June 02, 2013

Full Details Uncovered on Chinese Top Supercomputer

Nicole Hemsoth

At the end of May, an international group of high performance computing researchers gathered at the International HPC Forum in Changsha, China. One of the talks detailed the specs for the new Tianhe-2 system, which as we reported last week, is expected to rather dramatically top the Top500 list of the world's fastest supercomputers.

Artist's rendering of the system as it will look once finally implemented at its final destination.

As noted previously, the system will be housed at the National Supercomputer Center in Guangzhou and has been aimed at providing an open platform for research and education and to provide a high performance computing service for southern China.

Dr. Jack Dongarra from the University of Tennessee and Oak Ridge National Lab, one of the founders of the Top500, was on hand for the event in China and shared a draft document that offers deep detail on the full scope of the Tianhe-2, which will, barring any completely unexpected surprises, far surpass the Cray-built Titan.

The 16,000-node Inspur-built Tianhe-2 is based on Ivy Bridge (32,000 sockets) and 48,000 Xeon Phi boards, meaning a total of 3,120,000 cores. Each of the nodes sports 2 Ivy Bridge sockets and 3 Phi boards.

According to Dongarra, there are some new and notable LINPACK results:

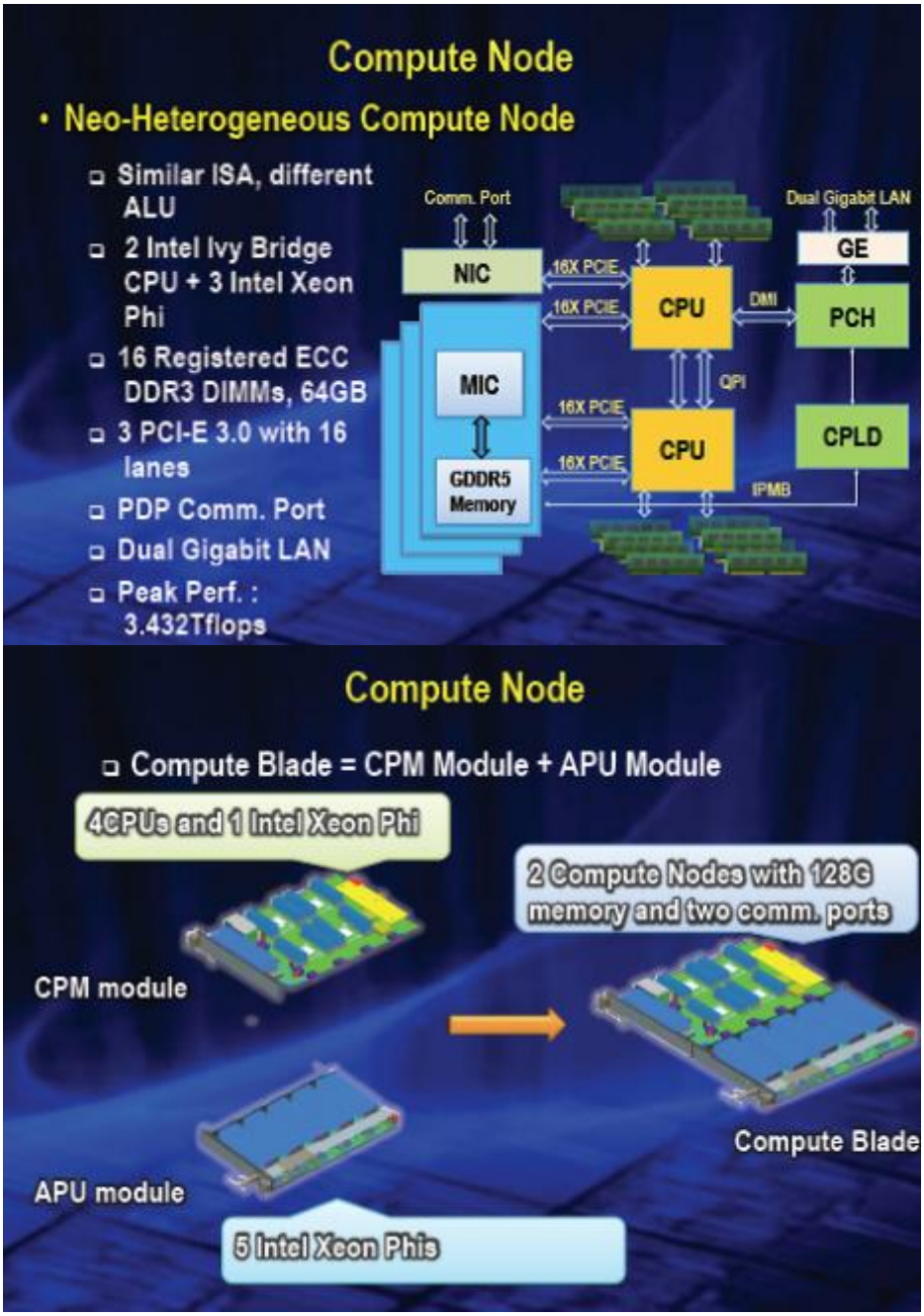
I was sent results showing a run of HPL benchmark using 14,336 nodes, that run was made using 50 GB of the memory of each node and ***achieved 30.65 petaflops out of a theoretical peak of 49.19 petaflops, or an efficiency of 62.3% of theoretical peak performance taking a little over 5 hours to complete.****The fastest result shown was using 90% of the machine. They are expecting to make improvements and increase the number of nodes used in the test.*

