Executive Summary

America's nuclear stockpile of weapons is aging, yet they must be certified as effective without physical testing. That is the challenge of the Department of Energy's National Nuclear Security Administration (NNSA). Without underground testing, scientists must test, validate, and certify the nuclear stockpile using a combination of focused above-ground experiments, computational science and simulations that run on large computing clusters. The Commodity Technology System 1 (CTS-1) deployment has over 15 Penguin Computing, Inc. supercomputer clusters consisting of 83 Tundra Scalable Units (SU). These clusters incorporate Intel Xeon processors and many other Intel technologies, and Asetek DCLC direct chip cooling technologies. All clusters began deployment at the NNSA's tri-labs in early 2016, providing the computational power for scientists to carry out their missions. At the center of these powerful clusters is Intel Omni-Path Architecture (Intel OPA) connecting the large numbers of nodes together. This
paper looks at the need for these systems and why Intel OPA was chosen as a foundational component.

Protecting America—Safely

The U.S. Department of Energy (DOE) is responsible for maintaining, certifying, and developing America’s nuclear weapon stockpile without relying on live, underground nuclear testing, which the U.S. ended in 1992. Predicting the behavior of a nuclear weapon detonation—both those very old and new—requires understanding the behavior of a very complex system that possibly changes over years and decades of material storage. Instead of integrated underground nuclear tests, DOE scientists must now rely on a combination of focused ground tests and high performance computing to certify the stockpile and assure its continued reliability and safety. High Performance Computing (HPC) helps simulate effects on the ultimate performance of the weapons, which can no longer be tested, and it helps to interpret the results of focused experiments, which also provide validation for the simulation tools. These same simulation tools and HPC clusters are also being used in support of other Department of Defense missions, such as modeling of Navy railguns and blast effects on military assets from conventional weapons.

Stockpile Stewardship Through Computational Science

Under the DOE’s National Nuclear Security Administration (NNSA), and the NNSA’s Stockpile Stewardship Program (SSP), simulations are done using a range of supercomputing resources that are funded by the NNSAs Advanced Simulation and Computing (ASC) program and deployed at several national laboratories. Three of these laboratories (the tri-labs) are Lawrence Livermore National Laboratory (LLNL) in California, Los Alamos National Laboratory (LANL) in New Mexico, and Sandia National Laboratories (SNL), also in New Mexico and California.

The Commodity Technology Systems (CTS-1), the most recent of NNSA’s tri-lab commodity cluster procurements, are the daily workhorses for the NNSA’s SSP. These clusters are built by Penguin Computing on Intel® Architecture, including Intel® Xeon® processors E5 v4 Product Family and Intel® Omni-Path Architecture (Intel® OPA) fabric. (CTS is an evolution of earlier clusters called the Tri-Lab Linux Capacity Computing (TLCC) clusters deployed in 2007 and 2011.) CTS systems provide tens to hundreds of millions of core-hours each year to tri-labs' scientists.

Central to the computational science carried out largely on the CTS-1 clusters are uncertainty quantification studies for a wide range of parameters to assess the efficacy and safety of the stockpile in production and operation. These studies run several types of codes that have been developed at the national laboratories, from multi-physics and materials science codes, to others that are focused on a particular area, such as hydrodynamics, radiation transport, and electronic structure, among others. The process of uncertainty quantification requires iterating hundreds to thousands, and even tens of thousands, of the same simulation with slightly different parameters, since many of the parameters are not precisely known, to determine overall uncertainties. Hundreds of scientists run their parametric studies, requiring massive amounts of computational core hours throughout the year. Through these studies, scientists can better understand characteristics and properties of the physics and focus efforts on the particular models that are identified as the largest contributors to uncertainties in an effort to reduce those uncertainties.

The Commodity Technology Systems

The work done on the CTS-1 systems is a stepping stone to larger simulations. The CTS-1 parametric studies help scientists fine tune these larger simulations, which are run on massive supercomputers, such as Sierra, being sited at LLNL, and Trinity at LANL.

But, the CTS-1 systems are very powerful HPC clusters, able to deliver as much as 3.2 petaFLOPS. Since the first installation of CTS-1 in April of 2016, another 14 clusters of various sizes have been installed across all of the tri-labs. They range in size from 32 nodes to a few thousand nodes.

All the CTS-1 systems have been built by Penguin Computing headquartered in Fremont, California U.S.A., using Intel Xeon Processors (codenamed Broadwell architecture) and Intel OPA fabric for HPC. Many of the CTS-1 systems have taken top-ranked positions in the Top500.org’s list of the fastest supercomputers in the world. As of June 2017, Livermore has Jade #45 at 2.6 petaFLOPS, Quartz #46 at 2.6 petaFLOPS, Topaz #224 at 0.8 petaFLOPS; Los Alamos has Grizzly #77 at 1.5 petaFLOPS, Fire #105 at 1.1 petaFLOPS, Ice #106 at 1.1 petaFLOPS, and at Sandia, Cayenne is #97 at 1.2 petaFLOPS, Serrano #98 at 1.2 petaFLOPS, and Dark Ghost #223 at 0.8 petaFLOPS.

