THE CLEVER MACHINATIONS OF LIVERMORE'S SIERRA SUPERCOMPUTER

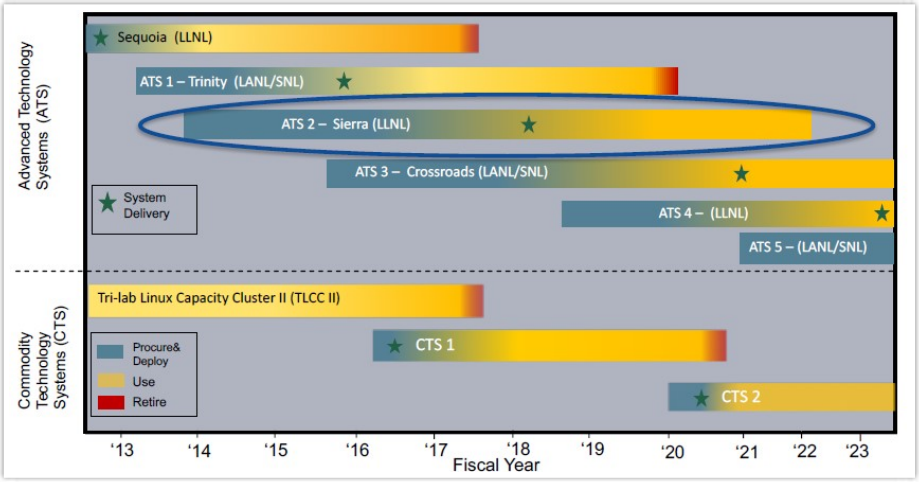
October 5, 2017 Timothy Prickett Morgan

The potent combination of powerful CPUs, floating point laden GPU accelerators, and fast InfiniBand networking are coming to market and reshaping the upper echelons of supercomputing. While Intel is having issues with its future Knights massively parallel X86 processors, which it has not really explained, the two capability class supercomputers that are being built for the US Department of Energy by IBM with the help of Nvidia and Mellanox Technologies, named "Summit" and 'Sierra" and installed at Oak Ridge National Lab and Lawrence Livermore National Laboratory, are beginning to be assembled.

We have previously profiled the nodes in the Summit machine, which like the Sierra system will make use of IBM's impending "Witherspoon" Power S922LC for HPC server, which is said to be coming to market commercially before the end of the year. What has not been clear until now is the differences between the Summit and Sierra systems, and how the specific configurations of the systems, which originally started out being nearly identical, have diverged due to budgetary and technical pressures.

To get some insight into the Sierra system, *The Next Platform* had a chat with Bronis de Supinski, chief technology officer at Livermore Computing, the IT arm of the nuclear research lab that is one of the three key facilities (along with Sandia National Laboratories and Los Alamos National Laboratory) that are stewards for the US military's nuclear arsenal. The technologies deployed by the Tri-Labs has more far-reaching effects than this, of course, since the cutting edge always helps define the rest of the blade that the rest of the world comprises.

The Tri-Labs share the classified supercomputers that they commission, and they leapfrog each other with every other generation of systems, like this:



Back in 2012, Livermore installed Sequoia, the largest BlueGene/Q system massively parallel system that IBM ever built, which has 1.57 million PowerPC AS cores and delivers 20.1 petaflops of peak double precision floating point performance. That is not even the biggest BlueGene/Q that IBM could have built; the system was designed to scale to 256 racks with a peak of 53.6 petaflops, but no one (to our knowledge) ever bought such a beast.

Los Alamos and Sandia got the next big machine for the Tri-Labs, which was known as ATS-1 and named "Trinity" after the initial nuclear bomb tests in the United States during World War II.

