ATOS REJIGGERS SEQUANA SUPERCOMPUTERS, ADDS AMD ROME CPUS

November 29, 2018 Timothy Prickett Morgan

The Sequana line of supercomputers from the Bull division of Atos offers some of the highest compute density available

in the HPC realm. The initial Sequana X1000 machines made their debut back in late 2015, sporting "Broadwell" Xeon E5 and "Knights Landing" Xeon Phi processors from Intel, and over time the line has expanded to include server nodes that have Intel Xeon SP processors goosed by Nvidia "Pascal" and "Volta" Tesla GPU accelerators.

A little more than a year after after the debut of the Sequana X1000s, the Barcelona Supercomputer Center in Spain commissioned Atos to build its Mont-Blanc 3 system, and asked for Marvell/Cavium's ThunderX2 Arm server processors to be the main computing engines for the system, so Atos added these to the compute portfolio and others can benefit from that investment by BSC. Atos also added a compute blade that supported the current Intel "Skylake" Xeon SP processors in CPU-only nodes, and these will therefore support the impending "Cascade Lake" follow-ons that are expected to be announced by Intel soon since Skylake and Cascade Lake are socket compatible. It is not clear, unless you look at the details very carefully, that the "Cascade Lake AP" processor, which crams two 24-core Cascade Lake chips into a single socket, will plug into the same sockets. But it sure looks like it, as we explain below.

Companies looking at the Sequana X1000 machines today might want to take a gander at the forthcoming Sequana XH2000 systems, which have a more streamlined rack structure and which will also support AMD's future "Rome" Epyc server chips as well as 100 Gb/sec Ethernet interconnects, faster 200 Gb/sec InfiniBand interconnects, and current 100 Gb/sec BXI interconnects that Atos already supports in the Sequana X1000 systems.

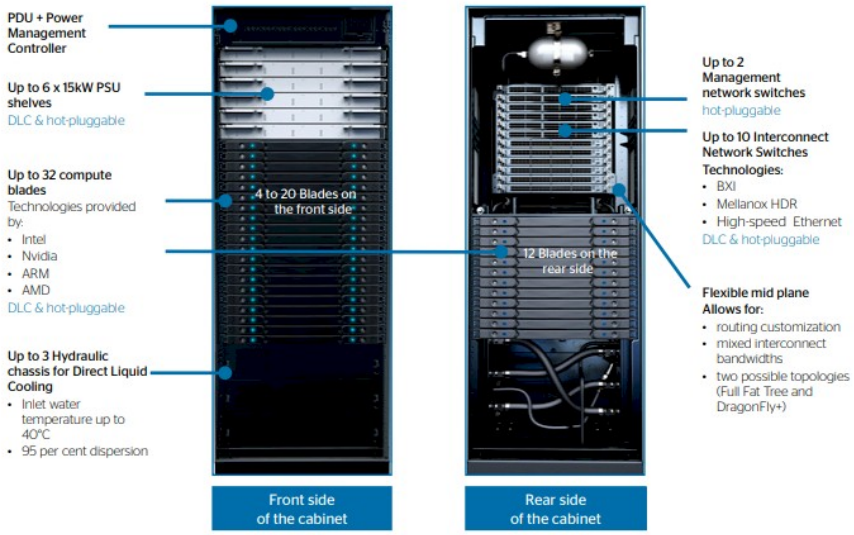
Before getting into all that, let's talk about the differences between the existing Sequana X1000 and the new Sequana XH2000. Here is what the Sequana X1000 cell, which includes compute, networking, power distribution, and cooling, looks like:

As you can see, the Sequana X1000 has two racks with 48 compute trays, which are 1U high and which contain three X86 or Arm server nodes per tray, for a total of 96 trays and 288 nodes. The compute racks have four 90 kilowatt power distribution units on the top of the compute racks and three water-cooling heat exchangers. The switch cabinet includes the switch fabric, which implemented a midplane to link the backend of the servers to the switching in the central cabinet. This made the Sequana X1000 a giant, rack-scale blade server, in essence. Bull built its own EDR InfiniBand switches and adapters from raw silicon acquired from Mellanox, and it took a dozen 36-port switches to link the 288 nodes to each other and another dozen 36-port switches to linking each Sequana X1000 cell to others and external storage.

