Jory Anderson (jba@uvic.ca)

Micheal Tom (mtom@uvic.ca)

Justin Bao (justinbao@uvic.ca)

Parm Johal (parmj@uvic.ca)

IBM Summit: A Comparative Analysis Amongst Supercomputers

CSC 350

Project 2

Spring 2019

I. IBM Summit (USA)

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 the American state of 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 ("November 2018"). The Summit utilizes a non-blocking dual-rail Mellanox interconnect, allowing for both storage transfer and inter-node processing (Biebelhausen). 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 (Rosenfield). For example, ORNL has found that the Summit was able to generate longer, more detailed simulations of supernovae events (Rosenfield), compared to ORNL’s legacy computational systems.

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

Fig. 1. The front chassis of the IBM Summit, located at ORNL. McCorkle, Morgan L. “ORNL Launches Summit Supercomputer.” *ORNL*, 8 June 2018, www.ornl.gov/news/ornl-launches-summit-supercomputer.

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) ("Summit User Guide"). These processors are split between two sockets per node; therefore, one CPU is directly paired with three GPUs. According to IBM (Biebelhausen), 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 as the theoretical maximum ("November 2018")). Each of the two sockets on Summit nodes utilize NVLink, a proprietary software maintained by Nvidia for the purpose of creating efficient communications between nearby CPUs and GPUs.

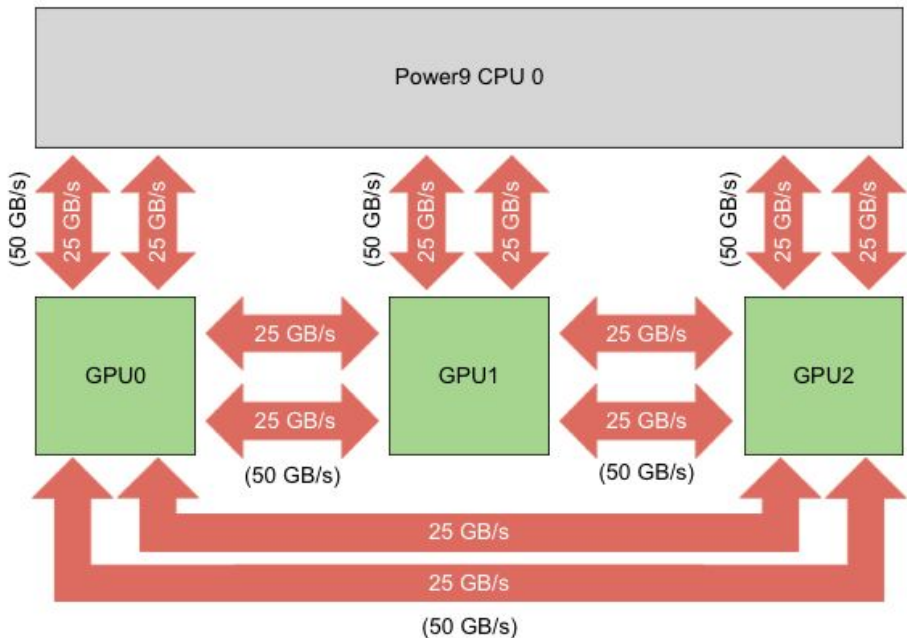

Fig. 2. NVLink connections between a Power9 processor and V100 GPUs in a single AC922 socket. "Summit User Guide." *Oak Ridge Leadership Computing Facility*, www.olcf.ornl.gov/for-users/system-user-guides/summit/summit-user-guide/.

From above, we can see NVLink providing 50 GB/s bandwidth in one-direction, thereby allowing 100 GB/s of bi-directional bandwidth ("Summit User Guide"). In conjunction with PCIe 4th generation buses, this improves throughput by a factor of 5.6 per CPU-GPU pairing, in comparison to 3rd generation PCIe ("IBM Power System AC922").

While there are four variants of the Power9 processor, the SMT4-type processors included in 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 is optimized for Linux systems (Morgan, see Fig 3), where the Summit uses Red Hat Enterprise 7.5 ("Summit User Guide"). In Fig. 3, you may notice a discrepancy between core counts. This is due to IBM shipping a 22-core variant for use with the Summit, instead of their standard 24-core processor.

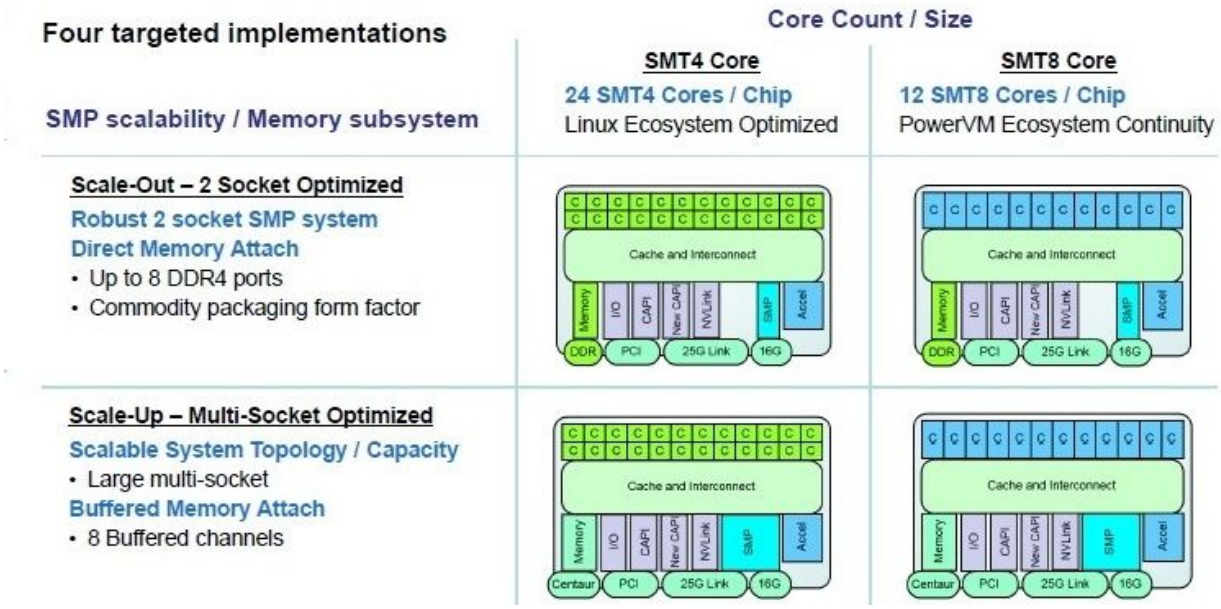

Fig. 3. The four IBM Power9 variants. Feldman, Michael. "IBM Ups Its Game with the Power9 Processor." *IBM Ups Its Game with the Power9 Processor | TOP500 Supercomputer Sites*, 25 Aug. 2016, 3:04, www.top500.org/news/ibm-ups-its-game-with-the-power9-processor/.

