IBM US Nuke-lab Beast ‘Sequoia’ is Top of the Flops (Petaflops, that is)

06.18.2012

Mi piace

By Timothy Prickett Morgan • Get more from this author

For the second time in the past two years, a new supercomputer has taken the top ranking in the Top 500 list of supercomputers – and it does not use a hybrid CPU-GPU architecture. But the question everyone will be asking at the International Super Computing conference in Hamburg, Germany today is whether this is the last hurrah for such monolithic parallel machines and whether the move toward hybrid machines where GPUs or other kinds of coprocessors do most of the work is inevitable.

No one can predict the future, of course, even if they happen to be Lawrence Livermore National Laboratory (LLNL) and even if they happen to have just fired up IBM’s “Sequoia” BlueGene/Q beast, which has been put through the Linpack benchmark paces, delivering 16.32 petaflops of sustained performance running across the 1.57 million PowerPC cores inside the box.

Sequoia has a peak theoretical performance of 20.1 petaflops, so 81.1 per cent of the possible clocks in the box that *could* do work running Linpack did so when the benchmark test was done. LLNL was where the original BlueGene/L super was commercialized, so that particular Department of Energy nuke lab knows how to tune the massively parallel Power machine better than anyone on the planet, meaning the efficiency is not a surprise. And the supercomputing lab was absolutely banking on the Sequoia machine’s power efficiency; the super-efficient beast burns only 7.89 megawatts to deliver that 16.32 petaflops of oomph.

The former top flopper on the list – the K massively parallel Sparc64-VIIIfx machine built by Fujitsu for the Japanese government, which shifts down to number two – had a sustained Linpack performance of 10.5 petaflops against a peak of 11.3 petaflops, for an impressive efficiency of 93.2 per cent. But this monster Sparc box sucks down 12.7 megawatts when it is running, or a mere 830 megaflops per watt. Sequoia is 2.5 times as energy efficient as K, at least when running Linpack.

Thus far, the largest hybrid CPU-GPU machine on the list, the Tianha-1A super at the National Supercomputing Center in Tianjin, China – which uses a mix of Intel Xeon X5760 processors and Nvidia Tesla M2050 GPU coprocessors – wastes 45.4 per cent of its aggregate number-crunching capability because of the hybrid programming model and the latencies in talking between the CPU and the GPU. The Tianha-1A, which delivers 2.57 petaflops of performance, only delivers 635 megaflops per watt. By contrast, the Sequoia machine is 3.25 times as energy efficient per unit of real work done – again, assuming that you consider Linpack indicative of real work.

The Top 500 list does not include the cost of the machines, which is also a very important factor. The BlueGene/Q machine costs millions of dollars per rack – IBM does not say how much precisely, since this is essentially a custom product – and can scale to 512 racks and up to 100 petaflops of aggregate peak performance. The trouble is, who has the estimated $1bn to build such a behemoth? Governments have a hard time coming up with that kind of cash these days, even if they do want to simulate nuclear weapons.

The point is that a real ranking of the world's supercomputers would look at sustained performance, computational efficiency, performance per watt, and bang for the buck. Three out of four ain't bad, but it also ain't enough. (Moreover, the power draw figures are not available for all of the machines on the Top 500 list, so it is more like two-and-a-half out of four.)

LLNL awarded Big Blue the contract to build Sequoia back in February 2009. The massively parallel machine is based on IBM's 18-core PowerPC A2 processor, which is a 64-bit chip that has one core to run the Linux kernel; one spare in case one goes dead; and 16 cores for doing compute tasks. One chip and 16GB of memory are packaged up on a compute card, while 32 cards are plugged into a node card – which has optical modules to link into the 5D torus that allows all the nodes to talk to each other. You put 16 of these node cards in a chassis with eight I/O drawers to make a half-rack midplane, and then stack two of these to make a rack.

The BlueGene/Q interconnect runs at 40Gb/sec and has a node-to-node latency hop of 2.5 microseconds. The logic for that 5D torus interconnect is embedded on the PowerPC A2 chips, which run at 1.6GHz, with 11 links running at 2GB/sec. Two of these can be used for PCI-Express 2.0 x8 peripheral slots. The 14-port crossbar switch/router at the center of the chip supports point-to-point, collective, and barrier messages and also implements direct memory access between nodes.