Trinity was supposed to be comprised entirely of Intel's "Knights Landing" Xeon Phi processors, but ended up with over 9,500 nodes employing Intel's "Haswell" Xeon E5 processors in 2015 and then had another 9,500 nodes using Knights Landing added in 2016 because of the delays in getting this massively parallel part, which included on-package MCDRAM, an on-chip mesh network, integrated 100 Gb/sec Omni-Path interconnects, and 512-bit AVX floating point units, among other innovations that made it tough. Trinity weighed in at more than 2 PB of total main memory, and delivers more than 40 petaflops of peak performance across its nodes. Significantly, it used a 3.7 PB burst buffer to smooth out storage I/O on an 80 PB parallel file system powered by Lustre – all within a 10 megawatt power envelope.

While Livermore has had GPU-accelerated systems in the past, with Sierra, the lab is going all in with a hybrid machine that marries IBM's Power9 processor with Nvidia's "Volta" Tesla accelerators, with the in-node components connected by the NVLink 2.0 interconnect from Nvidia and the nodes linked to each other through 100 Gb/sec EDR InfiniBand. (Oak Ridge and Livermore had both hoped to have 200 Gb/sec HDR InfiniBand from Mellanox, which was announced last fall, in the systems, but it was not quite ready when the trigger to build the systems was pulled.) Livermore has been vague about the feeds and speeds of Sierra until now, saying only that it would deliver somewhere around 120 petaflops to 150 petaflops of peak performance, with total memory on the order of 2 PB to 2.4 PB and within an 11 megawatts of power envelope. That's about five times as energy efficient as the Sequoia system, which is not a bad improvement for machines installed five to six years apart.

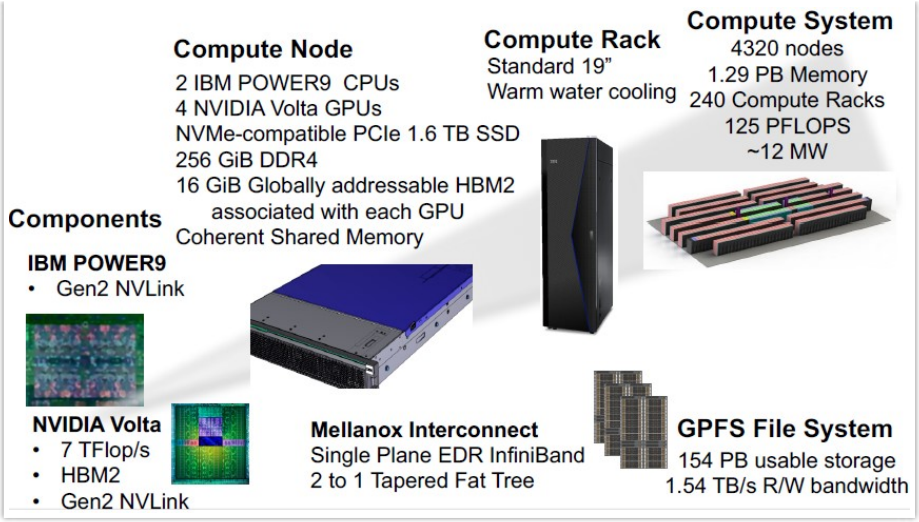
De Supinksi confirmed the configurations of the Sierra nodes, telling us that it did indeed have two Volta accelerators per Power9 processor within the Witherspoon server. (That leaves two SXM2 slots open, but these capability class supercomputers are not generally upgraded in the field.)

As it turns out, the massive increase in memory prices starting in late 2016 and throughout 2017 have had a dramatic impact on the configuration of the Sierra nodes and therefore the entire supercomputer. While the original idea was to have the Summit and Sierra machines have the same configuration, simplifying things for all parties involved, Summit ended up with six Volta accelerators and two Power9 chips with 512 GB of main memory per node, while Sierra backed off to four Voltas and two Power9 processors with 256 GB per node. Their network configurations are different, too. And De Supinksi explains why.