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 ("Power ISA Version 3.0 B"). 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 ("Power ISA Version 3.0 B"). While providing flexibility to the programmer, the new

instructions and additional decimal number formats provide faster conversion to binary, and vice-versa.

The Nvidia Tesla (codenamed “Volta”) V100 is Nvidia’s latest flagship processor. The V100 contains a 6MB L2 cache, 128kb shared memory, 21.1 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 (Smith). Each SM totals 64 CUDA cores, inferring a difference of 1536 CUDA cores between both cards (Smith).

Strengths / Parallelization

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 ("IBM Power System AC922"). 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 (beyond compute nodes) meant to facilitate access and job scheduling, named “Login” and “Launch” nodes respectively ("Summit User Guide"). Instead of using two 22-core CPUs and six GPUs, both nodes 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) ("Summit User Guide"). 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 ("Summit User Guide").

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 ("Summit User Guide").

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) called IBM Spectrum, that is accessible across multiple nodes ("Summit_FactSheet.pdf"). 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) ("Summit_FactSheet.pdf").

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 ("Summit User Guide"). 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 ("Summit User Guide").

With the release of Volta GPUs, Nvidia has released a new type of core called “Tensor cores” ("Summit User Guide"; Smith). 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 (Smith).

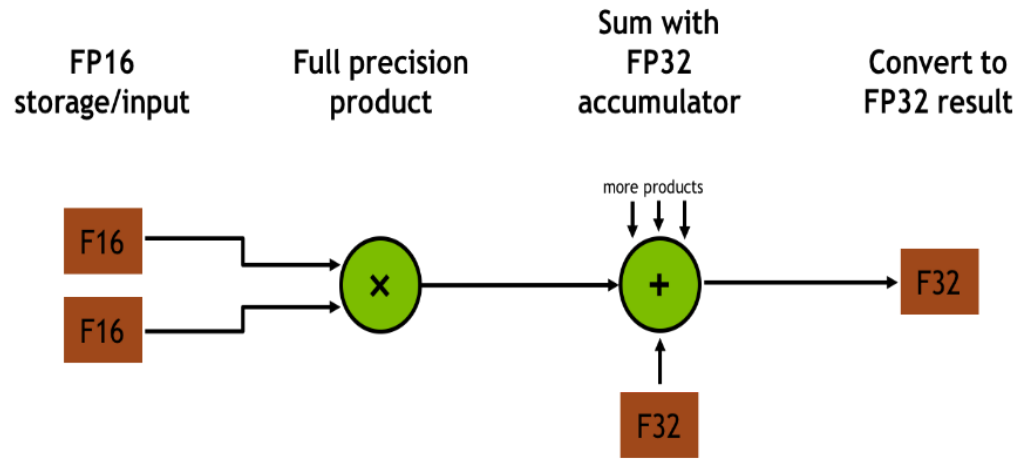

Fig. 4. (D = AB + C) Matrix-multiplication conducted by each of Nvidia’s Tensor cores. “Summit User Guide.” *Oak Ridge Leadership Computing Facility*, www.olcf.ornl.gov/for-users/system-user-guides/summit/summit-user-guide/.

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 message passing and interfacing APIs to help reduce development time on

applications that require inter-node communication (Quintero, et al). A secondary in-house parallel development toolkit can analyze FLOP performance, trace MPI applications, communication patterns, I/O analysis, OpenMP tracing, and hotspot analysis.

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 such bad practices, there is an assumption of competence placed on the users of the Summit. Mismanagement could possibly cause data loss or disruption of services for other users.

In a real-world deep learning application done on the IBM Summit (Hemsoth), 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 manually 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 in respect to the Summit will add to the learning curve of fully utilizing 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 (Thompto). This may or may not be a downside, though one should ask why such a design decision was made. Currently, there doesn't seem to be a documented reason for this change.

II. Sunway TaihuLight (China)

Introduction

While the Sunway TaihuLight (known herein as "Sunway") once stood on top of the computing world, it has been overthrown by the recent IBM Summit supercomputer. However, before the IBM Summit supercomputer came into the picture, China was leading the world in supercomputing and was sitting at rank number one for two years. Only to be overthrown in June of 2018 with the Summit's debut (Erich, Dongarra and Simon).

The Sunway is located in Wuxi, China at the National Supercomputing Center and was developed at the same center with the support of the National High-Technology Research and Development Program of China (The National Supercomputing Center). To date, the Chinese supercomputer ranks third in the Top500 rankings with a peak performance of 125,435.9 TFlops/s, and an observed performance of 93,014.6 TFlops/s (Erich, Dongarra and Simon).

Architecture / System Overview

The Sunway is comprised of 10,649,600 cores, 1,310,720 GB of memory, 230 TB of SSD storage and is supported by the Chinese Sunway SW26010 260C processors (The National

Supercomputing Center). Figure 5 shows the general architecture of the Sunway. It is comprised of 40 cabinets, with each cabinet containing four super nodes, and each super node is made up of 256 nodes. One of the key components of the Sunway are the processors, with each processor having a total of 260 processing elements in just one CPU and a potential performance of just over three TFlops (Fu).

The processor architecture can be found in Figure 6 of which it can be seen that each processor is composed of four core-groups (CG's), with each core having its own management processing element (MPE), a computing processing element (CPE) and a memory controller (Fu). Each MPE has a 32KB L1 instruction cache, 32KB L1 data cache and a 256 KB L2 cache which gets split between both data and instructions. On the other hand, each CPE has a 16 KB L1 cache and a user-controlled scratch pad memory (SPM). The SPM is essentially a cache that the user can tweak for whatever needs that they have, making the processors more flexible with different types of computations. Furthermore, each CG is connected to a network along with a storage system. The many-core processor runs at 1.45 GHz and implements DDR3 technology to deal with its data intake.

