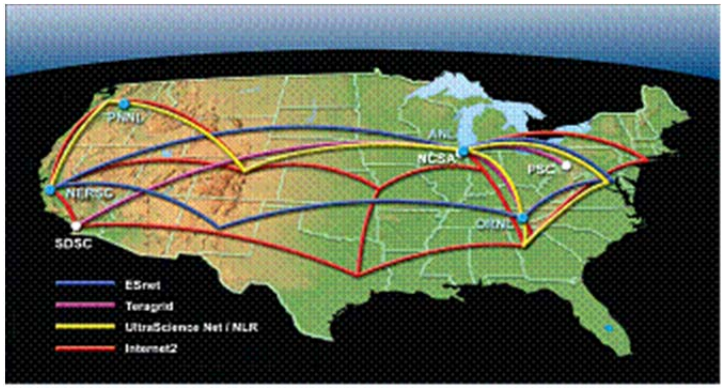
The Georgia Institute of Technology (GaTech) Keeneland Project is a 5-year, Track 2D grant awarded by NSF for the deployment of an experimental high-performance system. In 2010, the GaTech and its project partners, UTK and ORNL acquired and deployed the Keeneland Initial Delivery System (KIDS), a 201 Teraflop, 120-node HP SL390 system with 240 Intel Xeon CPUs and 360 NVIDIA Fermi graphics processors, with the nodes connected by an InfiniBand QDR network. KIDS was used to develop programming tools and libraries to ensure that important scientific and engineering applications can be productively accelerated on GPGPU architectures.

In 2012, the Keeneland Full Scale system (KFS) was accepted by the NSF and went into production. KFS is a 264-node cluster based on HP SL250 servers. Each node has 32 GB of host memory, two Intel Sandy Bridge CPU's, three NVIDIA M2090 GPUs, and a Mellanox FDR InfiniBand interconnect. The total peak double precision performance is ~615 TF. Keeneland is an XSEDE service provider, providing allocations to users through the XRAC. Both KIDS and KFS are located in the OLCF machine room and are administered by NICS staff.

2. Infrastructure

Physical and Cyber Security. ORNL has a comprehensive physical security strategy including fenced perimeters, patrolled facilities, and authorization checks for physical access. An integrated cyber security plan encompasses all aspects of computing. Cyber security plans are risk-based. Separate systems of differing security requirements allow the appropriate level of protection for each system, while not hindering the science needs of the projects.

Network Connectivity. The ORNL campus is connected to every major research network at rates of between 10 GB/s and 100 GB/s. Connectivity to these networks is provided via optical networking equipment owned and operated by UT-Battelle that runs over leased fiber-optic cable. This equipment has the capability of simultaneously carrying either 192 10-GB/s circuits or 96 40-GB/s circuits and connects the OLCF to major networking hubs in Atlanta and Chicago. Currently, 16 of the 10 GB circuits are committed to various purposes, allowing for virtually unlimited expansion of the networking capability. The connections into ORNL provide access to research and education networks including ESnet, XSEDE, and Internet2. To meet the increasingly demanding needs of data transfers between major facilities, ORNL participated in the Advanced Networking Initiative that provides a native 100 GB optical network fabric that includes ORNL, Argonne National Laboratory, Lawrence Berkeley National Laboratory, and other facilities in the northeast. This 100G fabric became the production network in December 2013.



ORNL network connectivity to university, national laboratory, and industry partners.

The EVEREST laboratory has been upgraded with dual power walls and 3-D capability.

The local-area network is a common physical infrastructure that supports separate logical networks, each with varying levels of security and performance. Each of these networks is protected from the outside world and from each other with access control lists and network intrusion detection. Line rate connectivity is provided between the networks and to the outside world via redundant paths and switching fabrics. A tiered security structure is designed into the network to mitigate many attacks and to contain others.

Visualization and Collaboration. ORNL has state-of-the-art visualization facilities that can be used on site or accessed remotely.

ORNL's **E**xploratory **V**isualization **E**nvironment for **RE**search in **S**cience and **T**echnology (EVEREST) facility is a scientific laboratory deployed and managed by the Oak Ridge Leadership Computing Facility (OLCF). The primary mission of this laboratory is to provide tools to be leveraged by scientists for analysis and visualization of simulation data generated on the OLCF supercomputers.

Three computing systems are currently provided in the laboratory. These consist of a distributed memory Linux cluster, a shared memory Linux node, and a shared memory Windows node. Access to the Linux computing resources requires an EVEREST account and an RSA Secure ID. Access to the Windows computing resources requires a standard ORNL UCAMS account and does not require a specific EVEREST account.

Two tiled display walls are provided. The primary display wall spans 30.5' x 8.5' and consists of 18 1920×1080 stereoscopic Barco projection displays arranged in a 6 x 3 configuration. The secondary display wall consists 16 1920×1080 Planar displays arrange in a 4 x 4 configuration providing a standard 16:9 aspect ratio.

There are four additional peripheral video inputs located on pop-out boxes in the conference table. Each input supports both digital DVI and analog VGA. Users of the laboratory are welcome to control either wall using personal hardware that is brought into the laboratory. Power outlets are provided at the conference table.

The laboratory instruments are controlled using a touch panel interface located at the control desk. All computing resources can be routed to any available display wall. User hardware using the video input ports on the conference table can also be routed via the touch panel.

High Performance Storage and Archival Systems. To meet the needs of ORNL's diverse computational platforms, a shared parallel file system capable of meeting the performance and scalability requirements of these platforms has been successfully deployed. This shared file system, based on Lustre, Data Direct Networks (DDN), and InfiniBand technologies, is known as Spider and provides centralized access to petascale datasets from all major on-site computational platforms. Delivering more than 1 TB/s of aggregate performance, scalability to more than 20,000 file system clients, and 30-petabyte (PB) storage capacity, Spider is one of the world's largest scale Lustre file system. Spider consists of 36 DDN SFA12KX storage arrays managing 20,160 2-TB Nearline-SAS drives served by 288 Dell dual-socket, quad-core I/O servers Metadata are stored on a NetApp E5500 storage array and are served by five Dell dual-socket, six-core systems with an aggregate of over 1 Terabyte of memory. ORNL systems are interconnected to Spider II via an

OLCF tape archive.

InfiniBand system area network, which consists of 3 Mellanox SX6506 Director Class IB switches, and 36 Mellanox SX6036 IB switches; with more than 3 miles of optical cables. Archival data are stored on the center's High Performance Storage System (HPSS), developed and operated by ORNL. HPSS is capable of archiving hundreds of petabytes of data and can be accessed by all major leadership computing platforms. Incoming data are written to disk and later migrated to tape for long term archiving. This hierarchical infrastructure provides high-performance data transfers while leveraging cost effective tape technologies. Robotic tape libraries provide tape storage. The center has six SL8500 tape libraries holding up to 10,000 cartridges. The libraries house a total of 24 T10K-A tape drives (500 GB cartridges, uncompressed), 32 T-10K-B tapes drives (1 terabyte cartridges, uncompressed), 64 T-10K-C tape drives (4 terabyte cartridges, uncompressed), and 32 T-10K-D tape drives (8 terabyte cartridges, uncompressed). Each drive delivers a bandwidth in excess of 120 MB/s. ORNL's HPSS disk storage is provided by DDN and NetApp storage arrays with nearly 2 petabytes of capacity and over 20 GB/s of bandwidth. This infrastructure has allowed the archival system to scale to meet increasingly demanding capacity and bandwidth requirements with more than 39 PB of data stored as of November 2014.