"First of all, let's talk about the difference in the number of GPUs," he says. "When we signed the original contract, IBM said we could have two or three GPUs per Power9. Oak Ridge had lots of GPUs already and were confident that their applications could use GPUs, so they chose three GPUs. We said on the other hand that while we did have machines with GPUs already, they are relatively small systems and they are primarily used for visualization. It is not true that none of our codes had run on GPUs before, but we had not had made our entire workload have to use GPUs. The cost-effectiveness of the system was therefore dependent on how ready our codes were for the use of GPUs and related issues for the configuration to make a balanced system. We don't know roughly three years before the system is going to be delivered how ready our codes will be. So we have a period of time to get our applications ready to use the system and to have a more informed decision about how many GPUs we should get. We put into the contract what the exact node architecture and therefore the number of nodes that our money would buy was a go-no go decision that would be made in March 2017 based on how ready our applications would be."

The way the CORAL procurement contract was structured, according to de Supinski, Livermore had target requirements and then, as technology and software development progresses over the years, it has a go-no go review. At at that point, given all of the data, the target requirements are turned into hard requirements.

"So when we put out the RFP, we know that the offer can be a little more aggressive because we know we can revise what we are actually going to get," says de Supinski. "And if a revision from the vendor is too significant, we can say no go because that is not what we signed up for."



The choice of the number of GPUs was a complicated one, and it does not always make sense to just cram as many as possible into a chassis. You have to balance CPU and GPU compute based on the expected workloads and you have to get the right ratios of memory bandwidth and latency in these elements relative to the performance of all of the components. There is no point in adding a compute device that can't be fed because of the nature of the data or program.

"We actually had a fair amount of discussion internally over two or three GPUs, and several of the owners of our physics engines, which are at the center of our applications, argued for getting three GPUs," de Supinksi recalls. "But in the end, we figured our workloads would be best served by getting nodes with two GPUs because the nodes cost less and therefore you can get more nodes."

The fact that the nodes were cheaper was important for another reason: The cost of memory is a lot higher than expected, and while flash is less expensive than Livermore planned, it is still not cheap enough to offset that memory increase. For this reason, the CORAL contract at Livermore has provisions where IBM and the lab share the risks on costs. This flexibility is vital, given how far out into the future these contracts go. If any supercomputer center asked any prime contractor to make a hard commitment on how much memory is going to cost, what they would do is take a very conservative prediction on what memory costs would be and therefore offer a much smaller system. So Livermore asked vendors bidding on the RFPs to make projections and then the vendor who won the deal and Livermore would see what the costs are when the system was built and adjust accordingly.

"As it turns out, in late December last year and early January this year, memory costs just started going up precipitously," de Supinkski explains. And the reason for that is that the mobile market has greatly increased their demand for memory. All of the phone manufacturers started recommending you get twice as much memory, and memory companies sell memory into the mobile market, and typically in the server market the margins are much lower. Because the companies cannot increase their supply, and the mobile market will pay the higher price, server memory had to meet the same margins. The problem is that, with the rise in memory prices, the cost of memory was more than 2X what our original projections were. And that did not fully compensate. So we had to either reduce the size of the system, or fill fewer memory slots, both of which would impact the performance of the system, and we did not want to do that. Or come up with something else."

	Sierra	uSierra
Nodes	4,320	684
POWER9 processors per node	2	2
GV100 (Volta) GPUs per node	4	4
Node Peak (TFLOP/s)	29.1	29.1
System Peak (PFLOP/s)	125	19.9
Node Memory (GiB)	320	320
System Memory (PiB)	1.29	0.209
Interconnect	2x IB EDR	2x IB EDR
Off-Node Aggregate b/w (GB/s)	45.5	45.5
Compute racks	240	38
Network and Infrastructure racks	13	4
Storage Racks	24	4

The feeds and speeds of the Sierra and unclassified Sierra (uSierra) systems. The uSierra system is not acquired through the CORAL contract.

So Livermore's techies looked at their applications, specifically the uncertainty qualification (UQ) workload that is a mission for the Sierra system and that uses the vast amounts of compute not to just do simulations, but to tweak variables in multiple simulations to see how sensitive they are to those tweaks to try to see how good or bad those simulations are.

