Whitepaper

Summit and Sierra Supercomputers:

An Inside Look at the U.S. Department of Energy's New Pre-Exascale Systems

November 2014

Contents

New Flagship Supercomputers in U.S. to Pave Path to Exascale with GPUs

U.S. Department of Energy and CORAL

The U.S. Department of Energy (DoE) has a long history of driving advances in high performance computing. Over the last two decades, half of the machines ranked number one on the Top500 list have been U.S. DoE systems used for large-scale research. CORAL (Collaboration of Oak Ridge, Argonne, and Livermore)[1] is the agency's latest procurement of pre-exascale supercomputing systems, which will continue its leadership in using the largest and fastest computers in the world to accomplish its demanding mission.

Under the CORAL collaboration, Oak Ridge National Laboratory (ORNL), Argonne National Laboratory (ANL), and Lawrence Livermore National Laboratory (LLNL) will each deploy powerful pre-exascale supercomputers. These systems will provide the computing power required to meet the mission of the Office of Science and the National Nuclear Security Administration (NNSA) of the U.S. DoE.

In November 2014, the U.S. DoE announced that ORNL and LLNL have selected NVIDIA® GPU-accelerated systems based on the IBM OpenPOWER platform. These systems, *Summit* and *Sierra*, will serve as the model for future exascale designs.

Introducing *Summit* and *Sierra*

The ORNL *Summit* system will be a leadership computing platform for the Office of Science. Delivered in 2017, *Summit* is expected to reach between 150 and 300 petaFLOPS and is positioned as a precursor to the U.S. DoE's exascale system.

As the lead federal agency supporting fundamental scientific research across numerous domains, the Office of Science is chartered to meet the insatiable need for computing resources by researchers and scientists. *Summit* will carry on the tradition set by *Titan*, ORNL's current GPU-accelerated supercomputer, which is among the world's fastest supercomputers today.

The LLNL *Sierra* supercomputer will be the NNSA's primary system for the management and security of the nation's nuclear weapons, nuclear nonproliferation, and counterterrorism programs. In support of the complex mission of its Advanced Simulation and Computing program, LLNL has had numerous top-5 supercomputers, including most recently *Sequoia*, an IBM Blue Gene/Q system. *Sierra* will replace *Sequoia* and is expected to deliver more than 100 petaFLOPS, over 5x higher compute performance than its predecessor.

In this white paper, we will explore key features of these new supercomputers and how those technologies, enabled by the Tesla® accelerated computing platform, will drive the U.S. DoE's push toward exascale.

Under the Hood of *Summit* and *Sierra*

Summit and *Sierra* are poised to surpass numerous technological and performance barriers, setting a historical milestone in the road to exascale computing. While the two systems will have unique system

[1] https://asc.llnl.gov/CORAL/

configurations based on their application requirements, they will share an efficient and scalable heterogeneous node architecture that tightly integrates IBM POWER CPUs with NVIDIA GPUs using NVIDIA NVLink™ high-speed coherent interconnect technology.

NVLink is an energy-efficient, high-bandwidth path between the GPU and the CPU at bandwidths of 80-200 GB/s, or 5 to 12 times that of the current PCIe Gen3 x16, delivering faster application performance. NVLink merges the CPU and GPU memory spaces, allowing developers to use the right processor for the right job while sharing data at high throughput.

Figure 1: *Summit* and *Sierra* supercomputers at a glance. The two systems share the same system and node architecture based on GPU-accelerated OpenPOWER platform.

Summit features more than 3,400 compute nodes, enough to deliver a peak performance between 150 and 300 petaFLOPS, and is expected to deliver more than five times the system-level application performance of *Titan* while consuming only 10% more power. Each compute node includes multiple next-generation IBM POWER9 CPUs and multiple NVIDIA Tesla® GPUs based on the NVIDIA Volta architecture. Each node is expected to deliver more than 40 TFLOPS, which is enough to outperform an entire rack of Haswell x86 CPU servers. In fact, just four nodes in *Summit* system would be powerful enough to qualify for the Top500 list, as of June 2014.

Each compute node in *Summit* will be equipped with over 512 GB of coherent memory, including large-capacity DDR4 system memory and ultra-fast HBM stacked memory on the GPU. All data will be directly addressable from either the CPU or the GPU, an important feature enabled by the NVLink interconnect. Extending the impressive amount of memory is an additional 800 GB of NVRAM per node, which can be configured either as a burst buffer or as extended memory. The system is interconnected with dual-rail Mellanox EDR InfiniBand, using the full, non-blocking fat-tree design. *Summit* will also include the GPFS parallel file system with 1 TB/s of I/O bandwidth and 120 PB of disk capacity.