As we explained before, the BXI interconnect has bandwidth that Atos says is comparable to 100 Gb/sec InfiniBand or Ethernet, and it is a commercialized version of the the Portals protocol that has been evolving under its development at Sandia National Laboratories for the past three decades. Various Sandia machines have used different implementations of Portals, including the Paragon parallel RISC system built by Intel in 1991, the ASCI Red parallel X86 system built by Intel in 1994, and the ASCI "Red Storm" XT system built by Cray. Unlike Intel's Omni-Path interconnect, which Atos did not support with the Sequana X1000 and somewhat like InfiniBand that it did, the BXI interconnect offloads all communication overhead for the interconnect to the network interface cards in the servers and the switches;; InfiniBand offloads somewhere around 85 percent of the

work, according to Mellanox, which has the goal of reaching 100 percent. This offloading leaves all of the CPU capacity available for doing compute rather than managing network overhead. The BXI interconnect switches and adapters are the same form factors as the InfiniBand variants, but BXI scaled a lot further with supporting up to 64,000 nodes in a three-tier network, with EDR InfiniBand supporting up to 11,644 ports. But, **with 200 Gb/sec Quantum InfiniBand, also called HDR,** Mellanox can scale a three-tier network to 64,000 ports running at 200 Gb/sec or 128,000 ports running at 100 Gb/sec with cable splitters.

Now take a gander at the forthcoming Sequana XH2000 setup:



The first thing you will notice is that the switching cabinet is gone and the network and Octopus midplane is being pulled into the compute rack. This means sacrificing some compute in the rack, but it also means getting ride of the non-standard cell rack size.

The Sequana XH2000 has six 15 kilowatt power shelves and the power distribution and management controllers at the top front of the rack, with up 20 compute trays in the front and the three direct liquid cooling units at the bottom of the cabinet. In the back, there is room for a dozen more compute trays, plus two management switches for the rack and up to ten interconnect switches. (More on this in a moment.)

The three existing compute blades that were supported on the Sequana X1000 supercomputers will slide right into the Sequana XH2000 machines, which is the point of having an architecture that spans many different generations. That would be the X1120 blade based on the Skylake Xeon SP processors from Intel, and the X1310 blade based on the Marvell/Cavium ThunderX2; both of these have three two-socket nodes on each blade. The X115 blade that has two Xeon SP processors in the host node and plus four Nvidia "Volta" Tesla V100 GPU accelerators that link to the host processors over PCI-Express 3.0 links. The new blade is based on the Rome Epyc processor from AMD, and it has two of these processors on each of the three nodes. Each of the nodes has its InfiniBand or BXI networking on a mezzanine card, with one port per slot. Details on the Ethernet switch were not given here, but presumably there is a one-port mezzanine card here, too, that snaps into the system board of each node on the blade. Here are the basic feeds and speeds of the Sequana compute blades, which we think will also snap into the older X1000 machines even though this isn't said anywhere in the product literature:

	BullSequana X H2410 AMD blade	BullSequana X 1120 Intel blade	BullSequana X 1115 GPU blade	BullSequana X 1310 Arm blade
				
Design	1U blade comprising 3 compute nodes side-by-side	1U blade comprising 3 compute nodes side-by-side	1U blade with 1 accelerated compute node	1U blade comprising 3 compute nodes side-by-side
Processors	3x2 AMD® EPYC Rome® Processor	3 x 2 Intel® Xeon® Processor Scalable Family	2 Intel® Xeon® processors Scalable Family 4 Nvidia® Volta V100 16/32 GB GPUs	3 x 2 Cavium® ThunderX2™ Armv8 processors with up to 32 cores
Architecture	3x1 motherboard	3 x 1 Intel® C620 chipset	1 Intel® C620 chipset	3 x 1 motherboard compatible with Cavium Borg reference platform
Memory	3x16 DDR4 memory slots (max 2048GB with 128 GB DIMMS)	3 x 16 DDR4 memory slots (max 1024 GB with 64 GB DIMMs) + 4 optional NVRAM DIMMS (NVRAM availability TBC)	12 DDR4 memory slots (max 768 GB with 64 GB DIMMs)	3 x 16 DDR4 memory slots (max 2048 GB with 128 GB DIMMs)
I/O slots	InfiniBand HDR 1 port mezzanine board PCIe gen4 BXI 1 port mezzanine board	InfiniBand HDR 1 port mezzanine board or BXI	InfiniBand HDR 1 port mezzanine board or BXI 1 port mezzanine board	InfiniBand HDR 1 port mezzanine board BXI 1 port mezzanine board
Storage	3x1 optional NVMe M2 format	3 x 1 optional SATA drive 3 x 1 optional NVMePCIe SSD drive via PCIe switch	1 optional SATA SSD drive	1 optional SATA SSD drive

