Stampede 2: The Evolution of an XSEDE Supercomputer

Dan Stanzione
dan@tacc.utexas.edu

Bill Barth
bbarth@tacc.utexas.edu

Niall Gaffney
ngaffney@tacc.utexas.edu

Kelly Gaither
kelly@tacc.utexas.edu

Chris Hempel
hempel@tacc.utexas.edu

Tommy Minyard
minyard@tacc.utexas.edu

Texas Advanced Computing Center,
The University of Texas at Austin

S. Mehringer
Cornell University
susan@cac.cornell.edu

Eric Wernert
The University of Indiana
ewernert@iu.edu

H. Tufo
The University of Colorado
tufo@cs.colorado.edu

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

P. Teller
The University of Texas at El Paso
pteller@utep.edu

ABSTRACT

The Stampede 1 supercomputer was a tremendous success as an XSEDE resource, providing more than eight million successful computational simulations and data analysis jobs to more than ten thousand users. In addition, Stampede 1 introduced new technology that began to move users towards many core processors. As Stampede 1 reaches the end of its production life, it is being replaced in phases by a new supercomputer, Stampede 2, that will not only take up much of the original system's workload, but continue the bridge to technologies on the path to exascale computing. This paper provides a brief summary of the experiences of Stampede 1, and details the design and architecture of Stampede 2. Early results are presented from a subset of Intel Knights Landing nodes that are bridging between the two systems.

KEYWORDS

High Performance Computing, Supercomputing, Xeon Phi

1 INTRODUCTION

The NSF[1] vision for cyberinfrastructure (CI) recognizes the need for high capability and capacity HPC systems in the national operational CI available to all open science research projects in the U.S. [1]. Over the last 4 years, Stampede, located at the Texas Advanced Computing Center (TACC) at the University of Texas at Austin, has been a leader in providing these services to the XSEDE[2] community. Over 10,000 users working on over 2,000 funded projects have run more than *eight million* simulations and data analysis jobs on Stampede in the first half of its production life. Stampede has been a vital capability enabling researchers in every field of scientific inquiry, from Astronomy to Zoology.

Based on the success of the Stampede project, TACC and its partners received a renewal award to deploy Stampede 2, a powerful new system that builds on the success of Stampede and will continue to enable groundbreaking science. Stampede 2 servers are provided Dell and Intel, and they use a mix of Intel's Xeon and many-core Xeon Phi technologies. Stampede 2 *doubles* the performance of Stampede in most of the dimensions relevant to the science and engineering research community. Stampede 2 will deliver:

- **Twice the performance from conventional Xeon processors** (>5PF) as the original Stampede, creating a smooth transition for today's science application code base.
- **Twice the performance from the many-core component** (>13PF), with a new stand alone Knights Landing processor, much simpler to use and program, and twice the total system performance (~18PF).
- **Double the available memory (>550TB) and double the storage bandwidth (~350TB/s)**, providing the capability to conduct next-generation science at unprecedented scales and broadening the system's relevance to the emerging data science community, while its appliance-based storage

PEARC17, July 09-13, 2017, New Orleans, LA, USA

ACM 978-1-4503-5272-7/17/07…$15.00
http://dx.doi.org/10.1145/3093338.3093385

architecture adds integrated declustered RAID capability to improve reliability.

- **An innovative experimental capability,** provided by a subset of nodes with non-volatile "3D Crosspoint" memory, to be deployed mid-life and explore novel computation approaches by creating very large memories or very fast storage, and to explore future memory hierarchies.

Stampede 2 was designed for effective user transition today while serving as a bridge to the compute paradigms of tomorrow, providing a uniquely balanced set of capabilities that support both capability and capacity simulation, data intensive science, visualization and data analysis. The next section looks back at Stampede 1, then following sections detail how that workload drove the design of Stampede 2.

2 STAMPEDE-1: An Overview and Retrospective

2.1 Stampede-1 Architecture

Proposed in 2011 and deployed into production in January of 2013, the base Stampede 1 system consisted of 6,400 compute nodes built by Dell, with dual-socket Intel Sandy Bridge processors, each with 8 cores and nominally clocked at 2.7Ghz, for a total of 102,400 cores, with 14PB of storage and an FDR Infiniband interconnect in a fat tree topology provided my Mellanox. A subsystem of nodes provided 128 Nvidia GPUs, and another subsystem provided 16 1TB large memory nodes.

2.2 A Look back at Stampede – Sustained Success in Science.

The best way to make the case for the science and engineering need and promise of Stampede 2 is to look at the success of the current Stampede. Individual jobs on Stampede have been successful even at the extreme scale, ranging to more than half a million cores (counting cores on the Xeon Phi), and have come from nearly all fields of science. Since deployment, 2,668 principle investigators working on 3,531 projects have made production use of Stampede. About 12,000 researchers from more than 400 universities, labs, and companies have accounts, of which >8,000 have run a production job through the job queues on the system. No one institution, not even UT-Austin as the host institution, approaches even 10% of the cycles on the system; Stampede is, in the truest sense, a national resource.

