The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems

***Abstract*—CORAL, the Collaboration of Oak Ridge, Argonne and Livermore, is fielding two similar IBM systems with NVIDIA GPUs that will replace the existing Titan and Sequoia systems. We discuss the designs of both systems, with a particular focus on their key differences. Our evaluation of the system highlights the following. Applications that fit in HBM see the most benefit and may prefer more GPUs; however, for some applications, the CPU-GPU bandwidth is more important than the number of GPUs. The node-local burst buffer scales linearly, and can achieve a 4X improvement over the parallel file system (PFS) for large jobs; smaller jobs, however, may benefit from writing directly to the PFS. Finally, several CPU, network and memory bound analytics and GPU-bound deep learning codes achieve up to a 11X and 79X speedup/node, respectively over Titan.**

I. Introduction

The Collaboration of Oak Ridge, Argonne and Livermore (CORAL) began in late 2012 with the goal of delivering three systems that would each improve delivered performance by 4-6X on a range of Department of Energy (DOE) benchmark applications as compared to the existing 20 PetaFlop DOE systems - Titan [1] at Oak Ridge National Laboratory (ORNL) and Sequoia [2] at Lawrence Livermore National Laboratory (LLNL). After extensive discussions with potential offerors, including formal responses to a Request for Information (RFI), the CORAL Request for Proposals (RFP) was released in early 2014. From the RFP responses, ORNL and LLNL selected IBM systems with a hybrid CPU/GPU architecture for their pre-exascale systems, Summit, located at the ORNL Leadership Computing Facility (OLCF) and the Livermore Computing (LC) Sierra system. Each system incorporates POWER9™ CPUs, NVIDIA Volta™ V100 GPUs, and Mellanox Enhanced Data Rate (EDR) InfiniBand (IB) network technologies.

The Summit and Sierra systems reflect a critical exascale architecture path in hybrid computing. While both systems must answer questions of national importance related to science, energy, environment, and national security, their mission needs and the expected usage differ. Specifically, Summit caters to the DOE Office of Science's (SC) workload, and will consist primarily of full-system jobs that expand our knowledge of the natural world through scientific inquiry and maintain a vibrant effort in science and engineering as a cornerstone of the nation's economic prosperity. While Sierra will also run full-system jobs, its workload will primarily consist of ensemble runs that provide uncertainty quantification (UQ) of issues related to the National Nuclear Security Administration's (NNSA) stewardship of the US nuclear stockpile. High-level differences between the two systems reflect these differences in their expected workloads.

Since the selection of the systems in late 2014, CORAL and its vendor partners (IBM, NVIDIA and Mellanox) have engaged in system co-design activities. Funding for Non-Recurring Engineering (NRE) has facilitated these activities that have shaped the design and development of the system hardware and software. NRE projects have spanned areas that include the I/O systems, messaging, cluster system management and the programming environment. Overall, these activities have greatly enhanced the resulting systems and their ability to meet key CORAL requirements.

We present the design, development, evaluation and the lessons learned from acquiring and deploying the pre-exascale Summit and Sierra systems. Specifically, our contributions are:

- The design and development of the next generation of GPU-based large-scale systems, including the motivations for their similarities and differences;
- An evaluation of these architectural designs through microbenchmarks and both proxy and real applications;
- The lessons learned from the design, deployment, evaluation and procurement strategy of the systems, including:
  - Shared fate and the complementary expertise of two major centers yield better procurements;
  - System requirements and design must carefully reconcile performance and transparency;
  - GPU memory (High Bandwidth Memory, HBM) and NVLink™ bandwidth are critical performance factors on these systems.

To begin the rest of the paper, Section II discusses the procurement process. Section III then motivates the design of the selected system architecture. Next, Section IV details the Sierra and Summit system architectures, interconnect and the I/O subsystem, and the key differences between the systems. We then provide a detailed evaluation in Section V, and conclude with the lessons learned from the design, procurement, deployment and evaluation of these novel systems.

II. CORAL Procurement Process

The CORAL procurement built upon the process used in several successful DOE procurements but included the novel aspect of targeting multiple leadership-class systems with one RFP. Similar to the LLNL Sequoia and Argonne Mira procurements, the RFP employed a set of target requirements that would not be converted into hard requirements until well after contract signing. This flexible model allows considerable risk sharing between CORAL and the selected offeror(s), which enables more aggressive responses that better reflect the likely technology that will be delivered.

A key aspect of any DOE large-scale system procurement is its mission need. While scientific computation provides a predictive tool for the design, analysis, and decision making for complex systems, current systems still fail to meet key DOE requirements. In particular, DOE's SC and NNSA have several critical mission deliverables, including annual stockpile certification and safety assurance for NNSA and future energy generation technologies for SC. Computer simulations play a key role in meeting these critical mission needs.

The CORAL systems target different mission needs. Summit will expand the boundaries of our scientific understanding of critical questions related to our fundamental physical processes and future energy generation technologies. This mission requires simulations at unprecedented scales that stress all aspects of a large-scale system, including network properties such as bisection bandwidth. Sierra will improve the confidence with which NNSA can certify the nation's stockpile through detailed ensemble calculations. These uncertainty quantification calculations involve several throughput-oriented jobs. While they stress the overall computational capability of the system, their network requirements are necessarily lower than those of scalable science runs. Importantly, while their primary mission needs will determine the majority of their workloads, both systems must run both throughput-oriented and scalable science jobs.

Reflecting the similarities in these mission needs, the CORAL high-level system requirements placed a particularly high importance on system performance. To capture the differences as well as the similarities, CORAL required offerors to provide performance results of the projected system for four scalable science benchmarks (LSMS, QBOX, HACC and NEKBone) and four throughput benchmarks (CAM-SE, UMT, AMG and MCB), and included five additional throughput benchmarks (QMCPACK, NAMD, LULESH, SNAP and miniFE). The target performance requirement was at least a 4-6X increase compared to performance on Titan or Sequoia.

In addition to directly capturing the mission needs through benchmark performance, the CORAL RFP included several other target requirements. These requirements, such as a total of 4 PB of memory available for direct application use to enable running larger problem sizes, would improve the extent to which the systems could meet their mission needs or their total cost of ownership, such as limiting system power consumption to at most 20 MW. Others reflect the anticipated evolution of the mission needs, such as performance on a set of data-centric benchmarks (Graph500, Hash, and Integer Sort), or anticipated system limitations, such as burst buffers that can enable smoothing of the file system workload and also serve as a small file system for short-lived files.

III. Motivations for System Selection

From the responses to the CORAL RFP, OLCF and LC selected an IBM system with NVIDIA GPUs for Summit and Sierra since, in their assessment, the proposed system would provide the best value in meeting the mission needs for those systems. The single largest factor in that assessment was the projected benchmark performance of the systems and the supporting documentation that indicated the project performance would extend to the intended system workloads. Modifications to benchmark source code convincingly showed that real applications could attain most of that performance through a directive-based approach for using the GPUs. Further, the system architecture reflects several design principles that support that conclusion, as this section details.

The lessons [3] of the Cray-1 [4] and CDC Star-100 [5] motivate one the most significant aspects of the system architecture design. The systems feature powerful IBM POWER9 CPUs in addition to the NVIDIA Volta GPUs that provide over 95% of the floating point capability of the systems. GPUs have many similarities to vector processors and the computations that run well on them reflect those similarities. However, also like vector processors, not all portions of scientific applications exhibit the characteristics that allow the performance potential of those GPUs to be realized. Thus, the availability of strong sequential performance in a highly capable CPU is essential to realizing high overall performance.

