

BUYER CASE STUDY

Texas Advanced Computing Center (TACC) and Sun Pioneer HPC Best Practices at Near Petascale

Richard Walsh
Steve Conway

Jie Wu
Earl C. Joseph, Ph.D.

IDC OPINION

The exponential growth in the capability of the world's most powerful high-performance computing (HPC) systems as tracked at the Top 500 Supercomputers Web site continues unabated. September 2008 marked the breaking of another compute capability barrier with the installation of the world's first system capable of delivering a Linpack petaflop, 1,000 trillion double-precision floating-point operations per second, at Los Alamos National Lab (LANL). Petaflop and near-petaflop capable systems are unprecedentedly complicated, often with more processors (>100,000 in some cases) to apply to their workloads than the population of a midsize city. Defining best practices for the reliable and serviceable operation of these systems is a work in progress today at a handful of sites with unique buyer-vendor partnerships. In IDC's opinion, interesting lessons can be learned by examining the experiences to date at some of these early petascale and near-petascale sites. In this first of several buyer case studies on petascale HPC datacenters, IDC visited the staff and users at the Texas Advanced Computing Center (TACC) and at TACC's system vendor partner, Sun Microsystems (Sun), to understand TACC's approach to running its near-petascale system, nicknamed Ranger, the first of the new National Science Foundation (NSF) "Path to Petascale" systems. TACC is funded by the University of Texas at Austin and several other external agencies, including the National Science Foundation. In 2006, TACC, Sun, and AMD, together, won NSF Track 2 funding for a \$30 million Sun Constellation System based on AMD quad-core processors and an additional \$29 million to run and support the system for four years. TACC's academic partners include the University of Texas Institute for Computational Engineering and Sciences, Arizona State University, and Cornell University. In production since February 2008, Ranger is the most powerful resource on the NSF TeraGrid and is among the most powerful HPC systems in the world (ranked number 4 as of June 2008), with 62,976 AMD processors (cores) and a peak performance of more than 579.4TFLOPS. Key points of interest on Ranger and its operation at near-petascale include its:

- Power, space, and cooling requirements and datacenter integration
- Cluster architecture, as it relates to both performance and reliability
- Support for storage and visualization systems
- Operating system and HPC management software
- Mix and scale of workloads, utilization, operating practices, and policies
- Early scientific achievements and the challenges its researchers face

IN THIS BUYER CASE STUDY

This IDC Buyer Case Study first presents an organizational overview of Texas Advanced Computing Center (TACC), including the size of the center and its available resources and objectives. We then focus on TACC's recent acquisition of Ranger, a near-petascale Sun Constellation System. IDC looks at how TACC and Sun have approached the challenges that have arisen on several fronts in deploying an HPC system capable of more than 500 teraflops for use in TACC's "open science" environment, which supports NSF and other researchers from across the country. IDC presents its findings to provide insight and best practice messages to those planning large or petascale HPC installations of their own.

SITUATION OVERVIEW

Organization Overview

TACC is an advanced computing center at the University of Texas at Austin, chartered with enabling discoveries that advance science and society through the application of advanced computing technology. TACC offers its resources (including technical documentation, training, and support) to faculty, students, and staff throughout the University of Texas system. As an NSF Track 2 research center, TACC's resources are also available to any researcher at a U.S. institution who submits a successful proposal describing the nature of his/her scientific work, explaining why a large HPC resource is required, and demonstrating his/her ability to use it. Allocations are made quarterly. Once an allocation is made, access to TACC from around the country is facilitated through the NSF TeraGrid project. Ranger system is currently the most powerful HPC resource on the NSF TeraGrid. TACC also offers 5% of its available resources to support industrial research and development and another 5% (amounting to 25 million CPU hours) to researchers from any institution of higher education in the State of Texas.

In addition to providing advanced computing infrastructure and support, TACC staff conduct research to develop new computing techniques and technologies. Current areas of research in computing technology include:

- Evaluating and modeling the performance characteristics of HPC systems and of algorithms and codes on those systems
- Exploring the impact of large displays and immersive techniques on data analysis and knowledge discovery
- Developing new visualization tools for collaborative and remote visualization
- Building reliable high-performance commodity clusters for HPC simulations and scientific visualization (SciVis)
- Developing computational grid software to seamlessly integrate scientific compute, visualization, and data storage systems and to link these systems with systems at partner institutions

In service to its mission, TACC currently provides the following advanced computing resources:

- ☒ High-performance computing server systems
- ☒ Advanced scientific visualization resources
- ☒ Massive data storage and archiving systems
- ☒ High-speed interconnecting networks
- ☒ HPC and SciVis applications, libraries, and tools

TACC's Ranger Sun Constellation System is one of the latest and most powerful high-performance computing resources in the world and is described in detail below. In addition to the Sun Constellation System Ranger, the focus of this study, TACC also supports the following HPC compute systems:

- ☒ **Lonestar:** A Dell-Linux cluster that contains 5,840 Intel Xeon cores (2.66GHz), 24 Dell PowerEdge 1850 compute-I/O server nodes, and 2 Dell PowerEdge 2950 login/management nodes. Storage includes a globally accessible 103TB shared, Lustre parallel file system and an aggregate of 106TB of local storage. All compute blades have 8GB of local memory and are connected via an SDR IB network fabric.
- ☒ **Stampede:** A Dell-Linux cluster consisting of 217 compute nodes, two login nodes, and a dedicated file server attached to one of the login nodes. Each compute node has two quad-core Intel Clovertown processors, 8GB of memory, and 600GB of local disk space (of which 520GB is available to the user) and 536GB of shared disk. Stampede's interconnect is based on Gigabit Ethernet technology, and every node has the Lonestar parallel file system mounted on it.