This certainly seems to confirm that this will indeed be the top system on this June's list. But let's take a closer look at some architectural elements to put those numbers in context...

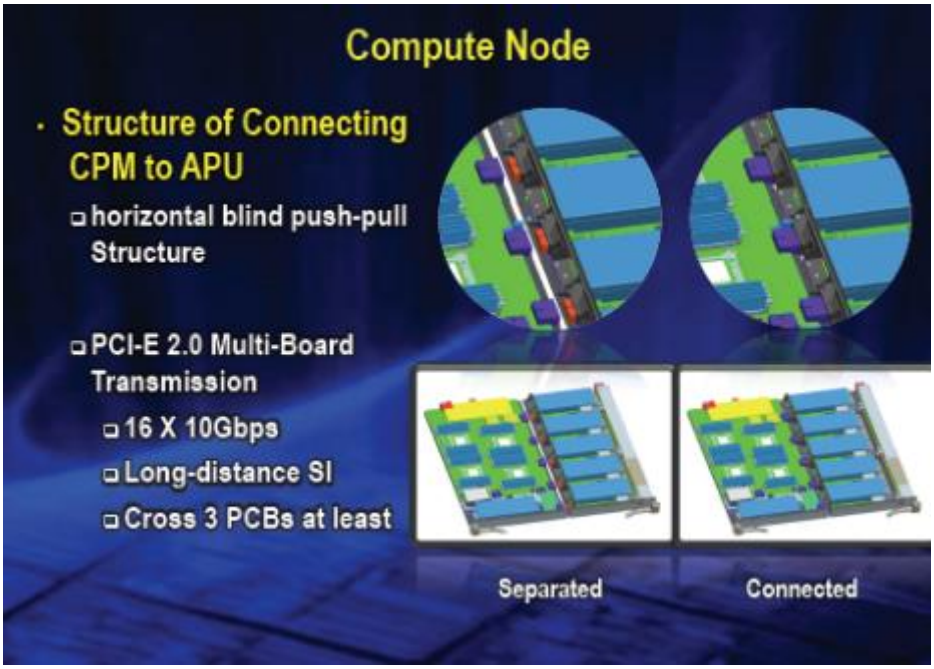
Interestingly, each of the Phi boards have 57 cores instead of 61. This is because they were early in the production cycle at the time and yield was an issue. Still each of the 57 cores can boast 4 threads of execution and each thread can hit 4 flops per cycle. By Dongarra's estimate, the 1.1 GHz cycle time produces a theoretical peak of 1.003 teraflops for each Phi element.

Each of the nodes is laden with 64 GB of memory, each of the Phi elements come with 8 GB of memory for a total of 88 GB of memory per node for a total of full system memory at 1.404 petabytes. There is not a lot of detail about the storage infrastructure, but there is a global shared parallel storage system sporting 12.4 petabytes.