Like K and its "Tofu" 6D torus/mesh interconnect, this flagship BlueGene/Q super is no slouch on any dimension you want to measure. Fujitsu has commercialized the K super as the PrimeHPC FX10 line, which has a 16-core Sparc64-IXfx processor and which scales to 23 petaflops. The only problem is the all-out FX10 machine with 1,024 racks – that's 98,304 compute nodes and 6PB of main memory – burns 23 megawatts and costs $655.4m at list price. That's a big number, even for the HPC racket. (And no, it cannot play *Crysis*, and neither can BlueGene/Q. A Windows-based *ceepie-geepie* certainly could, so there is that to consider.)

IBM takes five out of the top 10

IBM is having a very good Top 500 this time around, with five of the top 10 systems bearing its stripey moniker.

Number three on the list behind the K super is another BlueGene/Q machine called "Mira," which is installed at Argonne National Laboratory. This machine is essentially half of Sequoia.

Number four on the list is SuperMUC, another IBM box, but this one is based on Intel's latest Xeon E5-2680 processors, which are plopped into IBM's iDataPlex dx360 M4 rackish-bladish servers. SuperMUC was built by IBM under contract from the Partnership for Advanced Computing in Europe (PRACE) for the Leibniz-Rechenzentrum (LRZ) located in Germany. The SuperMUC contract was awarded in January 2011, and the neat thing about this box is that there are water blocks on processors and main memory on the iDataPlex system boards, and a closed-loop water-cooling system uses relatively warm water (up to 45 degrees Celsius, which is 113 degrees Fahrenheit) to keep these active components from overheating. (We'll be taking a separate look at SuperMUC later.) SuperMUC cost $110.9m to build and operate over five years, according to the contract; the machine currently has 147,456 Xeon cores and delivers just under 2.9 petaflops of sustained performance on the Linpack test with a computational efficiency of 91 per cent. That's pretty good, and is no doubt helped by the 56Gb/sec FDR InfiniBand network linking those iDataPlex nodes together. But the machine does burn 3.42 megawatts, so it only delivers 847 megaflops per watt. LLNL's Sequoia is 2.44 times as energy efficient.

Number five is the Tianhe-1A machine that *ceepie-geepie* weighs in at 2.57 petaflops and which was the fastest machine in the world back in November 2010 and signaled the arrival of China as a contender in the exascale supercomputing arms race.

The "Jaguar" massively parallel supercomputer ranks sixth on the list and is installed at Oak Ridge National Laboratory, another nuke lab controlled by the US Department of Energy. Jaguar is in the process of being upgraded to the 20-petaflops "Titan" super *ceepie-geepie*. But this process is only just beginning, with nodes being upgraded by Cray to the latest Opteron 6274 processors and boosted with the latest "Gemini" XE interconnect and some of the nodes getting Nvidia Tesla M2090 coprocessors.

As it now stands, Jaguar has 298,592 cores and delivers 1.94 petaflops of sustained oomph (at a computational efficiency of 73.9 percent across those CPUs and GPUs), but Jaguar consumes an incredible 5.14 megawatts of electricity as it runs. That works out to only 377.5 megaflops per watt. At this point in the transformation from Jaguar to Titan, Sequoia is 5.5 times as energy efficient. That gap will close considerably as "Kepler" Tesla K20 GPUs are added to the Titan machine this fall and Oak Ridge takes advantage of GPUDirect and all ofthe funky innovations that Nvidia has put into these GPU coprocessors.

The rest of the best

Numbers seven and eight on the June 2012 Top 500 supers ranking are both BlueGene/Q machines. Number seven is nick-named "Fermi" and is installed at CINECA, a consortium of 54 universities in Italy that has a long history of buying IBM and Cray supers. The Fermi super has 163,830 cores and delivers 1.73 petaflops of sustained Linpack performance. Number eight is called "JuQueen" and is at Forschungszentrum Juelich (FZJ) in Germany. This machine is a 131,072-core BlueGene/Q that delivers 1.38 petaflops sustained on Linpack.

Groupe Bull's "Curie" thin node machine – based on the Bullx B510 server nodes with Xeon E5-2680 processors and using 40Gb/sec InfiniBand interconnect – had 120,640 cores and delivered 1.27 petaflops of double-precision matrix math performance. This machine has a computational efficiency of 81.5 per cent, which is not bad, but it only delivers 603.7 megaflops per watt, which is not all that great in terms of energy efficiency. (But, if your code is tuned for x86 and InfiniBand, then maybe this is what matters more.)