Challenges and Solutions

Petascale Space, Power, and Cooling Demand Dense Designs

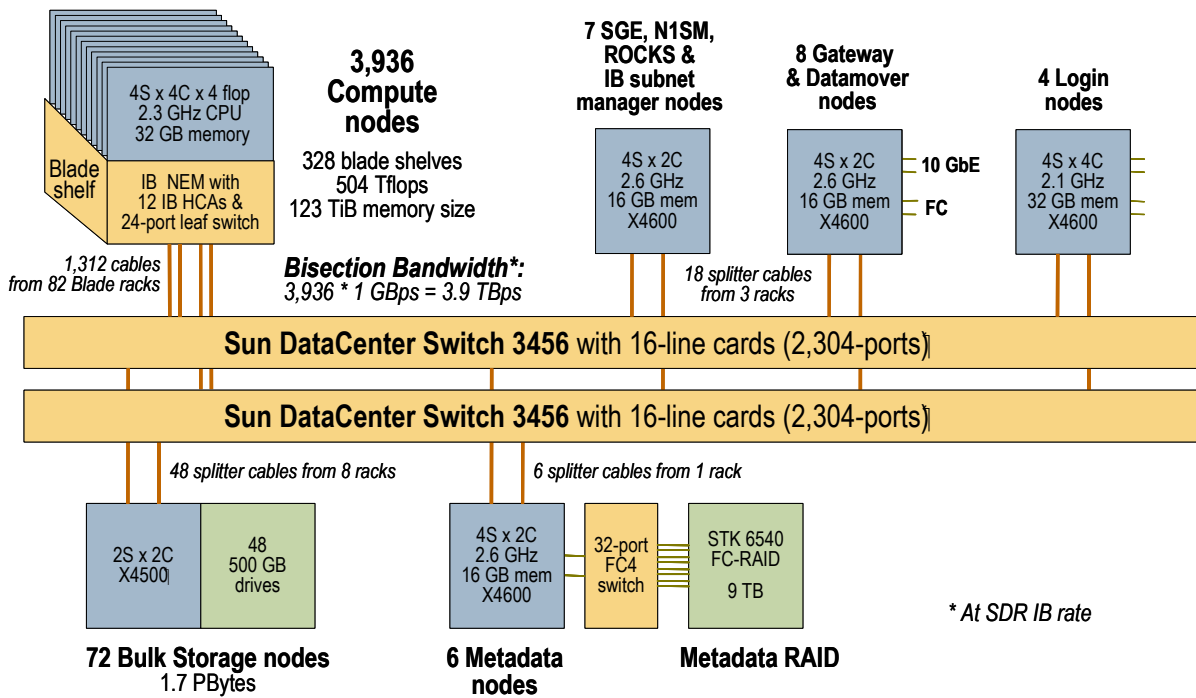
While TACC had just completed constructing a new datacenter for Ranger, TACC knew that compared with national lab sites such as Sandia, Los Alamos, and Oak Ridge it had less power, space, and cooling capacity. To provide a near-petascale system at its facility, TACC had to work within site space, power, and cooling budgets and, in particular, limit the amount of floor space consumed by Ranger to approximately 2,000 sq ft (about half a basketball court). As such, system density and power efficiency were essential elements in the review and selection process.

TACC chose to bid Sun's Constellation System, which offered the highest performance within TACC's site constraints at the time. The Sun Constellation System's density features begin with AMD's quad-core Opteron processors, Sun's quad-socket series 6000 blade, and Sun's Blade 6048 chassis that can hold up to 48 blades. These fully populated racks hold 768 Opteron cores (4 x 4 = 16 per blade) and with 82 server racks in total, Ranger provides 62,976 AMD Opteron cores each running at 2.3GHz. Each blade also provides 32GB of memory (2GB per core), giving the system about 125TB of

memory and making Ranger one of the most memory-rich systems in the world. Ranger's nodes perform well on standard synthetic benchmarks like STREAM (21GBps) and DGEMM (86% of peak). Ranger's core- and memory-dense Sun 6000 blades offer users the flexibility to run larger memory jobs by idling some cores on the node and also to run mixed parallel programming model work (OpenMP and MPI). Figure 1 provides a schematic of the basic system architecture of Ranger.

FIGURE 1

TACC's Ranger, a Sun Constellation System



Source: Sun, 2008

Ranger's high compute and memory density drive high per rack power requirements. Sun's 6048 chasses were designed for this and are rated for up to 28kW of power, roughly double that of more conventional racks. While in operation, Ranger consumes roughly 2.4MW of power with an additional 1.0MW drawn to support its cooling and storage systems. While meeting TACC's spatial and compute density requirements was a challenge, meeting its power distribution requirements was still more difficult. Twenty-one power distribution units (PDUs) are needed to supply Ranger with power, and they effectively double the floor space that Ranger occupies to approximately 4,000 sq ft. Connecting Ranger to its PDUs under floor in such a small space required more than 1,300 circuits. Sun's 6000 series blade and chassis was clearly engineered to support petascale computing, but the special requirements of systems of this size push themselves out into the datacenter. More dense and/or rack-row-integrated PDU

designs perhaps represent an issue-opportunity pair for enterprising vendors to improve best practices for petascale systems expected to be installed over the next decade.