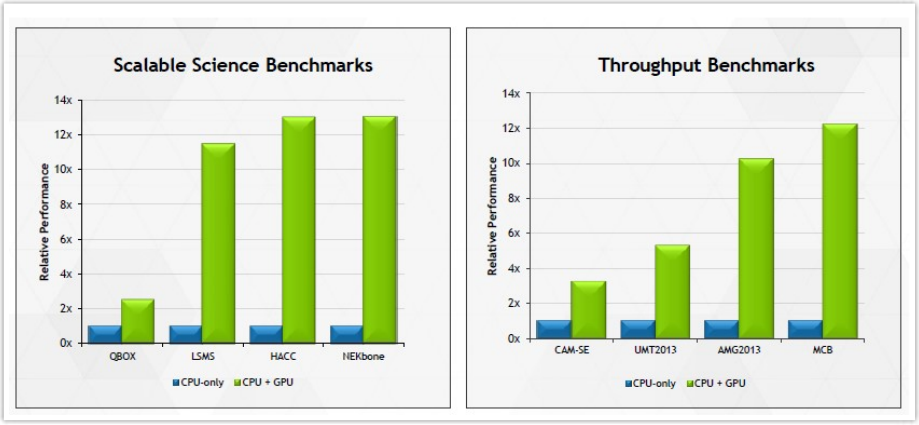
Unlike Oak Ridge, which could come up with the extra money to keep the memory configuration at 512 GB per node, Livermore could not. So Livermore asked IBM to shift from the 32 Gb DDR4 memory chips used in the original configuration to cheaper and less capacious 16 Gb memory chips. That cut the memory capacity on the Sierra nodes in half to 256 GB, but it kept the memory bandwidth the same because the memory chips clocked at the same speed and filled the same number of memory slots. (We presume it is 2.4 GHz DDR4 memory, but we don't know.)

Livermore realized that for the bulk of what its applications, it could not only cut the memory in half, but it could also taper the InfiniBand network, and actually have little to no performance difference. The InfiniBand network is a fat tree interconnect and has two 100 Gb/sec adapters per card and one card per node, just like with the Summit machines at Oak Ridge. So the connections and bandwidths out of the servers and into the top of rack switches is the same in the Summit and Sierra machines. But Livermore cut the number of InfiniBand director switches in the two layers higher in the network fabric in half, which cuts the bi-section bandwidth in half between the top of rack switches the director fabric. For most jobs, says de Supinksi, the impact of the network change is under 1 percent performance hit; it is higher for most jobs, but across all jobs it is around 1 percent. This network change more than compensated for the change in memory prices, and so as a result Livermore got 5 percent more nodes.

Thanks to the "Titan" supercomputer installed at Oak Ridge back in October 2012, which was built by Cray using a balanced number of AMD Opteron CPUs and earlier "Kepler" generation Tesla GPU accelerators, that lab had lots of experience with GPUs and it is not surprising it boosted the GPU count in its IBM Witherspoon nodes to the max at six. But Livermore was dealing with more uncertainty, particularly three years into the future with the code and the hardware both needing to change at the same time. The cost-effectiveness of the system was therefore absolutely dependent on how ready Livermore's codes would be for the GPUs, and because of the fixed budget – we estimate that Sierra represented $125 million of the $325 million of the CORAL procurement based on a flat flops cost between the two machines, with Summit being the remaining original $200 million in spending – the configuration to make a balanced system was not trivial, given all of these changes. To be sure, flash was less expensive than planned, so that helped, and sticking with 100 Gb/sec InfiniBand instead of waiting for 200 Gb/sec InfiniBand probably did, too, although de Supinksi did not comment on that.

The point is, it is no simple thing to just configure up a machine like Sierra three or four years in advance. The good news, says de Supinksi, is that Livermore's code is making the jump from the massively parallel Power-based Sequoia to the hybrid Sierra.

