1. **Introduction**

As of November 2018, the IBM Summit supercomputer (known herein as “Summit”) is the world’s most powerful supercomputer. Located at the Oak Ridge National Laboratory (ORNL) in Tennessee, the Summit was the first system to surpass an exaflop (amounting to 1.435 exaflops, or 143,500 trillion floating-point operations per second) outside of theoretical performance [1]. The Summit utilizes a non-blocking dual-rail Mellanox interconnect, allowing for both storage transfer and inter-node processing [2]. All nodes are capable of coordinating via Mellanox’s dual-rail network, which provides immense potential for parallelization across nodes.

The Summit was created with multi-core processing in mind, specifically to find solutions for problems related to graph analytics, artificial intelligence, and machine learning [1]. For example, ORNL has found that the Summit was able to generate longer, more detailed simulations of supernovae events [1], compared to ORNL’s previous computational systems.

The overall system weighs 340 tonnes and spans 5600 square feet [3], which is roughly equivalent to one-tenth of a football field.



Figure 1: The front chassis of the IBM Summit, located at ORNL [1].

1. **Architecture/System Overview**

The Summit uses 4,608 independent AC922 servers (or “compute nodes”) for computation, where each node is composed of two 22-core IBM Power9 processors and six NVIDIA Tesla V100 GPU accelerators (i.e., auxiliary GPUs with nearby memory) [5]. These processors are split between two sockets per node, therefore one CPU is directly paired with three GPUs. According to IBM [2], each node performs at approximately 42 teraflops (TFlops) per second. This brings the Summit to a theoretical total of 193,536 TFlops/s (Note: Top500 lists 200,795 TFlops/s [1].)Each of the two sockets on the compute node utilize NVLink, proprietary software maintained by Nvidia for the purpose of creating efficient communications between nearby CPUs and GPUs.

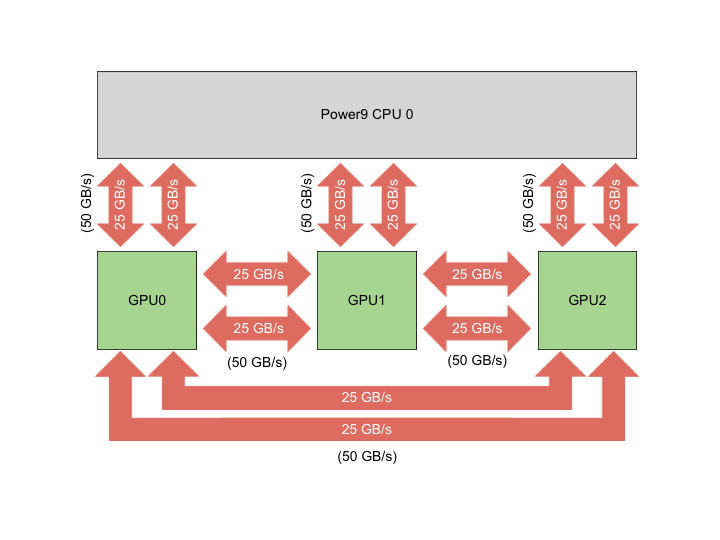


Figure 2: NVLink between a Power9 processor and Nvidia V100 GPUs on a single AC922 socket [5].

From above, we can see NVLink providing 50 GB/s bandwidth in one-direction, therefore allowing 100 GB/s of bi-directional bandwidth [5]. In conjunction with a PCIe 4th generation buses, this improves throughput by a factor of 5.6 per pairing, in comparison to 3rd generation PCIe [6].

While there are four variants of the Power9 processor, the type (SMT4, Scale-Out) included with each AC922 compute node is a DMA-enabled 24-core chip capable of 120 GB/s memory access time, is designed for use with two sockets, and optimized for Linux systems [7] (Summit uses Red Hat Enterprise 7.5 [5]).