These statistics substantially undercount the real number of users impacted by the system. "Science Gateways" – web portals that make use of Stampede to run jobs such as Galaxy, NanoHub, Cipres, and iPlant – typically appear as a single user to Stampede. Thus, in the count of 7,000 active users above, Galaxy (35,000 users) and iPlant (18,000 users) count for a total of 2. The total number of members of the national science and engineering community whose work is critically enabled by Stampede is easily in the tens of thousands. The degree of the user community's desire to make use of Stampede is evident in the gap between allocation requests and available resources, detailed in figure 1. Impact, however, is not completely represented by numbers. The quality of the computational research is also crucial. A handful of selected accomplishments give a sense of the impact Stampede has had, for both simulation and data intensive science.

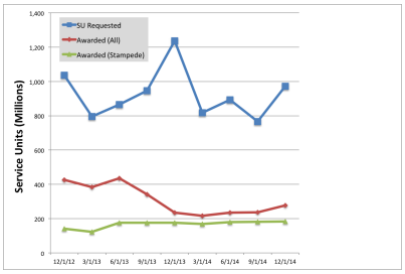

Figure 1: XSEDE Requests and Allocation over the life of Stampede

During commissioning, Stampede was used by the Southern California Earthquake Center to predict the frequency of damaging earthquakes in California for the latest Uniform California Earthquake Rupture Forecast (UCERT3). The results of the simulations will be incorporated into USGS's National Seismic Hazard Maps, which are used to set building codes and insurance rates. PI Tom Jordan stated, "*We do a lot of HPC calculations, but it's rare that any of them have this level of potential impact".*

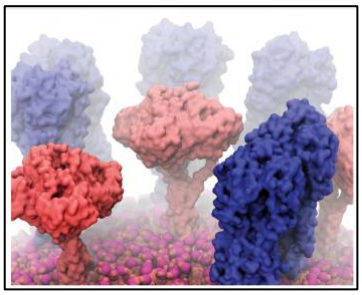

Figure 2: Influenza protein "stalks" modeled by UCSD researchers on Stampede.

Dr. Rommie Amaro, at UC-San Diego, used Stampede and its Xeon Phi processors to perform whole-viron simulation of influenza, modeling surface protein "stalks" which control interactions with cells and drugs, in a 200 million atom "all-atom" model (See Figure2). Dr. Amaro's team is one of many that make use of NAMD. NAMD now leverages the system's Xeon Phi manycore processors, and is developed by Dr. Klaus Schulten's team at the University of Illinois at Urbana-Champaign (UIUC), also large users of Stampede. Dr. Schulten's team uses Stampede to unravel how a newborn protein folds and to design novel enzymes to produce second-generation biofuels. Dr. Schulten commented, "*We are extremely excited about the strong computational power of Stampede. It is the fastest machine we have experienced right away and we have performed a lot of interesting scientific computational experiments on the system."*

One of the top users in climate modeling is Fuqing Zhang at Pennsylvania State University. Recently, this team undertook a hindcast study of all tropical storms from 2008-2012 using WRF and ensemble Kalman filtering to incorporate airborne radar data. This approach reduced forecast errors by 15 to 40%, and received the AMS 2014 Banner Miller award (visualization shown in Figure 3). A reduction in forecast error of this magnitude has the potential to save lives and protect property in coastal regions.

Stampede has been used extensively in data intensive applications. An exemplar of this class of usage is work done by an international team led by Dr. Stephen Wong at Houston Methodist Research Institute, with collaborators at Harvard

Medical School and around the world. Dr. Wong's team uses a systems biology approach to uncover new understanding and connections in the development and management of cancer. This project improved data coverage by 1,000-fold, and sequenced tumor genes from the Cancer Genome Atlas and the Alzheimer's Disease Neuroimaging Initiative to uncover pathways in common between Alzheimer's and a type of Brain Cancer. Commenting on this work, Dan Gallahan, the Deputy Director of NIH, stated *"This work of Dr. Wong's is quite exciting in that it shows connections between two of the most intractable diseases in modern society."*

Data intensive analysis is a hallmark of another large class of users where the Stampede project has had success. The Stampede team has built successful collaborations with the Advanced LIGO project, the Large Hadron Collider (LHC) community, the Hobby-Eberly Telescope Dark Energy Experiment (HETDEX), and the Cyverse project. With each of these projects, Stampede is being used by large instruments and large users for data analysis.