According to Dongarra, there are "2 nodes per board, 16 boards per frame, 4 frames per rack, and 125 racks make up the system." He says that the compute board has two compute nodes and is composed of two halves—the CPM and APM. The CPM portion of the board contains the 4 Ivy Bridge processors, memory and 1 Xeon Phi board while the CPM half contains the 5 Xeon Phi boards.



There are also 5 horizontal blind push-pull connections on the edge; connections from the Ivy Bridges to each of the coprocessors are made via PCI-E 2, which has 16 lanes and are 10 Gbps each. Dongarra points out that the actual design and implementation of the board is for PCI-E 3.0 but the Phi only supports PCI0E 2. There is also a PCI-E connection to the NIC.

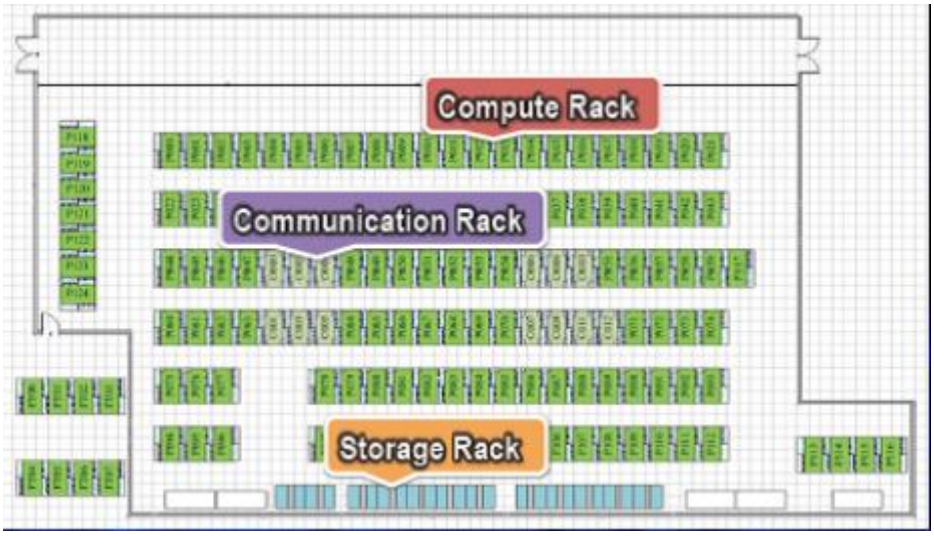


We already knew that this was a system from the Chinese IT company, Inspur. According to Dongarra, "Inspur contributed to the manufacturing of the printed circuit boards and is also contributing to the system installation and testing." At this point, the system is still being assembled and tested at the National University of Defense Technology before being installed at its permanent home.

As we know from the original Tianhe-1A system, NUDT has been hard at work on their own interconnects. On the TH-2, they are using their TH Express-2 interconnect network, which taps a fat tree topology with 13 switches, each with 576 ports at the top level.

As Dongarra notes, "This is an optoelectronics hybrid transport technology and runs a proprietary network. The interconnect uses their own chip set. The high radix router ASIC called NRC has a 90 nm feature size with a 17.16x17.16 mm die and 2577 pins."

He says that "the throughput of a single NRC is 2.56 Tbps. The network interface ASIC called NIC has the same feature size and package as the NIC, the die size is 10.76x10.76 mm, 675 pins and uses PCI-E G2 16X. A broadcast operation via MPI was running at 6.36 GB/s and the latency measured with 64K of data within 12,000 nodes is about 85us.



Dongarra says that the 720 square meter footprint means a rather confined space and isn't optimally laid out. However, this is just temporary since when it arrives in its permanent home in

Guangzhou it will be laid out more efficiently, as seen in the artist's rendering of the system at the top of the article.

The peak power consumption under load for the system is 17.6 MWs, but this is just for the processors, memory and interconnect network. When the closely-coupled chilled water with customized liquid water cooling unit operations are added in, the total consumption is 24 MWs. Dongarra says that it has a high cooling capacity of 80 KW and when installed at its home site, it will use city water as its source. Power load is monitored by a series of lights on the cabinet doors.

For far more details about these and other aspects of the Tianhe-2 system, check out Dr. Dongarra's extensive report...

http://www.netlib.org/utk/people/JackDongarra/PAPERS/tianhe-2-dongarra-report.pdf