Ranger's cooling system is another well-conceived aspect of TACC's dense HPC solution. TACC had just finished building a new datacenter, and therefore had what would normally be considered ample computer room air-conditioning (CRAC) resources, but its NSF Track 2 win and Ranger brought dramatically higher cooling requirements. A more efficient cooling system than one based on standard ambient air techniques was required. TACC decided on a multifunctional approach for Ranger that includes support from its CRAC units, enclosed hot aisles and, most importantly, the InRow RC cooling units from American Power Conversion (APC). Ranger's APC units stand side by side with Sun's high-density 28kW racks as shown in grey in Figure 2. The InRow cooling units draw hot air from enclosed hot aisles on the rack row's backside and return cool air to a cold aisle at the front of the rack row. The heat captured is expelled from the datacenter via a chilled water circuit that is part of the InRow units. The remaining ambient heat is handled by TACC's CRAC units. It is no surprise that traditional ambient-air-only, CRAC-based, cooling systems are not adequate for many petascale and near-petascale installations. Vendors supplying systems at the high end, including Sun, Cray, IBM, SGI, and HP, have all introduced cooling technologies that take advantage of the higher heat exchange efficiency of liquids.

Petascale systems will put extraordinary demands on the basic inputs of power, space, and cooling in the datacenter. As a consequence, datacenter upgrade costs will push their way into the buyer's total cost of ownership (TCO) calculations. The incentive to avoid steep, datacenter-driven climbs in TCO will be very high and further increase the pressure to make HPC systems more efficient and more dense.

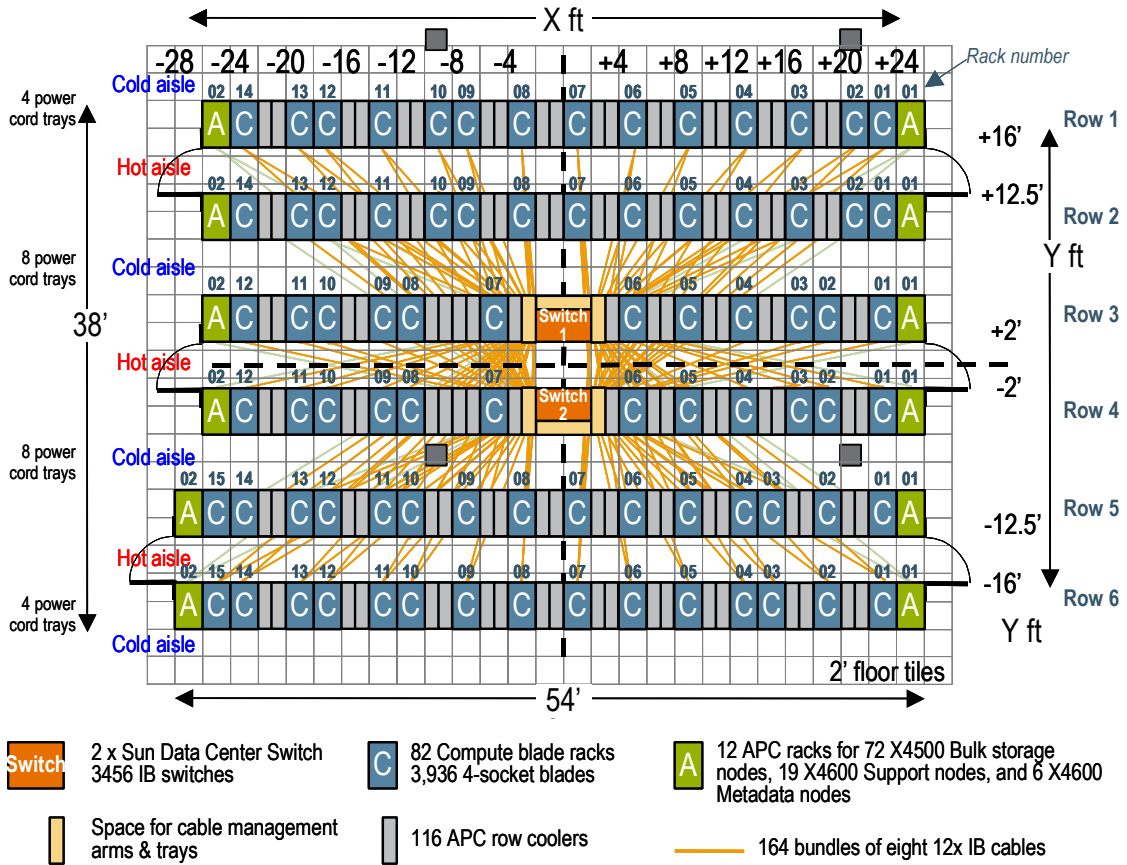
TACC's Petascale Density Best Practice Messages

IDC sees several density-related, petascale best practice messages in the Ranger implementation of Sun's Constellation System at TACC:

- ☒ Node density, through higher socket and core counts, will be used to avoid datacenter upgrade costs.
- ☒ Fat petascale nodes require more memory per node and the best available intranode, memory bandwidth.
- ☒ Memory- and processor-dense nodes will force power and cooling delivery systems to become more efficient and denser. The use of liquid cooling, and rack and row, integrated and consolidated, power and cooling delivery systems will increase.

FIGURE 2

TACC Ranger Floor Plan



Note: The Metadata RAID is located in an additional rack to the lower-right of the system

Source: Sun, 2008

Petascaling Ranger's Interconnect

Special mention is reserved here for the important role played by the interconnect technology Sun developed to support Ranger or any petascale Sun Constellation System. Petascale systems such as Ranger, even when constructed from so-called fat nodes, must manage complicated switched or meshed interconnect architectures. Switched IB interconnects at this scale, such as Ranger's, must be multitiered (two or more) because until recently IB-capable switch chips were limited to 24 ports. As a consequence, rack-mountable switches have been limited to $12 \times 24 = 288$ ports, which means IB interconnects supporting petascale HPC clusters require at least 12, 288-port spine switches and 288, 24-port leaf switches. Such a two-tiered architecture could interconnect 3,456 nodes in a non-blocking fat tree, but racking and cabling this design is very labor intensive, consumes space on the backplane, is hard to troubleshoot, and is severely limited by the bend radius and length restrictions of typical off-the-shelf CX4, SDR-capable cabling. In TACC's case, such an interconnect architecture was unacceptable.

