



# News | TOP500 Supercomputer Sites


<http://top500.org/blog/category/feature-article/feeds/rss>


Are you the publisher? [Claim](#) or [contact us](#) about this channel

 **Embed this content in your**

**HTML**

 [Search](#)

Report adult content: 

click to rate 

Account: ([login](#))

**Browsing the Latest Snapshot**

**Browse All Articles (217 Articles)**

**Live Browser**

 **Channel Description:**  
TOP500 News


 **04/27/18--03:14: UK Commits a Billion Pounds to AI Development**  0  0  


The British government and the private sector are investing close to £1 billion pounds to boost the country's artificial intelligence sector. The investment, which was announced on Thursday, is part of a wide-ranging strategy to make the UK a global leader in AI and big data.

Under the investment, known as the "AI Sector Deal," government, industry, and academia will contribute £603 million in new funding, adding to the £342 million already allocated in existing budgets. That brings the grand total to £945 million, or about \$1.3 billion at the current exchange rate. The UK government is also looking to increase R&D spending across all disciplines by 2.4 percent, while also raising the R&D tax credit from 11 to 12 percent. This is part of a broader commitment to raise government spending in this area from around £9.5 billion in 2016 to £12.5 billion in 2021.

The UK government [policy paper](#) that describes the sector deal meanders quite a bit, describing a lot of programs and initiatives that intersect with the AI investments, but are otherwise free-standing. In some cases, the text appears to be contradictory: In one passage the authors stress the need to be "strategic and focused: recognising the increasing convergence of technologies and focusing on the areas where we can compete globally," while just a few paragraphs before in a discussion of the AI/big data opportunity, the paper states the "the UK can lead the world for years to come."

The latter sentiment is at least partly based on the fact that a number of well-regarded AI businesses are already based in the UK, with Deepmind, Swiftkey, and Babylon offered as the most prominent examples. However, it's worth

 **More Channels**


 **Showcase**

[RSS Channel Showcase 1586818](#)

[RSS Channel Showcase 2022206](#)

[RSS Channel Showcase 8083573](#)

[RSS Channel Showcase 1992889](#)

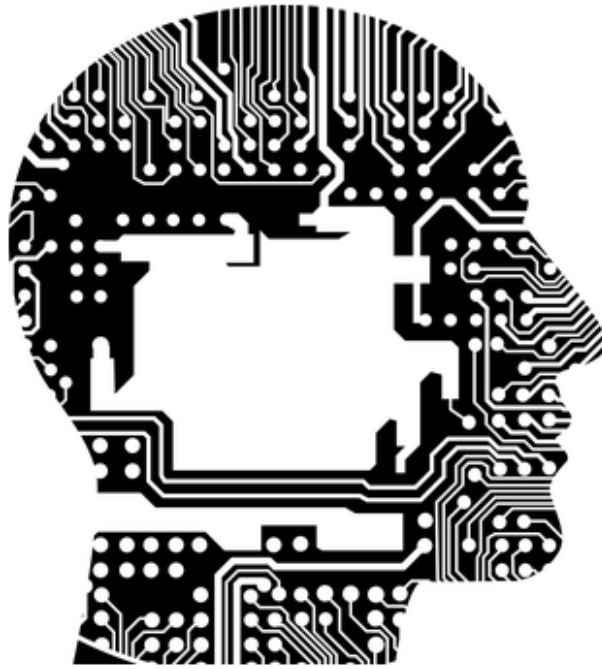
 **Channel Catalog**

[Subsection Catalog](#)

 **Articles on this Page**

(showing articles 1 to 20 of 20)

noting  
that



[04/27/18--03:14: UK Commits a Billio...](#)

[05/01/18--01:43: Google Expands Its ...](#)

[05/08/18--22:31: CERN Prepares for N...](#)

[05/14/18--23:18: Did Google AI Just ...](#)

[05/21/18--00:41: Chip Startup Unveil...](#)

[05/23/18--21:13: Intel Lays Out New ...](#)

[05/29/18--23:13: NVIDIA Brings HPC a...](#)

[06/06/18--01:51: As Moore's Law Wind...](#)

[06/08/18--17:21: Summit Up and Runni...](#)

[06/18/18--04:46: Sandia to Install F...](#)

[06/23/18--03:55: Thomas Sterling Tal...](#)

[06/24/18--17:37: US Regains TOP500 C...](#)

[06/25/18--11:58: Distortions, Trends...](#)

[06/25/18--21:43: CERN's Maria Gi...](#)

[06/26/18--09:09: New GPU-Accelerated...](#)

[06/29/18--06:39: Cloud Computing in ...](#)

[07/02/18--09:23: Benchmarks in Hand,...](#)

[07/05/18--13:27: European Program to...](#)

[07/16/18--09:18: A Tightening Exasca...](#)

[07/20/18--10:16: AMD May Be About to...](#)

(showing articles 1 to 20 of 20)

Deepmind is owned by Google, while Swiftkey is part of Microsoft. And even though Babylon is still operating independently, some of its investors are based in Sweden (Vostok New Ventures and KinnevikAB) and the US (NNC Holdings).

More to the point though, from a commercial perspective, the UK is not a major force in AI and is unlikely to be anytime soon. The biggest players are hyperscale companies based in the US (Google, Amazon, Microsoft, IBM, Facebook, and a few others) and China (Tencent, Baidu, and Alibaba). And that points to a larger challenge for the UK, namely, it has a relatively small population from which to draw the kind of data that drives most of the deep learning models that currently underlie AI. To develop a larger presence, it will have to establish at least one big multinational that can tap into foreign markets – a Barclays of AI, if you will.

Having said that, the overall rationale for a targeted investment in this area is well-justified, especially considering that the AI sector – something that barely existed five years ago – is [projected to add more than £630 billion to Britain's economy by 2035](#). And given the R&D investments in artificial intelligence taking place in China, the US, Europe, Japan, and elsewhere, it behooves the UK to keep pace.

The growing importance of the AI sector also means government policies, and agencies implementing those policies, will have to be established to develop and manage this market. For example, the government plans to establish a new Office of Artificial Intelligence, as well as a Center for Data Ethics and Innovation. Working with those entities will be a new AI Council, an advisory group of business and academic leaders in the field. The government will also initiate public-private programs to attract AI talent, businesses, and investments to the UK.

Some of this investment is already happening organically. Google and Amazon have a substantial AI presence in the UK; Google has three offices in London

and another one planned for King's Cross, while Amazon has a total workforce of 24,000 in the country and plans to open two new robotics-enabled fulfillment centers in the near future. Meanwhile, Hewlett Packard Enterprise (HPE) recently partnered with ARM, and SUSE, and the universities of Bristol, Edinburgh and Leicester to deploy [three ARM-powered supercomputers](#) to be used for AI and big data work.

A number of international companies are also making AI-related investments in the country, either by partnering with UK organizations or by opening offices in the country. These include Elemental AI (Canadian AI research lab), Beyond Limits (American autonomous software developer), Ironfly Technologies (Hong Kong advanced financial analytics company), Astroscale (Japanese space debris removal tech company), Chrysalix (Canadian venture capital firm targeting AI and robotics), and Global Brain (Japanese venture capital firm with a focus on AI, blockchain, and robotics). In addition, [Bloomberg is reporting](#) that Microsoft, IBM and Facebook are "making undisclosed commitments, along with consulting firm PwC and pharmaceutical company Pfizer Inc."

On the academic side, the University of Cambridge Research Computing Service is making its latest £10 million supercomputer available to AI technology companies. This is almost certainly the [Wilkes2](#), a 1.2-petaflop (Linpack) P100 GPU-accelerated cluster that is the UK's most powerful academic supercomputer. It's supported by a consultancy team at the university.