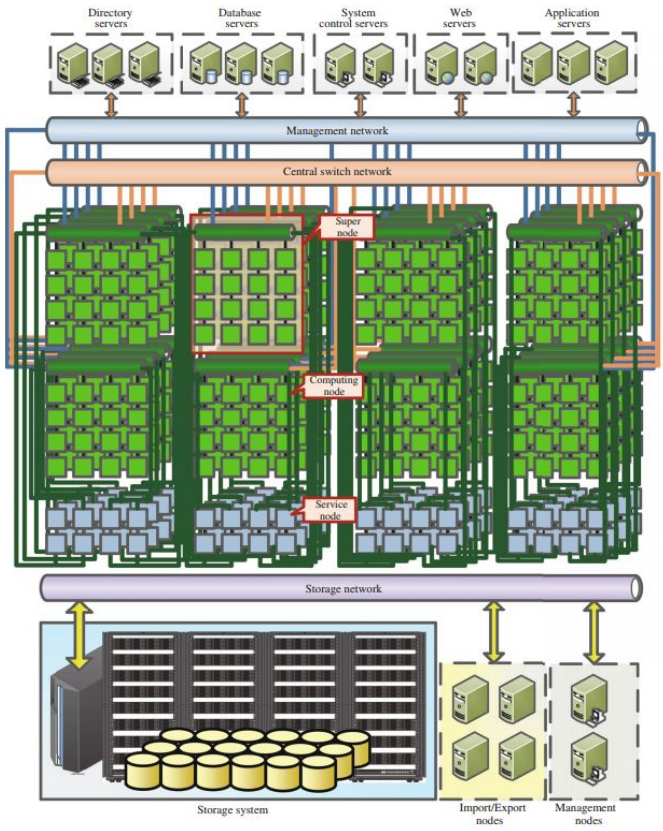

Fig. 5. *General architecture of the Sunway TaihuLight system***.** Dongarra, Jack. *Report on the Sunway TaihuLight System*. Technical Report. University of Tennessee: Department of Electrical Engineering and Computer Science, 2016.

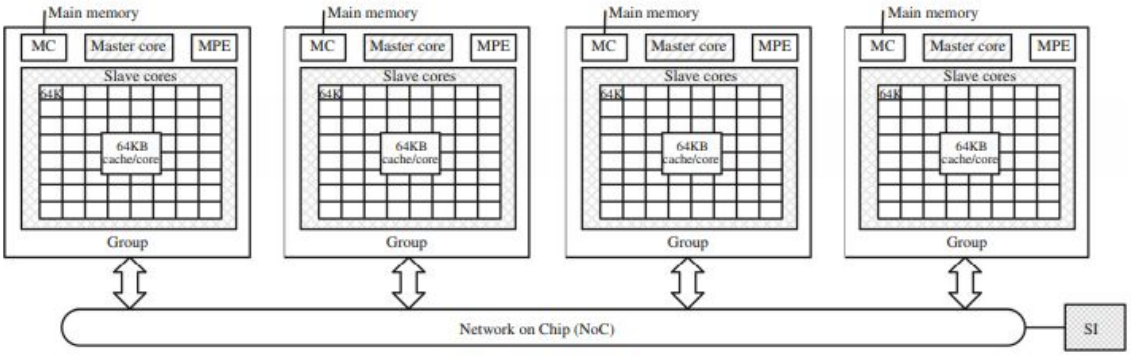

Fig. 6. *General architecture of the new Sunway processor.* Fu, HaoHuan, et al. "The Sunway TaihuLight supercomputer: system and applications." *Science China Information Sciences 59.7* (2016): 072001.

Strengths / Parallelism

Many of the strengths for the Sunway comes from the compatible software that's available with the supercomputer. Since the main applications of the Sunway are research-based there is an abundance of data and applications that the computer must deal with. The following are some of the major areas that the Sunway is used for (Dongarra):

- Atmosphere modelling
- Weather forecasting
- 3D Parallel Numerical Simulation
- Atomic simulation of silicon nanowires
- Big data analytics

The Sunway has done its best to optimize it's Sunway OpenACC 2.0 tool to allow for better data management and fast computation times. The software was optimized such that the algorithms use the data on-board the processors as much as it can as opposed to using the main memory, and allows for a significant increase in run-time. To add to this, the algorithms that have been optimized for the Sunway make difficult, large, and complex computations like those of the atomic simulation of silicon nanowires or atmosphere modelling run smoother with more cores simultaneously working on the problem due to the help from the OpenACC 2.0 tool (Fu).

Another key strength comes from the processor. Despite running at 1.45 GHz, the efficient utilization of multiple cores is where the Sunway really shines. Thanks to the parallelism that will be discussed shortly, the Sunway is able to tackle previously thought to be extremely difficult simulations with relative ease.

The Sunway's main source of parallelism comes from its node-level design. Sunway uses the basic C/C++ languages, fortran compilers, an automatic vectorization tool, and basic math libraries (Fu). Using these in conjunction with the Sunway's compiling tool OpenACC allow for parallelism at the node level.

The Sunway OpenACC 2.0 syntax targets the CG's in the processor to allow for better computation. For example, one of the functions the Sunway was optimized for was highly-effective global surface-wave numerical simulation with ultra-high resolution (Fu), which, up until the Sunway, had some difficulties computing with such vast amounts of data. However, Sunway's research teams optimized an algorithm that used features of the processor, such as the MPE and CPE as well as the data that was available on the processor. Figure 7 displays an example of the processors working simultaneously, synchronizing up, and returning an output. Initially all data is broken up into chunks and assigned to different blocks, then the processors grab the data they require to do computations. Once done the computations, the processors set a sync flag using the scratch pad memory that is available to each processor. The processors all wait until everything is synced up before moving on, and by using the storage on the processor they are able to move faster since they don't have to reach to the main data (Fu).