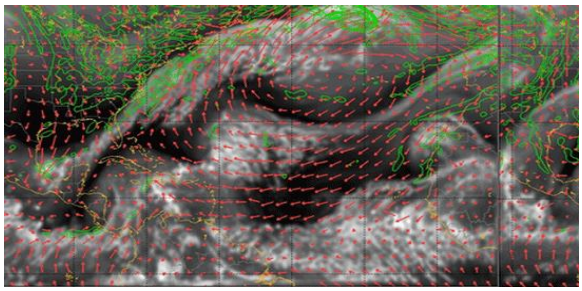

Figure 3: A snapshot of the real-time hurricane analysis and forecast system Credits: Fuqing Zhang and Yonghui Weng, Meteorology, Penn State University

2.3 Lessons Learned

Stampede has been by most measures a great success – delivered on time and on budget in its proposed configuration, and highly utilized for impactful science. However, while getting users to make use of the system, getting them to make the most *effective* possible use of the system has remained a challenge. As so many systems have demonstrated, getting programmers to move to a new technology is a difficult culture change, and it requires time and effort – even for relatively incremental changes. Adoption of multi-core proved difficult for small teams and those using code "out of the box". With the proliferation of cores in each node, as well as the diversity of types of nodes, users would make mistakes in how to configure their jobs. The rise of many new kinds of software systems that made different assumptions about hardware (particularly Hadoop, which boomed in Stampede's early years) also caused users to make poor assumptions about how to use resources. A suite of monitoring tools deployed during Stampede's lifetime were critical in helping rapidly identify users who made poor use of resources and assisting them to use the system more effectively (typical problems might be requesting more cores or nodes than their job could employ, or confusing when to use local versus shared filesystems).

The tremendous use of Stampede, the types of applications that ran on it, and the change in user behavior over time all provided a tremendous amount of data to drive the design of the renewal system.

3 Designing a new system: User Needs

The evolution from terascale to petascale systems—accompanied by improvements in memory, network, I/O, storage, and visualization—have provided the hardware capabilities to make modeling and simulation possible for complex phenomena in fields as diverse as astronomy, biology, climate, and the social sciences. At the same time, the petascale era has presented new opportunities to mine, analyze, and interpret rapidly expanding observational and model-generated data. However, despite the growth in HPC capabilities, the demand for greater processing power has accelerated. As models have increased in fidelity, underlying algorithms have become more scalable, community software has become more robust and widely available, user groups have broadened, and a new generation has been trained.

Science and Engineering Needs for HPC

In response to the growing demand for high-end computational resources, a number of blue ribbon panels over the past several years have produced reports that articulate the critical need for substantial growth in HPC resources available to the U.S. scientific community to sustain advances in scientific discovery, engineering innovation, U.S. competitiveness, and security [3][4][5][6][7]. The findings of all of these reports strike a common chord: to fully realize the opportunities that lie ahead, massive increases in available computing resources—in capability and capacity—are critically needed.

To assess the HPC needs of current large-scale science and engineering applications, TACC analyzed utilization patterns on Stampede over the sixteen-month period from September 2013, to the end of January 2015. During this period over 1.5B Service Units (SUs) were consumed, 1.06B SUs of which were jobs executed using TACC-developed tools that permit detailed analysis. Sorting this data by application, and limiting the analysis to only those applications with more than 2M SUs consumed during the period, leaves a population of 60 applications, the top 34 of which are responsible for 50% of all SUs consumed during the sample period. Of the .5B SUs consumed by these applications, 47% of usage is from applications that solve PDEs, 45% is from molecular dynamics (MD) applications, 6% is from lattice QCD, and 2% is from n-body applications. The largest single application on Stampede during this period is NAMD (a popular MD code). NAMD alone accounts for 13.5% of all SUs consumed. Despite this clustering at the top, the tail of the utilization curve is very long: although only 60 codes consume 50% of the SUs during the sample period, the remaining 50% is divided among more than 5,000 individual applications. Interactions with Stampede's many users lead us to believe that the bulk of this long tail is PDE-based applications, but the population size is too large to study in detail.

Examination of the performance characteristics of the four dominant application categories (PDE, MD, lattice QCD, and n-body problems) informed the architecture of Stampede 2. Most parallel PDE applications employ explicit solution methods that scale well on distributed memory systems such as Stampede. Implicitly-solved PDEs are less common among the largest projects, but are an important class for which a global solve is required, and for which effective memory use is a significant challenge. Nonetheless, as mentioned above, results in the HPC community reveal that PDE applications can scale effectively to tens of thousands of cores.

The other large applications that dominate the portfolio—MD, lattice gauge QCD, and n-body problems—also have potential to make excellent use of distributed memory systems with accelerators. Lattice gauge QCD methods resemble PDEs in their execution of sparse, local matrix-vector products, but with a 4D grid. Despite the challenges, good performance is now being achieved on accelerator systems [8]. On the other hand, the class of MD, n-body, and many-body problems 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. As with PDEs, a low-latency/high-bandwidth network is the key to scalability. The dense local interactions mean that with sufficient effort, good performance is also possible on accelerators, as demonstrated by the inclusion of a Xeon Phi-accelerated application in the finalists for the 2014 Gordon Bell prize [9].