Can this strategy propel the UK into global AI dominance? Unlikely, given the current drivers of the AI market, the dominance of the US and China, and the size of the investment Britain can justify. But the country can certainly leverage its home-grown AI intellectual property and boost imports and exports of the technology. And that can form the basis of a vibrant industry that propels local businesses and increases productivity across all sectors of the economy.

🚩 **05/01/18--01:43: [Google Expands Its GPU Cloud Options](#)**

0



0



Google has announced it is offering NVIDIA Tesla V100 GPUs for its HPC and machine learning cloud customers. But how will the company square this with its TPU cloud offering?

Google's deployment of the V100 follows that of Amazon, IBM, and Microsoft, who have offered this GPU in their respective clouds for some time. Amazon was the first provider to make it available to cloud customers, [when it rolled out its V100 instances in October 2017](#). The V100 is currently NVIDIA's most advanced accelerator, offering 7.5 teraflops of double precision performance for HPC and 125 teraflops of tensor mixed precision performance for machine learning.

According to a [blog](#) posted on Monday by NVIDIA product managers Chris Kleban and Ari Liberman, the V100 solution will initially be offered in beta.

Customers will be able put up to eight of the GPUs, along with 96 vCPUs and 624 GB of memory in a single virtual machine. With that configuration, a user can tap into 60 teraflops of double precision or one petaflop of tensor performance. Kleban and Liberman write that HPC and deep learning workloads will realize a 40 percent performance boost with these latest GPUs.

In conjunction with the V100 launch, Google is also moving its Tesla P100 GPU cloud offering from beta to general availability. This one has almost the same configuration as the V100 offering, with up to 96 vCPUs and 624 GB of high bandwidth memory, but with a maximum GPU count of four. Given the lower relative performance of the P100 – 5.3 teraflops for HPC and 21.2 teraflops for machine learning – the P100 is a good deal less powerful than its younger sibling. The older K80 GPU, which Google has offered for some time and continues to support, is even less powerful than the P100. The company has priced them all accordingly, as can be seen from the table below.

Google Cloud GPU Type			VM Configuration Options		
NVIDIA GPU	GPU Mem	GPU Hourly Price**	GPUs	vCPUs*	System Memory*
V100	16GB	\$2.48 <i>Standard</i> \$1.24 <i>Preemptible</i>	1,8 (2,4) coming in beta	1-96	1-624 GB
P100	16GB	\$1.46 <i>Standard</i> \$0.73 <i>Preemptible</i>	1,2,4	1-96	1-624 GB
K80	12GB	\$0.45 <i>Standard</i> \$0.22 <i>Preemptible</i>	1,2,4,8	1-64	1-416 GB

Source: NVIDIA

Currently, the V100s will be available to Google customers in its Western and Central US regions, as well as parts of Western Europe. The P100 has a wider distribution and is available in Google's Western, Central and Eastern US regions, Eastern Asia, and Western Europe. The price table above applies only to US regions.

Google is ostensibly aiming the V100 and P100 at HPC and machine learning customers. In the blog announcement, two customers are mentioned: LeadStage, which is using both the V100 and P100 for optical character recognition on handwritten documents, and Chaos Group, which is employing the V100s for V-Ray Cloud rendering.

As we've reported before previously, Google has a divergent strategy with regard to machine learning in the cloud. Back in February, the company [launched its Tensor Processing Units \(TPUs\) into its public cloud](#). The TPUs were custom-designed specifically for machine learning workloads and up until February had only been used internally by Google for its own applications like web search and language translation. By offering the TPUs to the general public, the web giant seemed to be probing customer interest in an alternative machine learning platform.

Google is renting its TPU board at a price of \$6.50 per hour. For that you get 180 machine learning teraflops and a minimal software stack based on the TensorFlow framework. Now for \$2.48 per hour, you can get a V100, which delivers 125 teraflops and comes with a much more extensive ecosystem of libraries and tools. It should be noted that the TPU boards, which contains four TPU chips, comes with 64 GB of high bandwidth memory (16 GB per chip), versus 16 GB on a single V100 device.

Given all that, the V100 would seem to be the better deal. However, a few months ago Elmar Haußmann, cofounder and CTO of RiseML, [compared the performance of four V100s in AWS against Google's four-TPU cloud board](#), running an image classification training application (ResNet-50 on ImageNet). The idea was to match up both processor count and memory capacity. He found that the application performance for the TPU board and four-V100 setup was very similar, with the V100 only showing an advantage with smaller batch sizes. Moreover, Haußmann found that the TPU's performance per dollar was significantly better than the V100 based on the AWS pricing. Since Google's standard rate for its V100 for only about a third of the cost of its TPU board, Google's custom silicon would still come out on top in price-performance – at least for this particular application. That said, the preemptible pricing for the V100 is only about 20 percent that of the TPU, which would give the NVIDIA hardware the price edge.

For other machine learning applications, or perhaps even for ResNet 50 if a different software stack was employed, the V100 might be able to demonstrate a much clearer advantage. That would have to be determined on a case-by-case basis. Of course, if a customer is already invested in CUDA and NVIDIA machine learning libraries, the choice is pretty obvious. And certainly, for more conventional HPC applications, the V100 or any of the other GPUs would be the way to go, given the unsuitability of the TPU for non-machine learning software.

For the time being, Google is likely resigned to the fact that only its most adventurous customers will opt for the TPU platform. In the longer term though, the web giant would probably like to see this solution become a more standard option for its cloud customers, which would spread the cost of its chip design, production, and software support across a larger user base. In the meantime, Google seems willing to offer a somewhat confusing choice to its customers.

 **05/08/18--22:31: [CERN Prepares for New Computing Challenges with Large Hadron Collider](#)**



Thanks to the discovery of the Higgs boson in 2012, CERN's Large Hadron Collider (LHC) has probably become the most widely recognized science project on the planet. Now almost 10 years old, the 27-kilometer ring of superconducting magnets is the world's largest and most capable particle accelerator. As such, it enables physicists to push the envelope of particle physics research.

Less well-known is the computing infrastructure that supports this effort – that of the The Worldwide LHC Computing Grid (WLCG), a network of more than 170 computing centers spread across 42 countries. Because LHC experiments can involve processing petabytes of data at a time, the computational, networking, and storage challenges for the project are immense. And when the next-generation High-Luminosity LHC (HL-LHC) is launched in 2026, these challenges will become even more formidable.

To find out more about what this entails, we asked Dr. Maria Girone, CERN's **openlab** CTO, to describe the high performance computing technology that undergirds the LHC work and talk about what kinds of hardware and software are being considered to support the future HL-LHC machine. Below is a lightly edited transcript of our conversation.

**TOP500 News: Can you outline a typical computing workflow for an LHC application – for example, the workflow that resulted in the discovery of the Higgs boson particle?**



**Maria Girone:** Workflows in high-energy physics typically involve a range of both data-intensive and compute-intensive activities. The collision data from the cathedral-sized detectors on the Large Hadron Collider needs to be filtered to select a few thousand interesting collisions from as many as one billion that may take place each second. The search for new phenomena is like looking for needles in enormous haystacks.

Once interesting collision events have been selected the processing-intensive period begins. The particles from each collision in the detectors are carefully tracked, the physics objects are identified, and the energy of all the elements are measured with extreme precision.

At the same time, simulation takes place on the Worldwide LHC Computing Grid, the largest collection of computing resources ever assembled for a single scientific endeavor. The WLCG produces a massive sample of billions of simulated beam crossings, trying to predict the response of the detector and compare it to known physics processes and potential new physics signals. In this analysis phase, the data from the detector is examined against predictions based on known background-only signals. When the data diverges statistically significantly from the background-only signals, we declare a discovery.

