Frontera: The Evolution of Leadership Computing at the National Science Foundation

Dan Stanzione
John West
R. Todd Evans
Tommy Minyard
dan@tacc.utexas.edu
john@tacc.utexas.edu
rtevans@tacc.utexas.edu
minyard@tacc.utexas.edu
Texas Advanced Computing Center,
The University of Texas at Austin

Omar Ghattas
omar@oden.utexas.edu
Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin

Dhabaleswar K. Panda
panda@cse.ohio-state.edu
The Ohio State University

ABSTRACT

As part of the NSF's cyberinfrastructure vision for a robust mix of high capability and capacity HPC systems, Frontera represents the most recent evolution of trans-petascale resources available to all open science research projects in the U.S. Debuting as the fifth largest supercomputer in the world, Frontera represents a robust and well-balanced HPC system designed to enable large-scale, productive science on day one of operations. The system provides a primary compute capability of nearly 39PF, delivered completely via more than 8,000 dual-socket servers with conventional Intel 8280 ("Cascade Lake") processors. A unique configuration of both desktop GPUs and advanced floating units from NVIDIA enables both machine learning and scientific workloads, and the system delivers nearly 2TB/s of total filesystem bandwidth with 55 PB of usable Lustre disk-based storage and 3PB of all flash Lustre storage. A Mellanox InfiniBand (IB) interconnect provides very low latency with 100Gbps to each node, and 200Gbps between switches in a fat tree topology with minimal oversubscription for efficient communication, even in jobs that use the full system with complex communication patterns. The system hardware is complemented by a robust set of software services, including Application Programmer Interfaces (APIs) to support an evolving user base that increasingly demands productive access via science gateways and automated workflows, as well as a first-of-its-kind partnership with the three major cloud service providers to create a bridge between "traditional" HPC and the cloud infrastructure upon which research increasingly depends.

CCS CONCEPTS

• **General and reference** → **Design**; **Performance**; • **Computer systems organization** → **Parallel architectures**.

PEARC '20, July 26–30, 2020, Portland, OR, USA

ACM ISBN 978-1-4503-6689-2/20/07…$15.00
https://doi.org/10.1145/3311790.3396656

KEYWORDS

cyberinfrastructure, HPC, supercomputer, system design

ACM Reference Format:
Dan Stanzione, John West, R. Todd Evans, Tommy Minyard, Omar Ghattas, and Dhabaleswar K. Panda. 2020. Frontera: The Evolution of Leadership Computing at the National Science Foundation. In *Practice and Experience in Advanced Research Computing (PEARC '20), July 26–30, 2020, Portland, OR, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3311790.3396656

1 INTRODUCTION

The NSF vision for cyberinfrastructure (CI) recognizes the need for a robust mix of high capability and capacity HPC systems available to all open science research projects in the U.S. [2]. In 2017 the NSF called for proposals for a system to replace its then-current petascale resource, Blue Waters, run by the National Center for Supercomputing Applications at the University of Illinois. The new system was to have at least a two- to three-fold time-to-solution performance improvement over Blue Waters, and would serve as the starting point for an upgraded design of a leadership-class system and the physical facility that would host it. That Phase 2 design is expected to have ten-fold or more improvement in capability relative to the initial system.

TACC's response resulted in deployment of a system named Frontera, currently operational and serving its second cohort of NSF petascale users. Frontera debuted at the number five position on the TOP500 list. Design highlights include:

- A primary compute capability of nearly 39PF, delivered completely via conventional processors, consisting of more than 8,000 dual-socket servers with Intel 8280 ("Cascade Lake") processors, in nodes provided by Dell. This base computing capability represents a more than 5.5x increase over the Blue Waters base processors, and even greater memory capacity than the 22,000 nodes of that system.
- Primary compute is augmented with an additional GPU computing capability that addresses both single- (particularly relevant in deep learning and molecular dynamics) and double-precision requirements, delivered via a mix of NVIDIA Quadro and Volta cards.