"Looking at our workloads, we were actually quite happy with how ready they are to use GPUs," he says. "We have made a lot of progress, in no small part due to the tight relationship we have had with IBM and Nvidia thanks to non-recurring engineering, or NRE, funding and through a center of excellence we have in place, which has IBM and Nvidia engineers working directly with our application programmers."



Projected application speedup on the Sierra system, CPU versus CPU+GPU

Unlike Oak Ridge, Livermore does not use OpenACC to parallelize and dispatch workloads to the GPUs. "We have a thing that is being developed called RAJA that uses C and C++ features, and Sandia has something similar called Kokkos that allows programmers to easily change parallelism using CUDA or OpenMP or whatever. Most of our applications will be using OpenMP, some are using CUDA. OpenMP 4.0 and 4.5 provide a way to use devices to target accelerators through device constructs, and that is what a lot of what we will be using. Most of our applications are C++, and there is a fair amount of C and a tiny amount of Fortran. Most of our applications are actually multi-language, but we have a few holdouts that are primarily Fortran codes. Sandia is similar. Most of the Los Alamos applications are still Fortran."

There is of course storage behind this Sierra beast. The original plan called for a 120 PB Spectrum Storage (GPFS) cluster with 1 TB/sec of write and 1.2 TB/sec read bandwidth. This is pretty fast, but there are clustered file systems already hitting this performance. So IBM and Livermore started tweaking, and came up with changes to GPFS that boosted the read and write performance to 1.54 TB/sec and fatter disk drives that delivered 154 PB, all for the same price in the CORAL contract. The metadata performance for GPFS was also improved, and this was a big concern for Tri-Labs.

What we wanted to know is, looking ahead, will Livermore do a hybrid CPU-GPU system again. And the answer is as non-committal as it has to be, with the ATS-4 contract being four years or so out into the future.

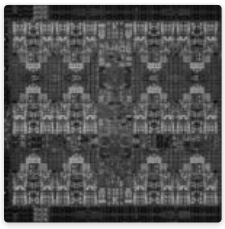
"The fact is, it will be an open RFP and procurement, and we will see what types of systems we get bid," says de Supinksi. "If you would ask me what I think is the most likely type of system we would select for ATS-4, I would say that it would have GPU acceleration. But it really depends on what is offered and what we think will provide the best cost/benefit to our applications. Things can change a lot as we go about seeing what is available and what is being bid. Four years ago, everybody from Livermore, including myself, was saying that we would never go with a GPU-based system. Then we looked at what was available, and we decided that the best value for the ASC program is the Sierra system that we are getting. And I feel pretty comfortable saying that we made the correct decision there."

This time around, Livermore is perhaps more fortunate than Argonne National Laboratory, which was trying to move from BlueGene/Q to a future system from Cray built using Intel's "Knights Hill" processors called "Aurora" that was due in 2018 and that has been pushed out to 2021 to a new system with an architecture unknown to anyone outside of Argonne, Intel, and Cray. But it looks like Argonne is going to get to exascale first in the United States, so that is something.

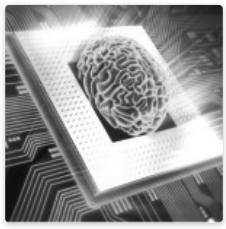
Sierra will be under construction and go through the usual testing and qualification process at Livermore, and is expected to be up and running in under a year from now.

SIMILAR VEIN

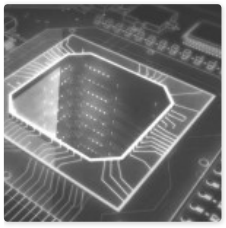
Details Emerge On "Summit" Power Tesla AI Supercomputer



The Power9 Rollout Begins With Summit And Sierra Supercomputers



Thinking Through The Cognitive HPC Nexus With Big Blue



The Year Ahead for GPU Accelerated Supercomputing

BSC Keeps Its HPC Options Open With MareNostrum 4

NVLink Shines On Power9 For AI And HPC Tests