**TOP500 News: Do any of the experiments in the LHC project currently use deep learning or some other form of AI?**

**Girone:** Neural networks and machine-learning techniques have been used in high-energy physics for many years. Optimization techniques, such as boosted decision trees, have been widely used in analysis. The field is now looking to expand the use of deep-learning and AI techniques based on the progress made by industry in these areas.

There is potential for applications throughout the data-selection and processing chain, which could increase the efficiency and performance of the physics searches. Other areas we are exploring include object identification based on 3D image-recognition techniques, improved simulation using adversarial networks, better monitoring via anomaly-detection techniques, and optimized resource use through machine-learning algorithms.

**TOP500 News: How are the computing challenges for the Large Hadron Collider different from typical HPC simulations?**

**Girone:** The computing challenges of the LHC differ from typical HPC applications in the structure of the problem, the time-scale of the program and the number of contributors to the code. Whether processing a data event or producing simulations, each collision event can be treated independently. This means that the application lends itself to simple parallelization across many nodes.

With the LHC program running over multiple decades, there is need for software to be continuously improved. Occasionally, big components are reworked entirely, and there is also a lot of legacy code and services to support.

For many applications used in high-energy physics application, it is the case that several hundred people may well have contributed to the code base over many years. Traditional HPC simulations are often developed by much smaller groups of contributors, with more specific expertise in this area. Our use of

code developed by very large numbers of contributors makes it challenging to reach the level of optimization often achieved with other HPC codes.



**TOP500 News:** Given that the computing grid is spread around the world, what types of challenges are encountered with regard to sharing the large datasets associated with LHC work?

**Girone:** Data management has been a consistent area of development in LHC computing. We move petabytes per day and all the data needs to be monitored for consistency. We have become leaders in moving data using global networks. In the last few years, we have augmented the traditional techniques for moving and replicating data to provide real-time remote access to data files across the globe. Our global data-access model has helped to optimize the use of processing and storage resources, as well as making it possible to use commercial cloud and HPC resources in an opportunistic manner.

**TOP500 News:** Can you describe what the High-Luminosity LHC will be able to do that cannot currently be done with the present-day LHC?

With the HL-LHC, about five to ten times more beam crossings will take place compared to today. Each of these crossings will result in about five times as many individual proton-proton collisions. This increase will help us to search for rarer signals and more precisely measure rare phenomena.

**TOP500 News:** As you look ahead to the High Luminosity LHC, what emerging technologies in hardware and software do you think are the most promising?

**Girone:** Looking forward to the HL-LHC, there are many interesting new technologies. The continued improvements in networking technologies will help us to continue distributing data efficiently. The progress made with various types of accelerators is being explored too. We have programs with GPUs and FPGAs that have the potential to dramatically improve the performance of the computing systems.

For software, better optimization and code modernization also hold great promise. Finally, new techniques like advanced data analytics and deep learning have the potential to change how analysis and reconstruction are performed, thus enabling us to process more data more efficiently.



*Dr. Girone will present greater depth on the subject of computing challenges at CERN during the opening keynote of the [2018 ISC High Performance Conference](#), which will take place on June 24-28 in Frankfurt, Germany. Her keynote address will take place on Monday, June 25.*

*Images: Maria Girone; CERN IT center. © CERN*

📅 **05/14/18--23:18: [Did Google AI Just Pass the Turing Test?](#)**

0



0



Google has demonstrated an artificial intelligence technology that represents the most sophisticated example to date of a computer engaging in natural conversation with a human. Upon hearing the interaction, some listeners felt the software had convincingly passed the Turing test.

Even though it was developed in 1950, the [Turing test](#) is perhaps the most well-recognized way of measuring an AI system's capacity to demonstrate human intelligence. Developed by legendary computer scientist Alan Turing, the idea was to have a computer program converse with someone at a level where the person would be unable to tell if they were talking to a computer or a human. The test actually encompasses a good deal more complexity than that, but the gist of it is to prove whether or not a computer can pass as human.

Before we get too far into this, you need to watch the five-minute demonstration of the technology, known as Google Duplex, presented by Google CEO Sundar Pichai at last week's Google I/O 2018 event. The demo represents two phone conversations with different people in which Duplex successfully navigated some challenging exchanges. It's kind of mind-blowing, to the point you almost forget one of the participants is a computer.

As Pichai noted, the key to the technology is its ability to “understand to nuances of conversation.” However, Duplex can’t converse about everything. In a [blog](#) posted by Google Duplex lead Yaniv Yaniv Leviathan, Google Duplex lead, and Matan Kalman, engineering manager on the project,, being able to pull this off necessitated constraining the models to particular “closed domains” in order to develop the extensive conversational networks required for specific tasks. At this point, the technology is not sophisticated enough to produce a general-purpose AI conversationalist. In that sense, it might fail the Turing test once the conversation strayed into unsupported subject areas.

But the demonstration does illustrate how sophisticated those models are for the selected domains. Not only was Duplex able to converse naturally with the people on the phone, it was able to react appropriately when problems were presented – especially in the second phone call, when the person led the conversation astray. Leviathan and Kalman say the technology is also able to extract the meaning from context when ambiguities are presented. For example, the phrase “OK for four” could refer to 4 people or 4:00, depending on the conversation that preceded it.

The other thing that is immediately apparent is how well the technology has advanced for basic speech input and output. On the input side, the poor quality of the call on the first exchange and strong accent on the second exchange did not appear to trouble the Google software a bit. As far as Duplex’s own voices, they appears to be based on the company’s [WaveNet technology](#), which has advanced speech generation to the point where it is all but indistinguishable from a real person. The addition of filler words like ‘umm’ and ‘uh’ and colloquialisms like “mmm-hmm” and “gotcha” is also a nice touch, adding some extra authenticity.


In the blog write-up, Leviathan and Matias offer a few details on the underlying technology, which they encapsulate thusly:

*“At the core of Duplex is a recurrent neural network (RNN) designed to cope with these challenges, built using TensorFlow Extended (TFX). To obtain its high precision, we trained Duplex’s RNN on a corpus of anonymized phone conversation data. The network uses the output of Google’s automatic speech recognition (ASR) technology, as well as features from the audio, the history of the conversation, the parameters of the conversation (e.g. the desired service for an appointment, or the current time of day) and more. We trained our understanding model separately for each task, but leveraged the shared corpus across tasks. Finally, we used hyperparameter optimization from TFX to further improve the model.”*

No word on what hardware was used or how long the training took for any particular domain. According to Pichai, Duplex has been in the works for years, so presumably was developed over two or three generations of Tensor Processing Units (TPUs) and possible other hardware. As we reported last week, Google used this same I/O event to [unveil its third generation TPU](#), which will be used to develop bigger and better neural networks for the web giant’s internal needs. Special mention was made of using TPU 3.0 to improve the AI behind Google Assistant, which also happens to be the initial platform for Duplex.

In this case, the idea is to be able to tell Google Assistant to schedule something on your behalf by phone – a haircut appointment or restaurant reservation in the examples above. The app then does this phone magic offline via Duplex and notifies you when it completes the task. Ironically, this initial application is most useful for interacting with low-tech businesses that have yet to embrace modern online tools for managing appointments and reservations. But the underlying technology seems destined to expand into more lucrative areas like automated technical support, human intelligence gathering, or essentially any type of expert system that relies on personal interaction.

Google plans to start testing Duplex in Google Assistant this summer.

 **05/21/18--00:41: [Chip Startup Unveils Processor That Aims to Scale the Datacenter Power Wall](#)**

0



0



Tachyum, a Silicon Valley startup has unveiled a new processor that the company says can tackle a broad range of workloads in HPC, data analytics, artificial intelligence, and web services, while using a fraction of the power of existing chips.

In fact, the company claims that the new chip, known as “Prodigy,” can deliver ten times the performance per watt of conventional processors across these application domains. And supposedly it will be able to do so in as little as 1 percent of the datacenter space. As a result, Tachyum predicts Prodigy will be



reduce datacenter costs by a factor of four.