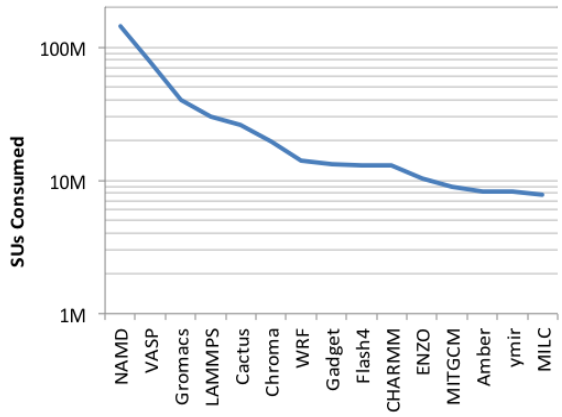

Figure 4: Top 15 Applications, SU Usage on Stampede

In addition to large HPC projects identified through system usage analysis, we considered new and emerging uses of advanced computing resources for which Stampede 2 will be critical. Prominent among these are data-driven and data-intensive applications. The NSF Advisory Committee for Cyberinfrastructure (ACCI) recommended that the NSF recognize the field of computational and data-intensive science and engineering as a distinct discipline [10]. Data-driven/intensive science fundamentally uses observed/collected digital data for either data assimilation (inverse problems), in which observational data is used to infer uncertain states or uncertain parameters, or statistics, informatics, and analytics methods in which the data, not the physics, is the primary driver of discovery. These data applications often require non-traditional software (e.g. parallel R and Matlab) and lead to development of new data applications in which I/O read performance is even more important than write performance. Nonetheless, we believe it is important to provide a well-balanced petascale HPC system with comprehensive capabilities that can serve a wide range of simulation-based and data driven-driven science.

The design of an HPC system that is effective for this wide-ranging science and engineering workload is a complex challenge. Acquisition cost, operational cost, performance, and user productivity must be balanced to create a system that best satisfies these (often conflicting) constraints. A system that offers outstanding potential performance that is difficult to achieve in practice is a poor choice if the "default" performance is poor. Likewise, the potential of outstanding application performance is irrelevant if the chosen system is unavailable to users because of unplanned outages. Another significant consideration is the ease of migration from the current to the future platform: applications should run well on the future platform with little change, but users with the expertise and resources to tune their application for most effective use of platform should be rewarded.

In designing Stampede-2, we assessed a wide range of options in the HPC market against the above criteria, ultimately limiting ourselves to 3 options – a conventional CPU system, and ones with combinations of either GPUs or Xeon Phi nodes. We determined that a Xeon/Xeon Phi configuration offers the most potential performance over the widest range of general science and engineering workloads, with a programming model that allows users to achieve that potential in practice while, at the same time, placing them on a high confidence path to exascale.

Our assessment of GPUs is informed by our experience running several hundred GPU nodes within TACC's diverse computational infrastructure. For codes in computational chemistry and molecular dynamics in particular (two of our leading user communities) GPUs offer performance advantages. The emerging software base machine learning also uses a number of libraries whose best implementations are on GPUs While some of our largest applications would likely excel on this architecture, only about 30% of our user requests are in that category. We also noted when we developed our proposal that Blue Waters makes a large number of GPUs available to the community, and will continue to do so through the first half of Stampede-2's operational life – likewise GPUs are available in SDSC's Comet system, PSC's Bridges system, as well as TACC's Maverick platform within XSEDE.

This led us to an x86-based design, and a decision about the balance between Xeon and Xeon Phi processors. While great strides have been made in enabling users to make effective use of the Xeon Phi accelerators in Stampede today, adoption remains far from universal (this is a challenge with both GPUs and Intel's Xeon Phi). However, we believe that two related technology trends make the inclusion of Phi in the final configuration both *desirable* from a performance perspective and *necessary* to

provide today's trans-petascale HPC users with a viable path to the likely exascale future. First, all likely technology paths to exascale trend toward a "manycore" future. If Stampede 2 is to serve as a bridge for the scientific community from 2017-2021, embracing a manycore approach will leave the user community much better prepared for exascale than continuing solely with conventional CPUs. Second, the Xeon and Xeon Phi platforms are *converging*. Clock rate, core count, and vector length are all substantially closer between Xeon and Xeon Phi than they were for Stampede 1 (about 1.5 to 1 vs. 3 to 1). Software investments made to improve performance on Xeon will improve Phi performance as well (and vice versa). For the major community applications we forecast of a Xeon Phi node and the projected 2017 performance of a high-bin dual-socket Xeon node are very similar – and the Xeon Phi node *costs substantially less* than the Xeon node as well as using *substantially less power*. In Stampede 2, we will not view the Xeon Phi as an "accelerator", rather as *a more cost and power efficient processor for achieving similar node performance*.

Despite these advantages, we recognize that programming complexity is increasing; this is a significant challenge for all architectures today, and a challenge facing all of scientific computing. However, the complexities inherent in the manycore future appear to be inevitable, at least through the initial round of exascale deployments. The Stampede 2 architecture offers users a manageable transition from the programming model of today while placing them on a viable path to an exascale future that is designed to enable – and reward – incremental performance investments in existing software applications.