- Approximately 2TB/s of total storage bandwidth, 55PB of usable Lustre disk-based storage, and 3PB of all flash Lustre storage providing the capability to conduct next-generation science at unprecedented scales and broadening the system's relevance to the emerging data science community.
- A Mellanox InfiniBand (IB) interconnect that provides very low latency with 100Gbps to each node and 200Gbps between switches in a fat tree topology with minimal oversubscription.

The hardware is augmented by an extended investment in software and web services, which includes partnerships with commercial cloud providers to provide Frontera users with access to additional high-integrity storage, sustainable archive options, and regularly refreshed novel computing technologies.

Frontera provides a uniquely balanced set of capabilities that supports both capability and capacity simulation, data intensive science, visualization, and data analysis, as well as emerging applications in AI and deep learning. Large CI users find a familiar programming model and tools in a system that serves as a bridge to future exascale systems that will have many more cores and deeper memory hierarchies.

2 SCIENCE DRIVERS FOR LARGE-SCALE CYBERINFRASTRUCTURE

One lesson that we have learned well from previous generations of supercomputing is that, while we can anticipate that increased computing will yield better solutions to the problems we know about today, we almost always fail to foresee the truly revolutionary discoveries that will be possible as new computing power opens previously unanticipated areas of study. One of the most effective ways to understand the science and engineering requirements for Frontera is to examine the success of current supercomputers. Stampede and its successor, Stampede 2, have each enabled research by tens of thousands of scientists in over three thousand projects spanning many disciplines of science and engineering. Individual jobs on these systems have been successful even at the extreme scale, ranging to more than half a million cores, and have come from nearly all fields of science.

Of the projects that relied on the resources of the Stampede supercomputer, the Nobel-prize winning discovery of gravitational waves by the NSF-funded Laser Interferometer Gravitational-Wave Observatory (LIGO) in 2015 is one of the most important. Researchers used roughly seven million core hours to analyze the first detected gravitational waves. *"The collaboration with TACC computing experts and the computing cycles provided by Stampede both supported the first direct detection of gravitational waves 100 years after Albert Einstein first predicted their existence in his theory of General Relativity,"* said Stuart Anderson, a research manager for LIGO based at Caltech.

An application with immediate relevance to public health and safety is hurricane prediction: getting the track and intensity right could make the difference between life and death for those near the storm. Penn State University researcher Fuqing Zhang used over 22 million core hours on Stampede in a study of 100 tropical storms between 2008-2012 that reduced forecast intensity errors by 25 to 28 percent when compared to the corresponding official forecasts issued by the National Hurricane Center. Zhang describes the impact of his work this way: *"We developed techniques that enable better prediction and understanding of severe weather, including hurricanes and severe weather which certainly have profound significance to society. Much of the research outcome has been adopted or implemented by the research community, federal agencies, and operational weather prediction agencies that will eventually benefit the general public."*

DNA strands use electrostatic attraction or repulsion to fold together or come apart. This property enables cells to store genetic information, replicate and repair that information, and regulate how that information is expressed. Computational physicist Aleksei Aksimentiev of the University of Illinois at Urbana–Champaign used 24 million core hours to explain for the first time how DNA, a highly-charged polymer, is assembled into compact structures at the cell's nucleus in apparent defiance of the law of repulsion of same-sign charges. "*. . . we set out to figure out the molecular mechanism of the same sign charge attraction just to find out that our modeling approach was not accurate enough to account for existing experimental data. After a painful and lengthy recalibration effort, we arrived with the most advanced molecular force field to describe DNA-DNA interactions, which allows us to accurately evaluate forces within compact DNA structures,"* Aksimentiev said.

The success of these projects has been tremendous for large numbers of users at every scale, for capability and capacity runs, for data intensive and simulation-driven science, for individual researchers and large communities. As discussed in the next section, the design of Frontera is informed by our experiences with, and deep data insights about, many years of successful scientific computing at very large scale.