The performance-per-watt story is a bit subtler than that, however. What the company is actually promising is computational power and efficiency on par with that of GPUs, but with a programming model that more closely resembles that of a CPU. Tachyum overall pitch is that Prodigy is a universal processor that can handle compute-intensive task like those found in HPC and machine learning, as well as more mainstream datacenter workloads, such as web services and analytics. And it can do so more efficiently than any of its competition.

According to Tachyum co-founder and CEO Radoslav 'Rado' Danilak, the energy efficiency of the chip and the financial benefits that accrue from is key to Prodigy's value proposition. The example he uses to illustrate the importance of this efficiency is Google. "They spent more than 10 billion dollars last year on the datacenter," he said recently. "With our technology they can easily save 6 billion, potentially 7 billion dollars, every year."

All of this is being accomplished without customized circuitry for specific applications. "Rather than build separate infrastructures for AI, HPC and conventional compute, the Prodigy chip will deliver all within one unified simplified environment, so for example AI or HPC algorithms can run while a machine is otherwise idle or underutilized," said Danilak. "Instead of supercomputers with a price tag in the hundreds of millions, Tachyum will make it possible to empower hyperscale datacenters to produce more work in a radically more efficient and powerful format, at a lower cost."

Even though the company is aiming Prodigy at a broad swathe of datacenter applications, it seems to be particularly focused on the burgeoning AI space, claiming that the processor's general-purpose nature will give it an advantage in tackling the various computational models being used in that domain. Tachyum is also pitching the chip as "an ideal tool" for the Human Brain Project. That effort is expected to require something on the order of 10 exaflops in order to support a real-time simulation of human brain. The company estimates it will take about 256 thousand of its processors to

accomplish this feat. The implication is that such a machine can be built by 2020.

According to Danilak, Prodigy's power efficiency and computational density will also make it suitable for powering containerized datacenters. The advantage here is that these containers can be located in the midst of large population centers, such that latency-sensitive applications like virtual reality-based gaming could be practical. Danilak also thinks that the chip could be employed in edge computing environment, like for example, cell phone towers, where they could be used by telcos to deliver novel types of media services.

Not a whole lot has been revealed about the specifics of Prodigy's design. In presentations, Danilak has outlined the architecture in general terms, noting that it will support some number of 100 Gbps Gigabit Ethernet and PCIe ports, as well as the capability for flash storage processing. From an instruction perspective it will have data path for things like AI and a universal control path for more general computation. That suggests they have found some way to combine the data parallelism of a GPU with the scalar model of a conventional CPU.

But the real critical piece of technology appears to be related to a solution of the "the wire problem" – the divergence between transistor speeds and the wires that feed them. As transistors get smaller the wires connecting them must get smaller as well, which slows down their transmission speed, hindering overall performance. The company has apparently attacked this problem at a fundamental level by changing the wire model. In a [video](#) recorded last year, Danilak says they have devised an architecture where the wires become "incredibly short," which gets around the problem by making wire speed "almost irrelevant."

At this point, Tachyum's business strategy is unknown. Even though the chip is now officially unveiled, the company has yet to announce system partners or early customer trials. The release date for Prodigy is also still up in the air. That said, if even half of these claims prove to be true, Tachyum could easily get acquired before the first chip leaves the fab. There is certainly no shortage of computer companies that would be interested in such technology.

 **05/23/18--21:13: Intel Lays Out New Roadmap for AI Portfolio**

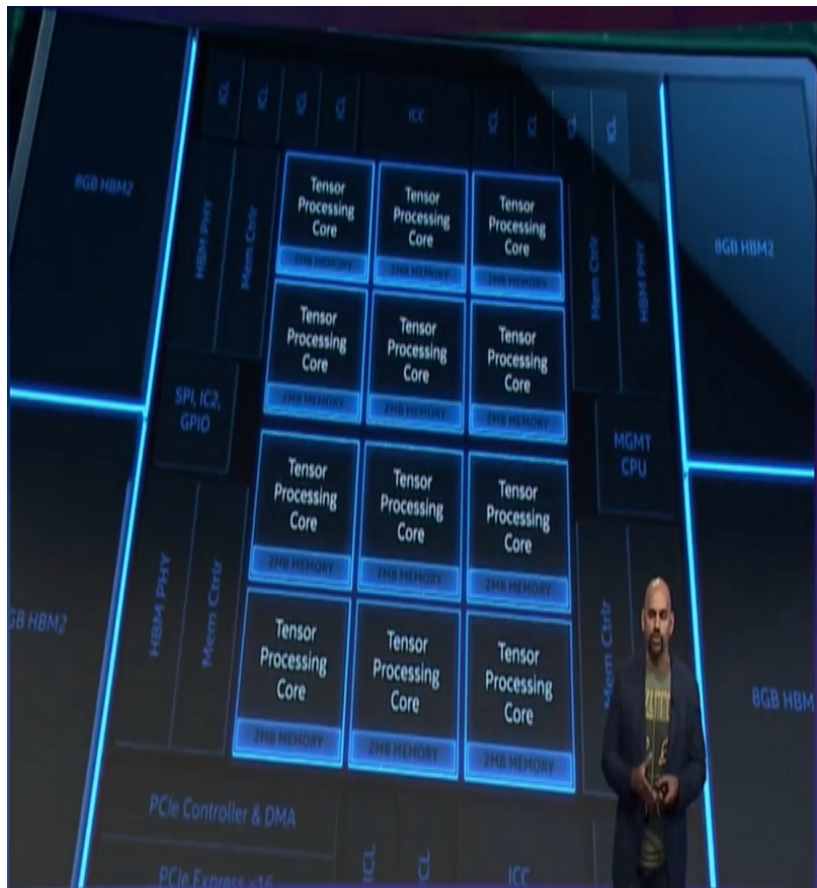
0



0



At Intel's inaugural AI DevCon conference this week, AI Products Group chief Naveen Rao updated their roadmap for its artificial intelligence chips. The changes will impact the much-anticipated Neural Network Processor, and to a lesser degree, its general-purpose products like Xeons and FPGAs.



*Naveen Rao presenting Lake Crest processor at AI DevCon. Source: video screen capture*

During the conference's opening keynote, Rao began by noting that most machine learning currently runs on Xeon processors, and does so with respectable performance. "You may have heard that GPUs are 100x faster than CPUs," he told his developer audience. That's simply false."

The wide use of Xeon processors for these applications stems from the fact that the largest volume of machine learning computation, by far, is on the inferencing side, not the training side. Training is done once to build a model, while inferencing is performed innumerable times to query it. General-purpose CPUs like Xeon processors do inferencing reasonably effectively since it tends to be less computationally demanding than training and is more concerned with delivering a response in a reasonable of time to users.

To make those Xeon processors even more attractive for this work, Intel has added instructions in the latest Skylake processors that speed up neural network processing. In addition, the company has added software support and optimizations for these types of workloads, which have also increased performance. That work is ongoing.

Facebook's head of AI infrastructure, Kim Hazelwood, was brought out on stage to bolster that argument. She told the audience that the company currently runs a large segment of its inferencing work in areas like speech recognition, language translation and news ranking using regular CPU-powered servers. The main reason for this is that Facebook has a lot of

different types of applications to support and using general-purpose hardware for all of them across its vast datacenter empire is just easier-- and presumably less expensive. "Flexibility is really essential in this type of environment," said Hazelwood. What was left unsaid is that Facebook uses NVIDIA GPUs to do much of their offline AI training work.

In some cases though, CPUs are more attractive for training since the datasets are so large that the more limited local memory available on a GPU accelerator becomes too limiting. Right now, the largest local memory configuration for an NVIDIA V100 GPU is 32 GB.