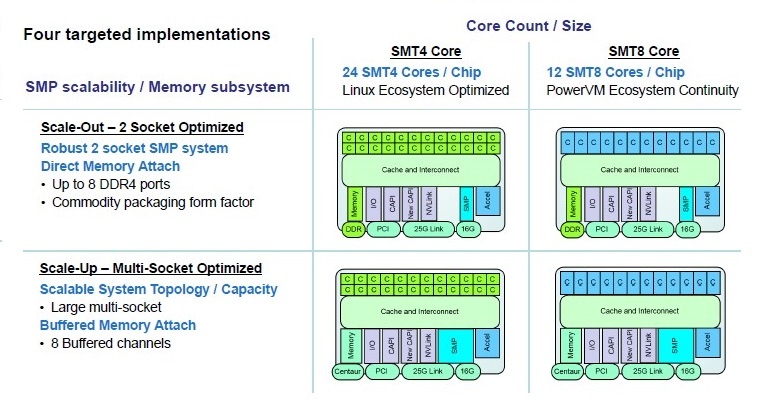


Figure 3: The four IBM Power9 variants [8]

The Power9 CPU uses IBM’s Power 3.0 instruction set architecture (ISA). The newest version of the ISA includes support for a piece of functionality named “Instruction Fusion”, which allows for the reinterpretation of certain sequences of instructions to improve performance [9]. Further, the ISA provides a modulus operator, and several new instructions which range from binary/decimal conversion and random number generation to instructions for string and character processing. Among other changes, there is improved support for multi-threaded programming (wait instruction, conditional branching), and a system call for passing one or more locations for program execution. The ISA can also support 128-bit signed integers, 128-bit floating-point integers (as per IEEE-754-2008 standard) and binary-coded decimal [9]. While providing flexibility to the programmer, the added instructions and formats for decimal numbers provide faster conversion to binary, vice-versa.

The Nvidia Tesla (codenamed “Volta”) V100 is Nvidia’s latest flagship processor. The V100 contains a 6MB L2 cache, 128kb shared memory, 21.1B (billion) transistors and 84 streaming multiprocessors (SM), while Nvidia’s former top-of-the-line card (the GP100) retains 15.5B transistors and 60 SMs [10]. Each SM totals 64 CUDA cores, inferring a difference of 1536 CUDA cores between both cards [10].

1. **Strengths / Parallelism**

Mentioned earlier, each compute node within the Summit is an IBM AC922 equipped with two 22-core Power9 CPUs and six Nvidia V100 GPUs, yet the server can be configured to use 16-core CPUs instead [6]. In order to service the plethora of applications maintained by the U.S government and ORNL research team, there are two other types of nodes (read: not compute nodes) meant to facilitate access and job scheduling, named “Login” and “Launch” nodes respectively [5]. Instead of using two 22-core CPUs and six GPUs, both nodes instead use two 16-core CPUs and four Nvidia GPUs. The login node gives users a space to compile their data and work and submit jobs to launch nodes. The launch nodes handle job scheduling (ideally in batches), which allow the compute nodes to purely focus on the given job(s). The benefit of having identical hardware architecture across all node types, is projects need only be built once (on the login node) [5]. In all, most nodes in the Summit supercomputer are compute nodes, where each house two 22-core SMT4 processors, which allows for a maximum of 176 hardware threads per compute node [5].

When working with multiple threads, each physical core of the Power9 can utilize hardware threading in slices of four and can delegate work amongst cores with three simultaneous multithreading modes: SMT4, SMT2, and SMT1. SMT4 has slices working independently (via OpenMP), SMT2 has slices paired, and SMT1 has all slices on a core working on the same execution stream or thread [5].

Since the main application of IBM’s Summit pertain to deep learning, data processing, graph analytics, and deep learning, a heavy emphasis is set on efficient reading and writing on a shared and adequately-sized storage system. To address this, IBM has built a “general parallel file system” (GPFS) that is accessible across multiple nodes [11]. This file system totals 120PB of space and is capable of 1 terabyte per second of I/O bandwidth. In addition, Summit is connected to ORNL’s high performance storage system for project and data archival, and Summit nodes include an extra 800GB of non-volatile random-access memory (NVRAM) [11].

Since each compute node has work parallelized across multiple types of processors, it is important for each component within a node to be able to share particular memory spaces and resources. Luckily, programmers can allocate a “unified memory” space for all processors within a node to access [5]. Normally, the entirety of the memory space is passed to the accessing processor, however, post-Pascal GPUs are capable of page faulting, and can instead load the desired pages instead. A combination of GPU page faulting and virtual addressing can allow for memory overcommitment. The result is programs using unified memory having access to the entire memory space, and the capacity to process extremely large data sets. Volta GPUs also support access counters and can be tuned to take advantage of spatial locality [5].

With the release of Volta GPUs, Nvidia has released a new type of core called “Tensor cores” [5,10]. These types of cores are essentially a large collection of ALUs reserved for matrix multiplication operations and can deliver four times the number of FLOPs for this operation [10].