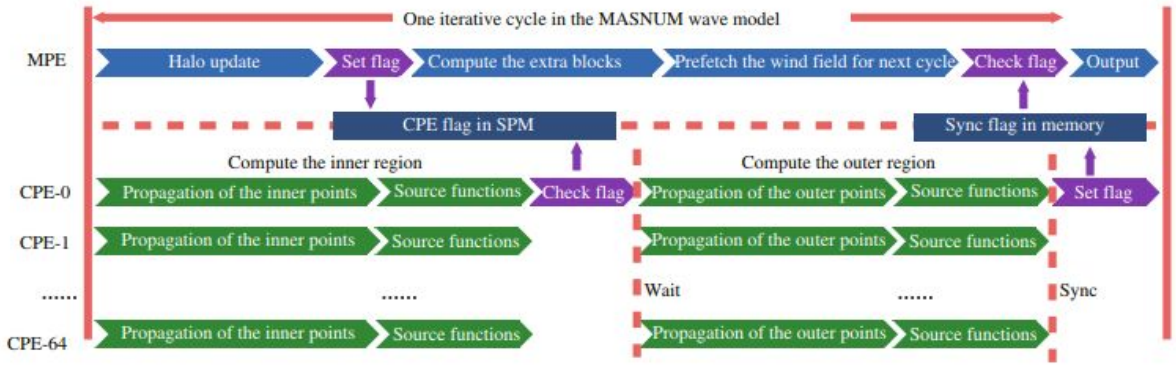

Fig. 7. *An optimized algorithm for one iterative cycle of the MASNUM wave model. The halo update, wind input, and result output are handled by the MPE. Through setting and checking the CPE flag in SPM, CPE-0 knows the status of the halo update. A flag in memory is used for synchronizing the MPE and CPE-0. All CPEs compute the propagation and source functions for the inner points first, and for the outer points when the halo update is finished. All CPEs except CPE-0 communicate with CPE-0 only*. Fu, HaoHuan, et al. "The sunway TaihuLight supercomputer: system and applications." *Science China Information Sciences 59.7* (2016): 072001.

Weaknesses

Some of the weaknesses in the Sunway would include the amount of power required to machine. It takes approximately 15,371 Kw to power the Sunway and when you compare that to some of the other supercomputers like the Summit supercomputer which runs at 9,783 Kw, or Sierra which runs at 7,438 Kw, they are able to run at nearly half the power consumption and are more computationally powerful than Sunway. Even looking at some of the weaker supercomputers behind the Sunway, Switzerland's Piz Daint only requires 2,348 Kw mind you that it is also about 70 TFlops slower (Erich, Dongarra and Simon). Additionally, despite having nearly 8 million more cores compared to the Summit supercomputer, the Sunway is nearly 50 TFlops weaker. Not only are there hardware flaws but there are also flaws within the software that was developed for the Sunway. Though it was optimized from their previous software, it still has some flaws with how fast the data can be read and stored by the processor.

Lastly, not all of the software developed for the Sunways was completely finished. With some room for further improvement and optimization, its left additional work for the programmers to deal with.

Comparative Analysis **(to the IBM Summit)**

As mentioned earlier, the Sunway used to sit as number one on the Top500 charts before being overthrown by the IBM Summit supercomputer. With the Summit running at a peak performance of 193,536 TFlops and the Sunway running at 125,435.9 TFlops/s the Sunway is clearly outclassed and perhaps a bit behind the times, taking significantly more power to run and using more cores to produce less. However, it should be considered that the Summit uses GPU units while the Sunway does not which might account for the difference in cores. One important feature to also consider is that both the computers were made with difference uses in mind. Lastly, even though both the Summit and Sunway deal with large amounts of data, the way of which they are used differ in their respective areas of focus. Since the Sunway is dealing with mostly simulations and atmosphere modelling it's algorithms are optimized for simulation speed while Summit is focused on Graph Analytics, deep learning and machine learning, which opts for better prediction and synthesis of data.

III. Cray Piz Daint (Switzerland)

Introduction

Located at the Swiss National Supercomputing Centre, the Piz Daint is the flagship supercomputer for the Swiss HPC(high performance computing) service. In 2016, it was

declared the fastest supercomputer in Europe, with a peak performance of up to 7.787 petaflops. Since then, Piz Daint has been combined with Piz Dora's CX40 system into a hybrid XC50/XC40 Cray system that has speeds of up to 27 petaflops. The XC50 computation nodes contain a GPU and CPU while the XC40 nodes only have a CPU. All of these nodes are connected with a Aries routing and communications ASIC, with Dragonfly network topology ("Piz Daint").

Fig. 8. Front end cabinets of Piz Daint "Piz Daint." *CSCS*. N.p., 2019. Web. 2 Apr. 2019.

Architecture / System Overview

Piz Daint is two different supercomputer systems(CX50 and CX40) merged together into one. It contains 7517 total compute nodes. 5704 of these are CX50 nodes and 1813 are CX40 nodes. Each CX50 node is composed of one 12-core Intel Xeon E5-2690 v3 CPU and one NVIDIA P100 Pascal GPU. Conversely, each CX40 node contains two 18-core Intel Xeon E5-2695 v4 CPUs and no GPUs ("Piz Daint").

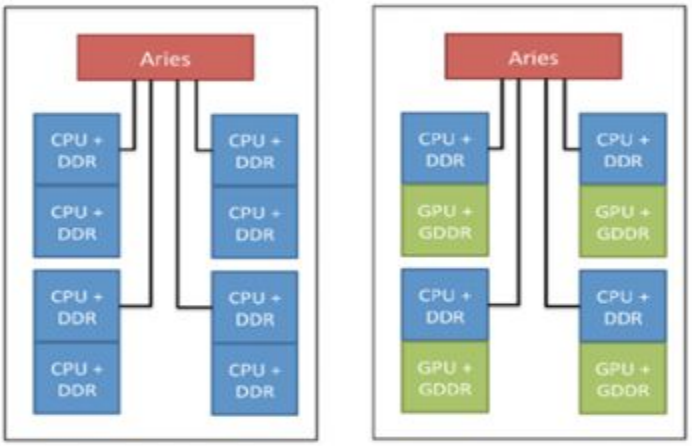

Fig. 9. Diagram of the Cray CX40 and CX50 nodes, respectively. "Systems-Level Configuration And Customisation Of Hybrid Cray XC30." *Research Gate*. N.p., 2014. Web. 2 Apr. 2019.

The highest theoretical performance of the CX50 system is 27154 teraflops/s. For the CX40 the highest performance is at 2193 teraflops/s. If we look at performance per compute node, the CX50 has 4.76 teraflops/s per node, while the CX40 system has 1.209 teraflops/s per node("Swiss National Supercomputing Centre (CSCS) | TOP500 Supercomputer Sites"). Each CX50 node utilizes NVIDIA NVLink along with PCIe 3rd generation busses to maintain high speed connections between the CPU and GPU (Brueckner Rich). This helps the Intel Xeon E5-2690 v3 CPU achieve a maximum memory bandwidth of 68GB/s and the Intel Xeon E5-2690 v4 CPU achieve 76.8 GB/s.

Piz Daint uses the Nvidia P100 GPU accelerator. It features 15.5 billion transistors and 54 streaming multiprocessors(SM) to help run multiple processes. It also has 3584KB of shared memory and a 4096KB L2 cache. Compared to the previous Nvidia M40 GPU, the P100 has over 500 more total CUDA cores and 7.3 billion more transistors .