To meet TACC's requirements and those of very large HPC systems generally, Sun condensed the typical IB architecture outlined above at every tier. Refer back to Figure 1, in the center of which two of Sun's DataCenter 3,456 Switches (12 x 288 = 3,456 ports) are visible. They each condense an entire spine (12, 288-port switches) described above into a single chassis. TACC uses two switches for failover and to accommodate a total of $82 \times 48 = 3,936$ nodes, which is slightly more than a single switch's 3,456 maximum port capacity. Further consolidation is designed into Sun's cables, which combine the equivalent of three 4x SDR IB cables into one. The cables split out for Sun's rackmounted leaf switches, called network express modules (NEMs). The cable count is reduced by a factor of three and the occupied backplane by a factor of six because Sun's 12x cable is 50% smaller in diameter than the three it replaces. Finally, the NEMs (four per rack) are 24-port, blade-dense, leaf switches augmented with 12 Gigabit Ethernet ports. In combination, these three Sun interconnect components (the DataCenter Switch, the 12x cable, and NEM) custom engineered to meet SDR IB standards are essential to Ranger's dense architecture that provides TACC with near-petascale capability in about 2,000 sq ft.

TACC's Petascale Interconnect Best Practice Messages

The petascale, best practice messages to be drawn from Ranger's Sun Constellation System interconnect experience include:

- ☒ Interconnects for petascale systems can be built around standard protocols, but general-purpose interconnect components (cables and leaf and spine switches) will need to advance before they can be comfortably used at petascale.
- ☒ Switch chips with higher port counts (36-port IB silicon switches have been announced), long-range flexible cabling with suitable connectors and smaller backplane footprints (such as optical cabling available from Intel and Luxtera), and port- and space-dense switching technologies will be needed to avoid heroic measures to interconnect petascale systems.
- ☒ Petascale systems will demand interconnects with maximum bandwidth to service the backplanes of fat petascale nodes with more memory and higher processor counts, and as accelerators find wider use.

Petascale Approaches to Storage and Visualization at TACC

It is clearer than ever that to produce scientific results quickly on HPC systems of the size of Ranger, the stages preceding and following the compute piece of the HPC work cycle must meet certain minimal performance standards or they will become bottlenecks, reducing productivity significantly. Just as individual HPC servers with balanced designs deliver better sustained performance, datacenters with petascale systems must have a design balanced to process petascale inputs and outputs. Petascale datacenter storage and visualization capability should match its petascale server capability.

Ranger's Storage System

TACC's partnership with Sun included provisions for storage and visualization capabilities designed to serve Ranger. Ranger's storage system includes node-local disks totaling 31TB for provisioning, a node-shared Lustre file system of 1.7PB on 72 Sun Fire X4500 storage servers (i.e., Thumpers), and a back-end, archival storage

system (called Ranch) on Sun StorageTek hardware running SAM-FS to manage its disk-to-tape hierarchy. The back-end tape library system offers up to 10PB of capacity. The workhorse supporting Ranger's computation is the shared Lustre file system. TACC staff noted for IDC that Ranger's I/O capability matches Ranger's compute capability much better than on their other systems. Lustre runs natively over Ranger's IB interconnect, split across its two Sun DataCenter 3,456 Switches. Each of its 72 Thumper storage servers houses 48, 500GB SATA drives. Lustre data rates have been measured at 700MBps for client writes and 45GBps in aggregate. This architecture is expected to hold 500 million files and 500TB at steady state. TACC plans to add storage policies to its scratch partition as needed, but to date has set only maximum file sizes of 10TB. Quotas and file size limits are enforced on its smaller, transactional home and work partitions. At this stage in Ranger's climb to steady-state operation, its file system is meeting its requirements, but more time will be required to define figures of merit for petascale storage systems from the operations at TACC.

TACC's Visualization Capabilities

Sun has worked with TACC to bring remote visualization to its TeraGrid researchers wherever they are working. TACC's recently installed Spur visualization cluster, powered by Sun Fire x64 servers and NVIDIA GPUs, allows TACC researchers to interact with and visualize their results from their desktop or laptop, regardless of their location. Spur utilizes Sun's Scalable Visualization System 1.1 software, which gives users the ability to view their results from their simulations without having to copy large amounts of data across the network.

Previously, TACC worked with Sun to prototype this technology on Maverick, a visualization system based on Sun Fire E25K technology. On Maverick, valuable lessons were learned, and Sun's Visualization System software was tuned to meet TACC's needs. It was installed and ready to run when Spur went live on October 13, 2008. Spur consists of eight fat nodes (one Sun x4600 and seven Sun x4440 servers) with a total of 1TB of memory and 128 cores. Sun's Scalable Visualization software allows Spur to render large data sets in parallel across the entire system. Spur's visualization nodes are clustered with Ranger's compute nodes via Ranger's two Sun Datacenter 3,456 Switches. Spur also includes eight NVIDIA Quadro Plex GPUs. Spur's shared visualization capability can be reserved like other Ranger resources using Sun Grid Engine's (SGE's) advanced reservation system.

Supported by collaborative grants from Dell, Intel, Microsoft, and Cisco, TACC has also recently completed a transformation of its 2,900 sq ft ACES visualization lab, which (along with the Spur installation above) has expanded its visualization capabilities dramatically. ACES includes:

- ☒ State-of-the-art graphics workstations
- ☒ Stallion, a 24-node Dell visualization cluster with 47 NVIDIA graphics cards, powering a 20.5 x 7.5ft, tiled display consisting of 45 Dell 30in. flat-panel monitors that provide 184 million pixel resolution
- ☒ A Sony Projection System with a flat screen, 20in. x 11in. display (4096 x 2160 resolution) driven by a Sony SRX-S105 overhead projector and a high-end Dell workstation.