The memory system reflects several key design motivations. Many MPI applications, particularly DOE SC and NNSA multiphysics applications, require at least 2 GB of memory per MPI process. Thus, the system design must include enough main memory to fit the data of multiple physics packages of several MPI processes. However, experience with Titan and its 6 GB of GDDR memory per GPU has proven that high application performance requires that the application working set must fit into fast GPU memory. Thus, each GPU should have a large capacity of HBM that has a high bandwidth path to the node's main memory. Further, while explicitly staging large data structures between those two memories can support overlap of data movement with computation, hardware coherence between the memory systems and the GPUs and CPUs that access them simplifies code correctness. Thus, the architecture features sixteen DDR4 DIMMs, 16 GB of HBM2 per GPU and a second generation NVLink protocol that provides the desired coherence support.

The system architecture centers around *fat nodes* with dual CPU sockets, each with multiple NVLink-connected GPUs, which reduces the number of system interconnect endpoints. Applications can then use fewer, larger memory footprint, MPI processes to reduce scaling difficulty. Further, the reduced node count allows the use of a fat-tree network topology at a reasonable cost. This topology provides several application benefits, including lower runtime variability and performance that is largely independent of the placement of MPI processes.

Each compute node includes an NVMe-device that provides high bandwidth local storage. These SSDs provide a burst buffer solution that significantly reduces the time for file I/O. Local storage, as part of a multi-level solution [6] allows checkpoints to be written locally and discarded once the next checkpoint is completed, This solution is ideal for applications that perform N-N (i.e., file per process) I/O, which analysis of OLCF workloads indicates accounts for more than 90% of application runs. While the portion of SC and NNSA

TABLE I: Overall system characteristics of Summit, Sierra, and Titan.

Overall System Characteristics			
	Summit	Sierra	Titan
Node Count	4,608	4,320	18,688
Peak Performance	200 PF	125 PF	27 PF
Total GPU Memory	442 TB HBM2	277 TB HBM2	112 TB GDDR
Total DDR Memory	2.4 PB	1.1 PB	598 TB
Total Combined Memory	2.8 PB	1.4 PB	710 TB
Interconnect Bi-Section BW	EDR IB 115 TB/s	EDR IB 54 TB/s	Gemini 112 TB/s
Topology	1:1 Fat Tree	2:1 Fat Tree	3D Torus
Burst Buffer Capacity	7.4 PB	6.9 PB	NA
Burst Buffer Bandwidth	9.7 TB/s	9.1 TB/s	NA
File System Capacity	250 PB	150 PB	30 PB
File System Bandwidth	2.5 TB/s	1.5 TB/s	1.0 TB/s

applications that use N-1 traffic patterns require more complex software to exploit this node-local storage, the partitioned, phased nature (i.e., write-only epochs during checkpointing and read-only during restart) of that access pattern can still benefit substantially from it.

IV. A Modern GPU-Based HPC Architecture

Table I summarizes the key aspects of the Summit and Sierra system architectures. The compute nodes of both systems have two IBM Power9 (P9) CPUs and several NVIDIA Volta V100 GPUs and are connected with a Mellanox EDR IB network. Differences in GPU count per compute node (six on Summit and four on Sierra), network topology and system budgets lead to different compute node counts on Summit (4,608) and Sierra (4,320). Summit's 256 racks provide a system peak of around 200 petaflops at approximately 13MW while Sierra's 240 racks achieve a system peak of over 125 petaflops at approximately 12MW. While both systems employ a fat-tree network topology, Sierra has approximately two-to-one tapering after the top-of-rack (TOR) switches. The following subsections detail the node, interconnect and I/O designs.

A. Node Design

The compute nodes of both Summit and Sierra have two IBM 3.07 GHz POWER9 CPUs, each with 22 cores. As discussed in Section III, the powerful CPU cores reduce the impact of inherently sequential code regions and smooth application transitions from homogeneous architectures. The sockets are connected by IBM's X-Bus™, which provides 64 GB/s of coherent access between the sockets. Each socket has eight memory channels connecting 256 GB DDR4 on Summit and 128 GB DDR4 on Sierra, both providing 340 GB/s of peak memory bandwidth per node. Each node also includes a 1.6 TB Samsung NVMe SSD for use as a write cache (i.e., burst buffer) that can also be used for local-scratch storage, cached libraries, or extended memory via `mmap`.

TABLE II: Node Characteristics of Summit, Sierra, and Titan

Node Characteristics			
	Summit	Sierra	Titan
CPU	2 POWER9	2 POWER9	1 AMD Opteron Interlagos
Cores	44 (22 per P9)	44 (22 per P9)	16
Memory	512 GB	256 GB	32 GB
Memory Bandwidth	340 GB/s	340 GB/s	51.2 GB/s
SMP Bus	X-Bus 64 GB/s	X-Bus 64 GB/s	NA
CPUs:GPUs	2:6	2:4	1:1
GPU	6 Volta V100	4 Volta V100	1 Tesla K20x
SM	480	320	14
GPU DP Flops	42 TF	28 TF	1.31 TF
GPU Memory	96 GB HBM2	64 GB HBM2	6 GB GDDR
GPU Memory Bandwidth	5.4 TB/s	3.6 TB/s	250 GB/s
NVLink BW	50 GB/s/GPU	75 GB/s/GPU	NA
SSD Capacity	1.6 TB	1.6 TB	NA
SSD Write BW	2.1 GB/s	2.1 GB/s	NA
Interconnect Injection BW	2x 12.5 GB/s EDR	2x 12.5 GB/s EDR	1x 5.5 GB/s Gemini

Each node has four (Sierra) or six (Summit) NVIDIA Volta V100 GPUs that each has eighty 1.333 GHz Streaming Multiprocessors (SM) for approximate total peak performance of 7 TF double-precision and 14 TF single-precision. The GPUs also include 112 TF TensorCores that perform a 4X4 matrix multiply on half-precision inputs with single-precision accumulations. Each POWER9 has 150 GB/s of NVLink connectivity [7], shared by three (Summit) or two (Sierra) GPUs. The design also provides NVLink connectivity between GPUs connected to the same CPU socket at 50 GB/s for Summit and 75 GB/s for Sierra. In either case, NVLink enables faster intranode data movement, which will benefit GPU-enabled applications compared to the limited PCIe bandwidth on Titan.

In addition to improved latency and bandwidth with NVLink 2, the nodes provide a single coherent address space that includes system memory and the GPU HBM2 memory. Applications (or the OpenMP runtime) can explicitly manage memory by using `cudaMalloc()` and `cudaMemcpy()` to move data between system memory and GPU memory or they can use `cudaMallocManaged()` to let the CUDA™ runtime manage the copies. The single address space allows applications to use `malloc()` and then pass the pointer to a GPU kernel, which greatly eases application porting.

Table II summarizes key features of the Summit, Sierra, and Titan nodes. The block diagram in Figure 1 shows half of the Summit node on the left and half of the Sierra node on the right (both systems include the NVM device connected to socket 0 shown on the right). The key differences are that the Summit node has twice the system memory and 3 GPUs per POWER9 while the Sierra node has two. Both sockets in both systems connect directly to the Mellanox InfiniBand HCA.