4 STAMPEDE 2 ARCHITECTURE

In its final form, Stampede 2 will consist of 5,932 total compute nodes, a 28PB storage subsystem and 24 additional login and management servers on an Intel Omni-Path interconnect. The compute nodes are based on two different processors with 1,728 nodes with dual-socket Intel Xeon Skylake (SKX) processors and 4,204 nodes with an Intel Xeon Phi Knights Landing (KNL) bootable processor, combining 3700 new nodes from this proposal with 500 nodes anticipated in the Stampede upgrade. These compute nodes are configured to meet TACC's requirements for the solicitation and provide a unique, innovative, and scalable system to support the broad user community.

The compute nodes will be housed in 107 racks in the system, 75 racks will contain 56 KNL compute nodes each and the remaining 32 racks will hold 54 SKX compute nodes. Each compute rack will also include two Omni-Path 48-port leaf switches and two 48-port gigabit Ethernet switches with 10-gigabit uplink to aggregation switches. Eight more racks will hold login, data transfer, management, and storage servers described below. Six wider racks will house the core Omni-Path Director Class Switches and cabling. Total system power for the complete system will be just over 4MW.

4.1 Intel "Knights Landing" Processor

The Knights Landing (KNL) processor that will be used for this system has 68 cores. This processor has four hardware threads per core (like KNC), and two 512-byte vector units per core. The vector units will use the AVX512 instruction set that will be in use in Xeon, offering full ISA compatibility with the latest Xeon processors. The package, including the processor and on package memory (see below) consumes a maximum of 200W. The "peak" performance is approximately 3 teraflops. KNL features two kinds of memory. On the KNL package itself, 16GB of high performance MCDRAM will be incorporated with a bandwidth in excess of 500GB/sec. In addition, the processor supports six channels of 2400MHz DDR-4 RAM. In Stampede 2 we populated each of the six channels with a single 16GB DDR-4 DIMM for maximum performance, for a total of 96GB of DDR-4 RAM and bandwidth exceeding 102GB/s. These two memories can be configured to use in one of three modes:

- Cache Mode: The MCDRAM acts as a direct mapped cache, to the DDR-4, 96GB is available to the application.
- Flat Mode: The MCDRAM is a separate, user-managed high speed portion of the address space. All 112GB of RAM is available to the application.
- Hybrid mode: A fraction of the MCDRAM (25 or 50%) is set up as cache, with the rest available as user-managed address space.

We support all three memory modes on the new Stampede 2 system (though Cache Mode is most popular to date, with a significant minority of highly tuned applications using Flat Mode).

4.2 Intel "Skylake" Processor

The next generation Intel Xeon processor, codenamed Skylake (SKX), builds upon the already successful Intel Xeon processor family including the Sandy Bridge, Haswell, and Broadwell lines of processors. Manufactured on Intel 14nm process technology, this new processor provides another leap in performance and energy efficiency beyond its predecessors [[11]]. With tight integration of all cores on a single chip and unified memory controller, significant performance improvements including more memory bandwidth and faster buses can be achieved. The new processor also doubles the peak FLOPS per cycle compared to currently available processors through a combination of fused multiply-add (FMA3) instruction and Intel Advanced Vector Extensions 3.2 (AVX512), an extension to previous AVX and AVX2 instructions. Note that the instruction set for this new processor is compatible with the KNL processor except for a few specialized uncommon instructions, so binary executables will be able to run on both processors.

The specific SKX processor selected for Stampede 2 is the 145W processor thermal design power (TDP) part.. Each node will have a two-socket motherboard with 192GB DDR-4 RAM, local hard drive, and Omni-Path PCIe card along with a shared data/management Gigabit Ethernet interface.

4.3 High-Performance Interconnect

Tightly coupled scientific applications require a high-bandwidth, low-latency network, and Stampede 2 system provides a 100 Gb/s Intel Fabrics Division Omni-Path network to support all inter-node application communications (e.g. MPI messages and shared file system transfers). The Omni-Path network is the Intel follow-on to the proven QLogic TrueScale InfiniBand technology acquired by Intel. Architecturally, the interconnect is a fat tree topology using six 768-port Director Class core switches, with each capable of more than 75 TB/s of bandwidth. Two 48-port Omni-Path leaf switches are installed into every compute rack, each with 28 nodes connected to it and the remaining 20 ports uplinked to the core switches for a marginal oversubscription of 7:5. The I/O servers have full non-blocking connectivity to ensure maximum network bandwidth is available to the storage subsystem. The remaining login and support nodes will also connect with non-blocking connectivity to the core switches, however, no MPI communication traffic will run across these uplinks.