3 A DATA-DRIVEN APPROACH TO DESIGNING A NEXT GENERATION HPC SYSTEM

Frontera is intended to address challenging science, irrespective of whether studies are done with a single large simulation (a "capability" job) or ensembles of smaller simulations ("capacity" jobs). Although the Stampede project was designed for a broader range of problems and communities, there is nonetheless a significant overlap in the types of science that are expected on the two systems, and the ways in which the expected computational science communities will make use of supercomputing to support their research. We, therefore, found it instructive to derive a portion of the design input for Frontera from a study of Stampede's performance and workload data.

The system design team analyzed utilization and application performance for calendar year 2016 gathered on the Stampede supercomputer using TACC Stats [5] and XALT [4]. During 2016 over 750,000 jobs were run in total, of which about 387,268 jobs were run using TACC- developed tools that permit detailed analysis. While this is only about half of the jobs, it represents 79% of all node hours consumed during 2016. This population comprises roughly 2,300 unique application names, but utilization was dominated by the top 15 most commonly used applications which together consume 63% of the SUs during this period; the remaining 2,285+ applications are in a very long tail to the right. Of the 21 million node hours

consumed by the top 15 applications, 56% are from applications that solve PDEs (we include integrodifferential equations in this category) and 41% are from molecular dynamics (MD) applications. The largest single application is VASP, with 22% of utilization during the period, followed by NAMD with 16%, and GROMACS with 13%; the remaining applications (including the weather code WRF, the astrophysics code FLASH4, and the CFD code OpenFOAM, among others) are less than 10% of SUs each, reflecting the diversity of science done on this machine. MPI represents more than half of all parallel computations, with the balance used by MPI+OpenMP primarily, and finally, a small fraction represented by on-node parallelism only (such as pthreads or OpenMP). More than 95% of the jobs analyzed required 32 GB of memory or less, with the remaining jobs evenly distributed between 33 and 1024 GB.

Of the 21 million node hours consumed by the top 15 applications, 56% are from applications that solve PDEs (we include integrodifferential equations in this category) and 41% are from molecular dynamics (MD) applications. The largest single application is VASP, with 22% of utilization during the period, followed by NAMD with 16%, and GROMACS with 13%; the remaining applications (including the weather code WRF, the astrophysics code FLASH4, and the CFD code OpenFOAM, among others) are less than 10% of SUs each, reflecting the diversity of science done on this machine. MPI represents more than half of all parallel computations, with the balance used by MPI+OpenMP primarily, and finally, a small fraction represented by on-node parallelism only (such as pthreads or OpenMP). More than 95% of the jobs analyzed required 32 GB of memory or less, with the remaining jobs evenly distributed between 33 and 1024 GB.

Analysis of the Blue Waters workload over 3 years [3] also reveals a handful of dominant applications (the top 10 consume $\frac{2}{3}$ of node hours), including a mix of PDE, MD, DFT, and lattice QCD applications, though with greater emphasis on LQCD.

Finally, to gauge the characteristics of extreme-scale applications we analyzed the 54 DOE INCITE allocations awarded in 2017, which were divided equally between ORNL's Titan and ANL's Mira. The analysis indicates the same set of applications types as on Stampede and Blue Waters, but heavily tilted toward PDEs (at 48%) relative to MD (17%), quantum Monte Carlo (11%), LQCD (7%), density functional theory (DFT, 6%), and all others (11%).

This examination of the dominant Stampede, Blue Waters, and INCITE applications — PDE and MD, with lattice QCD and many-body/n-body problems representing important additional application categories — motivates specific decisions about the architecture most likely to effectively support users of Frontera. PDE simulations have historically been large consumers of high-end computing cycles and we can expect that to continue over the operational life of Frontera. Many high-end PDE applications employ static-grid explicit solvers, which feature local interactions and hence scale well on highly-parallel systems. Rarer than explicit solvers, but still important, are semi- or fully-implicit PDE solvers for which a global system solve is required, usually requiring sophisticated preconditioners to be effective. The challenge of global system solves (and explicit methods that employ adaptive mesh refinement) is the complex and possibly dynamic memory access and communication patterns. But these challenges to large-scale parallelism are also being overcome. The 2015 Bell Prize winning application [7] shows that, in principle, there is no reason why implicit and adaptive PDE solvers cannot scale with nearly ideal parallel efficiency to CPU-based systems with greater than million-way parallelism—provided sufficient effort is expended in algorithmic and software engineering. More challenging is obtaining good node performance, stemming from the low arithmetic intensity of typical stencil or sparse matrix operations associated with PDE solvers. But reasonable node performance can be achieved with requisite effort if local interactions are sufficiently dense and structured, which is often the case for highly nonlinear or multiphysics problems, or high order discretizations. With their complex dynamic underlying algorithms and modest arithmetic intensity, and the considerable effort required to extract performance, GPU adoption has been slow. The most productive and performant option for most PDE solvers remains a CPU-based system combined with a fast, low-latency network.