- ☒ A large suite of rendering software (Paraview, AVS, Ensight, Chromium, etc.)

Stallion makes use of many standards-based components and is responsible in part for limiting the cost of the visualization center. The final stage of almost all modern HPC computing work cycles is processing and imaging the results. On petascale systems with terabytes of memory and petabytes of storage, quickly converting results files to the high-resolution, visual images required to draw conclusions and determine the next steps in an HPC research program can be the rating limiting step if the visualization resource is not balanced to match the computing resource.

TACC's Petascale Storage and Visualization Best Practice Messages

TACC has understood that a petascale HPC resource is a pipeline that is incomplete without balanced high-performance storage and visualization systems at both its front and back ends. Taking advantage of collaborative partnerships and a supportive funding environment, TACC has assembled a resource to meet its results-oriented productivity goals that will serve as a benchmark for others. The storage and visualization best practices messages TACC offers include:

- ☒ High-speed, parallel, network-attached storage and visualization clusters matched to the input and output requirements of petascale servers yield a balanced and consequently more productive resource.
- ☒ Remote, zero-copy visualization capability provided by the right combination of visualization software and clustered, network-attached visualization hardware is an essential element in petascale system productivity.
- ☒ Beyond size and speed, petascale storage systems must meet unique uptime and reliability requirements sustained through underlying capabilities of the hardware and file system software.

Scaling Open Source Software to Petascale: Can It Be Done?

As TACC's management and support team considered what Ranger's software environment should look like, they were attracted to the challenge of designing a software stack based on open source components that could run and be maintained at petascale, a notion sometimes challenged by those selling proprietary HPC system software. This goal also conforms to the "open science" vision at TACC in that a system whose operating software is based on an open source model should stimulate and support open science in the long term. TACC was also looking for those that understood its vision for petascale computing and would function more as partners than simply as vendors.

Ranger's open source HPC software stack starts with the Community Enterprise Operating System (CentOS 4.4, Linux kernel 2.6.9) version of Linux, which is based on Red Hat Enterprise Linux. This freely available Linux is generally thought to track Red Hat Enterprise Linux most closely, has been around since 2003, and has adopted and integrated several other smaller Linux OSs with the same mission. The use of CentOS on Ranger at this scale was a first. TACC's system programming and administration team has had to work through a number of scale-related issues and continues to tune its CentOS configuration on Ranger to optimize performance. A key

requirement is to minimize the size of the kernel on the compute nodes and minimize the frequency of interrupts from the remaining services necessary for compute node operation. Suboptimal operation here will generate skew-related load imbalance and potentially limit peta-scalability. TACC has chosen widely used cluster installation and management tools and augmented them for Ranger to provide a consistent and scalable stack that can serve its mix of HP compute and infrastructure requirements. TACC chose to use NPACI ROCKS, an open source tool, to perform an initial baseline provisioning of Ranger nodes. ROCKS generates node-specific Kickstart files dynamically from tree-based, XML-formatted, control files. This allows system administrators to build multiple role-specific node configurations (graphs) from software description files call rolls and nodes. Once the nodes were deployed, TACC's internally developed cluster management framework took over to handle all software management duties in a scalable fashion. To accommodate potentially frequent updates for OS components, development libraries (e.g., compilers, MPI stacks), third-party software, and more, TACC's framework adopts a "provision once, update often" approach, using package manager-derived utilities to keep individual components in sync.

TACC had intended to use Lustre (a high-performance, open source, multivendor, parallel file system) on Ranger before Sun's acquisition of CFS, Lustre's parent company, in October 2007. Sun's purchase added another key piece of Ranger's software to its responsibilities at TACC. Already in use on some of the world's largest HPC systems and delivering world record performance (serving 100,000+ clients, multiple petabytes of storage, and 100GBps aggregate bandwidth), Lustre has needed little in the way of special configuration to meet TACC's near-petascale requirements. However, the purchase has put TACC in a position to both drive and benefit from Lustre software developments, including future support for improved recovery, Kerberos security, hierarchical storage management (HSM), and ZFS-based Lustre file servers. TACC is pleased with the performance of its Lustre-based file system, which delivers about 700MBps to a single node and up to 45GBps in aggregate across TACC's entire scratch file system.

TACC's working relationship with Sun in putting together and winning the Ranger bid led it to chose SGE, an open source workload manager written and supported by Sun, over the one it was already using and more familiar with. It was not an easy decision, but TACC negotiated with Sun for certain features it required, including advanced reservation and system health capability. Both of these features have now been added, and the TACC-Sun SGE partnership has led them to identify and fix a number of petascale-related performance issues within SGE that have been fed back into the source. These include adding and resetting daemon tuning parameters, minimizing large job initiation message traffic, and reducing SGE start-up time after reboot. TACC and Sun continue to work on very large job efficiency (>32,000 cores), to improve interactive job support, and to reduce SGE's memory and processor-time footprint. These petascale-related improvements are being fed back into the SGE source for the benefit of all.

For the user and runtime components of the software stack, TACC has again emphasized open source supporting MVAPICH and OpenMPI for MPI over Ranger's SDR IB interconnect, the GNU compilers, and a suite of open source libraries and tools (PETSc, FFTW, Kojak, PAPI, and so forth). There have been challenges here.

One has been working through MPI over IB software stack issues with their counterparts at the Open Fabrics Alliance and Ohio State. When the price has been right, and TACC has found a vendor with the time and resources to partner on development, TACC has been willing to augment its open source offerings with some commercial applications. It has a close relationship with Allinea, a vendor of HPC cluster debugging and profiling software (DDT and OPT), and it supports the PGI compilers as well as those from Intel.

