

2.1 Summit and Sierra: Designing AI/HPC Supercomputers

James A. Kahle¹, Jaime Moreno², Dan Dreps³

¹IBM Research, Austin, TX

²IBM Research, Yorktown Heights, NY

³IBM Systems and Technology, Austin, TX

The Summit and Sierra Supercomputer Systems, deployed in 2018 at the Department of Energy (DOE) National Laboratories, Oak Ridge (ORNL) and Lawrence Livermore (LLNL), provide a significant increase in computing capability relative to the predecessors, and represent a major step in the path to Exascale computing. Arising from the DOE CORAL program (<https://asc.llnl.gov/CORAL/>), Summit and Sierra were listed as #1 and #2, respectively, in the November 2018 version of the Top500 list (<https://www.top500.org/lists/2018/11/>).

Supercomputers such as Sierra and Summit have unique design challenges in network bandwidth, communication patterns, synchronous coordination and parallel computation not found in traditional cloud deployments. While Supercomputers and Data Centers are both large collections of interconnected compute nodes installed in a single location, they differ in the style of computing they perform. In particular, Supercomputers are characterized for synchronous computation on “bare metal” resources, coordinated parallel work, message passing among parallel tasks, intensive computation periods that can last many hours or even days so that job duration may be longer than the reliability window. In contrast, general Data Centers use virtualized images, interacting components driven by external events, information exchange used to drive work, leading to massive short-lived computation pieces and software structures designed for failures.

Driven by the high efficiencies demanded by large supercomputer machines and the slowdown of Moore’s law of scaling, the CORAL systems address the design challenges with a heterogeneous approach. Applications have a diverse set of needs that can be addressed with strong CPUs for analytics capabilities and powerful GPUs for massively parallel sections. Examples of analytics capabilities well suited to CPUs are exemplified by complex codes with data-dependent paths, lots of indirection and pointer chasing, and dependency on latency of the memory subsystem, as found in oil and gas reservoir simulation, AI and graph analytics. Examples of massively parallel compute which are well suited to GPUs are exemplified by simple kernels, dense FP operations, and simple data access patterns as found in oil and gas seismic simulation, financial value at risk and image analytics. Tying diverse compute engines together through a high bandwidth interconnect and state-of-the-art memory systems was key in meeting application needs.

To address the explosion of data and associated computational demands, IBM has created a data centric systems approach to these demands with four guiding principles: 1) Minimize data motion, 2) enable compute at all levels of the system hierarchy, 3) modularity of system components, 4) application-driven design. With OpenPOWER [1], IBM created an environment for modularity with its partners: NVIDIA supplied the GPU for highly parallel work, and Mellanox provided the high-speed interconnect for the system.

Figure 2.1.1 describes the scale and compute power of these two systems, whereas Fig. 2.1.2 provides an overall overview of the components in them. To illustrate the data movement capabilities of the larger Summit system, the data starts in a GPFS parallel file system comprised of 32,648 10TB hard disks in 39 Racks. Data flows to the compute nodes via a network with a cross sectional bandwidth of 115.2PB/s. For each compute node (see Fig. 2.1.3), there is 25GB/s of injection bandwidth feeding two POWER9 processors, each with 150GB/s of memory BW to 512 GB of DRAM. A high speed NVLink 2.0 with 150GB/s of bandwidth feeds the Tesla V100 GPUs that have 900GB/s of B/W to the 16 GB of HBM2 memory.

The POWER9 processor [2-3], depicted in Fig. 2.1.4, supports high single-thread computation, runs the operating system, and coordinates data movement for the system. The high-speed POWER bus supports memory movement between the system DRAM and the NVLINK. The heterogeneous mix between the CPU and GPU allows a diverse set of applications to efficiently run on the system.

To keep the processor node at a suitable temperature, a cold-plate technology was deployed [4]. Warm water first enters the system through rack-level connectors, flows over the POWER9 processors and then to the Tesla V100 GPUs. The rest of the components in the node are cooled with forced air that can be further cooled with a rear door heat exchanger. The CPU and GPUs consume 300W each, leading to a 2.4KW 2U drawer, 58.5KW rack, for a total system power under 15MW.

To maintain optimal performance with the 300W power budget, the POWER9 processor implements dynamic clock and voltage control with an embedded

PowerPC 405 on-chip controller and 20 additional microcontrollers distributed around the chip. Multiple sense points monitor for any voltage droop and can lengthen the clock pulse within a few cycles to avoid any timing conditions. While maintaining minimal timing margins, the processor achieves consistent performance at an application basis.