The differences in GPUs per node reflect the maturity in the transition to heterogeneous architectures between the

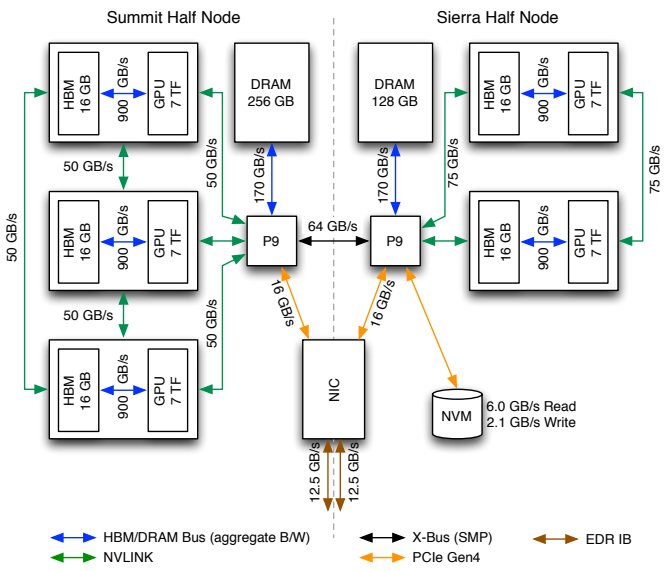


Fig. 1: Block Diagram of Summit and Sierra Half Nodes

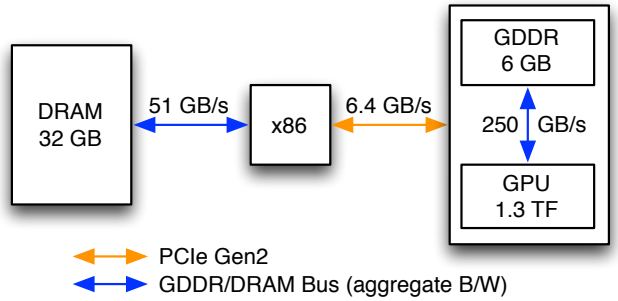


Fig. 2: Block Diagram of Titan Node (Omits NIC Links)

workloads of the two facilities and an expectation that NNSA multiphysics applications will achieve better performance with more system memory bandwidth per GPU. Code porting at LLNL while preparing for Sierra showed that some NNSA applications (e.g., Quicksilver and the setup phase of AMG [8]) run better or as well on the CPU as the GPU. Further, multiple NNSA applications, such as, UMT [8] are NVLink bound so additional GPUs do not result in additional performance. The POWER9 processor has a fixed number of *bricks* (i.e., bundles of NVLink lanes) that are divided between the GPUs. Thus, the Summit node provides 50 GB/s from the CPU to each GPU while the Sierra node provides 75 GB/s.

Figure 2 shows the block diagram for the simpler Titan node. Although a Titan node has 32 GB of system memory, most applications only use 6 GB per node (i.e., the capacity of the GPU's GDDR memory), probably because the PCIe bandwidth is too low. The ratio of GDDR memory bandwidth to system memory bandwidth is approximately 5:1. However, the PCIe bandwidth raises that ratio to about 39:1 by restricting the realizable memory bandwidth to 6.4 GB/s. Summit's six GPUs have an aggregate of 5.4 TB/s of HBM2 bandwidth, for a ratio of 18:1, given the 340 GB/s of peak system memory bandwidth and the 300 GB/s of NVLink bandwidth. Sierra's four GPUs have an aggregate of 3.6 TB/s of HBM2 bandwidth for a ratio of 12:1.

Between contract signing and system delivery, the cost of memory doubled. OLCF chose to pay the increased cost while LC chose to reduce capacity to 256 GB per node rather than reduce overall node count. LC's choice reflects estimates of the total memory footprint of multiphysics application based on the number of physics packages and the HBM2 capacity.

B. Interconnect Design

The Mellanox EDR fat-tree network topologies use Switch-IB™ 2 switches and ConnectX™-5 host channel adapters (HCAs). The compute racks have 18 compute nodes that each has one dual-port Mellanox ConnectX-5 HCA. Each node's HCA ports connect to a separate TOR switch (two per compute rack) and provide non-blocking connectivity within each compute rack. All TORs connect into one set of core switches thus creating a single InfiniBand subnet. Summit's fabric is fully non-blocking between TORs and core switches, while Sierra's fabric is 2:1 tapered at this level. LLNL's system tradeoff analysis between nodes and networking resources indicated that the tapered network produced the best overall cost neutral system architecture for the projected Sierra workload.

With support from NRE funding, Mellanox, IBM, and NVIDIA have significantly enhanced network performance. New or improved features include, adaptive routing, switch-based collective offload (SHARP™), Dynamic Connected Transport (DCT), tag-matching in the HCAs, Verbs-bypass, additional remote atomics, and GPU Direct™. These features combine to provide maximum traffic throughput, lower latency communication, enhanced collective performance, and improved GPU to GPU communication.

C. I/O Design

The Summit and Sierra system architectures include a two-tiered I/O subsystem that consists of on-node burst buffers (BBs) [9] and a parallel file system (PFS) in order to meet the system I/O requirements. These tiers are presented as two separate name spaces. OLCF and LLNL worked with IBM to improve the I/O subsystem design through several NRE activities. These activities include the implementation of a data movement mechanism across the BB and capacity tiers (BB-API), and purpose-built distributed file system (BSCFS) that is layered on top of node-local SSDs to support N-1 I/O access patterns. Other NRE activities increased PFS file system scaling and single directory shared metadata performance.

While other systems, including Cori at NERSC [10], have previously deployed BBs, the IBM system architecture will be the first at an extensive scale to use node-local devices. Compared to a shared BB tier, the node-local option requires less infrastructure (e.g., servers, network), and can reduce data movement (drain only every n^{th} checkpoint) and increase performance for the N-N use case. While shared BBs simplify

the N-1 use case, node-local software can exploit phased parallel I/O access patterns to supply similar functionality.

Also already discussed, each compute node includes a 1.6 TB NVMe drive that provides up to 2.1 GB/s write and 5.8 GB/s read I/O performance. Thus, Summit has an aggregate capacity of 7.4 PB (2.5X total system memory) and a write performance of 9.7 TB/s while Sierra 6.9 PB (5X total system memory) and a write performance of 9 TB/s. Each SSD has an endurance characteristic of 5 drive writes per day (DWD), which an analysis of I/O patterns on current systems and the compute capabilities of the systems indicates is sufficient.

Applications can interface with the BB tier at different levels of abstraction and ease of use: an XFS file-system built on the node-local SSDs, created at job allocation; directly through the BB-API or through BSCFS (for shared files); or through SCR [6] or other libraries that are built on top of BB-API. Hardware assisted NVMe over Fabrics (NVMeoF) enables the Mellanox HCA to move data to or from the BB directly without compute-node CPU overhead. QoS and throttling mitigate network impact. Applications can pre-stage data on to the BB prior to job dispatch. They can checkpoint to it and immediately resume computation phase while the data is asynchronously drained to the PFS.

Checkpoints are typically written as multiple files (i.e., N-N) or as one shared file (i.e., N-1). While the file-per-process I/O maps directly to the node-local BB, the shared file use case is more challenging. BSCFS is a per-job, distributed log-structured file system that supports N-1 checkpoints. BSCFS caches the shared files in the SSDs and then reconstitutes them in the PFS without further intervention from the application.

The capacity tier uses IBM's SpectrumScale GPFSi™ product with NRE enhancements such as larger file system block sizes, more file system sub-blocks, reduced GPFS lock contention and also improved scalability of file system tools such as parallel fsck. Spectrum Scale single-directory shared-metadata create performance has been increased 10X to 50,000 metadata operations per second. Further, use of its "scatter" mode, which randomizes block allocations, improves random I/O performance, which better reflects application I/O patterns, and reduces performance degradation as the PFS ages.