TACC Petascale Software Best Practice Messages

The petascale, best practice software messages to be drawn from TACC's Ranger experience to date include:

- ☒ Configuring a petascale resource with a largely open source software stack appears possible when combined with a talented support staff.
- ☒ Petascale-related configuration problems and missing features are still to be expected, but open source software improvements stimulated at pioneer sites such as TACC will accrue to benefit future petascale sites.
- ☒ TACC's open source success has depended on its careful selection of software and strong partnerships to minimize, identify, and correct shortcomings. It has chosen mature community-supported components, such as ROCKS and CentOS Linux that come with strong reputations, and favored open source communities with many stakeholders such as the Open Fabrics Alliance (OpenMPI). It has also benefited from its unique partner relationship with Sun (SGE and Lustre).

Petascale Parallelism and Utilization Through Policy and Practice

TACC views itself as the world's largest open science HPC research center and Ranger as the largest "open science" system. It has not placed such a large resource at the disposal of such a varied user population without some careful thought about policies and practices that will ensure that it is used effectively, but has allowed its user community to experiment to discover what best practice means on a near-petascale system. The policies of TACC administrative staff and the practices of its user community have to gradually adjust to the sheer size of Ranger.

TACC grants medium (less than 1 million CPU hours) and large (greater than 1 million CPU hours) resource allocations on Ranger on a major-minor quarterly grant cycle through the NSF TeraGrid. Initially, grant applications were under-sized as applicants failed to grasp the size of the resource they were requesting time on. One can easily use 1 million hours in a month when running on 4,000 or more cores. Still, adjusting one's grant application and one's imagination to suit Ranger's size is far from all that TACC staff want to do. They also want to push utilization up into the high 80% or 90% range.

Striking a balance between the need to encourage users to scale up their jobs and the need to use the resource efficiently, TACC has instituted a policy whereby users wishing to run jobs on more than 4,000 cores must demonstrate that their code scales efficiently. Below that size, efficiency concerns are left to the user. TACC staff is available to assist researchers in improving their code's scaling efficiency. This policy

has begun to push up the average processor count on Ranger, but it has had another interesting, petascale-specific beneficial effect. Users have discovered the value of larger parametric studies. Ensembles of jobs, each of which scales well into the thousands of cores (a maximum on most other HPC systems) but not well past the 4,000-core efficiency test barrier on Ranger, are run with different initial conditions and the results interpolated. TACC's mission is to stimulate breakthrough science, whatever the approach. Its mixed user population and Ranger's 62,976 cores make it the right place to test the return on investment of large parametric studies, something likely to be of interest to other petascale HPC centers.

This somewhat novel development in parametric simulation has not stopped TACC from targeting petascale performance for its application mix. WRF, a notoriously difficult-to-scale weather code, is now running on 8,000 cores after some work on message aggregation, topology-aware job scheduling, and node level optimization. More detail is provided in the next section on TACC's early successes with petascale applications.

TACC intends to improve throughput and utilization on Ranger through better resource scheduling and a plan to avoid overallocating the Ranger resource. Making better use of an advanced reservation feature added to SGE by Sun is part of this. Currently, 100–200 jobs are typically running on Ranger at any given time with 100–200 in the queue. Jobs approaching 4,000 cores are being run routinely, but TACC would like to raise the core count average. Ranger's resource-rich, high-performance design (high memory bandwidth, large memory, fast interconnect, and fast storage) makes the scheduling problem easier, but the workload at TACC is mixed and includes requirements for interactive and debugging runs. Improving the scheduling of Ranger's mixed workloads is an ongoing effort.

TACC Petascale Operating Policy Best Practice Messages

While Ranger is still seeking an optimal production state, several best practice points on petascale operating policy can be drawn from TACC's experience with Ranger:

- ☒ Prepare your petascale user community to think in terms of the full capability of the resource before it arrives to ensure a quick ramp-up.
- ☒ Define an efficiency requirement and assessment protocol for applications intended to run above a certain number of cores, and offer training, support, and encouragement to those users targeting petascale performance.
- ☒ Understand the potential for success of larger ensemble or parametric studies across your applications' mix. Prepare users with hard-to-scale applications to make use of this approach.
- ☒ Filling and efficiently scheduling a petascale resource will demand a lot from your workload manager, potentially including custom parameter tuning and even code modifications. Chose your workload manager with this in mind.

Getting Petascale Systems to Produce Peta-Science

Ranger is already running applications from a variety of disciplines, including geoscience, CFD/CAE, and weather forecasting, among others. TACC is eager to build up its collection of applications that make efficient use of Ranger at extreme scale.

One of the early users of Ranger is Dr. Clint Dawson, a professor at the University of Texas at Austin, whose research focus is hurricane forecasting and storm surge prediction. Dawson's application had run reasonably well on TACC's Lonestar system using 512 cores. Initial runs on Ranger at 1,024 and 2,048 cores revealed bugs that had not been detected at smaller scale. Adjustments to the solver and the size of the time step helped resolve them. Ranger's large memory and faster interconnect have allowed Dawson's group to refine their applications mesh and use larger data sets. Dawson's weather code is now running on 8,000 cores, with higher core counts targeted. Current simulations use 10 million to 20 million elements and have 30 million to 60 million degrees of freedom, but how has the scaled up application changed Dawson's results? Dawson can now complete storm surge simulations that used to take 3–6 hours in 20–30 minutes. To run at higher core counts and reduce turnaround further, work is being done with dynamic meshing to improve load balancing