The other large class of applications that dominate the portfolio – MD, along with lattice gauge QCD and n-body/many-body problems – can also make excellent use of highly-parallel distributed memory systems. The class of MD, n-body, and many-body problems often feature potentials that incorporate non- local interactions. In most cases these can be coarse-grained, treated by FFTs, or neglected beyond some cut-off radius, and effective highly-parallel solvers have been designed that exploit this feature. Dense localized interactions lead to high arithmetic intensity in these expensive kernels, meaning that high node performance is achievable. Lattice gauge QCD methods resemble PDEs in their execution of sparse, local matrix-vector products (but on a 4D grid). Despite the challenges of parallelizing these applications at large-scale, good performance is now being achieved on high-throughput systems. In all of these application categories and similar to PDEs, a low-latency/high-bandwidth network is the key to scalability. Overall, we conclude that a CPU- based primary system with powerful nodes and a fast network is the best choice to allow a broad spectrum of users with diverse needs to do their science at scale. For certain application classes that can make effective use of GPUs, such as MD and ML, a single-precision GPU subsystem provides a cost-efficient path to high performance.

In addition to large HPC projects identified through system usage analysis, it is important to consider emerging uses of advanced computing resources for which Frontera will be critical. Prominent among these are data-driven and data-intensive applications, which the NSF Advisory Committee for Cyberinfrastructure (ACCI) recommended be acknowledged as a distinct discipline [1]. There are two fundamental categories of data-driven science. In data assimilation and inverse problems, observational or experimental data are used to infer uncertain states or parameters in the (PDE, MD, etc.) model, which ultimately results in solution of the forward (PDE, MD, etc.) problem numerous times while (intelligently) exploring parameter space, and thus the analysis in the preceding paragraphs applies (with added opportunity for task parallelism). In the second class of data-driven science, statistics, informatics, and analytics are used to learn directly from the data without recourse to the physics. These data applications often require non-traditional software (e.g. parallel R and MATLAB) and lead to the development of new data applications in which I/O read performance is even more

important than write performance. A nascent but growing class in this second category is ML. The coming decade will see significant efforts to integrate physics-driven and data-driven approaches to learning. We believed it is important that Frontera be designed with the capability to address very large problems in these emerging communities of computation, and that it is essential to provide a well-balanced petascale HPC system with comprehensive capabilities that can serve a wide range of both simulation-based and data driven-driven science.

4 FRONTERA ARCHITECTURE

The final system comprises 8,210 total compute nodes split between two compute subsystems – a primary system based on Intel 8280 Xeon processors, and both single- and double-precision GPUs for jobs that can make effective use of them, a 70+PB storage subsystem with almost 2TB/s of aggregate bandwidth, and 26 additional service nodes on a Mellanox InfiniBand interconnect. The system includes interfaces for both direct user access, and automated access via application or API, as well as a bridge to the commercial cloud, for access to sustainable high performance archiving and other cloud services. The conceptual view of the complete ecosystem is shown in 2. The complete system is housed in 114 racks and has a max system power of just over 5.5MW.