Apparently, this was the critical factor for Novartis, which is using use Xeon Skylake gear to train image recognition models being used for drug screening. The images in question were so large and numerous that the model used 64 GB of memory – a third of the server's memory capacity – for this particular application. Moving from a single node to an eight-node cluster, Novartis was able to reduce training times from 11 hours to 31 minutes.

Having said all that, Rao said they are primarily positioning the Xeon platform for environments that mix training with other workloads or for dedicated setups for large-scale inferencing. The latter use case appears to overlap Intel's positioning of its FPGA products, but at this point Intel certainly realizes that most inferencing deployments are going to be CPU-based. Microsoft is the notable exception here, having already deployed Intel FPGAs at cloud scale on its Azure infrastructure for AI inferencing and a handful of other workloads.

Here, it's worth noting that Rao never brought up Knights Mill, [the Xeon Phi variant built for machine learning](#). Three SKUs of the product were quietly launched in the fourth quarter of 2017 and are still listed on Intel's website, but their omission here suggests Intel has given up on using Xeon Phi as a vehicle for this market, and, as we've suggested before [has likely has given up on Xeon Phi altogether](#).

Which brings us to Lake Crest, Intel's first-generation Neural Network Processor (NNP) custom-built for training neural networks. Based on technology Intel acquired in the 2016 [Nervana acquisition](#), Lake Crest was [supposed to debut last year](#), but for whatever reason, is only now seeing the light of day. The chip is equipped with 12 cores, each of which have two deep learning math units. The device also contains 24 MB of local memory – 2 MB per core – backed by 32 GB of on-package high bandwidth memory (HBM2).

Lake Crest devices can be connected via an I/O link that delivers up to 2.4 terabits/second at less than 790 nanoseconds of latency. This high-speed communication link fulfills one of Intel's primary goals for this platform, namely that a high level of parallelism can be supported using multiple NNP processors.

The other principle design goal for the NNP platform is to provide high utilization of the available computational power. According to Rao, Lake Crest will deliver approximately 40 peak teraflops of deep learning performance, which is a good deal less than the 125 teraflops of its principle rival, NVIDIA's V100 GPU. But Rao's contention is that Lake Crest achieves a much better yield of its available flops than that of the V100 GPU. We should point out here that Rao actually never mentioned NVIDIA or the V100 explicitly, referring to the competing platform as "Chip X."

During his keynote, he threw up a chart comparing utilization rates that illustrated the semi-fictional Chip X could only deliver 27 teraflops on General Matrix-Matrix Multiplication (GEMM), a key operation common to many deep learning algorithms. On this same benchmark, Lake Crest was able to achieve 38 teraflops, which represents a 96 percent yield of available flops. Rao also noted that Lake Crest achieves nearly the same GEMM yield on two connected chips versus a single NPP. Whether that scales up to even larger number of processors – say eight or more – remains to be seen.

Of course, machine learning codes encompass a lot more than GEMM operations, so utilization rates are going to vary from application to application. (And we're sure NVIDIA will have something to say about this.) But it's certainly plausible that a custom-built machine learning chip would be more efficient at these types of codes than a more general-purpose processor like a GPU.


Lake Crest draws less than 210 watts, which compares to 300 watts for an NVLink V100 and 250 watts for the PCIe version. That will make these chips marginally easier to squeeze into datacenter servers than their more power-demanding GPU competition.

But apparently Lake Crest will never get the chance. The product won't be generally available since Intel only intends to release it in limited quantities as a "Software Development Vehicle." Rao says the first broadly available NNP processor will be Spring Crest, whose product designation is NNP-L1000. It's scheduled for general release in late 2019 and is anticipated to be three to four times faster than Lake Crest. That should put it in the range of 120 to 160 peak teraflops per chip.

Spring Crest will also include support for bfloat16, a numerical format that essentially squeezes a standard 32-bit floating value into a 16-bit float customized for tensor operations. It uses the same 8 bits for the exponent as a standard 32-bit float but allocates only 7 bits for the mantissa, which the AI gods have deemed to be enough for deep learning computation. The more compact format means bandwidth can be effectively doubled as data is shuttled around the system. It also enables chip architects to design smaller multiply units, which means more of them will fit onto a die.

Intel, Google, and perhaps others, are hoping bfloat16 becomes a standard numerical format for processing neural network. Over time, Intel plans to support this format across all their AI products, including the Xeon and FPGA lines.

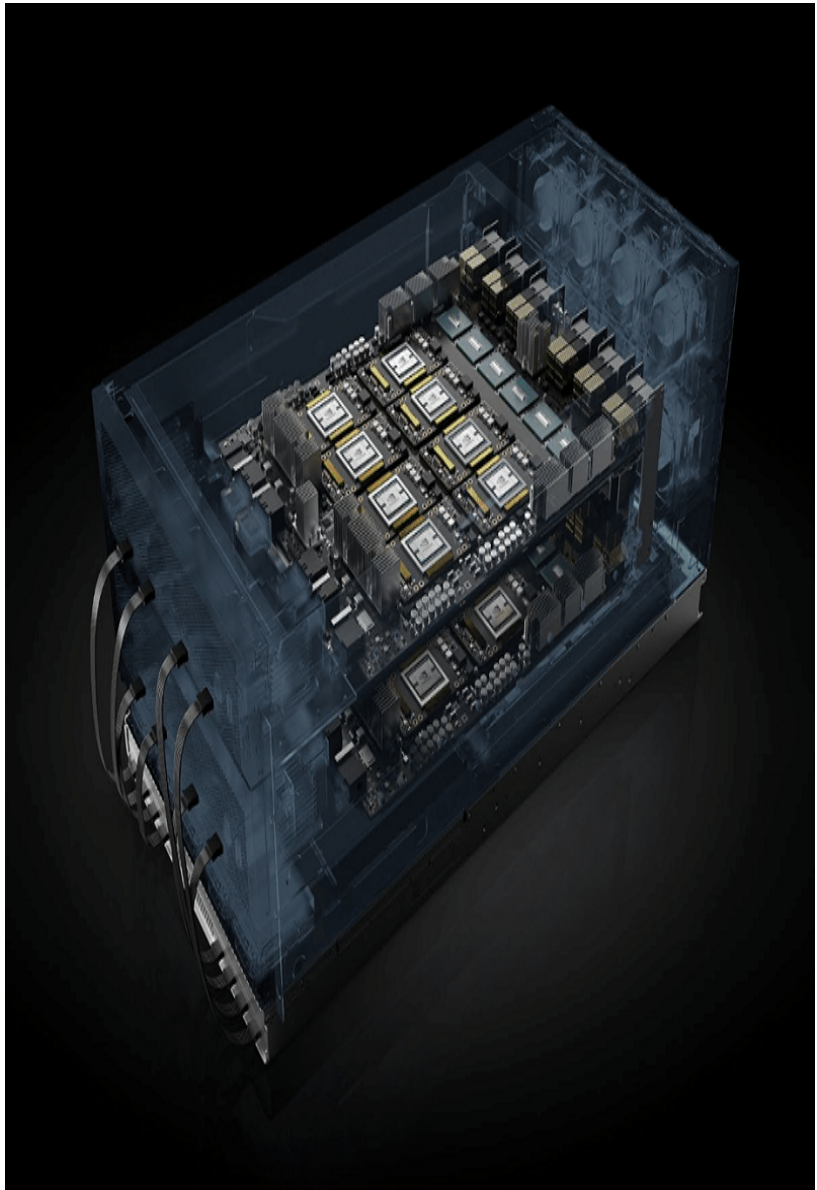
Rao wrapped up the hardware roadmap portion of his presentation by revealing that Intel is working on a discrete accelerator for inferencing, the idea being to achieve the best possible performance per watt for. The AI chief wasn't able to share any details of the future chip, not even its code name. "Look for announcements coming up soon," he said.