Rounding out the top 10 is the "Nebulae" *ceepie-geepie* built by China's Dawning for the National Supercomputing Center in Shenzhen. Nebulae pairs Xeon X5690 processors from Intel with Tesla M2050 GPU coprocessors from Nvidia. This machine came into the number two position on the June 2010 Top 500 list and is unchanged from that time. This box has 120,640 cores in total and delivers 1.27 petaflops of performance but burns a staggering 2.58 megawatts. Nebulae has a computational efficiency of only 42.6 per cent and delivers only 492.6 megaflops per watt.

Sequoia, as you would expect from a giant redwood tree, is being true to its name and setting the performance and efficiency bars pretty high in the HPC arms race.

Incidentally, the United Kingdom nearly edged into the top 10, with the 114,688-core BlueGene/Q machine nick-named “Blue Joule” at Daresbury Laboratory, weighing in at 1.21 petaflops and giving the lab the number 13 spot in the world Linpack rankings this time around. That’s a little shy of the 1.4 petaflops it was expecting and the number 10 ranking Daresbury was hoping for. The November 2012 list will have the “Blue Waters” XK6 hybrid CPU-GPU super from Cray on it as well as the fully upgraded Titan machine at Oak Ridge, so if nothing else changes, Daresbury needs to add a few hundred teraflops to make the top 10 this autumn.

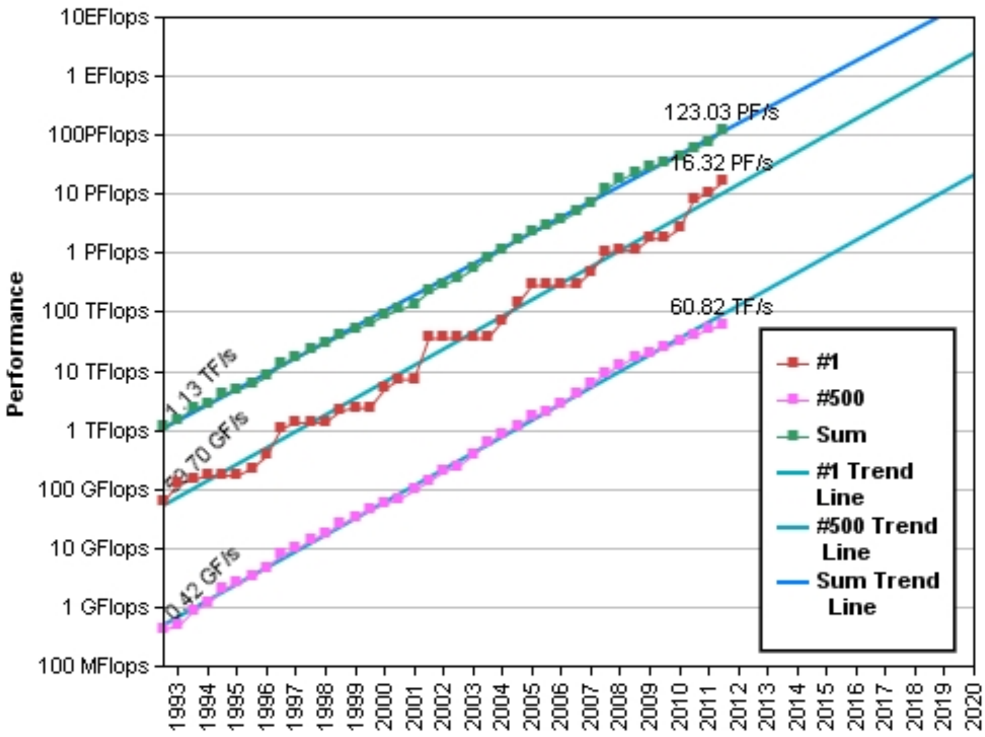
Power on the rise, x86 slipping a bit

The Top 500 list is put together twice a year by Hans Meuer of the University of Mannheim; Erich Strohmaier and Horst Simon of Lawrence Berkeley National Laboratory; and Jack Dongarra of the University of Tennessee. It is not meant to be a performance benchmark on which to base acquisition decisions, but it is useful for seeing trends in system design and projecting how they might be more widely adopted in the broader HPC market in the near term and in the mainstream systems market over the long haul.

To get onto the Top 500 list this time around, a machine needed to hit at least 60.8 teraflops. The aggregate performance of the entire list comes to 123.4 petaflops, a 66.3 per cent increase from the aggregate 74.2 petaflops on the November 2011 Top 500 ranking and more than double the 58.7 petaflops of a year ago. This time around, there are 20 machines with 1 petaflops or more of floating point power, and they are clearly bringing up the class average. But so is the addition of more powerful machines, many with GPU coprocessors, lower down in the list.