Fig. 10. A Nvidia P100 SM unit NVIDIA TESLA P100. Web. 2 Apr. 2019.

Additionally, each SM contains 32 double precision CUDA cores to process double precision arithmetic faster. This allows for more accurate results in HPS applications such as quantum chemistry (NVIDIA TESLA P100).

Piz Daint runs a Cray Linux environment along with the Nvidia CUDA toolkit v8.0. This includes various math libraries and optimization tools to take advantage of the CUDA cores, along with support for a variety of languages such as C, C++, FORTRAN, and Python (Marsella, Luca).

Strengths / Parallelization

Piz Daint was designed to analyze large volumes of data for use in data sciences and large scale high-resolution simulations. In order to return results in a reasonable time, the CX50 portion of the system utilizes parallelization. Every Intel Xeon E5-2690 v3 component in the CX50 portion has access to a Nvidia P100 GPU. This CPU has 12-cores, each of which can process 24 threads, for a maximum of 288 threads that can access up to 3584 CUDA cores on the GPU ("Intel® Xeon® Processor E5-2690 V3 (30M Cache, 2.60 Ghz) Product Specifications."). This allows the CPU to parallelize on a large scale, drastically reducing computational time. Additionally, Piz Daint has a 10-core Intel Xeon E5-2690 v3 CPU as a Login Node ("Piz Daint"). This node handles all user input and program compilation, allowing for the rest of the nodes on the CPU to focus on calculations.

Due to the large amount of data the Piz Daint processes, one important performance aspect is memory bandwidth speed. To address this, PIz Daint utilizes a high bandwidth memory (HBM) RAM interface (Brueckner, Rich). HBM contains stacks, containing 4 DRAM dies, each with 2 independent memory channels. This means for each stack there are 8 independent memory channels to access DRAM (O'Connor, Mike). These channels allow for multiple computational nodes to access the 8.7 PB of memory at once, reducing bottlenecks and possible stalls in the parallelization process.

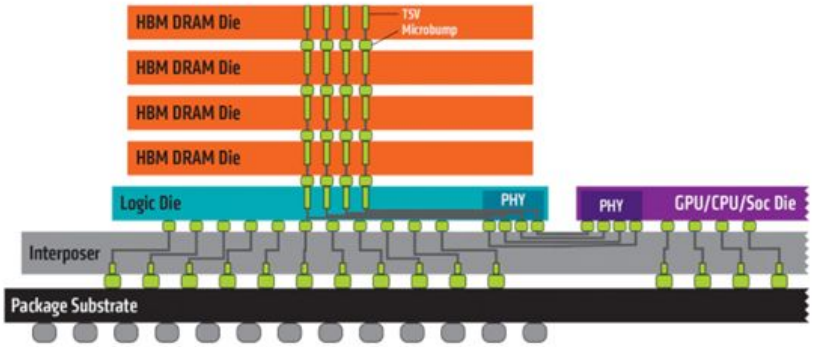

Fig. 11. High Bandwidth Memory Diagram. "High Bandwidth Memory | AMD." *Amd.com*. Web. 2 Apr. 2019.

Weaknesses

Although the Intel Xeon E5-2690 v4 CPU components in the CX40 portion of the system are slightly better than the Intel Xeon E5-2690 v3 CPU components within the CX50, the CX40 contains only CPU components. Since CPU's are primarily optimized for single threaded performance, this means CX40 system has far less parallelization potential compared to the CX50, making it far slower at performing large calculations. Since the Piz Daint is primarily used for large scale data analysis, most of it passes through the CX50 nodes with GPU accelerators, as they process much faster.

Another weakness is the Pascal GPUs in the CX50 system cannot page fault. This means any active processes must acquire large amounts of memory space they might not necessarily need, this creates more bottlenecks for other actives processes which in turn slows down the overall speed of the system, especially on a large scale.

Comparative Analysis (to the IBM Summit)

Piz Daint contains 7517 total compute nodes while Summit has 4608 nodes. However, only 5704 Piz Daint nodes are hybrid architecture. This means the rest of the nodes on Piz Daint cannot parallelize on a large scale. Each Piz Daint hybrid node has one 12-CPU and one GPU, on the other hand, each Summit compute node has 2 22-core CPUs and six GPU accelerators. Although both supercomputers utilize Nvidia GPU's with NVLink technology, the GPU's used by Summit are Nvidia V100's, which are superior to the Nvidia P100's on the Piz Daint. One aspect of this is that the V100 has 24 more streaming multiprocessors, meaning it has 1536 more CUDA cores per GPU and is able to run processes faster using more parallelization. Additionally, the Nvidia V100 can page fault, allowing processes on the Summit to load more specific sections of memory, reducing bottlenecks.

Both Summit and Piz Daint have employed solutions to clear bottlenecks related to memory access speed. However, Summit has smaller bottlenecks due it using more advanced technology. For instance, Piz Daint utilizes PCIe 3rd generation data busses, which are slower than the 4th generation busses on the Summit. Additionally, the hybrid CPUs on Piz Daint have a maximum bandwidth capacity of 68 GB/s, while Summit has 100GB/s.

IV. SuperMUC-NG (Germany)

Introduction

Ranking in at number eight of the world's fastest supercomputers, Germany's SuperMUC-NG is not only the top supercomputer in its country, but also one of the top

supercomputers in Europe. It is located in Garchung at the Leibniz Supercomputing Center (LRZ) and developed by LRZ with the help of Lenovo and Intel. With the purpose of helping scientists and other users in research fields such as astronomy, physics and other engineering/science applicable fields, this supercomputer is capable of handling giant amounts of data from many simulations and experiments.

Architecture / System Overview

The main features of the hardware design includes 48 cores per node, totaling to 311,040 cores over 6,480 nodes divided up into 6,336 thin nodes and 144 thick nodes. The total memory held is 719 terabytes with a parallel file system performance of 50 petabytes at 500 gigabytes per second. With a focus on energy efficiency and the innovative ideas of implementation given by the company Fahrenheit, the system is enveloped with a cooling and heat-reusing infrastructure that is unique amongst its competitors. Other notable hardware components can be seen in Fig. 12.