 **05/29/18--23:13: [NVIDIA Brings HPC and AI Under Single Platform with HGX-2](#)**





At Taiwan's GPU Technology Conference this week, NVIDIA founder and CEO Jensen Huang announced the HGX-2, a 16-GPU reference design aimed at some of the most computationally demanding HPC and AI workloads. As a reflection of its tightly integrated design, Jensen characterized the platform as the "the world's largest GPU."

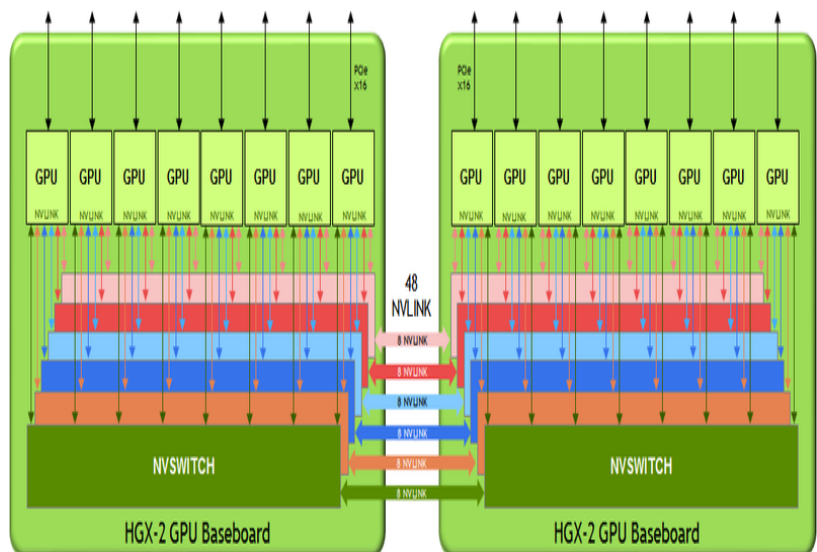


*HGX-2. Source NVIDIA*

According to him, such an architecture can deliver significant price-performance advantages for both high performance computing and machine learning environments compared to conventional CPU-based servers. The V100 processor, upon which the HGX-2 is based, offers both specialized Tensor Cores for deep learning acceleration, as well as more conventional IEEE floating point hardware to speed more traditional high performance

computing applications like physics simulations and weather modeling. "NVIDIA's HGX-2 with Tensor Core GPUs gives the industry a powerful, versatile computing platform that fuses HPC and AI to solve the world's grand challenges," Jensen declared. The HGX-2 is actually the design behind NVIDIA's own DGX-2 server product, which the company [launched in March at GTC in California](#). That product was aimed primarily at machine learning users who wanted to be first in line to use the latest 32GB V100 GPUs in a tightly connected 16-GPU configuration.

Connectivity is the key word here. The DGX-2, and now the HGX-2 upon which it is based, is comprised of 12 NVLink switches (NVSwitches), which are used to fully connect the 16 GPUs. Although that sounds simple enough, the design is a bit more involved. Briefly, the platform is broken up into two eight-GPU baseboards, each outfitted with six 18-port NVSwitches. Communication between the baseboards is enabled by a 48-NVLink interface. The switch-centric design enables all the GPUs to converse with one another at a speed of 300 GB/second -- about 10 times faster than what would be possible with PCI-Express. And fast enough to make the distributed memory act as a coherent resource, given the appropriate system software.



HGX-2 architecture. Source: NVIDIA

Not only does the high-speed communication makes it possible for the 16 GPUs to treat each other's memory as its own, it also enables them to behave as one large aggregated GPU resource. Since each of the individual GPUs has 32 GB of local high bandwidth memory, that means an application can access 512 GB at a time. And because these are V100 devices, that same application can tap into two petaflops of deep learning (tensor) performance or, for HPC, 125 teraflops of double precision or 250 teraflops of single precision. A handful of extra teraflops are also available from the PCIe-linked CPUs, in either a single-node or dual-node configuration (two or four CPUs, respectively).

That level of performance and memory capacity is enough to run some of the largest deep learning models and GPU-accelerated HPC simulations, without the need for a multi-server set-up. The HGX-2 can also be clustered into larger systems via 100Gbps network interface cards, but at the expense of being forced into a distributed computing model.

Of course, NVIDIA could just have just sold its own DGX-2 to serve these HPC and AI customers. But the company is not really in the system business and is not set up to build and deliver these machines in the kind of quantity it envisions. Making the HGX-2 design available to OEMs and ODM means it can reach a much larger audience, and theoretically sell a lot more V100 GPUs, not to mention NVSwitches.

In particular, making the platform available as a reference design will enable the HGX-2 to be deployed in cloud and other large-scale datacenter environments at volumes and price points that would have been impossible with the \$399,000 DGX-2. Lenovo, QCT, Supermicro and Wiyynn have announced plans deliver HGX-2-based servers later this year. In addition, ODMs Foxconn, Inventec, Quanta and Wistron revealed they are designing HGX-2 systems for “some of the world’s largest cloud datacenters.” Those systems are also expected to be available later in 2018.

Although NVIDIA is positioning the HGX-2 as a dual-use platform for HPC and AI, it’s likely that most of the deployments, especially the ones that end up in cloud datacenters, will be primarily running deep learning workloads. That was certainly the case for HGX-1-based installations at Amazon, Facebook and Microsoft. At this point, only a small percentage of HPC work is currently being performed in the cloud and only a fraction of those applications are GPU-enabled.

On the other hand, if Lenovo, Supermicro or one or more of the other system providers manage to sell an HGX-2-based machine to an HPC customer, there is certainly the possibility that it will be used for both traditional simulation work and machine learning. As we’ve reported before, the use of such mixed workflows appears to be on the rise at nearly all large HPC installations and that trend is expected to continue. The major limiting factor here is that not all HPC applications are GPU-ready, but the presence of a super-sized virtual GPU in the HGX-2 makes that limitation easier to overcome.

Looking more broadly, HGX-2 appears to be part of a larger strategy at NVIDIA to provide acceleration technology that is flexible enough to serve both HPC and AI. In that sense, the HGX-2 is really just a server-level extension of the V100 itself. While there is something of [a gold rush to develop custom-built AI chips and servers](#), the supercomputing world would seem to be best served by a unified solution.

Intel briefly flirted with such a product last year, with its AI-tweaked Knights Mill Xeon Phi processor, but the chipmaker has since [reversed course](#). For a time, AMD seemed intent on bringing its CPU-GPU accelerated processing unit (APU) to the datacenter with the Opteron X series, but that product line appears to be stalled. The [recently unveiled Prodigy processor](#) from startup Tachyum could theoretically give NVIDIA a run for its money, but even if the technology lives up to its claims, bringing that chip to market and building a software ecosystem around it is going to take many years.

For the time being, that leaves the GPU-maker essentially unchallenged in the unified accelerator space. And that means HPC sites looking to support traditional supercomputing, with a little machine learning on the side, will have a pretty clear idea of which chips they'll be buying. Which is probably just the way NVIDIA likes it.