Each PFS uses IBM's new GL4™ Elastic Storage Server (ESS) (two per GL4), which achieves high storage density through 10 TB hard drives in 106-drive enclosures. Thus, Sierra has 150 PB of usable capacity in 56 GL4 units that provide 1.5 TB/s of I/O bandwidth while Summit has 250 PB usable capacity in 77 GL4 units that provide 2.5 TB/s of I/O bandwidth. In both systems, each ESS has two dual port Mellanox IB EDR cards that connect to TORs. At OLCF, two of these connections serve Summit while the other two serve other systems to provide a center-wide file system. At LLNL, all PFS links connect directly into Sierra's compute fabric; gateway servers provide access to other LC systems.

V. Early Evaluations

Our experiments compare Summit and Sierra, or alternatively Peak and Butte, architecturally identical 18-node 1-cabinet test and development systems. Where possible we use matching software versions; unfortunately the early deployment status of the systems often leads to small software differences that complicate analysis of performance differences.

TABLE III: CPU stream rates on Peak and Sierra (GB/s)

system cores	Peak/ci 40	Peak/ci 42	Peak 40	Peak 44	Sierra 40	Sierra 44
Copy	272.9	273.5	273.1	274.6	277.3	278.3
Scale	269.6	270.6	269.5	271.4	274.4	275.7
Add	268.8	269.8	268.7	270.6	273.5	274.9
Triad	273.0	273.9	273.5	275.3	277.7	279.0

A. Memory Interconnect Evaluation

We first present results for several memory microbenchmarks. The stream code, compiled with GCC measures CPU memory bandwidth under OpenMP threading. Table III shows the best result in 1,000 trials for Peak with core isolation (ci) (its normal operating mode), Peak without core isolation, and Sierra (without core isolation). Performance is similar for both systems and slight differences between Peak and Sierra may be partly due to slightly different system memory configurations [11] or inherent performance variability. Multiple benchmark trials reveal runtime variation as high as 9%. Also, performance was up to 4% higher if the benchmark was run after the POWER9 was idle for several minutes prior to the experiment. While we are investigating this issue, one possible explanation is aggressive frequency throttling.

We use a variant [12] of stream to measure GPU HBM2 bandwidth on all GPUs of 15 Peak and Butte nodes. Best values for Peak were 789 (Copy), 788 (Mul) and 831 (Add and Triad) GB/s; values for Butte differed by less than 1%, confirming the expected result that node architecture differences do not impact GPU memory performance. These figures represent 88% and 92% of the 900 GB/s peak, a much higher fraction of peak than Titan GDDR memory. Most trials in 1,000 vary in performance less than 10%, although a few outliers were up to 16X slower than the best case.

We measure achieved CPU-GPU NVLink rates with a modified bandwidthTest from the NVIDIA CUDA Samples. As described earlier, peak NVLink rates between a CPU and a directly connected GPU are 50 GB/s (Summit, Peak) and 75 GB/s (Sierra, Butte). Table IV shows host to device (htod), device to host (dtoh) and bidirectional (bidir) transfer rates between core 0 and each GPU. Multiple trials show little variability. On-socket (Peak GPUs 0, 1 and 2; Butte GPUs 0 and 1) unidirectional and bidirectional bandwidths are 92% and 86% of theoretical peak, although bidirectional bandwidth to the final GPU of the socket is unexpectedly about 10% lower compared to the other on-socket GPUs when accessed from core 0. We are currently investigating possible affinities between cores and each GPU. Unsurprisingly, off-socket bandwidths are significantly lower, due to the intervening X-Bus. Thus, we expect users to avoid off-socket GPU access.

Table V shows the more typical use case of multiple MPI processes evenly spread between CPU sockets each simul-

TABLE IV: Single Node Single GPU NVLink Rates (GB/s)

GPU	0	1	2	3	4	5	(peak)
Peak htod	45.93	45.92	45.92	40.63	40.59	40.64	50
Peak dtoh	45.95	45.95	45.95	36.60	36.52	35.00	50
Peak bidir	86.27	85.83	77.36	66.14	65.84	64.76	100
GPU	0	1		2	3		(peak)
Butte htod	68.64	68.47	—	40.44	40.47	—	75
Butte dtoh	68.33	68.69	—	36.85	35.63	—	75
Butte bidir	128.98	114.99	—	64.79	64.60	—	150

TABLE V: NVLink Rates with MPI Processes (GB/s)

MPI Process Count	1	2	3	4	5	6
Peak htod	45.93	91.85	137.69	183.54	229.18	274.82
Peak dtoh	45.95	91.90	137.85	183.80	225.64	268.05
Peak bidir	85.70	172.59	223.54	276.34	277.39	278.07
Butte htod	68.66	137.39	206.05	275.47	—	—
Butte dtoh	68.91	137.48	203.80	271.12	—	—
Butte bidir	126.06	255.47	270.72	283.08	—	—

taneously using one GPU. Multiple trials exhibit run-to-run variability under about 3%. For a saturated node with the largest MPI process count, for the unidirectional case the expected NVLink rate (300 GB/s peak, $6 \times 46 = 276$ GB/s actual on Peak, $4 \times 69 = 276$ GB/s actual on Butte) nearly matches the CPU stream performance of about 275 GB/s, thus CPU memory bandwidth does not limit the transfers. However, attainable bidirectional bandwidth is reduced by 46% compared to the sum of rates for individual GPUs (600 GB/s peak, $6 \times 86 = 516$ GB/s actual on Peak, $4 \times 129 = 516$ GB/s actual on Butte), due to bandwidth limits of CPU memory. Thus, overlapped host-device and device-host transfers (as opposed to in sequence) will provide little performance benefit in some cases. In either case, since attainable NVLink speeds for a saturated node are roughly the same for both systems, Summit's additional GPUs may provide little performance benefit for applications highly bound by NVLink bandwidth.

Table VI shows NVLink transfer rates between GPUs (within a socket and across them), using p2pBandwidthLatencyTest from CUDA Samples. We show the average of ten trials on a single node; the maximum deviance across different trials and GPU-GPU connections was 8.7%. The peer-to-peer (P2P) access feature yield performance that approaches NVLink theoretical peak bandwidth; results are much lower without it (no P2P). Predictably, cross-socket bandwidth is much lower than that between GPUs attached to the same CPU socket. GPUs on socket 1 without peer-to-peer access underperform compared to socket 0, possibly due to the benchmark running on socket 0 controlling GPUs attached to socket 1. Socket 1 peer-to-peer bidirectional performance on Butte is also low by about 12%. Otherwise, on-socket performance with peer-to-peer access enabled is roughly 93% of theoretical peak.

B. EDR Interconnect Performance Evaluation

We evaluate MPI bisection bandwidth, collectives, and messaging rate performance at scales up to 2,048 nodes on the pre-GA Sierra and Summit systems. In addition to baseline

TABLE VI: NVLink Rates for GPU-GPU Transfers (GB/s)

	no P2P			P2P			
	socket 0	socket 1	cross-socket	socket 0	socket 1	cross-socket	(peak)
Peak unidir	33.18	25.84	30.32	46.33	46.55	25.89	50
Peak bidir	54.48	27.91	49.02	93.02	93.11	21.63	100
Butte unidir	41.27	24.72	31.04	69.49	69.49	31.05	75
Butte bidir	58.63	25.55	49.17	139.15	124.30	49.15	150