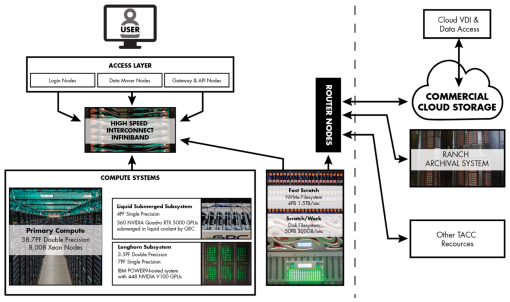

Figure 1: The Frontera project's computing ecosystem.

4.1 Intel Cascade Lake Processor

The primary compute system consists of 8,064 dual-socket server nodes based on the Intel "Cascade Lake" (CLX) Xeon processor, a successor to the shipping "Skylake" (SKX) processor, both in the current 14nm process. While this node count is only 30% larger than the Stampede2 system, based on our analysis that having the most powerful cores in the most powerful socket is the single greatest enabler of scientific productivity, we are using a much more powerful model of the processor than has typically been used in HPC systems. The Platinum 8280 model Xeon we have chosen (an update on the current 8180), has a more than 20% higher clock rate, 16% more cores, and 10% more memory bandwidth than the Platinum 8160 processors in Stampede2. The resulting difference means the 8,064 compute nodes of Frontera deliver more than double the theoretical peak performance of the 5,936 compute nodes of Stampede 2, and more than five times the performance of the 22,000 compute nodes of Blue Waters. While theoretical peak is an increasingly misleading indicator, in this case, from the SKX to the CLX processor we are scaling not only the peak frequency, but the more relevant AVX and turbo frequencies as well, and the actual productive throughput of the system is 2x and 5x the compute power of Stampede 2 and Blue Waters, respectively.

The specific CLX processor selected for Frontera is the 205W processor thermal design power (TDP) part with a projected theoretical peak performance of 2.4 TFLOPS at a 2.7GHz clock frequency and 28 cores. Each node has a two-socket motherboard with 192GB DDR-4 RAM. Liquid cooling to each node is required to dissipate the heat from both sockets, memory, and voltage regulators within each node.

4.2 Accelerated Computing

As previously noted, the primary capability of Frontera comes from the CPUs. However, for two reasons, we included an accelerated component to Frontera. First, to accommodate those applications that already have significant performance advantages on GPUs, even when adjusting for cost – particularly in artificial intelligence and molecular dynamics. Second, we wished to keep a GPU capability in case the application mix changes, and to track application performance and adoption with an eye towards future systems.

Accelerated computing on Frontera is divided into two subsystems, a single-precision optimized subsystem that focuses on cost-performance, and a double-precision GPU system that adds full NVLINK capability.

The single-precision subsystem comprises four NVIDIA RTX-5000 cards in each of the 90 Intel-based nodes, for a total of 360 cards. The nodes are cooled in oil using an immersion-based cooling process delivered by Green Revolution Cooling. These GPU nodes are integrated directly into the Frontera system fabric and are directly accessible to users logged in to Frontera.

The double-precision system consists of 112 IBM Power9 nodes, each with four NVIDIA V100. Because these are highly capable cards we felt it was important to include NVLINK-enabled compute nodes to make it possible to achieve sufficient bandwidth and keep data flowing to these cards. For that reason, this portion of the system is provided by IBM, using nodes similar to what is provisioned in the *Summit* and *Sierra* systems at the Department of Energy. The system consists of 112 Power9-based nodes, each with 256GB of RAM and four Tesla V100 GPUs, for a total of 448 GPUs. This system also has its own local parallel filesystem and InfiniBand fabric. Given that it will also support other users, we are operating it under its own system name ("Longhorn"). Frontera home and scratch filesystems are visible on Longhorn, albeit at a lower performance than users achieve on Frontera. As the binaries are different between the x86-based Frontera and Power9-based Longhorn, we believe providing users separate home and scratch directories will lead to less confusion and fewer errors.

4.3 High-Performance Interconnect

Tightly coupled scientific applications require a high-bandwidth, low-latency network. Frontera provides a Mellanox InfiniBand network to support all inter-node application communications (e.g. MPI messages and shared file system transfers), at 100Gb/s to compute nodes and 200Gb/s with HDR between switches in the fabric. The I/O servers have full non-blocking connectivity to ensure maximum network bandwidth of 3TB/s is available to the storage

subsystem with multiple dual-port EDR cards per storage controller and server. The core switches are also liquid-cooled at 22kW per switch and provide N+N redundant power supplies.