06/06/18--01:51: [As Moore's Law Winds Down, Chipmakers Consider the Path Forward](#)

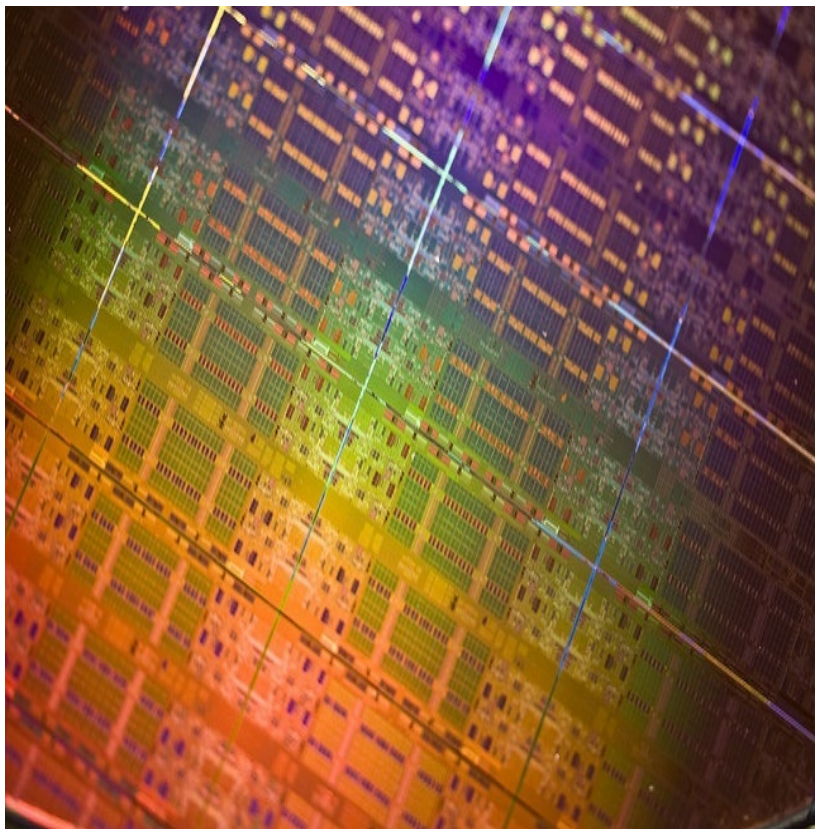
0



0



At this month's ISC High Performance conference, representatives from Intel, NVIDIA, Xilinx, and NEC will speak about the challenges they face as applications like machine learning and analytics are demanding greater performance at a time when CMOS technology is approaching its physical limits.



The four vendors will make their presentations during a 90-minute focus session called [Pushing Digital Computing to the Limits](#), in which each representative will outline their company's strategy to advance their processor platforms over the next 12 years. During this period, Moore's Law, the observation that transistor size and density doubles every couple of years, will slowly come grinding to a halt.

The session will be led by Andreas Stiller, a veteran tech journalist who has covered processor hardware and technology for more than 30 years. Now mostly retired, Stiller continues to work as a freelancer and chairs conference sessions when the opportunities arise. We recently spoke with him to get his thoughts on where the chip industry is headed, especially as it relates to high performance computing, and what it means for the industry.

When talking with Stiller, one thing that becomes apparent is that he seems more optimistic about the longevity of silicon-based semiconductor technology than your average industry watcher. "I really see classical CMOS silicon for the next 20 or 25 years," he says. Stiller thinks the underlying technology will still be practical until we reach transistor geometries in the 1.5nm range.

Beyond that, he says, the silicon-based lattice structures simply won't support electron flow. Chipmakers could perhaps squeeze another half nanometer or so with more exotic materials like molybdenum compounds and carbon nanotubes, but in short order, they will bump up against their own physical limits.

That's the good news. The bad news is that we won't get to enjoy such 1.5nm transistors until around 2030. That's because the increasing difficulty of shrinking transistors is slowing the rate of Moore's Law. Thus, instead of going from 10nm to 7nm in one jump, chipmakers like Intel are looking to develop multiple iterations of the same node (i.e., 10nm, 10nm+, 10nm++) over the same timeframe, with each version incorporating additional refinements, but with more or less the same transistor geometry.

That has real ramifications for the industry. For example, the delay of Intel's 10nm technology, which was originally scheduled for 2015/2016, has now been pushed all the way into 2019 for volume production. That delay probably had a role in [dooming the "Knights Hill" Xeon Phi processor and likely the entire product line](#). The abandonment of Knights Hill resulted in the cancellation of the Department of Energy's original Aurora supercomputer, which will reappear in exascale form in 2021 with yet-to-be-determined processors.

When chipmakers can no longer count on better price-performance and performance-per-watt every two years by shrinking transistors, they need to do something else. In general, that something else means relying a lot more on design improvements. Of course, such improvements were always integral to processor design, but without the additional performance of moving to the next process node, architectural innovation becomes much more critical.

One indication that Moore's Law, and to an even greater degree, Dennard scaling, are no longer driving processor advances, is the growing importance of alternative chip architectures like GPUs and FPGAs. The fact that vendors like NVIDIA and Xilinx literally have a place at the table in Stiller's session alongside traditional chipmakers like Intel and NEC, is a reflection of the breakdown of these fundamental trends.

Stiller has become particularly interested in a new FPGA design Xilinx recently [announced](#), which they call their adaptive compute acceleration platform (ACAP), which is a completely heterogeneous design that employs an FPGA fabric linking distributed memory, programmable DSP blocks, a multicore SoC, and one or more programmable compute engines, all connected via an on-chip network. Xilinx is promising performance and performance-per-watt that is

“unmatched by CPUs or GPUs.” Xilinx CTO Ivo Bolsens plans to elaborate on the ACAP technology during his talk at the ISC session.

NVIDIA has perhaps benefited most from the declining fortunes of transistor shrinkage. The highly parallel architecture of the GPU has turned out to be a particularly good match for applications like machine learning and other types of advanced analytics. Since such a design is inherently less reliance on fast transistors, the company has carved out a sweet spot for itself, especially in the HPC and AI markets. Steve Oberlin, former Cray engineer and now CTO for NVIDIA's Tesla business, will speak about his company's plans to evolve their GPUs into the next decade. Maybe he'll even offer some hints about the company's post-Volta architecture.

Meanwhile NEC is looking to make a comeback with its latest vector processor, the SX-Aurora TSUBASA. Once thought too specialized, NEC is hoping the vector platform will get new life in a post-Moore's Law world. Given the inherent performance advantages of this architecture for HPC codes, vector chips may indeed become more attractive as semiconductor advances become less of a factor. Rudolf Fischer, who heads up the HPC technology group in Europe, will make the case for the future of vector processors.

Given Intel's current dominance of the HPC market with its Xeon processors, the company is the elephant in any room where processor technology is the topic. During the session, Al Gara, who was the chief architect at IBM in charge of the Blue Gene platform and now occupies a similar position in Intel's Data Center Group, will speak to the future technologies the chipmaker will bring to bear on the HPC market. Stiller is hoping (as are we) that Gara will elaborate on how the company plans to heal the Xeon/Xeon Phi divide as it moves its datacenter silicon into the exascale era.

Unfortunately, no one from OpenPower or ARM will be present to represent those two architectures. The 90-minute time allotment for Stiller's session isn't long enough to encompass the current level of chip diversity in the HPC chip space.

In addition, there will be no advocates for more exotic platforms like quantum, neuromorphic, and optical processors, all of which offer compelling alternatives for HPC work. Stiller expects each of these will come into play over the next 10 to 15 years, and of these, he thinks quantum computing offers the most exciting prospects. That said, he believes none of them will replace conventional digital processors any time soon. “For the foreseeable future,” he says, “we'll still have CPUs and GPUs.”

*For those attending ISC High Performance in Frankfurt Germany this month, this session will take place on Tuesday, June 26, from 8:30am to 10:00am, in the Panorama 2 room of Frankfurt Messe.*

📅 **06/08/18--17:21: Summit Up and Running at Oak Ridge, Claims First Exascale Application**

0



0



The Department of Energy's 200-petaflop Summit supercomputer is now in operation at Oak Ridge National Laboratory (ORNL). The new system is being touted as "the most powerful and smartest machine in the world."

And unless the Chinese pull off some sort of surprise this month, the new system will vault the US back into first place on the TOP500 list when the new rankings are announced in a couple of weeks. Although the DOE has not revealed Summit's Linpack result as of yet, the system's 200-plus-petaflop peak number will surely be enough to outrun the 93-petaflop Linpack mark of the current TOP500 champ, China's Sunway TaihuLight.