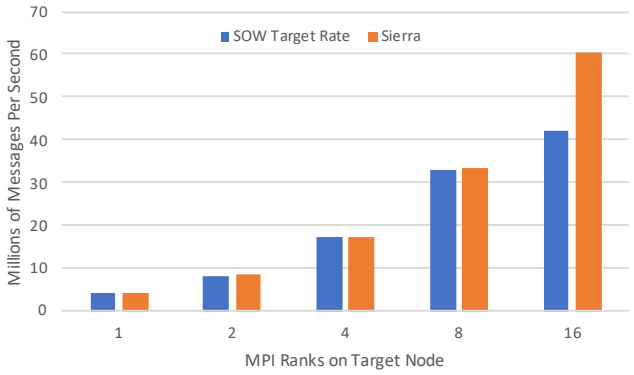


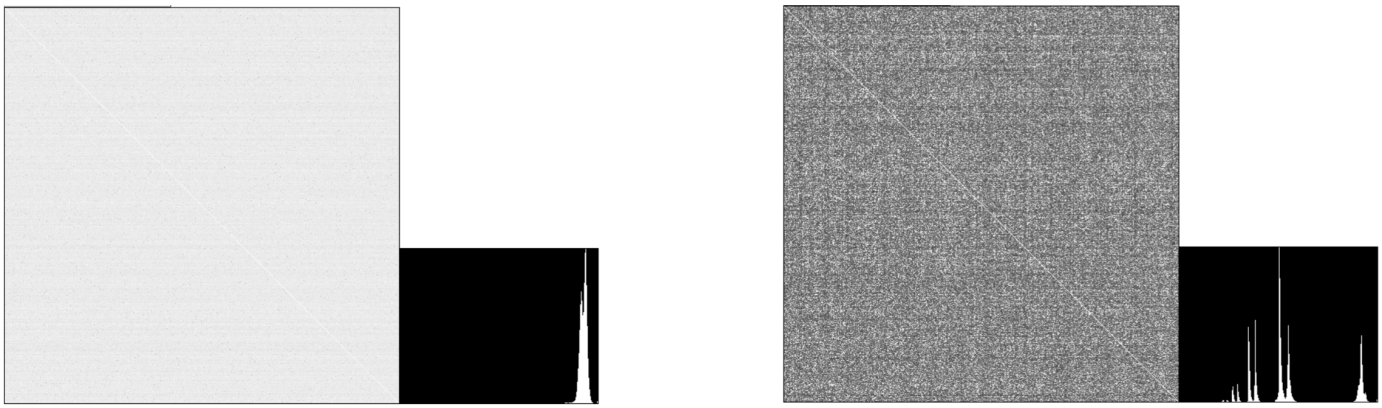
Fig. 3: SQMR Target vs. Measured Rates on Sierra

performance, we explore the impact of hardware support for adaptive routing (AR) and collective offload.

The Sequoia Messaging Rate (SQMR) benchmark measures the MPI messages per second. We scale these tests from 2-17 nodes, increasing the processes per node (Figure 3). Careful placement using mpibind yields rates on the actual systems that outperform the original target rates. We found that the use of Reliable Connected (RC) queue pairs over Dynamic Connected Transport (DCT) improves messaging rate significantly. DCT is a new transport type added by Mellanox to reduce the memory footprint when communicating with many partners. Large applications that benefit from DCT's reduced memory use need to consider RC's potential performance benefit.

CORAL's bisection bandwidth benchmark, based on mpiGraph [13], explores the bandwidth between possible MPI process pairs. Figure 4 shows Summit mpiGraph results using one of the two HCA ports with AR and with the commonly used single-path static routing. Lighter colors represent less variation and congestion and uniform color is better than patterned results. The histograms next to each image show the observed bandwidth distribution. The single cluster with AR indicates all pairs achieve nearly maximum bandwidth while single-path static routing has nine clusters as congestion can limit achieved bandwidth substantially. The single-port AR results demonstrate an average performance of 11.8 TB/s or 96% of the maximum bandwidth measured. In contrast, the single-path static routing results achieve an average bandwidth of 10.2 TB/s or only 80% of the maximum measured bandwidth. Results on Summit in Figure 5 confirm that adaptive routing achieves higher overall bandwidth.

The Mellanox EDR network also includes support for Mellanox Scalable Hierarchical Aggregation and Reduction



(a) With Adaptive Routing

(b) Without Adaptive Routing

Fig. 4: Summit's MpiGraph Output

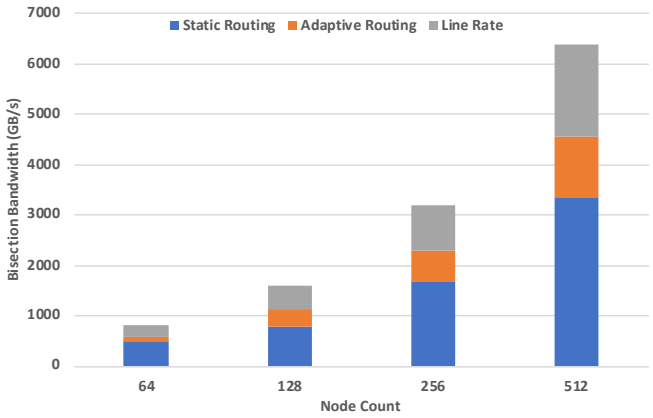


Fig. 5: CORAL Bisection Bandwidth Using One Port per Node

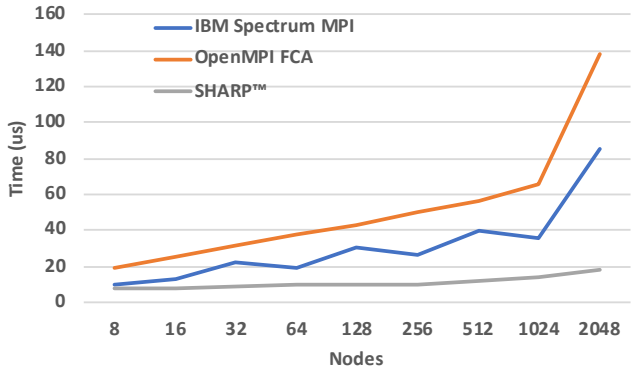


Fig. 6: OSU All Reduce 2 KiB Test on Summit

Protocol (SHARP) collectives. To study this we use one MPI process per node and evaluate collective performance with OSU benchmarks [14]. We focus on MPI_Allreduce in Figure 6 as SHARP is designed to optimize its performance. Our results use Mellanox's Fabric Collective Accelerator (FCA™) software, IBM's software optimized collectives and SHARP. The results indicate that IBM's software based collectives scale better than FCA. However, while the FCA results are relatively smooth across node scales, the IBM results have a sawtooth behavior, which may indicate that the library alternates between two algorithms. SHARP performs even better with an average performance of 74% faster than IBM's. SHARP also shows a relatively flat performance profile from 8 to 2,048 nodes. Thus, SHARP collectives can significantly improve the performance of applications that use MPI_Allreduce. SHARP barriers demonstrate the same scaling and performance impact. At 2,048 nodes the OSU Barrier test shows ~8µs for SHARP compared to ~35µs for Spectrum MPI™.

C. I/O Subsystem Evaluation

We evaluate a single GPFS GL4 unit (with two ESS servers) using multiple clients with the `IOR` benchmark for file I/O performance and `mdtest` for metadata performance. The Summit and Sierra systems (including their PFS) are in deployment and acceptance phases, which prevented running large scaling tests. The Summit PFS is formatted with a 16 MiB block size. IOR tests use 12 client nodes, 8 MPI processes per client and a single GL4 unit in the file-per-process (N-N) mode with a 4 GiB block size and a 32 MiB transfer size for a test duration of 20 minutes. The total amount of data written was 40 TB. For sequential I/O workloads, a single GL4 unit performs writes at 36 GB/s and reads at 40 GB/s. For random I/O, we observe 36 GB/s and 42 GB/s for writes and reads, which is consistent with the sequential rates. An early scaling result with 16 GL4s on the Sierra PFS yielded 370 GB/s for sequential writes. These results suggest that further optimization and scaling will allow the Summit and Sierra PFSs to attain their 2.5 TB/s and 1.5 TB/s of I/O rates.