4.4 Disk I/O Subsystem

Frontera's storage subsystems include a highly capable flash-based filesystem for those users that require high-bandwidth I/O for their large problems augmented with multiple large filesystems that provide capacity storage with very good performance for the less demanding I/O jobs. In addition, a pool of Non-volatile Memory Express (NVMe) storage is available as NVMe over InfiniBand for users that need access to their own very fast storage on the nodes beyond what the local SSD can provide, especially those running ML applications. The flash-based filesystem and NVMe pool are composed of 75 DataDirect Networks (DDN) IME240 storage appliances capable of delivering 1.5TB/s of I/O bandwidth with 150 100Gb/s IB ports using 1,500 2TB NVMe drives for a total raw capacity of 3PB. The home and two capacity-based scratch filesystems are composed of four DDN ES18K EXAScaler appliances each with four object storage servers, 64 90-drive SS9012 enclosures, and 4,960 14TB drives to provide more than 300GB/s of I/O bandwidth and 68PB of raw capacity, 55PB usable with Declustered RAID. All of these filesystems are run as scalable Lustre parallel filesystems.

4.5 External Cloud Services

The modern scientific computing landscape is changing rapidly, and the Frontera computing ecosystem reaches beyond the standard center-hosted supercomputer by integrating services from commercial cloud providers that complement modern scientific workflows. Through partnerships with Amazon, Google, and Microsoft we enable users of Frontera to consume specific integrated services from these providers via their allocations on Frontera (up to a consumption cap). The commercial cloud provides several services that we could not otherwise provide:

- A "sustainable" archive – users have the option to mirror their archive data to the cloud provider of their choice to create a high-integrity, multi-copy archive.
- Access to additional node types – our Frontera deployment made static technology selections, but many additional options will appear steadily through the period of operations ending in 2024. Through our partnerships, we make available nodes with every GPU and CPU vendor and option these three providers will offer, for both benchmarking and production workloads.
- Virtual Desktop Interfaces – if users opt to store their archive data in a commercial cloud, many additional data access interfaces and services can be offered. For instance, in the Microsoft cloud, data stored can be accessed via their VDI interface by users.

4.6 The Science Environment

Frontera's operating system is based on the latest release of the Linux kernel series and the Red Hat Enterprise Linux (RHEL)/CentOS Linux distribution. To optimize performance for all applications, TACC tunes the compute-node OS configuration to minimize the number of underlying services competing for processor cycles.

Overall cluster management and provisioning are managed by TACC's Linux operating System Framework (LoSF), which combines bare-metal provisioning via Cobbler with an integrated software/configuration toolkit that is used on all TACC HPC resources [6]. TACC has significant prior experience configuring, modifying, and regression testing HPC software to support large-scale production clusters through RPM software builds integrated by LoSF to maintain a consistent, traceable software environment.

C, C++, and Fortran compilers and high-performance MPI stacks for these languages represent the dominant development library requirements for application scientists. TACC supports multiple MPI compiler families including MVAPICH2 and Intel MPI, as well as CUDA and OpenCL for programming GPUs. For debugging, optimization, and analysis purposes, TACC provides Intel Debugger, ARM's DDT and MAP parallel debugging and performance tools, Perfexpert, PAPI, and TAU for users. Additionally, a broad range of scientific, mathematical, and profiling/optimization libraries are deployed on Frontera, including popular libraries such as PETSc, Trilinos, and others.

Although Frontera is designed for the subset of NSF CI users with the largest requirements, we support an incredibly wide array of scientific applications, including those from popular container repositories. Using technologies from the cloud community, with techniques we have developed and refined on Stampede, Stampede 2, Jetstream, and other systems, we provide seamless, integrated support for the use of Singularity containers, both from users and standard repositories. To make the experience seamless, we inject mount points and environment variables into the container to match the system environment – the $SCRATCH, $WORK, and $HOME filesystems all will be identical to what a user would see natively on any Frontera node.