4.4 Disk I/O Subsystem

The storage subsystem for Stampede 2 is based on Seagate's ClusterStor product, originally developed by Xyratex. The most important new features for Stampede 2 are GridRAID, parity declustered RAID, to provide better performance to a single target and improved rebuild times after drive failure and automatic active/active failover of the servers for high availability. The storage subsystem will consist of six meta-data servers 35 ClusterStor SSUs, with each SSU providing 82 10TB drives and two SSDs for external journals and RAID bitmaps with a total capacity of 28PB raw. The storage subsystem will be divided into two filesystems, $HOME and $SCRATCH. The $HOME filesystem will consist of two MDS and two SSUs with a usable capacity of almost 1.2PB and 20GB/s of aggregate bandwidth. $SCRATCH will use four MDS and leverage Lustre's distributed meta-data feature, DNE, along with 33 SSUs to provide a usable capacity of 20PB and overall aggregated bandwidth of 330GB/s with enough meta-data capacity to support four billion files. All servers will be connected into the Omni-Path fabric with non-blocking connectivity to provide maximum bandwidth to the rest of the system.

In addition to the $HOME and $SCRATCH filesystems, the TACC global filesystem, Stockyard, will be mounted as the $WORK filesystem on Stampede 2. This 25PB DataDirect Networks Lustre-based filesystem is already in use on the current Stampede and delivers more than 100GB/s of bandwidth.

4.5 Non-volatile memory

The innovative component in Stampede, the first-generation Xeon Phi, became a part of the proposed system with the introduction of KNL as the primary processor in Stampede 2 compute nodes. In this new system, while there won't be an innovative component at the scale of the original Stampede, we still will deploy a new experimental capability at small scale. We will introduce the use of Phase Change Memory in the memory/storage hierarchy of Stampede 2 through Intel's "Apache Pass" (AP) non-volatile DIMMs. These devices are persistent like disk (across power losses and reboots), but have performance much closer to traditional memory (DRAM).

On 50 of the Skylake compute nodes, we will add 4 512GB AP DIMMs (2 per socket) for a total of 2TB per node and 100TB in the system. The AP devices can be used in multiple modes; in some modes they look like normal memory, either using the conventional DDR DRAM as a cache or managed by the application (in much the same way as the MCDRAM on the KNL interacts with DDR DRAM). In other modes, the AP devices can be configured as a block device and treated as permanent storage.

This component will enable a number of use cases; using these nodes as very large memory nodes, as nodes with hyper-speed local storage, or for experiments in memory resilience or burst buffer capability. These nodes could serve several functions – for users with large memory requirements, that can't use distributed memory or fit in the 192GB per compute node, they can act as (albeit slower) 2TB large memory nodes. A number of users who make use of local disks on nodes for checkpoints or out-of-core file space will see substantial performance improvements. We also anticipate there will be a number of additional novel uses.

4.4 System Layout and Phased Deployment

The new system is being installed in the location where the current Stampede resides with a phased install plan. The entire system will consist of 121 total racks in six rows.

The first phase of deployment, in the spring of 2017 (ongoing during this writing), will install the KNL nodes and the new I/O capabilities. The second phase, in the fall of 2017, will add the SKX nodes and complete the management infrastructure. The final phase will add the non-volatile memory in 2018.

The phased deployment of Stampede 2 into the space currently occupied by Stampede will minimize the downtime and maximize the availability of the existing system. Stampede racks have been incrementally removed as new racks arrive leaving substantial portions of the existing system up and available to run jobs. As of the writing of this manuscript, space has been made for all of the phase 1 racks in the datacenter, but more than 60% of Stampede 1's capacity remains online. Phase 1 will be moved to full production before additional decommissioning begins to start phase 2 installation. Even during this phase, almost 30% of the original system will remain in production, meaning both Xeon and Xeon Phi nodes will always be available to users. The only expected full outage of both Stampede 1 and 2 will be a brief outage to move the UPS power connections from the I/O system of 1 to the new I/O system of 2, when users officially transition permanently to the new filesystems.

As of March 13, 2017, all but 9 of the phase 1 racks have been put in place on the datacenter floor, and cabling of these racks to the first two core switches has been completed. More than one thousand nodes have been installed with the software image, and MPI jobs have successfully run between these nodes across the Omnipath interconnect. The /home filesystem is also installed and available to the compute nodes. Early user testing is just a few weeks away, and phase 1 remains scheduled to go into production

in early June, 2017. Phase 2 activities should begin in July, with completion scheduled for October 1 of 2017.

5 STAMPEDE "1.5" A BRIDGE FROM 1 TO 2

The original Stampede plan called for an upgrade of a number of KNC cards in the compute nodes with second generation KNL cards. At the time that plan was crafted (early 2011), there was no expectation that the KNL would be available as a stand-alone processor, at least not in the Stampede 1 timeframe. When it was determined that the self-hosted version of Xeon Phi would be available with KNL, and that this version would be available before the PCI card version, we made the decision to reduce the number of KNLs in the planned upgrade, but deploy them in completely new servers as self-hosted compute nodes, rather than run new cards in the old nodes. In May of 2016, we ultimately deployed 508 new KNL nodes in lieu of the original plan for 1600 replacement cards. We believe this approach provided a number of advantages:

No offload model is required. The KNL is programmed as a standard processor. This removes substantial complexity from the programming model, including the need to account for the overhead of transferring data from the host to the coprocessor.