Table 1: Key features of the new *Summit* supercomputer.

System Features	*Summit* Supercomputer
Peak System Performance (FLOPS)	150-300 petaFLOPS
Peak Node Performance (FLOPS)	> 40 teraFLOPS
# of Nodes	> 3400 compute nodes
CPU	IBM POWER9
GPU	NVIDIA® Volta
Memory per Node	> 512 GB (DDR4 system memory + stacked memory)
NVRAM per Node	800 GB
Node Interconnect	NVIDIA® NVLink™
System Interconnect	InfiniBand Dual Rail EDR (23 GB/s)
File System	IBM Elastic Storage using GPFS technology, 120 PB
Peak System Power Consumption	10 MW (10% more than *Titan*)

Four Key Technologies Drive U.S. DoE Decision

The HPC industry faces a myriad of options when building a system: CPU architecture (ARM, POWER, or x86), accelerator type, the self-hosted model or the heterogeneous model, and an industry-standard or proprietary network. After analyzing various proposals, the U.S. DoE decided to build *Summit* and *Sierra* supercomputers based on the IBM OpenPOWER architecture and the Tesla accelerated computing platform.

An important set of technologies proved pivotal in the U.S. DoE's decision to build around the Tesla platform. These technologies are:

- Next-generation IBM OpenPOWER platform
- NVIDIA Tesla accelerator platform
- The Heterogeneous computing model
- NVIDIA NVLink high-speed GPU interconnect

The following sections analyze the latter two technologies in more detail.

Heterogeneous Computing Model

Today's most successful, scalable HPC applications distribute data and work across the nodes of a system and organize algorithms to operate as independently as possible on millions or billions of data elements. However, even simple applications can transition many times between periods of throughput-intensive parallel calculations and sections of latency-sensitive serial operations.

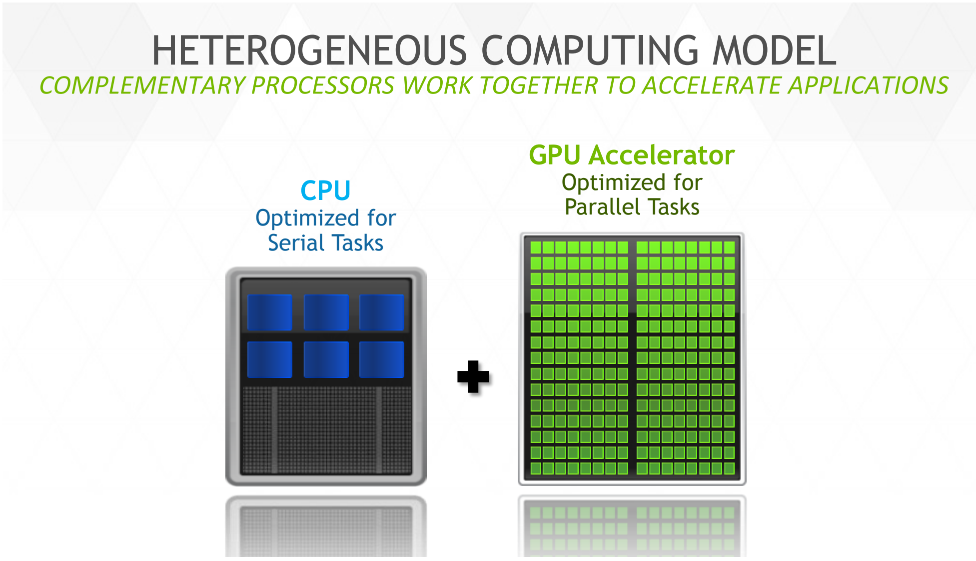

Figure 2: Heterogeneous computing model enables various workloads to run on the best-suited processor.

The ideal computing engine would be capable of optimal performance during both extremes of throughput-oriented and latency-sensitive computations, but the realities of physics make this impossible to achieve with a single processor architecture and technology design point. For uncompromising performance, a heterogeneous architecture coupling powerful latency-optimized processors with highly parallel throughput-optimized accelerators can significantly outperform non-specialized, homogeneous alternatives.