The use of containers greatly enhances the number of people who contribute to the Frontera software base, promotes portability with other resources, and frees up vast staff resources to focus on "core" applications while greatly expanding our supported catalog. While containers are not a panacea for performance at scale (though there is much progress in this direction), and we continue to maintain our "native build support." However, particularly for "discovery science for throughput computing" applications, life sciences applications, and the deep learning frameworks, the tremendous reduction in application support overhead, the ability to support portability from desktop to Frontera and all systems in between, and the ability to leverage a much broader community to rapidly build and support software stacks are invaluable.

Finally, we also support TACC's TAPIS REST APIs on Frontera. TAPIS serves as the foundation for supporting web-service enabled workflow tools, and a set of rich applications, including those providing more complex, automated, and real-time scientific workflows. We use both TACC-developed and standard APIs that we have successfully used in the past to support a large number of science gateways, and other large computing projects such as Open Science Grid and the LHC and LIGO collaborations.

5 PERFORMANCE RESULTS

The NSF solicitation prescribed a portion of the benchmark we used to assess the performance of Frontera and to ensure it passed the

requirement of 2-3x application performance on Blue Waters. The Sustained Petascale Performance (SPP) benchmark, which includes full application runs of the following codes: AWP-ODC (seismic), CACTUS (relativistic astrophysics), MILC (particle physics, lattice QCD), NAMD (molecular dynamics), NWChem (chemistry), PPM (astrophysics), PSDNS (fluid dynamics), QMCPACK (quantum chemistry), RMG (electronic structures), VPIC (particle physics), and WRF (weather). The benchmark itself was defined more than 13 years ago, and while most of the codes are still relevant, many of the problem sizes were small relative to the full capability of this new machine. We also added the RESNET benchmark, implemented in Caffe, to the SPP applications to capture machine learning requirements. In addition to the SPP, we also ran a complement of low-level hardware performance benchmarks (such as HPL, DGEMM, STREAM, and MPI latency and bandwidth tests) and system reliability tests.

Application	Acceptance Threshold[s]	Frontera Time[s]	% over Threshold	Improvement over Blue Waters	Threshold Node[#]	Frontera Node[#]
AWP-ODC	335	326	1.03	**3.2**	1366	1366
CACTUS	1753	1433	1.22	**3.3**	2400	2400
MILC	1364	831	1.64	**9.5**	1296	1296
NAMD	62	60	1.03	**4.0**	2500	2500
NWChem	8053	6408	1.26	**3.8**	5000	1536
PPM	2540	2167	1.17	**3.6**	5000	4828
PSDNS	769	544	1.41	**2.8**	3235	2048
QMCPACK	916	332	2.76	**5.5**	2500	2500
RMG	2410	2307	1.04	**3.2**	700	686
VPIC	1170	981	1.19	**4.3**	4608	4096
WRF	749	635	1.18	**5.2**	4560	4200
Caffe	1203	1044	1.15	**3.2**	1024	1024

Figure 2: Frontera application benchmark results.

All of the acceptance tests passed, some by a wide margin; on average we achieved 4.3x faster runtime than Blue Waters, exceeding our target of 3x Blue Waters performance on both individual runs and in total system performance. It is important to note that Frontera has 1/3 the nodes of Blue Waters, resulting in 9x more SPP throughput per dollar, and 4.7x more SPP throughput per watt (resulting in lower operational costs as well).

We did note an interesting result from our HPL runs, which we will illustrate by comparing the per node performance of Stampede. Stampede was deployed in November of 2012 and consisted of dual 8 core Intel Sandy Bridge at 2.7GHz each. We measured about 90% of peak performance using HPL on these nodes, or 331GF versus 346GF peak; since that time, it is much more common to achieve no more than 60-65% of peak on HPL.