Scalable Computing at the Tri-labs

In addition to the mission of the NNSA, the tri-labs have supported over the years other laboratory and collaborative projects with industry and academia that require HPC. Projects vary in size, and thus, the tri-labs' scientists require a range of computing capacity from project to project. Beginning with the TLCC systems, the tri-labs defined HPC clusters to be sized in Scalable Units (SU), providing a prescribed computing building block from which any size system could be configured and procured. Besides the SU, TLCC and CTS-1 systems share a common theme: being built from commodity components (servers, storage, memory, and fabric) that keep costs down while providing high levels of computing performance for the daily workloads these systems need to run. Using the SU building block concept and commodity components, the tri-labs can define their scalable, high-performance computing needs for each project, and then quickly deploy clusters to meet their needs at reasonable cost.
Building Scalable Systems with Intel® Omni-Path Architecture and Intel® Architecture

Scalable computing is a concept that Intel has used to help design and develop HPC systems for decades. Today, Intel® Scalable System Framework (Intel® SSF) integrates the components of scalable computing—compute, memory/storage, fabric, and software—with the Intel® Xeon® Scalable processors and multi-integrated core (MIC) Intel® Xeon Phi™ Processor, high-density Intel® Optane Memory, Intel® HPC Orchestrator software for HPC clusters, and the next-generation interconnect, Intel Omni-Path Architecture HPC fabric (see the sidebar “Intel Approach to Scalable Computing for HPC”). A key part of balancing the performance of the framework is the Intel OPA fabric.

Driving Next-Generation Fabric Performance and Reliability

HPC fabric technologies have been evolving for decades. InfiniBand® Architecture has dominated the HPC space for almost 20 years. But, it was not originally designed for HPC, which largely uses the Message Passing Interface (MPI) for high-performance parallel computing. Intel Omni-Path Architecture is a next-generation fabric—complete with host interface, switches, and cabling—that is specifically designed for HPC workloads. Intel OPA is designed to deliver high message injection rates and low latency, while providing full bisectional bandwidth across very large clusters. Compared to the latest evolution of 100 Gbps InfiniBand EDR, the 38 HPC systems on the June 2017 Top500 list with Intel OPA deliver 1.7 times the petaFLOPS of EDR—or a total of 67.1 petaFLOPS, which is ten percent of the total Top500 Rmax performance for June 2017.

48-Port Switch Silicon Enables Simple, Large, and Fast Clusters

Core to the fabric architecture is a 48-port radix switch silicon developed by Intel. This large port count component enables HPC architects to design large clusters that are less complex, with fewer switches and still able to deliver full bisectional bandwidth across more nodes than current alternative technologies allow. More ports and fewer switches allow designers to create networks with fewer devices to manage, which can reduce the cost of large clusters. Intel OPA has enabled many large, unique HPC configurations to recently be built.

Performance and Reliability for HPC

Intel OPA has shown to outpace InfiniBand EDR in Intel testing with lower latencies and higher message rates. Summarized results show that:

- Equal bandwidth—Both Intel OPA and InfiniBand EDR are capable of delivering nearly full wire rate of 100 Gbps.
- Dramatically higher message rate—Intel OPA has 64 percent higher message rate than InfiniBand EDR without message coalescing at the MPI level. This is a true hardware message rate test without message coalescing in software.
- Much lower MPI latency—Intel OPA latency is 11 percent lower than InfiniBand EDR, including a switch hop for both Intel OPA and InfiniBand EDR.
- Lower Natural Order Ring latency—Intel OPA has lower latency than InfiniBand EDR with 16 fully subscribed nodes using 32 MPI ranks per node.
- Significantly lower Random Order Ring latency—Intel OPA shows significantly lower latency at 16 fully subscribed nodes using 32 MPI ranks per node.

To deliver a next-generation HPC fabric that is enabling some of today’s and tomorrow’s fastest supercomputers, Intel OPA’s designers focused on the nuances of intelligent data movement—reducing latency, while ensuring packet and fabric integrity and scalability at speed. A key innovative approach is in Intel OPA's link layer, which is designed to deliver low and deterministic latency, penalty-free packet integrity, and resilient operation in the event of a lane failure. Additionally, Intel OPA architecture is built to intelligently determine the best method of data transport available to it—setting up an RDMA channel using the host adapter or using the fast and efficient resources of the Intel Xeon processor to transfer data. The objective is to feed the application without delaying the flow.