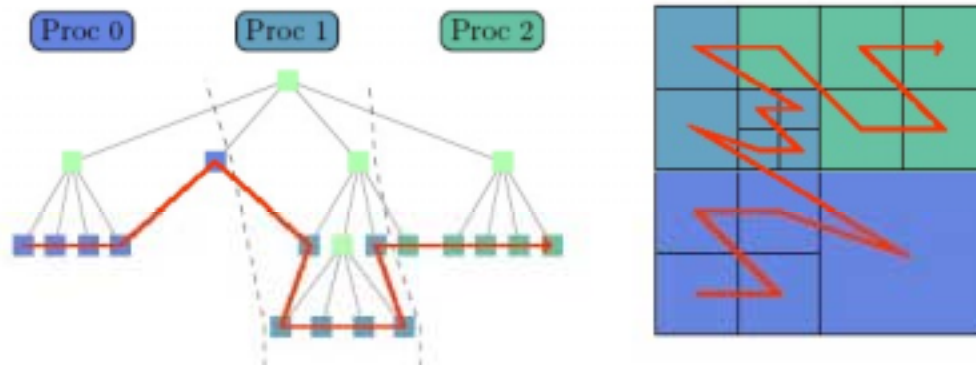
Dr. Omar Ghattas, director of the Center for Computational Geosciences and professor of Geological Sciences and Mechanical Engineering at the University of Texas at Austin, is targeting even more extreme scaling on Ranger. An expert in modeling earth mantle flow and convection, Ghattas and his team are developing an application designed to scale to 10^4 to 10^5 cores. Ghattas' group has chosen a problem with the data, algorithm, scaling, and practical application features that will test every dimension of the petascale science and computing problem. The application, which has both strong and weak scaling data modes, includes solvers with multiscale and multiphysics requirements, requires the use of Adaptive Mesh Refinement (AMR) and load balancing, demands error quantification code, and has some very practical applications — the computational prediction of earthquakes. The application is designed to create subsurface regional geophysical models for predictive purposes using the mathematics of inverse methods. In this mathematical technique, observed data (in this case, actual earthquakes) is collected and characterized. This data is then combined with parameterized, but theoretically accurate, geophysical equations to produce a model of the earthquake regional subsurface. It is a macro-variation on seismic modeling of earth subsurfaces to discover oil.

Ghattas reports some initial success in petascaling the AMR piece of his code to 32,000 cores on Ranger using an octree-forest algorithm developed by his team (see Figure 3). The simulation-driven movement of fine and coarse grids makes load imbalance a significant problem in this application. Its large mesh must be quickly reconstructed and rebalanced every few time steps. This requirement makes good use of the large memory, superior bandwidth, and low latency within and among the nodes on the Ranger system. Working on a system with the high-performance features of the Sun Constellation System eases the parallel programming burden

faced by Ghattas and other researchers using Ranger while making the idea of petascale science possible.

FIGURE 3

Adaptive Mesh Refinement Using Parallel Octrees



Source: Ghattas, 2008

TACC Peta-Science Best Practice Messages

Even in the short time that Ranger has been available, TACC researchers have begun to climb the barriers between where HPC is today and the future regime of sustained petascale computing and science. Some best practice messages to be taken from their experiences include:

- ☒ Researchers expecting to be among the first to obtain petascale performance and scientific results must be willing to accept the challenge of redesigning their applications and algorithms for these systems.
- ☒ Petascale computing and science will more often than not require multiscale, multiphysics, and adaptive mesh refinement techniques, each of which will need to petascale.
- ☒ Those HPC systems to first produce petascale results will offer better-than-baseline performance along multiple dimensions of the HPC system performance envelop. Large core counts by themselves will not be enough.

FUTURE OUTLOOK

The engine propelling the advancement of high-performance and technical computing continues to be the remarkable scientific results it delivers. As the Nobel Prize winning physicist Dirac noted more than 75 years ago:

The laws of physics and chemistry are now largely known. The difficulty arises in their application, which leads to equations too complex to be solved.

The whole of computational science pays homage to this fact. Each technical advance in HPC brings the solution of more complex problems into scope, and so it is with HPC's new petascale systems. TACC, in collaboration with Sun, AMD, and its academic partners, is pioneering the next wave of scientific problem solving at the leading edge of HPC, and, through its experience to date, offers those that follow something to benchmark their own plans against.

The difficulties and complexities of running an economically sustainable HPC resource at petascale are daunting but can be overcome with the support of large investments from national governments, industry, and the HPC user community over time. The interest and stakes in meeting the challenge are high as humanity struggles to sustain itself, improve the lives of its diverse populations, and satisfy its native curiosity about the world it finds itself in.

IDC expects the next several years to see multiple additional petascale and near-petascale HPC installations around the world. IDC encourages those planning these future installations to closely examine the practices (best and otherwise) of sites preceding them to refine their plans. IDC views TACC as a good place to start because of the diverse user population it serves, its growing experience as a path-to-petascale pioneer, and how it has challenged itself to work within local space, power, and software budgets.

ESSENTIAL GUIDANCE

IDC continues to watch petascale initiatives across the globe. Ranger's deployment marks another milestone in the development of next-generation, large-scale, high-performance, and technical computer systems. As plans for petascale system installations continue to come together, IDC's analysis of TACC's near-petascale operation offers the HPC community the following guidance:

- ☒ Benchmark your plans against preexisting petascale and near-petascale installations.
- ☒ Inventory the scientific results being targeted, the research teams available to the site required to achieve those results, and the application scaling challenges faced in each discipline by each team before budgeting for a petascale system.
- ☒ Determine what can be done with current technology within the constraints of the current datacenter to develop an incremental growth plan and avoid or alleviate the costs associated with building a new datacenter.
- ☒ Consortia planning petascale installations should look closely at siting decisions to minimize operating costs.