The per node theoretical peak performance of Frontera is 4,834 GFLOPS per node:

$$\begin{aligned} peak\ GFLOPS = {} & clock\ frequency(GHz) \times cores\ per\ socket \times \\ & sockets\ per\ node \times vector\ length \times FMA \times \\ & number\ simultaneous\ issues \\ = {} & 4,834\ GFLOPS \end{aligned}$$

The ratio peak FLOPS for Frontera to Stampede is thus about 14:1. The achieved per node HPL performance is 2,900GF on a Frontera node (about 60% of the theoretical peak), and 310GF on a Stampede node, for a ratio of only about 9:1 in realized performance on what is probably the best possible case, HPL. This translates to about 64% of the theoretical peak performance improvement for HPL in moving from Stampede to Frontera.

If, however, one computes the theoretical peak performance of Frontera using the AVX frequency rather than the headline clock rate, then the theoretical peak performance of Frontera changes to 3,222 GFLOPS per node, and the ratio of Frontera to Stampede peak performance becomes 9.3:1, much closer to the measured difference in performance between the systems, and much closer to the percentage of peak that we achieved in the Sandy Bridge timeframe. If we report the ratio of peak performance based on the practically achievable frequencies of the chips, it turns out the peak ratio is almost exactly predictive of the HPL speedup.

6 CONCLUSIONS

The Frontera system, now deployed and fully operational, is the evolution of the NSF's petascale computing program. The system is based upon the most common programming and was designed to allow users to achieve very high performance from day one without the need for extensive modifications to their codes. At the same time, the inclusion of both single- and double-precision GPU subsystems, as well as the POWER9 processors supporting NVLINK, provides an ideal environment for many machine learning applications, as well as a rich environment in which to evaluate the effectiveness on one of the primary exascale paths selected by DOE on the NSF workload. The system exceeds its designed performance goals, reach 4.3x improvement over Blue Waters versus a goal of 3x, and the integration with cloud resources provides a unique CI solution that is well-suited for evaluating the best path forward in the LCCF design.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation through the ACI-1134872 Stampede, OAC-1540931 Stampede2, ACI-1953575 XSEDE, and OAC-1854828 Frontera awards.

REFERENCES

[1] 2011. *NSF Advisory Committee for Cyberinfrastructure Task Force on Grand Challenges. Final Report.*

[2] 2016. *Future directions for NSF advanced computing infrastructure to support U.S. science and engineering in 2017-2020.* The National Academies Press. https://doi.org/10.17226/21886

[3] 2017. *Blue Waters Sustained Petascale in Action: Enabling Transformative Research, 2017 Annual Report.*

[4] K. Agrawal, M. R. Fahey, R. McLay, and D. James. 2014. User Environment Tracking and Problem Detection with XALT. In *2014 First International Workshop on HPC User Support Tools.* 32–40.

[5] Todd Evans, William L. Barth, James C. Browne, Robert L. DeLeon, Thomas R. Furlani, Steven M. Gallo, Matthew D. Jones, and Abani K. Patra. 2014. Comprehensive Resource Use Monitoring for HPC Systems with TACC Stats. In *Proceedings of the First International Workshop on HPC User Support Tools* (New Orleans, Louisiana) *(HUST '14).* IEEE Press, 13–21. https://doi.org/10.1109/HUST.2014.7

[6] Robert McLay, Karl W. Schulz, William L. Barth, and Tommy Minyard. 2011. Best Practices for the Deployment and Management of Production HPC Clusters. In *State of the Practice Reports* (Seattle, Washington) *(SC '11).* Association for Computing Machinery, New York, NY, USA, Article 9, 11 pages. https://doi.org/10.1145/2063348.2063360

[7] Johann Rudi, A. Cristiano I. Malossi, Tobin Isaac, Georg Stadler, Michael Gurnis, Peter W. J. Staar, Yves Ineichen, Costas Bekas, Alessandro Curioni, and Omar Ghattas. 2015. An Extreme-Scale Implicit Solver for Complex PDEs: Highly Heterogeneous Flow in Earth's Mantle. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis* (Austin, Texas) *(SC '15).* Association for Computing Machinery, New York, NY, USA, Article 5, 12 pages. https://doi.org/10.1145/2807591.2807675