Intel OPA in the Tri-labs—and Beyond

Prior to Intel OPA, Intel offered the Intel® True Scale adapters for InfiniBand, which included technologies that were later incorporated into the Intel OPA host fabric interface. The TLCC clusters incorporated Intel True Scale adapter and switch technologies. Thus, the tri-labs had early experience with the performance and phenomenal scalability of the technologies that were later incorporated into Intel OPA. According to Matt Leininger, Deputy for Advanced Technology Projects at LLNL, the tri-labs sought such continuing scalability, and, along with benchmarking of Intel OPA and consideration of technical and scheduling risks, Intel OPA was put at the forefront of choices for the fabric in the CTS-1 procurements.

With the new CTS clusters, Leininger expects to see a two- to four-fold performance improvement over the TLCC2 systems, running multiple jobs faster than the previous commodity clusters.

The beginning of 15 CTS-1 cluster installations across the tri-labs was among the first Intel OPA installations in 2016. Others include the Bridges supercomputer at Pittsburgh Supercomputing Center, a 1.3 petaFLOPS system; Stampede2 at the Texas Advanced Computing Center, a 1.5 petaFLOPS system, TX-Green at the Massachusetts Institute of Technology, a 1.7 petaFLOPS machine, and Oakforest-PACS at the Joint Center for Advanced HPC in Japan, a 25 petaFLOPS machine. Stampede2, TX-Green, and Oakforest-PACS are all based on Intel® Xeon Phi™ Processors. The massive Aurora supercomputer project at Argonne National Laboratory (ANL) will also run on Intel OPA.
Penguin Computing Designs for Density

With the early TLCC systems, an SU’s 162 nodes were configured in traditional racks. With CTS-1, the size of the SU increased to 192 nodes while the density of the racks changed considerably with Penguin Computing’s Tundra ES system configuration. Tundra ES is built on the Open Compute Project rack standard, which defines rack spaces that are a little wider and a little shorter, allowing more hardware into a rack. Penguin estimates they can increase density for HPC systems by as much as 50 percent, according to Sid Mair, Senior VP of their Federal Systems Division. And, with the Intel OPA 48-port radix silicon, they can connect 1,000 nodes on two layers of fabric instead of a more complex configuration with smaller switches for the same node density. That means Penguin can deliver CTS-1 systems that offer greater performance and consume less physical space.

OCP and Penguin’s Tundra ES also incorporate a 12-volt power system that disaggregates the power supplies from the equipment in the rack. Power shelves convert building power to 12-volt rack power, which is then distributed to all the shelves. A central power conversion and distribution architecture makes it much more efficient to manage and control consumption, which affects the demand for cooling. To accommodate Tundra’s 12-volt power rails, Intel designed a 12-volt version of Intel OPA Switches for the CTS-1 builds. These same switches would later be incorporated into Tundra ES clusters used in the Penguin Computing on Demand (POD) cloud HPC service.

Summary

The CTS-1 clusters at the tri-labs, built by Penguin Computing with Intel Omni-Path Architecture, are the daily workhorses of the NNSA’s scientists to help certify America’s nuclear stockpile. Intel OPA is helping to advance the performance and scalability of these HPC clusters. Intel OPA has been proven in the tri-labs and across other installations since it was first introduced in early 2016, taking more positions on the Top500 list and delivering greater petaFLOPS than the alternative 100 Gbps network technology.

Intel's further developments around Intel OPA will enhance the fabric and the processors used in HPC. For example, versions of Intel Xeon Processors and Intel Xeon Phi Processors will integrate the fabric interface into the chips, eliminating the need for a PCIe card and eliminating the latency an extra hop induces. Intel's commitment to HPC is manifested in its continuing developments to help move HPC to the exascale computing era.

The tri-labs scientists and the CTS-1 clusters are imperative to certifying and protecting America’s nuclear stockpile, as well as other nuclear and non-nuclear national security concerns. Using computational science on advanced HPC clusters, America can also help keep the world safe by eliminating physical testing and experimentation. HPC helps enable and fulfill the missions of the National Nuclear Security Administration’s Stockpile Stewardship Program.
Learn More

You may find the following resources useful:

- **Partner Company:**
  www.penguincomputing.com

- **Other Useful Resource:**
  Intel Scalable System Framework

---

1. www.top500.org/lists/2017/6
4. [https://computation.llnl.gov/computers/commodity-clusters](https://computation.llnl.gov/computers/commodity-clusters)

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more information go to [intel.com/technology](https://intel.com/technology) performance

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

Cost reduction scenarios described are intended as examples of how a given Intel® based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.


No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software, or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer, or learn more at [www.intel.com/hpc](https://www.intel.com/hpc).

Copyright © 2017 Intel Corporation. All rights reserved. Intel, Intel Xeon, Optane, Phi, and the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries.

* Other names and brands may be claimed as the property of others.