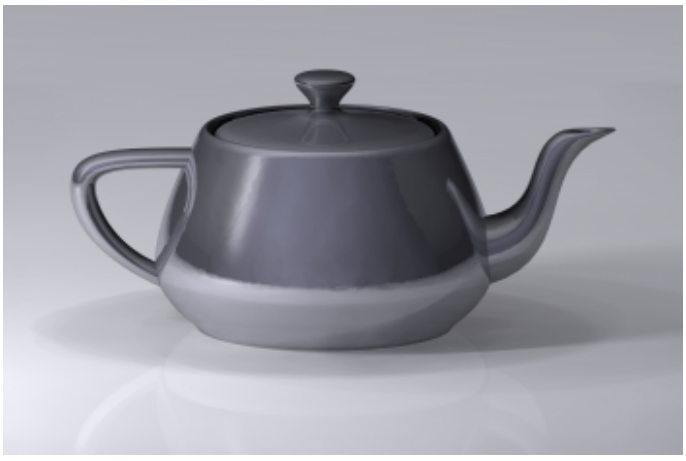
Is gaming tech the key to green HPC?

Graphical processing units have certainly helped today's high performance computers reach improved levels of energy efficiency, but will they be able to deliver similar power savings in the future to help overcome the looming 'power wall'?



The Utah teapot is a standard reference object in the computer
graphics community - appropriate given the location of this year's
SC12 conference. Image courtesy Dhatfield, Wikimedia commons

Posted on NOV 21 2012 5:38AM

Andrew Purcell
European editor

Share this story

It is unlikely to have escaped your notice that last week, during the SC12 conference, held in Salt Lake City, Utah, Titan was named the world's fastest supercomputer. However, you may not be aware that on Wednesday, it was also named the world's third most energy efficient supercomputer on the Green500 list. This energy-efficiency is largely attributable to the system's use of graphical processing units (GPUs). GPUs, which are primarily used for rendering 3D graphics for video games, are able to give Titan a ten-fold boost in computing power, despite the system only requiring slightly more power than a CPU-only system would.

Kirk Cameron, one of the lead investigators of the Green500 list, was on hand at SC12 to talk about improving energy efficiency in high performance computing and overcoming the looming "power wall". He spoke about the challenges involved in reaching the US Department of Energy's target of building supercomputers capable of 20 megawatts per exaflop by 2019.

While Cameron reported that the date for this target is likely to be bumped back to 2022, he still warned that the target may prove very difficult to achieve: "What we're seeing from 2007 to 2012 is a 6 fold increase in flops per watt. This means that by 2019/2020, we're going to be at about 15,000 megaflops per watt, but this is if and only if we have the same sort of efficiency improvements we've seen over the last five years. This is going to be tough, because the improvements over the last five years include GPUs. GPUs are very efficient from a flops per megawatt standpoint. So we're going to need something at least equivalent to the efficiency jump we got from using GPUs to happen again in the next five-to-seven years."

However, even with such a jump in efficiency occurring again, Cameron estimates that this will only get us to around 66 megawatts per exaflop. "This isn't bad, it's certainly better than we're doing now," he says. "But, 66 megawatts for an exaflop system, that's a little higher than 20 megawatts. It's out by three fold - and that's with the kind of improvements we've seen from GPUs. We need to get to 50,000 megaflops per watt for an exaflop at 20 megawatts by 2019."

Gaming's greening

Sarah Tariq, a senior developer and technology engineer in GPU computing at NVIDIA, was also present at SC12. She explains that her move from working in computer graphics to her current role at NVIDIA has given her unique insight into how GPUs can help improve high performance computing and argues that the technology still has plenty to give in the future, in terms of further increasing energy efficiency.

Tariq explains that GPUs have always been massively parallel machines, but that it's only more recently that both their processing power and their programmability have really evolved a long way. Today, a GPU may have around seven billion transistors, compared to just three million in 1995. "The functionality of GPUs has also improved," says Tariq. Originally, they were fixed function, which meant you couldn't really program them for

Tags

energy
green
High performance computing
High-performance computing
HPC
power wall
Salt Lake City
SC12
Supercomputing

things other than what they were meant for, which was graphics. But, over time, they've evolved programmability and, starting in 2006, GPUs are now completely transparent. So, if you want to do parallel computation, you don't need to know anything at all about graphics or how a GPU does graphics, you just program it like any general-purpose massively parallel machine."