Maximum Efficiency: The Heterogeneous Computing Model

The architectural emphasis on parallelism in GPUs leads to optimization for throughput, hiding rather than minimizing latency. Support for thousands of threads ensures a ready pool of work in the face of data dependencies in order to sustain performance at a high percent of peak. The memory hierarchy design and technology thoroughly reflect optimization for throughput performance at minimal energy per bit.

By contrast, latency-optimized CPU architecture drives completely different design decisions. Techniques designed to compress the execution of a single instruction thread into the smallest possible time demand a host of architectural features (like branch prediction, speculative execution, register renaming) that would cost far too much energy to be replicated for thousands of parallel GPU threads but that are entirely appropriate for CPUs.

The essence of the heterogeneous computing model is that one size does not fit all. Parallel and serial segments of the workload execute on the best-suited processor – latency-optimized CPU or throughput-optimized GPU – delivering faster overall performance, greater efficiency, and lower energy and cost per unit of computation.

ORNL and LLNL chose to build the *Summit* and *Sierra* pre-exascale systems around this powerful heterogeneous compute model using technologies from IBM and NVIDIA. IBM's POWER CPUs are among the world's fastest serial processors. NVIDIA GPU accelerators are the most efficient general purpose throughput-oriented processors on the planet. Coupling them together produces a highly efficient and optimized heterogeneous node capable of minimizing both serial and parallel sections of HPC codes. *Summit* and *Sierra* will utilize NVLink, a new on-node interconnect, as the node integration platform.

NVLink High-Speed GPU Interconnect

NVLink is an energy-efficient, high-bandwidth communications channel that uses up to three times less energy to move data on the node at speeds 5-12 times conventional PCIe Gen3 x16. First available in the NVIDIA Pascal™ GPU architecture, NVLink enables fast communication between the CPU and the GPU, or between multiple GPUs.

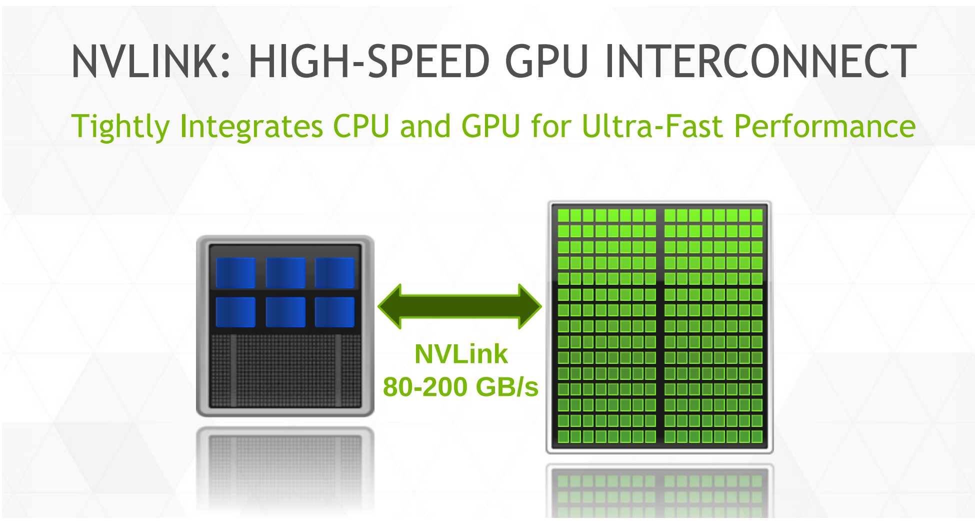

Figure 3: NVLink is a key building block in the compute node of *Summit* and *Sierra* supercomputers.

NVLink is a key technology in *Summit*'s and *Sierra*'s server node architecture, enabling IBM POWER CPUs and NVIDIA GPUs to access each other's memory fast and seamlessly. From a programmer's perspective, NVLink erases the visible distinctions of data separately attached to the CPU and the GPU by "merging" the memory systems of the CPU and the GPU with a high-speed interconnect. Because both CPU and GPU have their own memory controllers, the underlying memory systems can be optimized differently (the GPU's for bandwidth, the CPU's for latency) while still presenting as a unified memory system to both processors.

NVLink offers two distinct benefits for HPC customers. First, it delivers improved application performance, simply by virtue of greatly increased bandwidth between elements of the node. Second, NVLink with Unified Memory technology allows developers to write code much more seamlessly and still achieve high performance.