We also evaluate a single client's performance against a single GL4 with `IOR` and observe that a single client can perform read or write file I/O at about 18 GB/s.

Next, we measure the interactive metadata performance of a GL4, using `mdtest`, 12 clients and 48 processes. Each

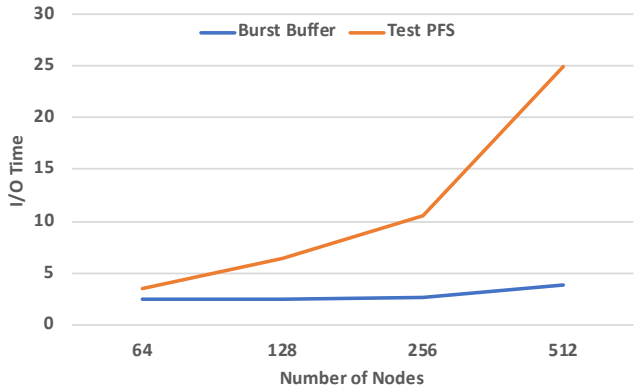


Fig. 7: Weak Scaling of GTC4 I/O Performance on Summit

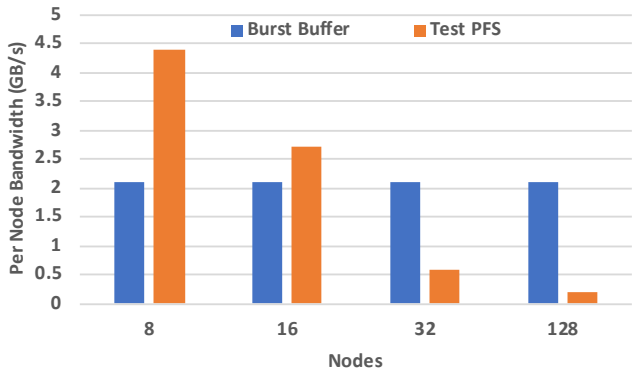


Fig. 8: FIO Synthetic Checkpoint Test

process creates 524,288 files in unique directories, writes 32 kiB into each file and closes the file. The GL4 performs around 21,000 file creates per second. Based on this result, we expect the overall interactive metadata performance for Summit system to meet its target of 1.05 Million file creates per second over 77 GL4 units.

D. Burst Buffer Performance

While deployment of full BB support is in progress, we cannot yet test the integration of the burst buffers and PFS. Instead, our measurements provide an understanding of the impact of the BB resources to applications. Our first experiments use the GTC fusion application, which uses the N-N checkpoint model. In these tests, we modify GTC input parameters to weak-scale the problem size to exploit power of 2 node-count increments and we contrast performance on Summit with the linearly scaling BB bandwidth to that with the PFS. Figure 7 shows that the BB checkpoints are significantly faster and achieve nearly constant scaling.

The BBs benefit GTC despite its relatively small checkpoints that likely hit in the compute-node PFS page-caches because a node-local file system also reduces metadata costs. Applications that write larger checkpoints will benefit from constant BB bandwidth scaling, as shown by synthetic FIO checkpoint tests with 6 processes per node that each write 10 GB files, directly to the BB with DirectIO. This data size exceeds the PFS page-cache so writes are flushed. To mimic large application checkpoints, we use a 1 MB block size. Figure 8 shows the BBs achieve 2.1 GB/s and scale linearly. This capability will reduce the checkpoint cost of a full system job on Summit by 4X compared to the PFS. The PFS provides better write performance for both tests at low node counts.

E. Application and Miniapplication Performance

We now present results for applications and miniapplications that represent aspects of the Summit and Sierra workloads. The NUCCOR kernels (scalable science) model the dense matrix triple product $D = A^T \cdot B \cdot C$ of the NUCCOR nuclear physics application [15]. Figure 9 compares average performance across ten trials on a single node of Peak and of

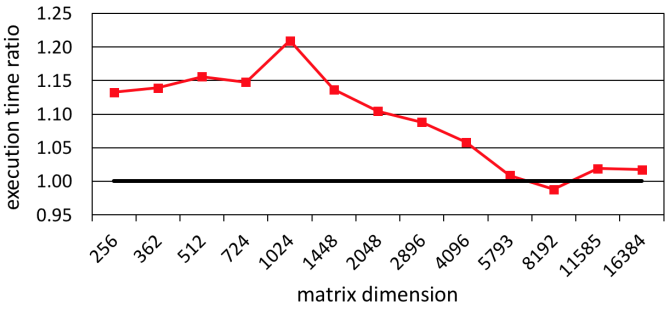


Fig. 9: Peak/Sierra Execution Time on NUCCOR Kernels

Sierra with one MPI process per GPU. We vary the (square) matrix size, ensuring that it fits in GPU memory. Each run consists of two cuBLAS DGEMM operations and four explicit matrix transfers with no overlap; more recent NUCCOR versions overlap some transfers with computation. For smaller matrices for which NVLink transfer costs dominate, compute time for Peak is 15-20% higher than Sierra, compared to a 50% runtime increase due to the different NVLink rates of the systems if transfer costs fully dominate. For the larger matrices, however, computation dominates transfers and per-GPU performance of the two systems is nearly equal as transfer costs are amortized by computation.

GTC is a scalable science application, modeling fusion reactors [16]. We use the PGI™ compiler 17.10 and CUDA 9.1.85, and SMT4 mode. The systems use Spectrum MPI version 10.02.00.00_PRPQ_2017.11.17-4-g78dfc805 release date 180110 but different CUDA driver versions (Peak 390.31, Butte, 387.25). Peak nodes have core isolation enabled (42 cores per node available); Butte does not. We first run identical cases at the same node count with four GPUs per node. Second, we run identical cases on the same number of GPUs, using all GPUs on a node (i.e., more Butte nodes for a given GPU count) for execution regimes typically used in production. Table VII shows similar results over ten trials on both systems, which are single-cabinet systems with identical interconnects (results omit initialization and I/O). NVLink costs are low since GTC leaves its data on the GPU during

TABLE VII: Fixed Node Count and GPU Count GTC Results

	Peak min	Peak max	Butte min	Butte max
1 node	5.08	5.17	5.11	5.30
2 nodes	8.89	9.14	9.09	9.47
4 nodes	13.13	13.46	13.86	14.28
8 nodes	15.50	15.92	15.97	16.56
12 nodes	17.74	17.94	18.19	18.59
12 GPUs	12.92	13.23	13.35	13.70
24 GPUs	14.02	14.35	14.75	15.02
48 GPUs	18.16	18.49	18.21	18.59

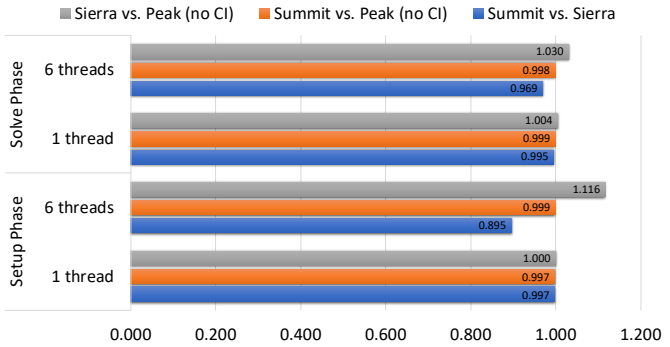


Fig. 10: AMG2013 Single Node FOM Ratios

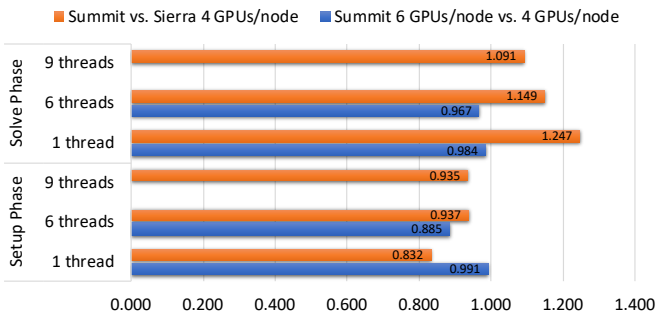


Fig. 11: AMG2013 FOM Ratios with 72 GPUs

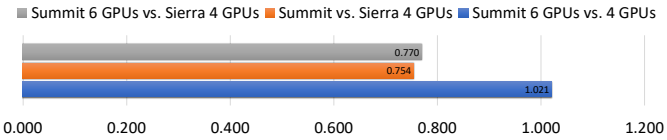


Fig. 12: UMT2013 FOM per GPU Ratios

the run. Previous efforts to optimize CPU-GPU transfer costs for applications on PCIe-2 based systems such as Titan will still provide benefits with newer NVLink-based systems.