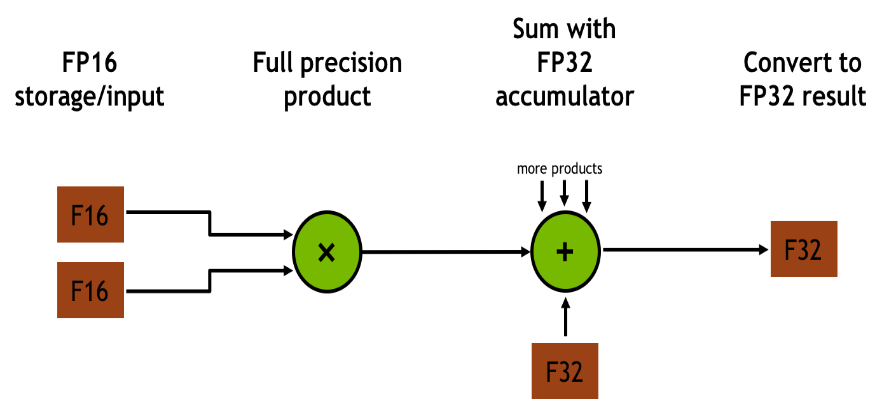


Figure 4: (D = AB + C) Matrix-multiplication conducted by each of Nvidia’s Tensor cores [5]

Introducing parallelism on this scale bestows a sizable burden on the programmer writing the application. Thankfully, IBM has released multiple toolsets to ease development and help conduct analysis. IBM has created a scalable math library that is capable of processing data on multiple nodes in parallel using SPMD (Single Program, Multiple Data) instructions. IBM has also developed a message passing and interfacing APIs to help reduce development time on applications that require inter-node communication [12]. Another in-house parallel performance toolkit can analyze FLOP performance, trace MPI applications, communication patterns, I/O analysis, OpenMP tracing, and hotspot analysis.

1. **Weaknesses**

One caveat with aforementioned login and launch node is that their resources are shared by other users, and any parallel or threaded jobs done on these nodes can disrupt other users. Unless there are safeguards to stop bad practices such as this, there is an assumption of competence placed on the users of Summit. Mismanagement could possibly cause data loss or disruption of services for users.

In a real-world deep learning application done on the IBM Summit [13], a research team determined a potential bottleneck when reading data from the GPFS via I/O. In a scenario where each GPU in the supercomputer was attempting to access a 20-40TB data set from the GPFS, the file system was unable to keep up. As a workaround, the research team distributed the needed data across several nodes (via NVRAM) before running their scheduled job(s). Currently, further research into scalable data stores is required in order to overcome this bottleneck.

Since this machine was created with parallelism specifically in mind, running serial jobs will waste potential computing power, costing precious time and therefore money. Learning how to appropriately parallelize your application will add to the learning curve of developing for the machine.

Despite supposedly using a 24-core version of the Power9 processor, the Summit instead uses a 22-core variant of the same processor [14].

**References:**

[1] <https://www.top500.org/lists/2018/11/>

[2] <http://www.mellanox.com/blog/2017/11/what-does-it-mean-to-summit/>

[3] <https://www.ibm.com/blogs/research/2018/06/summit/>

[4] <https://www.ornl.gov/news/ornl-launches-summit-supercomputer>

[5] <https://www.olcf.ornl.gov/for-users/system-user-guides/summit/summit-user-guide/>

[6] <https://www.ibm.com/ca-en/marketplace/power-systems-ac922/details>

[7] <https://www.nextplatform.com/2016/08/24/big-blue-aims-sky-power9/>

[8] <https://www.top500.org/news/ibm-ups-its-game-with-the-power9-processor/>

[9] <https://ibm.ent.box.com/s/1hzcwkwf8rbju5h9iyf44wm94amnlcrv>

[10] <https://www.anandtech.com/show/11367/nvidia-volta-unveiled-gv100-gpu-and-tesla-v100-accelerator-announced>

[11] <https://www.olcf.ornl.gov/wp-content/uploads/2014/11/Summit_FactSheet.pdf>

[12] <http://www.redbooks.ibm.com/redpieces/pdfs/sg248422.pdf>

[13] https://www.nextplatform.com/2018/10/09/hpc-file-systems-fail-for-deep-learning-at-scale/

[14] https://www.olcf.ornl.gov/wp-content/uploads/2018/12/summit\_workshop\_thompto\_smt.pdf