- ☒ Architect your system with your target applications and datacenter in mind; design in enough density and custom high-performance capability to minimize operating costs and petascale application core counts.
- ☒ Build up your expertise on liquid cooling alternatives and dense power distribution systems.
- ☒ Interconnects based on standard protocols may prove adequate, but they will require dense components using optical fiber, small connectors, and high radix switch chips that maximize backplane bandwidth.
- ☒ Producing scientific results at petascale sites will require a production pipeline that includes suitably sized and balanced network-attached storage and visualization cluster systems in addition to petascale HPC server systems.
- ☒ In designing the software stack, consider carefully what is possible with open source software as petascale-driven developments in open source begin to accrue to the benefit of future petascale system buyers.
- ☒ Key software components such as the workload manager, parallel software APIs, system installation and management software, and profiling and debugging tools will potentially require special features to make effective use of petascale systems. Good partner relationships with those providing these key applications will go a long way in ensuring efficient operation at scale.
- ☒ Design user training and development programs and operating policies that encourage users to think at petascale; provide freedom to explore operating protocols, but ensure efficient operation at large scale.
- ☒ Support from research teams with a record of accepting the challenge of the algorithm redesign required to achieve petascale performance — for mesh generators, solvers, and so forth — is an essential component in both selling and achieving site goals.
- ☒ Vendors building petascale systems need to determine whether to take the purpose-built or standards-based approach. Their choices will affect HPC system development — price, scalability, custom content, market penetration — at the high end.

LEARN MORE

Related Research

- ☒ *An Update on Nehalem, Intel's New Celebrity* (IDC #IcUS21405608, August 2008)
- ☒ *Government Directions in HPC: End-User Perspectives* (IDC #213741, August 2008)
- ☒ *HPC Storage Directions and Concerns* (IDC #213654, August 2008)

- ☒ *Intel Labels Larrabee a "Convergent Architecture," Details Reveal Significant HPC Design Content* (IDC #IcUS21402008, August 2008)
- ☒ *The Drivers of Rapid Growth in Technical Computing* (IDC #213626, August 2008)
- ☒ *Nvidia Celebrates Tesla's Second Year Birthday* (IDC #IcUS21298108, June 2008)
- ☒ *AMD Analyst Forum Highlights Quad-Core, and the Balance Between CPU and GPU* (IDC #212457, May 2008)
- ☒ *IDC's New 2008 HPC Market View and Technical Application Categories* (IDC #211507, May, 2008)
- ☒ *HPC Storage Market Outlook* (IDC #209882, April 2008)
- ☒ *Looking for Lost Love, SGI Rolls Out a New Visualization Product* (IDC #IcUS21183008, April 2008)
- ☒ *HPC User Forum Meeting Notes, October 2007: Stuttgart, Germany* (IDC #210903, March 2008)
- ☒ *New IDC HPC Market Segment Definitions and How They Apply to the 2007 HPC Market* (IDC #211019, February 2008)
- ☒ *IDC 2007 Worldwide HPC Market Revenue Results Show Continued Strong Growth* (IDC #210758, February 2008)
- ☒ *SGI Expands Its HPC Cluster Business By Acquiring Linux Networx* (IDC #IcUS21096808, February 2008)
- ☒ *Worldwide Technical Computing Server 2008 Top 10 Predictions* (IDC #210271, January 2008)
- ☒ *HPC User Forum Meeting Notes, September 2007: Santa Fe* (IDC #209924, December 2007)
- ☒ *HP Targets Technical Workgroups with "Supercomputer-in-a-Box"* (IDC #IcUS20955207, November 2007)
- ☒ *DataDirect Networks Introduces New High-Performance High-Capacity Storage Platform* (IDC #209645, November 2007)
- ☒ *IDC's Worldwide Technical Server Taxonomy, 2007: Updating the Comparison of Technical Servers with the Overall Server Market in Revenue, Systems, and Processors, Part 2* (IDC #209382, November 2007)
- ☒ *Worldwide High-Performance Technical Computing System 2006 Vendor Shares: HPC Market Census* (IDC #209251, November 2007)

- ☒ *First Supersized SiCortex Computer Scores with Argonne* (IDC #IcUS20909807, October 2007)
- ☒ *Appro Wins DOE/NNSA Tri-Lab Contract for Capacity Clusters* (IDC #IcUS20903207, October 2007)
- ☒ *Luxtera to Deliver Potentially Disruptive High-Performance Cabling Technology to the HPC Market* (IDC #209072, October 2007)
- ☒ *Mellanox Singing Duet — Increases I/O Performance to 25GBps* (IDC #IcUS20886307, September 2007)
- ☒ *Sun Brandishes Its HPC Sword Again: Giving Its HPC Effort "Lustre"* (IDC #IcUS20877007, September 2007)
- ☒ *AMD Launches Barcelona, Its Quad-Core Opteron Processor* (IDC #208721, September 2007)
- ☒ *SGI: Refocuses on HPC Technical Computing* (IDC #208625, September 2007)
- ☒ *In Pursuit of Petascale Computing: Initiatives Around the World* (IDC #208623, September 2007)

Copyright Notice

This IDC research document was published as part of an IDC continuous intelligence service, providing written research, analyst interactions, telebriefings, and conferences. Visit www.idc.com to learn more about IDC subscription and consulting services. To view a list of IDC offices worldwide, visit www.idc.com/offices. Please contact the IDC Hotline at 800.343.4952, ext. 7988 (or +1.508.988.7988) or sales@idc.com for information on applying the price of this document toward the purchase of an IDC service or for information on additional copies or Web rights.

Copyright 2008 IDC. Reproduction is forbidden unless authorized. All rights reserved.