AMG2013, a CORAL throughput benchmark [8], is a parallel algebraic multigrid solver for linear systems. Figure 10 shows the ratios of the Figures of Merit (FOMs) [17] from single node experiments executed on Summit, Sierra, and Peak (average of ten trials). AMG2013 has two primary phases: setup and solve. The setup phase runs exclusively on the CPU while the solve phase uses the GPUs. Our experiments use 4 GPUs per node with 4 MPI processes per GPU and use one or six OpenMP threads per process. Core-isolation was disabled on Peak to match the Sierra environment while it is enabled on Summit. Single node tests show that, for this case, the setup phase on Sierra results in a higher FOM value than on Summit, possibly due to differences in thread placement. Figure 11 shows the average of ten trials of a multinode case on both Summit and Sierra that uses 72 GPUs. On Summit, we use two layouts; one with 18 nodes and 4 GPUs per node, and another with 12 nodes with 6 GPUs per node. The former layout results in a setup FOM approximately 12% higher than the latter because more OpenMP threads can be used per GPU.

The UMT2013 throughput mini-application models deterministic radiation transport on unstructured meshes. The large benchmark problem size requires full use of the system memory and an overlapped batching strategy to move data on and off the GPU [18]. NVLink bandwidth bounds approximately 60% of the runtime, with the rest being CPU and network bound. We conduct multinode and single node benchmark runs on Summit (with 4 or 6 GPUs per node) and on Sierra, Figure 12 compares the average FOM [19] per GPU using SMT4 and 40 threads/GPU with 4 GPUs and 28 threads/GPU with 6 GPUs. Because Sierra has 50% higher CPU to GPU NVLink bandwidth, UMT achieves a 32% higher FOM per GPU on Sierra than on Summit. The comparison showcases that for applications with a significant NVLink dependence, such as those with a large memory footprint and low data reuse, Sierra can provide similar per node performance to Summit while using fewer GPUs per node. We also ran single node tests on Summit using one MPI process per each of 4 GPUs and varying the thread count. Table VIII shows the average FOM from 10 trials for each thread count (SMT off). As expected, more OpenMP threads per process increases the FOM per GPU, partly because UMT is about 40% CPU bound.

F. Machine Learning on the CORAL Node

The CORAL2 data science benchmarks [20] capture data analytics workloads, focusing on CPU, memory, and interconnect performance, and deep learning workloads, focusing on GPU, interconnect, and I/O performance. The CPU-only data analytics implementations include principle component analysis (PCA), K-Means and support vector machines (SVMs). The chosen input sizes and algorithms yield one memory bound technique (PCA), one CPU bound (SVM), and one network bound (K-Means). In contrast, the deep learning suite, which includes recurrent neural network (RNN), convolutional neural network (CNN), and application benchmarks from the Cancer Distributed Learning Environment (CANDLE) [21] project, uses GPUs extensively. The data analytics tests use randomly generated, in-memory data to weak-scale to arbitrary node counts while the deep learning benchmarks use a combination of random, in-memory data and data from disk and include a strong scaling benchmark on ImageNet data.

Figure 13 shows that Summit achieves speedups compared to all Titan baseline runs. The compute-bound benchmarks like SVM (on POWER9s) and CNN (on GPUs) achieve 11X and

TABLE VIII: UMT FOM on 1 Node

Configuration	Average FOM	Ave. FOM / GPU
4 GPUs, 1 thread	8.47E+08	2.117E+08
4 GPUs, 7 threads	2.40E+09	5.992E+08
4 GPUs, 10 threads	2.64E+09	6.609E+08
6 GPUs, 7 threads	3.36E+09	5.602E+08

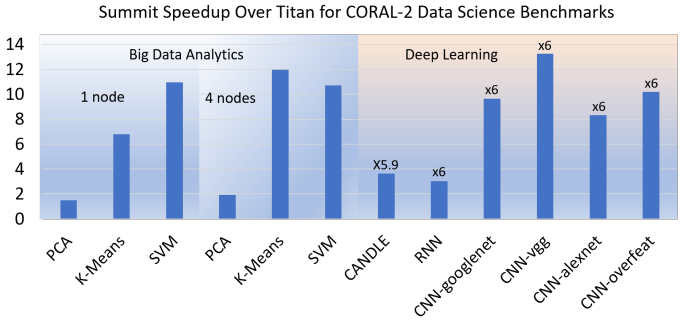


Fig. 13: Data Science Speedups (Titan baseline)

up to 79X speedup per node (13.2X/GPU x 6GPUs). Multinode runs of K-Means achieve an additional 1.8X speedup improvement on Summit, as a result of its high performance EDR interconnect. PCA achieves the lowest speedup among these benchmarks as its gain is consistent with the memory performance improvement. The RNN benchmark, which uses synthetic, in-memory inputs, achieves an 18X speedup compared to Titan (3X x 6GPUs). The CANDLE benchmarks, which load input data from the PFS, achieve near perfect throughput speedups since they greatly benefit from NVLink. For distributed training on ImageNet [22] data with ResNet-50 [23] based on Keras [24] (Tensorflow [25] backend) and Horovod [26], which stress the GPUs, intra- and inter-node interconnects, and node-local SSDs, an ideal scaling efficiency in terms of seconds per epoch is achieved on 64 nodes (384 Volta GPUs) of Summit (see Figure 14). These results are for the reference code [20] with default settings and are not optimized to the Summit architecture.

VI. Lessons Learned

Due to the nascent stage of the systems, distinguishing immaturity in the deployment and software from major negatives that require significant aspects to be reconsidered is difficult. Nonetheless, we discuss some lessons that we can already identify that would benefit future system designers.

From a procurement standpoint, CORAL has proven to be an extremely valuable partnership for all parties—the laboratories as well as the system providers. For the laboratories, CORAL has provided collective, often complementary, technical expertise at both sites, that improves the RFP and NRE co-design activities, as well as improving system deployment through shared fate. The broad commonalities in the chosen systems has enabled leveraged NRE investments that provide more capable systems. However, joint NRE investment requires compromises—as *the* big customer, a laboratory may have wielded more influence on some specific topics. Finally,

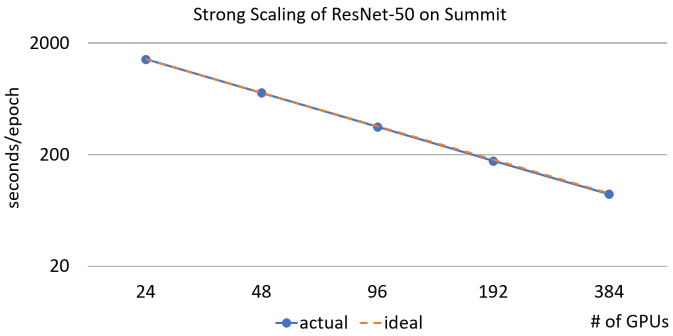


Fig. 14: Resnet-50 Scaling on Summit

while aligning the I/O subsystem acquisition with that of the system may provide a more integrated solution and mitigate risks associated with site integration, site-specific contract modifications proved necessary to meet diverse requirements.