NVLink also provides design flexibility, allowing for systems with different ratios of parallel and serial performance. This flexibility allows sites to design and build systems that are optimized for their expected workload in order to minimize the impact of Amdahl limits and maximize the system performance.

A recent study illustrates the impact of NVLink GPU-to-GPU connectivity on real-world applications, even if the CPU is connected via PCIe. The study assumes future-generation GPUs with performance higher than that of today's GPUs. Our application performance models project QUDA and AMBER to be 40-50% faster with NVLink, while algorithms that are communication-heavy will see a speedup of more than 2x.

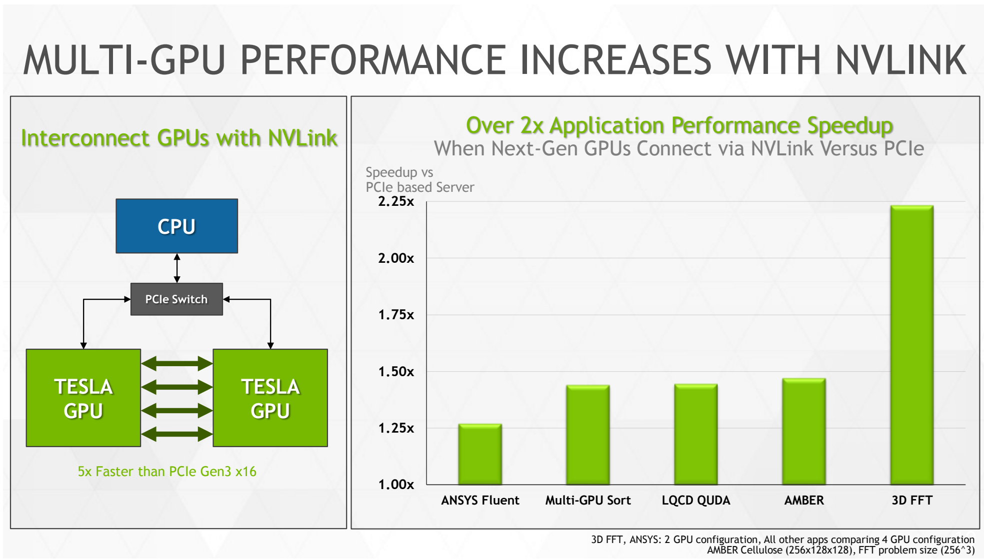

Figure 4: Multi-GPU applications benefit from NVLink-interconnected GPU-to-GPU communication.

CORAL Application Benchmarks

Paving the path to exascale is an important milestone, but accelerating the scientific work that ultimately leads to discoveries is the main reason why the U.S. DoE will deploy *Summit* and *Sierra*. Every year, computing facilities like *Titan* are oversubscribed by researchers who depend on supercomputing resources to further their science. Significant numbers of researchers and scientists are turned away every year due to the shortage of compute cycles.

The mission of *Summit* and *Sierra* is to fuel innovations in numerous research areas, from discovering new biofuels to developing new materials to managing the nation's nuclear program. To meet this goal, the U.S. DoE outlined selection criteria for vendors, which required them to submit an exhaustive study on the system-level performance of various application benchmarks, representing a wide range of scientific disciplines, such as quantum molecular dynamics, hydrodynamics, and radiation transport.

The chart below shows performance projections for benchmarks that were considered of highest priority to the CORAL procurement, also known as "TR-1" benchmarks. These projections are based on system-wide application performance estimates, comparing the proposed CPU+GPU system to a CPU-only system of similar cost. For the scalable science benchmarks, which test scalability at the full-system level, the CPU+GPU system is expected to deliver up to 13x higher performance. For the throughput benchmarks, which test the system's ability to run multiple ensemble jobs simultaneously, the CPU+GPU system is expected to deliver up to 12x higher throughput.