ComputeNodes	Thin Nodes	Fat Nodes	Total (Thin + Fat)
Processor	Intel Skylake	Intel Skylake	Intel Skylake
Cores per Node	48	48	48
Memory per node (GByte)	96	768	NA
Number of Nodes	6,336	144	6,480
Number of Cores	304,128	8,912	311,040
PEAK @ nominal (PFlop/s)	26.3	0.6	26.9
Linpack (Pflop/s)	TBD	TBD	TBD
Memory (TByte)	608	111	719
Filesystems			
High Performance Parallel Filesystem	50 PB @ 500 GB/s		
Data Science Storage	20 PB @ 70 GB/s		
Home Filesystem	256 TB		
Infrastructure			
Cooling	Direct warm water cooling		
Waste Heat Reuse	Reuse for producing cold water with adsorption coolers		
Software			
Betriebssystem	Suse Linux (SLES)		
Batchsystem	SLURM		
Paralleles Filesystem	IBM Spectrum Scale (GPFS)		
Programming Environment	Intel Parallel Studio XE GNU compilers OpenHPC Software Stack		
Message Passing	Intel MPI, (OpenMPI).		
Cloud Components			
Nodes with two Nvidia V100 GPUs Nodes without GPUs	32 32		

Fig. 12. Hardware components of the SuperMUC-NG. “300K-Core SuperMUC-NG System Launches at LRZ in Germany.” *InsideHPC*, 26 Oct. 2018, insidehpc.com/2018/10/300k-core-supermuc-ng-system-launches-lrz-germany/.

Since this supercomputer is a high-performance computing (HPC) system, each node represents a computer and are linked together, with each node being distinguished as either a thin node, thick node, or cloud node (“What Is a High-Performance Computer?”). The thin and thick nodes are where the majority of the computing is done for the system. The cloud nodes are a newer edition to the system as it presents user friendly benefits for access to resources outside the scope of LRZ, such as software, virtual machines, and virtual networks. The layout can be seen in figure 2, with compute nodes being “bundled into 8 domains (islands). Within one island, the Omnipath network topology is a 'fat tree' for highly efficient communication”

(Leibniz-Rechenzentrum, “SuperMUC-NG”). The Omnipath network is what connects all these islands together to interact with each other to finish tasks for a certain job.

Fig. 13. The SuperMUC-NG at the LRZ data center. “SuperMUC Petascale System.” *LRZ: SuperMUC Petascale System*, www.lrz.de/services/compute/supermuc/systemdescription/.

Strengths / Parallelization

When it comes to raising the computing speed of a computer, the negative impact is almost always overheating. One of the unique features of the SuperMUC-NG to counteract this is the adsorption-cooling technology designed by the company Fahrenheit. They state that it “offers enormous potential for energy savings in data centers, as all computing power is converted into waste heat, which can be used for environmentally friendly cooling” (Fahrenheit, “Design and Working Principle.”). Since cooling consumes so much power, reusing that same heat given off by the supercomputer to produce the cooling effect gives it a unique edge in energy efficiency. In the dual-chamber technology, there lies 2 adsorption modules, which is

shown in Fig. 14. One acts as an adsorber/desorber while the other works as an evaporator/condenser. This not only generates cold temperature continuously, but does so with minimum maintenance needed, maximizing reliability (Fahrenheit, “Data Centers Cooling Themselves: SuperMUC-NG Science Symposium at LRZ.”). Some users have criticized the design for the reason of it possibly flooding, which would be a major setback in the midst of large research-based computations.

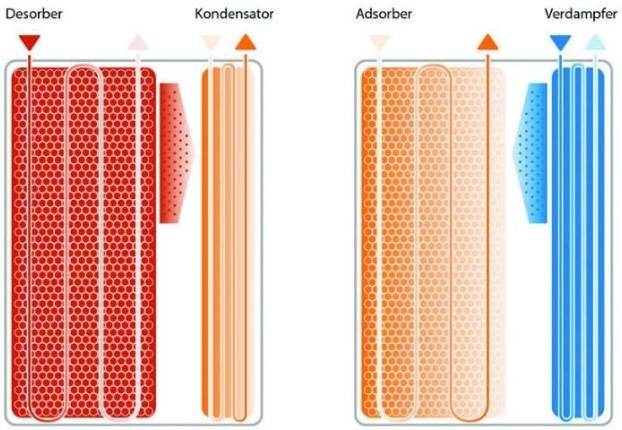

Fig.14. Structure of the adsorption modules in each chiller. Fahrenheit. “Design and Working Principle.” *FAHRENHEIT*, fahrenheit.cool/en/design-working-principle/.

The performance enhancement provided by parallelism is largely due to the IBM Spectrum Scale, a parallel computing file system designed to manage large amounts of data in an efficient manner. This software enables users to build their own storage infrastructure, which is useful because it accepts creativity from the user categorizing their data to be taken in and examined efficiently. Many users have their own way of organizing their data and the IBM spectrum scale is able to take any of these structured storages into account. A notable feature in this software is the Active File Management, also referred to as AFM. This feature is a caching layer integrated with the spectrum scale as it "creates associations from a local Spectrum Scale

cluster to a remote cluster or storage, defining the location and flow of file data and automating the data management" (Quintero et al. 6-7). With the cluster of nodes and Omnipath design, the overall structure of the SuperMUC-NG is able to accelerate parallel accesses for large amounts of data with the help of the AFM interconnecting all of the cluster nodes (as seen in Fig. 15). The parallel read and writes in the AFM help to access large data blocks within a single I/O operation by "striping" blocks of data from different files (7). This increases speed significantly when users bring in data broken up into many divided files to keep certain information separate from others.