Parallels

Tariq highlights a number of examples where the needs of GPUs in gaming and in high performance computing overlap. Medical imaging, for instance, is a clear example of a field which has benefited from GPUs. But there are other, less obvious links too, says Tariq. She cites similarities between hair simulation in gaming and particle simulation as a prime example of this. Equally, she says, convolution, which is used in high performance computers for seismic processing, is often used in gaming to produce lens effects, such as flaring and reduced depth of field, as well as sub-surface scattering. Sub-surface scattering is regularly used in gaming to create realistic-looking skin tones, by simulating the way light penetrates the upper surface of skin and bounces around before being reflected out. Lens flare can also be achieved in games using fast-Fourier transformations, which are commonly used in high performance computing for cryptography and molecular dynamics. Finally, solving partial differential equations - an important technique for fluid dynamics and thus used for things like aircraft design - is also popular in gaming. Here, it is used to create things like realistic plumes of smoke in war games. In fact, so computationally expensive is this technique, that it has only started to be used in video games over the last couple of years, explains Tariq.

"Gaming has been learning from high performance computing and high performance computing has been learning from gaming, but there's still a long way to go," says Tariq. "For high performance computing, you need higher and higher resolution and higher fidelity. But for games, there's also still a long way to go. We're nowhere near yet where we really want to be, which is completely realistic images. There's a way to go, both in terms of the amount of processing power we need, so GPUs need to get faster and faster, and also in terms of the quality of the algorithms that we use. That is the area in which gaming still really has lots to learn from high performance computing."

However, Tariq argues that the huge size of the gaming industry, means that companies are well placed to push this kind of development forward. "This $78-billion industry has been funding the research and development that we need to do to make massively parallel GPUs. This research and development and those economies of scale can now be translated to high performance computing." She also says that it's important to consider where the future of gaming lies, namely in mobile gaming. Since power consumption is so critical when it comes to mobiles, she believes the drivers are now in place to ensure that GPUs become ever more power efficient.

Increasing efficiency

"Comparing today's architectures: when we look at how many watts it takes to deliver a flop, it turns out GPUs are five times more energy efficient," says Tariq. "CPUs are really optimized for single-threaded performance; they're essentially trying to reduce the latencies of the single operations. CPUs are really good at reducing scheduling latency. But this means that they're spending fifty times more energy to schedule an instruction than to actually execute an instruction. They spend so much energy trying to ferry the data from one part of the chip to another, rather than actually doing the computation."

"By contrast, GPUs are optimized 'superclusters'," she says. "There are tons and tons of cores, but these are simple cores, so they have long latencies. If you miss in a small cache, then you suffer the latency of going all the way out to memory. But, it turns out, it's okay to have bad single-threaded performance, because there are so many threads they hide each other's latencies. The second thing that they're doing is that they're using data locality to reduce the amount of times that data is moved across the chip and that reduces energy."

The big picture

Read more from SC12 in our other articles in this week's issue.

"While the energy savings achieved in this manner are clearly impressive, SC12 general chair, Jeff Hollingsworth, sounds a note of caution against viewing GPUs as some sort of panacea: "I think GPUs, as they exist today, will have to make some substantial steps forward to fully address what is needed by high performance computing," he says. "Obviously, they're moving down the path of better bandwidth to the main processor and to the interconnection network. I think such streaming computing is going to be part of the solution, but I'm not sure if it's the entire solution."

Cameron expresses similar, cautious optimism: "There's some new stuff coming with GPUs, but it's not the same paradigm shift that we've seen in the past." Right now, they're high wattage - they're at about 60 or 70 watts, which is pretty inefficient. But, they could probably quite easily reduce this to just 20 or 30 watts. I expect to see the same sort of technology used in processors migrate to GPUs, to help them consume less power when they're not busy."

Despite this, Cameron warns: "I don't see a low power GPU in the future." He says that high power consumption in high performance computing is essentially "a systemic problem". "It's not going away. It prolific, it's embedded in everything that we do. It's endemic - we can't get away from it." As such, "a silver bullet is unlikely," he concludes.

Join the conversation

Contribute

Do you have story ideas or something to contribute? **Let us know!**

OUR UNDERWRITERS

Thank to you our underwriters, who have supported us since the transition from International Science Grid This Week (iSGTW) into Science Node in 2015. We are incredibly grateful.

View all underwriters

CATEGORIES

Advanced computing
Research networks
Big data
Tech trends
Community building

CONTACT

Science Node
Email:
editors@sciencenode.o
Website:
sciencenode.org

GIVE NOW