A key feature of CORAL nodes is the high-performance NVLink 2.0 bus [5] depicted in Fig. 2.1.5, which connects the CPU with the GPUs. NVLink 2.0 uses 25Gb/s coherent I/O, enabling seamless GPU/CPU integration, key to effectively achieving higher performance in a productive programming environment with full coherency between the CPU and GPU.

To enable the capabilities described above for NVLink 2.0, two copies of a 1.24Tb X24 bidirectional processor link macro were integrated into the POWER9 chip, in 14nm SOI, running at 25.8Gb per lane, using 6.3 mm² and consuming 4.3W. The X24 link has 24 lanes of RX and TX respectively. One low-noise LC PLL drives two resonant clock distributions to the RX bank and TX bank. This physical layout increases isolation of TX to RX. Both the RX and TX run off the same 1.0V supply, while the PLL has an AV_{DD} of 1.5V. Multiple copies of this core are used on a single die. The RX PHY is compatible with several short-reach industry standards. The core has full diagnostics, eye monitor capability and some enhanced features like TDR embedded.

The 25.8Gb/s transmitter achieves 1.3pJ/b at 1.0V supply with active pre-cursor and post-cursor feed-forward equalization (FFE), using a full rate source series terminated driver. The serializer utilizes a resonant clock to combine 16b of data into serial main data and FFE data. The FFE data can consist of pre-cursor, post-cursor, or both, depending on the needs of the product. FFE can be disabled if not needed by the channel, reducing power of the pre-drive buildup path. A simplified block diagram of the transmitter is depicted in Fig. 2.1.5.


Two classes of benchmarks were specified in the CORAL program: *scalable science* benchmarks, which represent single applications expected to run across the entire system (e.g., across 4200 or 4600 nodes) for which performance scalability is a major challenge, and *throughput* benchmarks, which represent the simultaneous execution of multiple instances of the same application or a set of different applications, each using 216 or 192 nodes.

Furthermore, two versions of each benchmark were specified in the CORAL program: *baseline*, which only allowed adding high-level code directives for performance (such as OpenMP or OpenACC program directives), and *optimized*, which allowed code restructuring, architecture-specific code optimizations, or any other coding technique to increase performance. *Baseline* is intended to give a metric of “out-of-the-box” performance of the systems, whereas *optimized* targets demonstrate the ultimate capabilities of the systems.

Figure 2.1.6 depicts performance measurements on the benchmarks and their geometric mean. These values are improvement ratios with respect to the prior-generation systems installed at the DOE labs, namely IBM BlueGene/Q at LLNL and Cray Titan at ORNL. The measurements collected show the *optimized* code performing 5.9-to-7.8× better than the reference systems (geometric mean across the benchmarks in each class), whereas the *baseline* code delivers 2.4-to-6.0× better performance, amply fulfilling the goals of the CORAL program. Additional performance improvements should be expected from further application tuning based on greater experience using the systems. Moreover, usage of early, partial, systems has already resulted in significant performance gains on scientific applications [6]. Moreover, early experiments leveraging NVLink in applications that require moving large data in/out from the GPUs are also reporting major performance gains, for example [7].

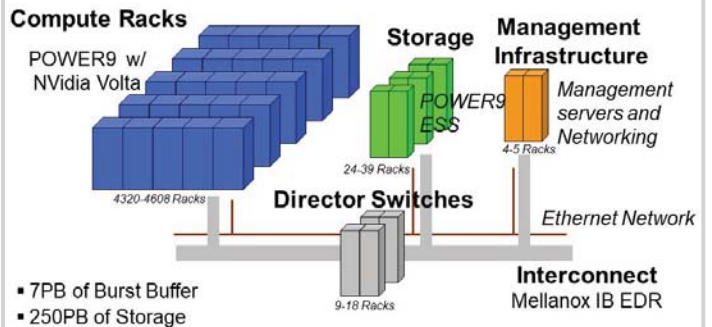
References:

- [1] <https://openpowerfoundation.org/>
- [2] C. Gonzalez, et al., “The 24-Core POWER9 Processor with Adaptive Clocking, 25-Gb/s Accelerator Links, and 16Gb/s PCIe Gen4,” *IEEE JSSC*, vol. 53, no. 1, pp. 91-101, 2018.
- [3] B. Thompto, “POWER9: Processor for the Cognitive Era,” *IEEE Hot Chips 28*, pp. 1-19, 2016.
- [4] S. Chun, et al., “IBM POWER9 Package Technology and Design,” *IBM J. of Research and Development*, vol. 62, no. 4/5, 2018.
- [5] IBM-POWER9-NPU team, “Functionality and Performance of NVLink with POWER9 Processors,” *IBM J. of Research and Development*, vol. 62/4-5, 2018.
- [6] T. Straatsma, “Early Application Results on Summit”, *ORNL Smoky Mountains Conference*, Sept. 2018.
- [7] <https://developer.ibm.com/linuxonpower/2018/07/27/tensorflow-large-model-support-case-study-3d-image-segmentation/>