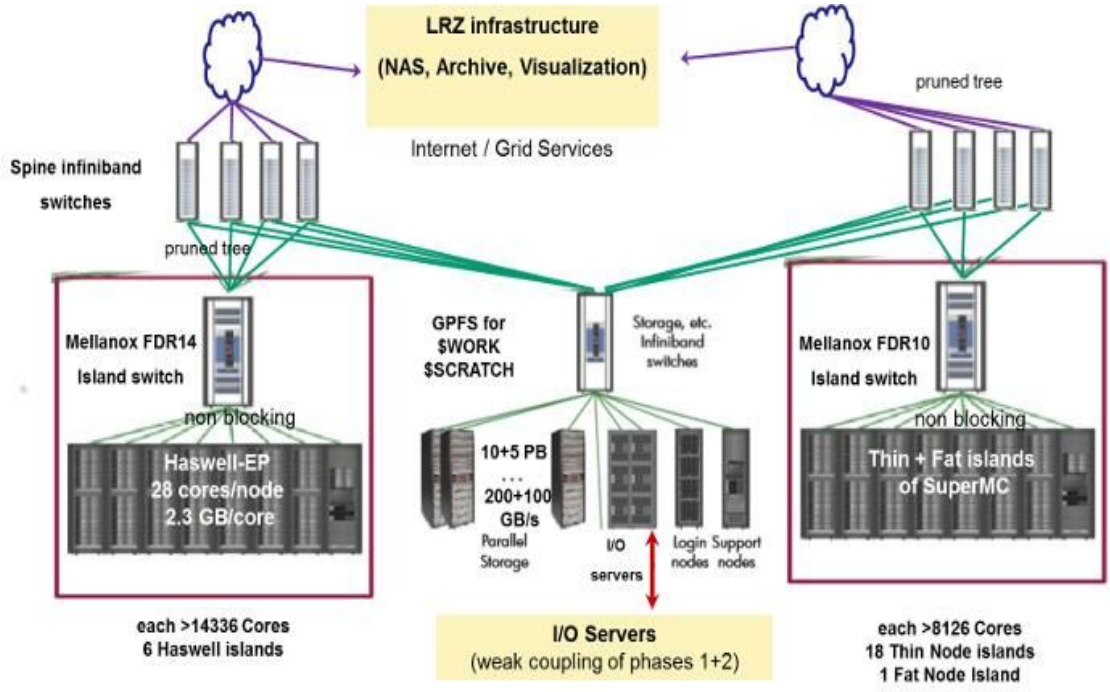

Fig. 15. Infrastructure/processes of the IBM Spectrum Scale. “SuperMUC Petascale System.” *LRZ: SuperMUC Petascale System*, www.lrz.de/services/compute/supermuc/systemdescription/.

Weaknesses

Although the SuperMUC-NG is renowned for its energy-efficient infrastructure, the supercomputer does have some weaknesses within its hardware, specifically in the chosen CPU.

The Intel Skylake processor has some notable issues, which are fixable, nevertheless should still be mentioned. The issue occurs during hyper-threading in the CPU, and the results include "application and system misbehavior, data corruption, and data loss" (Chirgwin). When short loops of less than 64 instructions are used in combination of certain registers, the CPU can lead the system to devastating drawbacks (Chirgwin). One fix for this issue is to disable hyper-threading in the CPU. This solution is able to help keep the large amounts of data intact, but at a cost of losing speed and efficiency.

Another limitation that the supercomputer has is in the parallel file system when handling large amounts of data in a certain hierarchical design. It is "not optimal for handling large quantities of small files located in a single directory with parallel accesses" (Leibniz-Rechenzentrum, "File Systems of SuperMUC-NG."). With an excessive amount of files per directory being accessed by parallel instructions and simultaneous jobs, the end result of the process will issue timeouts or crashes ("File Systems of SuperMUC-NG."). One approach to make users experience a more optimal performance is to increase the amount the subdirectories and limit the files in each subdirectory.

Comparative Analysis **(to the IBM Summit)**

When comparing the SuperMUC-NG to the IBM Summit, they both possess similar infrastructures. The node types on which they operate through are similar, using launch nodes, compute nodes, and cloud (or login) nodes for users to share information and processes simultaneously. Comparing stats alone, the IBM Summit is by far the superior supercomputer in terms of computing power and speed, having a far greater rate of floating point operations per

second and the total amount of cores on which it operates. In terms of power, the SuperMUC-NG is overlooked by the mighty IBM Summit, as it is capable of being implemented on deep learning applications such as artificial intelligence. However, with the cooling technology implemented by Fahrenheit, the SuperMUC-NG remains the superior supercomputer in regards to being energy efficient as it is able to reuse all of the waste heat that it produces from computing for output, reducing the power consumption and its carbon footprint.

V. Final Remarks

In conclusion, the Summit is a brilliant computer, more capable of parallelization than any other supercomputer available. Comparing the Summit to the Sunway, Summit employs GPU units while the Sunway does not, which allows for the Summit to efficiently perform simultaneous processes using GPU and CPU pairings while the Sunway utilizes only CPUs. Additionally, Nvidia's NVLink technology on the Summit allows for faster data transfers between processors, with the Sunway still using DDR3 and the Summit using DDR4 and HBM2.

Both Summit and Piz Daint have employed solutions to clear bottlenecks related to memory access speed. Additionally, the Summit continues to have issues regarding I/O throughput, which may hint to a commonality in issues among modern supercomputers. Additionally, the hybrid CPUs on Piz Daint have a maximum bandwidth capacity of 68 GB/s, while Summit has 100GB/s. Further, Summit nodes all have identical architectures (CPU-GPU pairings), which may result in faster compilation and job scheduling, whereas the Piz Daint utilizes both CPU pairings and CPU-GPU pairings, and may require additional coordination on the programmer's part.

The SuperMUC-NG and Summit were built using similar parallelizing technology, specifically the IBM Spectrum Scale, which is built off the Summit's GPFS. While the SuperMUC-NG is superior in terms of efficient energy consumption and innovative cooling technology, the Summit competes with raw power and processing ability.

IBM Summit (USA) References

Biebelhausen, John. *What Does It Mean to Summit?*, 13 Nov. 2017, www.mellanox.com/blog/2017/11/what-does-it-mean-to-summit/.

Feldman, Michael. "IBM Ups Its Game with the Power9 Processor." *IBM Ups Its Game with the Power9 Processor | TOP500 Supercomputer Sites*, 25 Aug. 2016, 3:04, www.top500.org/news/ibm-ups-its-game-with-the-power9-processor/.

Hemsoth, Nicole. "HPC File Systems Fail for Deep Learning at Scale." *The Next Platform* -, 17 Oct. 2018, www.nextplatform.com/2018/10/09/hpc-file-systems-fail-for-deep-learning-at-scale/.

"IBM Power System AC922." *IBM Power System AC922 - Details - Canada*, www.ibm.com/ca-en/marketplace/power-systems-ac922/details.

McCorkle, Morgan L. "ORNL Launches Summit Supercomputer." *ORNL*, 8 June 2018, www.ornl.gov/news/ornl-launches-summit-supercomputer.

Morgan, Timothy Prickett. "Big Blue Aims For The Sky With Power9." *The Next Platform* -, 26 Aug. 2016, www.nextplatform.com/2016/08/24/big-blue-aims-sky-power9/.

"November 2018." *November 2018 | TOP500 Supercomputer Sites*, Nov. 2018, www.top500.org/lists/2018/11/.

Power ISA Version 3.0 B, 29 Mar. 2017, ibm.ent.box.com/s/1hzcwkwf8rbju5h9iyf44wm94amnlcrv.