The thing to notice here is that for all of the two-socket CPU-only nodes, they all have sixteen memory slots per node. So the memory capacity is the same using identical capacity DDR4 memory sticks. The X1120 Intel blade is going to support the Cascade Lake Xeons because it says that it will support an optional four NVRAM DIMMs – that means Optane memory sticks. This suggests further that the machine us going to support the Cascade Lake AP variant

(which has two 24-core chips with a total of 12 memory controllers) in this modified X1120 blade because the way the math works, Atos said that it is using eight of the memory controllers on the Cascade Lake chip to drive DDR4 memory and that leaves four to drive the Optane DIMMs. This should mean that in this case at least, the modified X1120 blade will have the same DDR4 memory bandwidth as the X2410 Epyc blade and the X1310 Arm blade. (Versions of this X1120 blade using either plain vanilla Skylake or Cascade Lake processors would only support twelve memory slots, as the original Skylake compute node did, **which you can see here**. If this is how Intel is going to position the Cascade Lake AP, this is an interesting twist.

Now to the switching. Atos is supporting its homegrown 100 Gb/sec BXI interconnect, based on 48-port switches using custom switch and adapter ASICs code-named "Divio" and "Lutetia" respectively, in the new supercomputers. The faster HDR InfiniBand (again bought OEM from Mellanox) that delivers 40 ports per switch at that 200 Gb/sec speed or 80 ports running at 100 Gb/sec with splitters, is also available, and so is an unspecified "high-speed" Ethernet switch, which is a 48-port device running at 100 Gb/sec. (This Ethernet does not seem to come from Mellanox, which supports 32 ports running at 100 Gb/sec with Spectrum chips and 64 ports running at 100 Gb/sec with the Spectrum 2 chips.) Full fat tree and Dragonfly+ topologies are supported on all three types of networks.

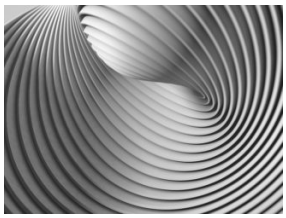
Obviously, getting rid of that switch cabinet in the middle does a few good things. For one, the system now has uniform rack sizes, which works well in tiled datacenters (those that are still done with raised floors) that have uniform tile sizes. By moving to higher radix switches – meaning there are more ports per switch – it takes fewer switches to link the nodes to each other. You will notice that the aggregate power distribution for the Sequana X1000 cell was 120 kilowatts (eight 15 kilowatt units), but with the Sequana XH2000, it has dropped down to 90 kilowatts (six 15 kilowatt units). More efficient switching is a big part of that. But now, the cell size can also be reduced down to one rack if need be, with compute and power and cooling in the front and more

compute and switching in the back. This is a far simpler and more elegant design, although the original Sequana did look neat. This change in form factor also allows for the aggregation layer of the supercomputer network to be pulled outside of the rack and put at the end of a row in the datacenter, as is common.

The net effect is that the old Sequana X1000 machine took up 40 percent more volume in the datacenter and offered 50 percent more compute across two compute racks than the Sequana XH2000. The Sequana XH2000 is, on first glance, actually a little bit less dense on the compute front if you just count nodes and sockets. However, AMD, Intel, and at some point Cavium with the "Triton" ThunderX3, are all going to pack much more computing oomph into each socket, so the net-net is that the newer Sequana XH2000 is going to make it up in core volume.

RELATED ARTICLES

ORACLE PUTS TOGETHER RDMA, BARE METAL FOR HPC



IN-DEPTH WITH THE NEXT PLATFORM: MACHINE LEARNING, HPC WITH UNIVA

OPENMP REACHES INTO THE PARALLEL UNIVERSE OF GPUS

BE THE FIRST TO COMMENT

LEAVE A REPLY