An immediate technical lesson involves requirements specification. While the laboratories knew that the BB must support the N-1 access pattern, IBM did not initially. Thus, BSCFS was added to the planned system software and may not fully meet the requirement. Further, while a node-local BB offers substantial performance benefits, the lack of a single mount point to the BB remains a concern for use cases other than checkpoints. To facilitate transparent BB use while still maintaining performance, the laboratories must develop software on top of BSCFS. In addition, a transparent interface atop the BB and PFS might be useful for easy adoption of the BB for some applications. Future procurements will better reflect these requirements. Also, BB and PFS results indicate that the BB can achieve linear scaling and a 4X speedup over the PFS for large jobs. Smaller jobs, however, benefit by writing directly to the PFS. While that choice might reduce overall system performance, the observation implies a need for considered use of BB resources. More generally, any system design must carefully reconcile performance and transparency.

The compute-node design and evaluation show that GPU memory (HBM) bandwidth remains a critical performance factor but CPU-GPU interconnect (NVLink) bandwidth can be equally important for applications with large memory footprints. Applications that can fit their working sets in HBM will see the most performance. System memory capacity and memory bandwidth can also limit performance despite the huge percentage of overall capability represented by GPUs. Further, steep memory pricing (DDR and HBM) can limit the ability to handle the emerging class of ML applications.

Although not discussed in this paper, both systems include a federated telemetry infrastructure to collect and to analyze system logs. We expect this facility will increase system reliability and provide a window into center operations. Future system designs should carefully plan log data collection from the early stages to glean insights and to optimize system and application performance [27].

REFERENCES

[1] Oak Ridge National Laboratory, "Titan - Oak Ridge Leadership Computing Facility," https://www.olcf.ornl.gov/olcf-resources/compute-systems/titan/ (visited March 2018).

[2] Lawrence Livermore National Laboratory, "Sequoia - Computation," https://computation.llnl.gov/computers/sequoia (visited March 2018).

[3] Wikipedia, "Cray-1," https://en.wikipedia.org/wiki/Cray-1 (visited March 2018).

[4] R. M. Russell, "The CRAY-1 Computer System," *Commun. ACM*, vol. 21, no. 1, pp. 63–72, Jan. 1978. [Online]. Available: http://doi.acm.org/10.1145/359327.359336

[5] P. B. Schneck, *The CDC STAR-100*. Boston, MA: Springer US, 1987, pp. 99–117.

[6] A. Moody, G. Bronevetsky, K. Mohror, and B. R. de Supinski, "Design, Modeling, and Evaluation of a Scalable Multi-level Checkpointing System," in *2010 ACM/IEEE International Conference for High Performance Computing, Networking, Storage and Analysis*, Nov 2010, pp. 1–11.

[7] IBM-POWER9-NPU team, "Functionality and Performance of NVLink with POWER9 processors," *IBM Journal of Research & Development*, vol. 62/4-5 - IBM POWER9.

[8] Collaboration of Oak Ridge, Argonne and Livermore (CORAL), "CORAL BENCHMARKS," https://asc.llnl.gov/coral-benchmarks/ (visited March 2018).

[9] N. Liu, J. Cope, P. Carns, C. Carothers, R. Ross, G. Grider, A. Crume, and C. Maltzahn, "On the Role of Burst Buffers in Leadership-Class Storage Systems," in *Mass Storage Systems and Technologies (MSST), 2012 IEEE 28th Symposium on*. IEEE, 2012, pp. 1–11.

[10] W. Bhimji, D. Bard, M. Romanus, D. Paul, A. Ovsyannikov, B. Friesen, M. Bryson, J. Correa, G. K. Lockwood, V. Tsulaia, S. Byna, S. Farrell, D. Gursoy, C. Daley, V. Beckner, B. Van Straalen, D. Trebotich, C. Tull, G. Weber, N. J. Wright, K. Antypas, and Prabhat, "Accelerating Science with the NERSC Burst Buffer Early User Program," Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), Tech. Rep., 2016.

[11] W. Joubert, "Performance of Variant Memory Configurations for Cray XT Systems," in *Cray User Group CUG2990 Proceedings*, 2009.

[12] University of Bristol, "BabelStream," https://github.com/UoB-HPC/BabelStream (visited March 2018).

[13] A. Moody, "Contention-Free Routing for Shift-based Communication in MPI Applications on Large-scale InfiniBand Clusters," LLNL-TR-418522, Lawrence Livermore National Laboratory.(LLNL), Livermore, CA (USA), Tech. Rep., 2009.

[14] Ohio Sate University Network-Based Computing Laboratory, "MVAPICH::Benchamrks," http://mvapich.cse.ohio-state.edu/benchmarks/ (visited March 2018).

[15] G. Hagen, G. R. Jansen, and T. Papenbrock, "Structure of ^{78}Ni from First-Principles Computations," *Phys. Rev. Lett.*, vol. 117, p. 172501, Oct 2016. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.117.172501

[16] W. Tang, B. Wang, S. Ethier, and Z. Lin, "Performance Portability of HPC Discovery Science Software: Fusion Energy Turbulence Simulations at Extreme Scale," *Supercomputing Frontiers and Innovations*, vol. 4, no. 1, pp. 83–97, 2017.

[17] Collaboration of Oak Ridge, Argonne and Livermore (CORAL), "CORAL Benchmarks: AMG2013 Summary," https://asc.llnl.gov/CORAL-benchmarks/Summaries/AMG2013_Summary_v2.3.pdf (visited March 2018).

[18] D. Appelhans and B. Walkup, "Leveraging NVLINK and Asynchronous Data Transfer to Scale Beyond the Memory Capacity of GPUs," in *Proceedings of the 8th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems*, ser. ScalA '17. New York, NY, USA: ACM, 2017, pp. 5:1–5:5. [Online]. Available: http://doi.acm.org/10.1145/3148226.3148232

[19] Collaboration of Oak Ridge, Argonne and Livermore (CORAL), "CORAL Benchmarks: UMT2013 Summary," https://asc.llnl.gov/CORAL-benchmarks/Summaries/UMT2013_Summary_v1.2.pdf (visited March 2018).

[20] ——, "CORAL 2 BENCHMARKS," https://asc.llnl.gov/coral-2-benchmarks/ (visited March 2018).

[21] Cancer Distributed Learning Environment, "CANDLE," https://candle.cels.anl.gov (visited March 2018).

[22] "ImageNet," https://www.image-net.org (visited March 2018).

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv:1512.03385*, 2015.

[24] "Keras," https://keras.io (visited March 2018).

[25] Google, "Tensorflow," https://www.tensorflow.org (visited March 2018).

[26] Uber, "Horovod," https://github.com/uber/horovod (visited March 2018).

[27] S. S. Vazhkudai, R. Miller, D. Tiwari, C. Zimmer, F. Wang, S. Oral, R. Gunasekaran, and D. Steinert, "GUIDE: A Scalable Information Directory Service to Collect, Federate, and Analyze Logs for Operational Insights into a Leadership HPC Facility," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, ser. SC '17. New York, NY, USA: ACM, 2017, pp. 45:1–45:12. [Online]. Available: http://doi.acm.org/10.1145/3126908.3126946