Quintero, et al. *IBM High-Performance Computing (HPC) Insights with IBM AC922 Clustered Solution*. Redbooks, 2019.

Smith, Ryan. "NVIDIA Volta Unveiled: GV100 GPU and Tesla V100 Accelerator Announced." *NVIDIA Volta Unvieled: GV100 GPU and Tesla V100 Accelerator Announced*, AnandTech, 10 May 2017, 20:00 PST, www.anandtech.com/show/11367/nvidia-volta-unveiled-gv100-gpu-and-tesla-v100-accelerator-announced.

"Summit User Guide." *Oak Ridge Leadership Computing Facility*, www.olcf.ornl.gov/for-users/system-user-guides/summit/summit-user-guide/.

Summit_FactSheet.pdf, ORNL, www.olcf.ornl.gov/wp-content/uploads/2014/11/Summit_FactSheet.pdf.

Quintero, et al. *IBM High-Performance Computing (HPC) Insights with IBM AC922 Clustered Solution*. Redbooks, 2019.

Thompto, Brian. *IBM POWER9 SMT Deep DiveSummit Training Workshop*, POWER Systems, 2018, www.olcf.ornl.gov/wp-content/uploads/2018/12/summit_workshop_thompto_smt.pdf.

Sunway TaihuLight (China) References

Dongarra, Jack. *Report on the Sunway TaihuLight System*. Technical Report. University of Tennessee: Department of Electrical Engineering and Computer Science, 2016.

Erich, Strohmaier, et al. *Top500*. November 2018. 1 April 2019.

Fu, HaoHuan, et al. "The sunway TaihuLight supercomputer: system and applications." *Science China Information Sciences 59.7* (2016): 072001.

The National Supercomputing Center. *Hardware*. 2019. 1 April 2019.

—. *Introduction*. 2019. 1 April 2019.

Piz Daint (Switzerland) References

Brueckner, Rich. "Creating Balance In HPC On The Piz Daint Supercomputer - Insidehpc." *insideHPC*. N.p., 2016. Web. 2 Apr. 2019. https://insidehpc.com/2016/08/creating-balance-in-hpc/

"High Bandwidth Memory | AMD." *Amd.com*. Web. 2 Apr. 2019. https://www.amd.com/en/technologies/hbm

"Intel® Xeon® Processor E5-2690 V3 (30M Cache, 2.60 Ghz) Product Specifications." *Ark.intel.com*. Web. 3 Apr. 2019. https://ark.intel.com/content/www/us/en/ark/products/81713/intel-xeon-processor-e5-2690-v3-30m-cache-2-60-ghz.html

Marsella, Luca. *Best Practices For Building Software On Piz Daint*. 2019. Web. 2 Apr. 2019. https://www.cscs.ch/fileadmin/user_upload/contents_userLab/building_software_piz_daint.pdf

NVIDIA TESLA P100. Web. 2 Apr. 2019. https://images.nvidia.com/content/pdf/tesla/whitepaper/pascal-architecture-whitepaper.pdf

O'Connor, Mike. *Highlights Of The High Bandwidth Memory (HBM) Standard*. 2014. Web. 3 Apr. 2019. http://www.cs.utah.edu/thememoryforum/mike.pdf

"Piz Daint" *CSCS*. N.p., 2019. Web. 2 Apr. 2019. https://www.cscs.ch/computers/piz-daint/

"Swiss National Supercomputing Centre (CSCS) | TOP500 Supercomputer Sites." *Top500.org*. Web. 2 Apr. 2019. https://www.top500.org/site/50422

"Systems-Level Configuration And Customisation Of Hybrid Cray XC30." *Research Gate*. N.p., 2014. Web. 2 Apr. 2019. https://www.researchgate.net/publication/285583469_Systems-level_Configuration_and_Customisation_of_Hybrid_Cray_XC30.

SuperMUC-NG (Germany) References

Chirgwin, Richard. “Intel's Skylake and Kaby Lake CPUs Have Nasty Hyper-Threading Bug.” *The Register® - Biting the Hand That Feeds IT*, The Register, 25 June 2017, www.theregister.co.uk/2017/06/25/intel_skylake_kaby_lake_hyperthreading/.

Hruska, Joel. “Major Hyper-Threading Flaw Destabilizes Intel Kaby Lake, Skylake CPUs.” *ExtremeTech*, 26 June 2017, www.extremetech.com/computing/251499-major-hyper-threading-flaw-can-destabilize-intel-cpus-based-kaby-lake-skylake.

Feldman, Michael. “Germany's Most Powerful Supercomputer Comes Online.” *Top500 The List*, 27 Sept. 2018, 18:11, www.top500.org/news/germanys-most-powerful-supercomputer-comes-online/.

Fahrenheit. “Data Centers Cooling Themselves: SuperMUC-NG Science Symposium at LRZ.”

FAHRENHEIT, 28 Nov. 2018, www.fahrenheit.cool/en/data-centers-cooling-themselves-supermuc-ng-science-symposium-at-lrz/.

Fahrenheit. "Design and Working Principle." *FAHRENHEIT*, www.fahrenheit.cool/en/design-working-principle/.

Leibniz-Rechenzentrum. "File Systems of SuperMUC-NG." *Leibniz-Rechenzentrum (LRZ)*, Leibniz-Rechenzentrum, www.doku.lrz.de/display/PUBLIC/SuperMUC-NG.

Leibniz-Rechenzentrum. "SuperMUC Petascale System." *LRZ: SuperMUC Petascale System*, www.lrz.de/services/compute/supermuc/systemdescription/.

Leibniz-Rechenzentrum. "SuperMUC-NG." *Leibniz-Rechenzentrum (LRZ)*, Leibniz-Rechenzentrum, www.doku.lrz.de/display/PUBLIC/SuperMUC-NG.

Quintero, et al. *IBM Spectrum Scale (Formerly GPFS)*. Redbook, 2015.

"SuperMUC Petascale System." *LRZ: SuperMUC Petascale System*, www.lrz.de/services/compute/supermuc/systemdescription/.

"What Is a High-Performance Computer?" *Introduction to High-Performance Computing: What Is an HPC System?*, Software Carpentry Foundation, www.epcced.github.io/hpc-intro/010-hpc-concepts/.

"300K-Core SuperMUC-NG System Launches at LRZ in Germany." *InsideHPC*, 26 Oct. 2018, www.insidehpc.com/2018/10/300k-core-supermuc-ng-system-launches-lrz-germany/.