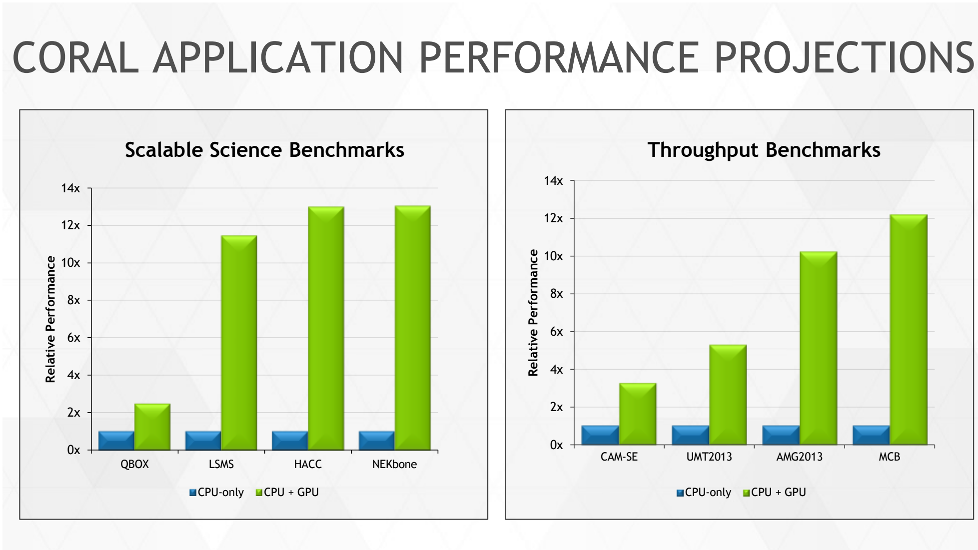

Figure 5: CORAL benchmark projections show GPU-accelerated system is expected to deliver substantially higher performance at the system level compared to CPU-only configuration.

The demonstration of compelling, scalable performance at the system level across a wide range of applications proved to be one of the key factors in the U.S. DoE's decision to build *Summit* and *Sierra* on the GPU-accelerated OpenPOWER platform.

Conclusion

Summit and *Sierra* are historic milestones in HPC's efforts to reach exascale computing. With these new pre-exascale systems, the U.S. DoE maintains its leadership position, trailblazing the next generation of supercomputers while allowing the nation to stay ahead in scientific discoveries and economic competitiveness.

The future of large-scale systems will inevitably be accelerated with throughput-oriented processors. Latency-optimized CPU-based systems have long hit a power wall that no longer delivers year-on-year performance increase. So while the question of "accelerator or not" is no longer in debate, other questions remain, such as CPU architecture, accelerator architecture, inter-node interconnect, intra-node interconnect, and heterogeneous versus self-hosted computing models.

With those questions in mind, the technological building blocks of these systems were carefully chosen with the focused goal of eventually deploying exascale supercomputers. The key building blocks that allow *Summit* and *Sierra* to meet this goal are:

- The Heterogeneous computing model
- NVIDIA NVLink high-speed interconnect
- NVIDIA GPU accelerator platform
- IBM OpenPOWER platform

With the unveiling of the *Summit* and *Sierra* supercomputers, Oak Ridge National Laboratory and Lawrence Livermore National Laboratory have spoken loud and clear about the technologies that they believe will best carry the industry to exascale.

Notice

ALL INFORMATION PROVIDED IN THIS WHITE PAPER, INCLUDING COMMENTARY, OPINION, NVIDIA DESIGN SPECIFICATIONS, REFERENCE BOARDS, FILES, DRAWINGS, DIAGNOSTICS, LISTS, AND OTHER DOCUMENTS (TOGETHER AND SEPARATELY, "MATERIALS") ARE BEING PROVIDED "AS IS." NVIDIA MAKES NO WARRANTIES, EXPRESSED, IMPLIED, STATUTORY, OR OTHERWISE WITH RESPECT TO MATERIALS, AND EXPRESSLY DISCLAIMS ALL IMPLIED WARRANTIES OF NONINFRINGEMENT, MERCHANTABILITY, AND FITNESS FOR A PARTICULAR PURPOSE.

Information furnished is believed to be accurate and reliable. However, NVIDIA Corporation assumes no responsibility for the consequences of use of such information or for any infringement of patents or other rights of third parties that may result from its use. No license is granted by implication or otherwise under any patent or patent rights of NVIDIA Corporation. Specifications mentioned in this publication are subject to change without notice. This publication supersedes and replaces all information previously supplied. NVIDIA Corporation products are not authorized for use as critical components in life support devices or systems without express written approval of NVIDIA Corporation.

Trademarks

NVIDIA, the NVIDIA logo, Pascal, NVLINK, and Tesla are trademarks or registered trademarks of NVIDIA Corporation in the United States and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

Copyright