What China's supercomputing push means for the U.S.

China is developing its own software and building its own infrastructure to create a tech industry, says a top computer scientist at the DOE's Argonne National Laboratory

By Patrick Thibodeau
June 10, 2013 06:00 AM ET

Computerworld - *China isn't just building supercomputers, it's creating an infrastructure to create a tech industry, according to Peter Beckman, a top computer scientist at the U.S. Department of Energy's Argonne National Laboratory and head of the DOE's exascale initiative.*

Peter Beckman, director of the Exascale Technology and Computing Institute at Argonne National Laboratory. (Photo: Argonne National Laboratory)

China's latest supercomputer, which may be officially cited as the world's fasted when Top 500 global rankings are released mid-month, is running Chinese- made interconnects and software. China-made CPUs are up next, Beckman said.

Tianhe-2 or Milkyway-2, which will have 3.1 million cores, has a theoretical speed of almost 55 petaflops. It's been tested so far at nearly 31 petaflops. A petaflop is 1,000 teraflops, or one quadrillion floating-point operations per second. An exascale system is 1,000 petaflops.

Beckman, in an interview, explains the significance of China's moves, and the power problem facing the push to exascale.

What does China's new system say about that nation's HPC technology development? It is a very clear statement of how serious they are with respect to scientific computing. If you look at their history of investment, this is just one data point in a much longer series of investments.

Was there anything about the home grown elements of this system that caught your attention? The network, and that's a pretty significant part. They have a front-end processor that's their own processor. They have slowly and incrementally woven in their own technology. They designed the interconnect from scratch, and they designed a software stack. They are taking their own approach on how to do parallelism. The two items that make the supercomputer super, the software and the interconnect, they are growing at home. The chips are well on their way. Once they have a chip that competes well it won't just be used for supercomputers.

Will China's next system have homegrown chips? I think so. I suspect that there is a national pride issue happening here as well. They will really work, in my opinion, to make a top machine that will be all (homegrown) tech from top to bottom -- the software, the interconnect and the CPU.

There is international cooperation in developing a software stack for an exascale system. Are the Chinese going their own way? They want to build their own components. They are not racing toward what is the most expedient, easiest way to deploy something. Inside the messaging layers, there were pieces that they were inventing, that they were doing over -- doing a different way. My impression is that their intent is obviously to collaborate and work with the community, but they really want to grow many of the components in-house.

Are they sharing any of this as open source? At this point it's pretty hard to see it. The software that the community is using, none of it is coming from China. It's hard to find, in some sense, on the Web. If you look for some of the pieces like the Kylin (Linux) operating system, it's not easy to find a community of people where this is being used or shared. It's certainly not prevalent yet. Maybe that's to come. I don't want, in some sense, to sell them short. It's very hard to document code in English if you're really writing essentially in Chinese. There may be language issues preventing them from doing this.

The new Chinese system will use 24 MW (megawatts) at peak when cooling is considered. What are your observations about its power use? That's an awful lot. The raw number is staggering when you think it's about $1 million per year per megawatt. That machine at peak would run $24 million a year in electricity. The goal for exascale is in the 20-30 MW range. In some sense, this shows that if we do nothing, we're stuck at this power rating.

Is there any agreement about how to lower power? There are several promising venues. One is the integration of memory on the chip. Right now, memory accounts for a healthy fraction of that power, and having it external to the CPU wastes power. Pulling it on to the CPU, that memory, with 3D chip stacking or other techniques, will make a big dent.

The other promising technology: Right now, all of our system memory is RAM, and RAM is very inefficient in terms of power. There are technologies that several companies are developing that could use NVRAM. It might not be quite as fast as RAM but the power difference is spectacular, so with that in mind, you can imagine developing systems in the future where some fraction of the memory is actually NVRAM, a smaller fraction of overall memory is RAM, and we get a big power savings. But the thing that we haven't tapped into at all really is managing power as a resource from the software. We just don't have a way right now to automatically move up or down the power in order to take advantage of processors being idle or not idle in a large HPC computation. So there are a lot of software changes that have to happen.

How will the power software management work? Google just wrote a paper, *The Tail at Scale*. When you do a Google search, it is searching several different servers for little bits of information that are then all pulled together, and that result is then sent back to you. So let's say that there are 20 machines that have to be touched, and a little bit of data from each of the search pieces is assembled and sent back to you. If one of those machines, and this is the part about the tail, comes back with an answer in a slightly longer time, the end result of the query is as long as the longest component. That's frustrating. We find that a little bit interesting, because [Google has] rediscovered what in high-performance computing we have known for a couple of decades, which is this concept of bulk synchronous computation, where you send out hundreds of thousands of tiny work objects to be done, one on each CPU, and if any one of those hundreds of thousands of chips runs slower, any one of them, then your result is as slow as the slowest one.

Supercomputer race

- Smartphone chips to power prototype supercomputer

- What China's supercomputing push means for the U.S.
- China surpassing U.S. with 54.9 petaflop supercomputer
- U.S. losing a Sputnik moment
- Smartphone chips could replace server processors in HPC, researchers say
- Cray offers a more modest supercomputer for the enterprise
- Dell working on ARM supercomputer prototypes
- Swiss supercomputer aims to predict mountain weather with help of GPUs
- IBM supercomputer takes on new role in health arena
- Supercomputers face growing resilience problems

More in Supercomputers

Let's say you paid $100 million for your machine, and you have all of those CPUs working hard on your problem, and one of them is slightly slower, then it's like degrading the value of your machine by 50% or more. That's how we do many of the computations right now. In terms of power management, the compiler, and the code, and runtime system have to cooperate in deciding when we can speed processors up and when we can slow them down. It can't be a self-deciding component.

What you try to do is make sure all the processors run at exactly the same speed, and they always return the answer at the same speed, so you don't have any lagging slowdown processor, or you try to cull [the laggards] out before they even run. Sometimes there are ways to determine that there are parts of a machine that aren't running as fast. But sometimes it's not so easy to do that.

With the size of memory that we have today, some part of your machine is likely to be correcting a single bit error at any given moment. Single bit errors can be detected and corrected automatically, well, it still takes a few CPU cycles so that means that that processor is still going to be late, just a fraction, to the computation because it had to clean up this fault. As we move to lower power, we also recognize that faults go up. The closer you are to operating at the jagged edge, the threshold of computing, the more noise there is in the system, and therefore the more faults there will be. This issue is quite a complex one.

What impact do you think China's new system will have, or should have on exascale development in the U.S.? My personal hope is it is a demonstrator of how hard work and investment in technology is important to China, and how that should be important to the U.S. as well.

It isn't just exascale. It's this notion that cutting-edge large science systems in computation drive a lot of research and lot of industry. Our investment in this space is really key to remaining competitive and being the innovators of this space. One of the things that's interesting about China's announcement, in my opinion, is they geared up this company, Inspur, to sell these machines inside China. They are building the infrastructure to churn out these systems within China and the question is then, who is next? Will they be shipping any to India? Will they eventually have the expertise to ship these to Brazil and to other countries?

So in sum, is it correct to say that China is accomplishing multiple things here: They are getting their science together, fueling a new IT industry, and are potentially creating new exports? It's exactly that. They are designing their own chips. They have geared up a set of students and professors, industry, and semiconductor companies to build this infrastructure. What about the software? They are not going to download software from around the world. They are designing teams to

build the software. Are they preparing to export this system? You bet. They aren't just building this in the university, they've included this company, and that company will then be able to make multiple versions of this.