	Summit	Sierra
Peak Performance	200 Petaflops	125 Petaflops
Number of Nodes	4608	4320
Node Performance	43 Teraflops	29 Teraflops
Memory per node	512 GB DDR4, 96 GB HBM2	256 GB DDR4, 64 GB HBM2
NV Memory per node	1600 GB	1600 GB
Total System Memory	11.1PB DDR4+HBM2+NV	9PB DDR4+HBM2+NV
Compute nodes	9,216 IBM POWER9™ CPUs 27,648 NVIDIA Volta™ GPUs	8640 IBM POWER9™ CPUs 17280 NVIDIA Volta™ GPUs
File System	250 PB, 2.5 TB/s, GPFS™	156 PB, 1.5 TB/s, GPFS™
Power Consumption	15 MW	12 MW
Interconnect	Mellanox EDR 100G InfiniBand	
Operating System	Red Hat Enterprise Linux (RHEL) version 7.4	

Figure 2.1.1: Summit and Sierra: next-generation AI and HPC platform.



- 7PB of Burst Buffer
- 250PB of Storage

Figure 2.1.2: The complete architecture build out.

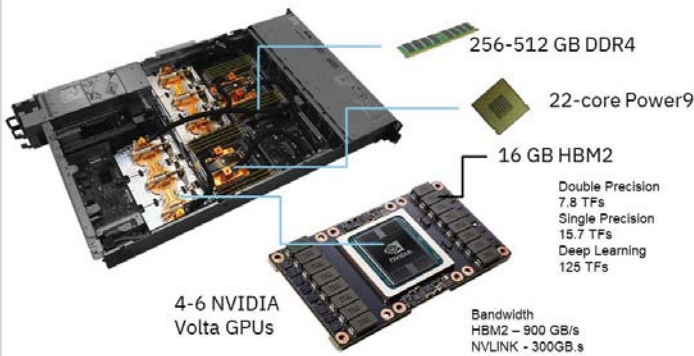


Figure 2.1.3: Compute planar – heterogeneous fat node.

- 24 cores Physical Design
 - 22 cores for yield optimization in CORAL
- 14nm FinFET SOI technology
 - 8 billion transistors
- PCI-Express 4.0
- NVLink 2.0
- Large Caches
 - L1/D: 32 KiB per core, 8-way set assoc.
 - L2: 258 KiB per core
 - L3: 120 MiB eDRAM, 20-way

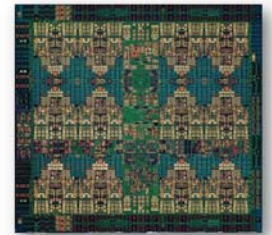


Figure 2.1.4: IBM POWER9 processor.

- Next generation of GPU/CPU integration
- Key to achievable acceleration performance
- High productivity programming environment

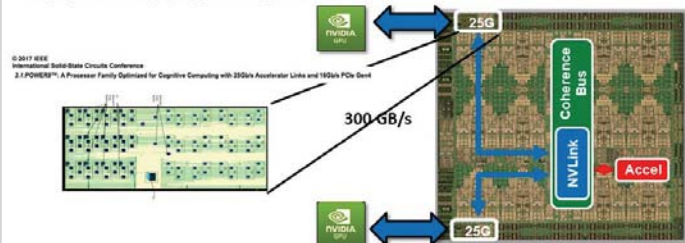


Figure 2.1.5: NVLink 2.0 25G coherent I/O.

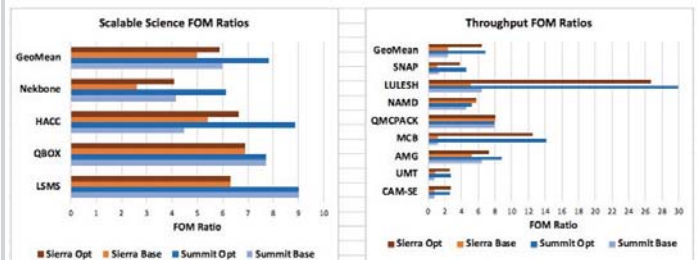


Figure 2.1.6: Performance results.