Removal of the memory limitation. The PCI card version of the KNC has only 8GB of local RAM; a substantial barrier to running in native mode for many users. With 16GB of on-package RAM and 96 additional GB of DDR4, the KNL will have about 1.5GB per core, as much RAM per core as traditional HPC systems. The small number of KNL nodes we deployed had substantially more total memory than the original upgrade would have contained.

No proxy structure to access the network. The current KNC cards require a layer of proxy software to relay requests to achieve good performance to the network and I/O capabilities through the existing host nodes. The stand-alone solution will remove this requirement, and provide native speed access to the network and the network-accessed filesystems.

No embedding in an old platform. The original upgrade plan called for the KNL processors cards to be installed in the original Stampede Sandy Bridge nodes, which at the time of installation would have been over three years old. In the revised plan, the KNL processors were deployed in a modern platform more matched to the capabilities of this new processor. These nodes had 4x the DDR RAM per node of the original host nodes, and the RAM is DDR4 at 2400Mhz vs. DDR3 at 1866Mhz. The Omni Path network deployed in the new nodes is twice the bandwidth of the FDR InfiniBand in the original nodes. New hardware also means higher reliability.

Less disruption for installation. The original plan called for inserting cards into one-quarter of the existing nodes of the system. While we would execute this in a rolling fashion, nonetheless there would have been significant downtime for portions of the current system. The revised plan allowed us to do the complete deployment without affecting existing nodes, leaving all of the current system production during deployment.

Existing nodes can be in use ***concurrently.*** In the Stampede 1 environment, use of the Xeon Phi processor occupies the entire node, either using the Xeon processor as the host for offload, or leaving it idle. Conversely, jobs using only the Xeon processors leave the embedded Phi coprocessor idle. The revised plan meant additional nodes, so KNL jobs can run while every Xeon (or KNC) in Stampede remained in use by a different job.

A Bridge to Stampede 2. Once it was determined that Stampede-2 would also make use of KNL nodes, the upgrade became a natural bridge between Stampede 1 and Stampede 2 – and to other leadership systems at DOE, in Japan, and in Germany that also are employing KNL as a self-hosted processor.

The deployment of Stampede "1.5" with 508 KNL nodes and the Omnipath network ultimately became risk mitigation for Stampede 2, allowing the operations team to build the software stack and build experience with these technologies, and also allowing the user community to begin the migration months before Stampede 2 would become available.

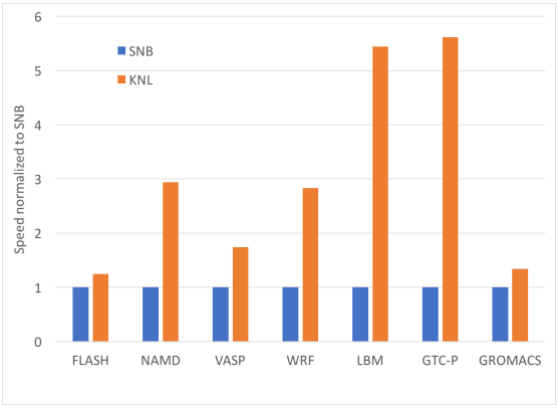

Figure 5: Stampede 1 and Stampede 2 code performance

4 PERFORMANCE RESULTS FROM KNL AND OMNIPATH

While Stampede 2 is still undergoing deployment, the availability of the 508 KNL nodes from Stampede "1.5" has allowed us to gather and present some basic performance data on both the Xeon Phi nodes as well as the Omnipath fabric. For the data in figure 5, we show the relative performance of Stampede 1 to Stampede 2 compute nodes for a number of the most popular community applications on the system – NAMD, VASP, WRF, and GROMACS. Each of these codes are run in an unmodified form as from their source repositories. We also show results for two (also unmodified) codes from the workload that aren't as widely used, but are known to have been optimized for both vectorization and hybrid OpenMP/MPI parallelization – LBM (A Lattice-Boltzman Method implementation) and GTC-P (the Princeton Gyrokinetic Toroidal code). What we see in the figure largely matches our expectations – for every code selected, with no modification the code will run successfully on a Xeon Phi node, and the user will see some performance benefit, ranging from as little as 20% to 2-3x for typical scientific codes. However, for codes that can truly exploit instruction, thread, and task level

parallelism, speedups of 5-6x are possible. As with Stampede, we expect that the system will be widely used with good results – but effort will continue to need to be applied to use it *effectively,* as the march continues towards ubiquitous massive parallelism in all processors. We hope that the continued increase of complexity of the conventional Xeon chip – as well as all other processors – will continue to push scientific codes to adopt the techniques that will benefit both Xeon and Xeon Phi.

Figure 6 shows a result from the Omnipath switch fabric in 1.5. Unlike Stampede 2, this fabric is built entirely from 48 port leaf switches, not the large core switches that will be in the finished machine. The figure shows a comparison of two MPI stacks, Intel MPI and MVAPICH, and the achievable bandwidth at different message sizes. For larger message sizes, the tests approach the theoretical 12.5GB/s achievable by the fabric. While space is not available in this work for a more comprehensive look at the fabric, we have in general found that Omnipath scales similarly to the Infinband fabrics of Stampede 1 and similar systems.