Speaking of coprocessors, there are now 58 machines on the Top 500 list that use accelerators of one kind or another – up from 39 machines in the November 2011 list. Of those 58 machines, 53 use Nvidia Tesla GPU coprocessors, two use Advanced Micro Devices’ Radeon graphics cards, and two use IBM Cell processors. A year ago, there were only 17 machines with GPUs. This is beginning to smell like a ramp, much like when Linux took over as the operating system for supercomputers (more or less) in the late 1990s.

But Intel is starting to get in the game now, too; an Intel research machine code-named “Discovery” is ranked number 150 on the list. Discovery was built with Xeon E5-2670 processors stoked with “Knights Corner” MIC x86 coprocessors. This machine weighs in at 118.6 teraflops sustained against 181 peak teraflops on the Linpack test, and delivers 1,176 megaflops per watt.



Top 500 performance over time – pushing to exaflops

On the CPU front, 372 of the machines, or 74.4 per cent of those on the list, are based on Intel Xeon or Itanium processors, down slightly from the 384 machines on the November list and obviously impacted by the addition of a slew of BlueGene/Q iron as well as the delay in the roll-out of the Xeon E5 processors from last fall to this spring. Oddly enough, there are 246 machines using Intel’s *prior* Xeon 5600 generation processors – this is up from 240 six months ago. That said, there are 44 machines based on the Xeon E5s, compared to 10 on the November 2011 list when the machines were built using pre-launch processors with the blessing of Intel.

The current Top 500 has 58 Power-based machines, up from 49 six months ago. There are 63 clusters based on AMD’s Opteron processors (some with GPU coprocessors, some not), and that is 12.6 per cent of the pool. It is also the same number as on the November 2011 ranking.

In terms of core counts on the CPU side of machines, 74.8 per cent of the machines on the list have six or more cores. The average system on the list has 26,866 cores, up from an average of 18,383 six months ago and 15,520 a year ago. Average power consumption of a machine on the Top 500 is now 671 kilowatts, up from 634 kilowatts last November and 543 kilowatts last June.

In another interesting turn, the number of InfiniBand-based machines now is larger than the number of Gigabit Ethernet machines on the Top 500 list. There were 208 InfiniBand machines driving 31.5 of aggregate petaflops compared to 207 Gigabit Ethernet machines driving 13.3 petaflops in total.

IBM had 213 systems on the June 2012 Top 500 list, which is 42.6 per cent of installed systems. Big Blue has 47.6 per cent of installed capacity as gauged in flops. Hewlett-Packard, which doesn’t pursue high-end machines generally, had 138 machines on this list, down from 141 six months ago – giving it 27.6 per cent of systems on the current list. Cray has 5.4 per cent of the base; followed by Appro International at 3.6 per cent; Silicon Graphics at 3.2 per cent; and Groupe Bull at 3.2 per cent as well.

IBM and HP pretty much have a lock on commercial HPC customers; together the two firms account for 247 of the 249 machines not going into government or academic labs. ®

This article originally appeared in The Register. It appears here in its entirety as part of a cross-publishing agreement.

Posted in HPC by Staff
0 comments

Share this with your friends. Mi piace

Like what you're reading? Come back every day for HPC news, or subscribe to email or RSS updates. Trackback URL: http://insidehpc.com/2012/06/18/ibm-us-nuke-lab-beast-sequoia-is-top-of-the-flops-petaflops-that-is/trackback/

Leave your own comment

Name*

eMail*(not published)

Website

Submit Comment

- **Related Stories**
  - » Panasas Secures Fifth Round of VC Funding
  - » Microsoft Releases Public Beta of Windows HPC Server 2008
  - » Slidecast: Violin Memory - Insanely Powerful
  - » Video: Patent Reform Needed for Job Creation
  - » Utility computing prices compared
- **News Navigation**

  « Video: HPC at the University of Colorado and the Student Cluster Competition
  Video: Achieving Ultra-low Latency in the Cloud: How Low Can We Go? »

HPC news for supercomputing professionals
Hardware Software Tools Visualization
Events New Installs Applied HPC Research
Enterprise HPC Cloud HPC Datacenter Ops System Mgmt
All Categories
More inside-Star Publications
inside-BigData inside-Cloud inside-Startups

Contact us Advertise2012 Media Kit Who we are

Search

Read the latest HPC news at insideHPC.com

Sitemap