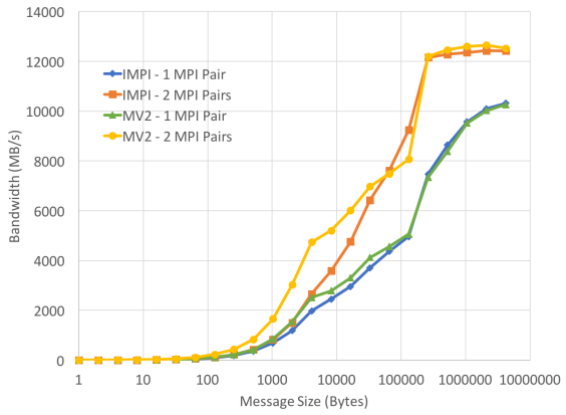

Figure 6: OmniScale Bandwidth for Intel MPI and MVAPICH vs. Message Size on Stampede "1.5' fabric.

4 CONCLUSIONS

As observed by Council on Competitiveness in a recent report on the critical dependence of the United States for a robust investment in advanced computing [[12]], "high performance computing is inextricably linked to innovation, fueling breakthroughs in science, engineering, and business." The nation's investment in HPC is a foundational element of efforts to secure US leadership in defense, science, and industry.

The Stampede project has been a key piece of this national investment, and continues to deliver value to scientists. Stampede enabled Aleksei Aksimentiev at UIUC to explore a cutting-edge method of DNA sequencing that uses an electric field to drive a strand of DNA through a small hole, or "nanopore," either in silicon or a biological membrane. Said Aksimentiev, "Stampede is by far the best computer systems my group has used over the past 10 years." Philipp Moesta and Christian D. Ott from Caltech succeeded in performing the first fully general-relativistic 3D magnetohydrodynamics simulations of progenitor stars that are believed to lead to energetic, jet-driven supernova explosions. Their findings show that the simulations behave very differently in full, unconstrained 3D compared to previous models. Said Ott, "Stampede really helped push our simulations to the limit. Our research would have been practically impossible without it."

Scientists from NREL are using Stampede to determine how certain enzymes break down cellulose (plant cell walls) to improve biofuel production. In a paper published in Proceedings of the National Academy of Sciences, they modeled a new enzyme that could significantly speed up the process. Said NREL Engineer Gregg Beckham: "Stampede has been an absolutely essential resource for our group to examine biological and chemical catalysts important for the production of renewable transportation fuels".

Stampede 2 will carry on in the tradition of Stampede 1, and continue to deliver science results to XSEDE users for the next four years – and continue to push the community to embrace new many-core technologies as we move closer to exascale.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation, through the Stampede (ACI-1134872), Stampede 2 (OAC-1540931), and XSEDE (ACI-1953575) awards.

REFERENCES

[1] NSF Cyberinfrastructure Council. National Science Foundation, January 20, 2006, http://www.nsf.gov/od/oci/ci_v5.pdf.

[2]John Towns et al, , "XSEDE: Accelerating Scientific Discovery", Computing in Science & Engineering, vol.16, no. 5, pp. 62-74, Sept.-Oct. 2014, doi:10.1109/MCSE.2014.80

[3] Community Input on the Future of High Performance Computing, NSF Workshop Report, December 2009.

[4] J. T. Oden, O. Ghattas, et al., Cyber Science and Engineering: A Report of the NSF Advisory Committee for Cyberinfrastructure Task Force on Grand Challenges, National Science Foundation, 2011 (to appear).

[5] J. Decker, et al., *Exascale Workshop Panel Meeting Report*, Department of Energy, January 19-20, 2010.

[6] Committee on the Potential Impact of High-End Computing on Illustrative Fields of Science and Engineering, *The Potential Impact of High-End Capability Computing on Four Illustrative Fields of Science and Engineerin*g, National Research Council, 2008.

[7] S. Glotzer, S. Kim, et al., *International Assessment of Research and Development in Simulation-based Engineering and Science*, World Technology Evaluation Center (WTEC) Panel Report, January 2009.

[8] R. Babich, M.A. Clark, B. Joó, "Parallelizing the QUDA Library for Multi-GPU Calculations in Lattice Quantum Chromodynamics", *Proceedings of SC10*, ACM/IEEE, New Orleans, LA, 2010.

[9] Heinecke, Barth, et al. 2014. Petascale high order dynamic rupture earthquake simulations on heterogeneous supercomputers. In *Proceedings of SC14*. IEEE Press, Piscataway, NJ, USA, 3-14. DOI=10.1109/SC.2014.6 http://dx.doi.org/10.1109/SC.2014.6

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

[11] Intel® Xeon Processor E5 Family, http://www.intel.com/content/www/us/en/processors/xeon/xeon-processor-e5-family.html

[12] Council on Competitiveness. "The Exascale Effect: the Benefits of Supercomputing Investment for U.S. Industry." October 2014.