Even though the general specifications for Summit have been known for some time, it's worth recapping them here: The IBM-built system is comprised of 4,608 nodes, each one housing two Power9 CPUs and six NVIDIA Tesla V100 GPUs. The nodes are hooked together with a Mellanox dual-rail EDR InfiniBand network, delivering 200 Gbps to each server.

Assuming all those nodes are fully equipped, the GPUs alone will provide 215 peak petaflops at double precision. Also, since each V100 also delivers 125 teraflops of mixed precision, Tensor Core operations, the system's peak rating for deep learning performance is something on the order of 3.3 exaflops.

Those exaflops are not just theoretical either. According to ORNL director Thomas Zacharia, even before the machine was fully built, researchers had run a comparative genomics code at 1.88 exaflops using the Tensor Core capability of the GPUs. The application was rummaging through genomes looking for patterns indicative of certain conditions. "This is the first time anyone has broken the exascale barrier," noted Zacharia.

Of course, Summit will also support the standard array of science codes the DOE is most interested in, especially those having to do with things like fusion energy, alternative energy sources, material science, climate studies, computational chemistry, and cosmology. But since this is open science system available to all sorts of research that frankly has nothing to do with energy, Summit will also be used for healthcare applications in areas such as drug discovery, cancer studies, addiction, and research into other types of diseases. In fact, at the press conference announcing the system's launch, Zacharia expressed his desire for Oak Ridge to be "the CERN for healthcare data analytics."

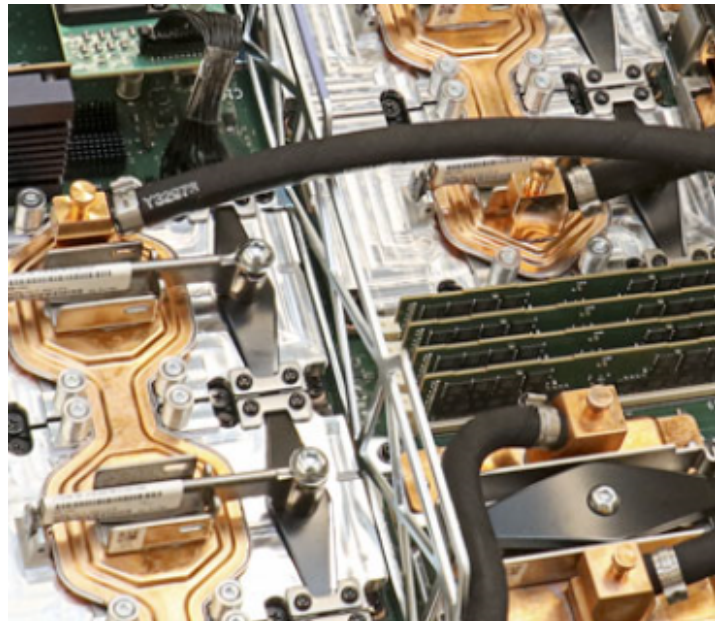
The analytics aspect dovetails nicely with Summit's deep learning propensities, inasmuch as the former is really just a superset of the latter. When the DOE first contracted for the system back in 2014, the agency

probably only had a rough idea of what they would be getting AI-wise. Although IBM had been touting its data-centric approach to supercomputing prior to pitching its Power9-GPU platform to the DOE, the AI/machine learning application space was in its early stages. Because NVIDIA made the decision to integrate the specialized Tensor Cores into the V100, Summit ended up being an AI behemoth, as well as a powerhouse HPC machine.

As a result, the system is likely to be engaged in a lot of cutting-edge AI research, in addition to its HPC duties. For the time being, Summit will only be open to select projects as it goes through its acceptance process. In 2019, the system will become more widely available, including its use in the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program.

At that point, Summit's predecessor, the Titan supercomputer, is likely to be decommissioned. Summit has about eight times the performance of Titan, with five times better energy efficiency. When Oak Ridge installed Titan in 2012, it was the most powerful system in the world and is still fastest supercomputer in the US (well, now the second-fastest). Titan has NVIDIA GPUs too, but these are K20X graphics processors and their machine learning capacity are limited to four single precision teraflops per device. Fortunately, all the GPU-enabled HPC codes developed for Titan should port over to Summit pretty easily and should be able to take advantage of the much greater computational horsepower of the V100.

For  
IBM,



Summit represents a great opportunity to showcase its Power9-GPU AC922 server to other potential HPC and enterprise customers. At this point, the company's principle success with its Power9 servers has been with systems sold to enterprise and cloud clients, but generally without GPU accelerators. IBM's only other big win for its Power9/GPU product is the identically configured Sierra supercomputer being installed at Lawrence Livermore National Lab. The company seems to think its biggest opportunity with its V100-equipped server is with enterprise customers looking to use GPUs for database acceleration or developing deep learning applications in-house.

Summit will also fulfill another important role – that of a development platform for exascale science applications. As the last petascale system at Oak Ridge,



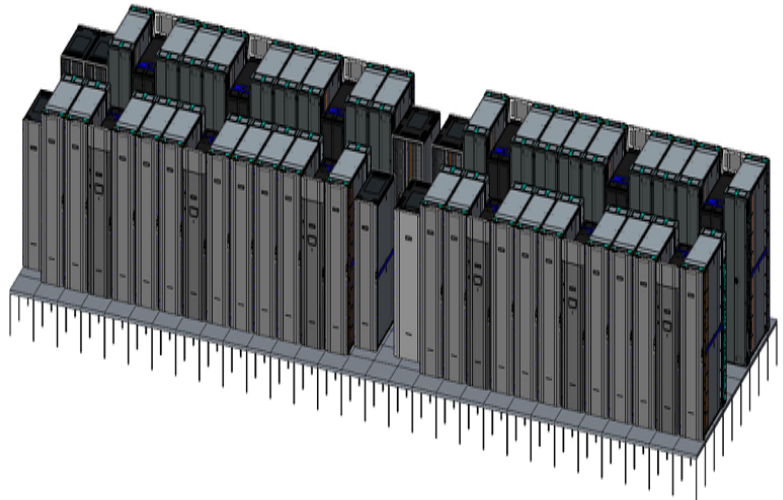
the 200-petaflop machine will be a stepping stone for a bunch of HPC codes moving to exascale machinery over the next few years. And now with Summit up and running, that doesn't seem like such a far-off prospect. "After all, it's just 5X from where we are," laughed Zacharia.

*Top image: Summit supercomputer; Bottom image: Interior view of node.*  
Credit: ORNL

**06/18/18--04:46: Sandia to Install First Petascale Supercomputer Powered by ARM Processors**



Sandia National Laboratories will soon be taking delivery of the world's most powerful supercomputer using ARM processors. The system, known as Astra, is being built by Hewlett Packard Enterprise (HPE) and will deliver 2.3 petaflops of peak performance when it's installed later this year.



*Astra rendering. Source: HPE*

"Sandia National Laboratories has been an active partner in leveraging our Arm-based platform since its early design, and featuring it in the deployment of the world's largest Arm-based supercomputer, is a historical moment not just for us, but for the industry as we race toward achieving exascale computing," said Mike Vildibill, vice president, Advanced Technology Group, HPE

Astra will be based on HPE's Apollo 70 system and will be comprised of 2,592 dual-socket nodes, containing 145,000 cores – by far the largest such system

the company has delivered. If it was up and running today, it would easily make it into the upper fifth of the TOP500 list.

Each node will be equipped with two 28-core Cavium ThunderX2 processors running at 2.0 GHz. These aren't the biggest or the fastest of Cavium's newest ARM processor, but represents something of a sweet spot in price-performance. In aggregate, the compute nodes will draw 1.2 MW of power, which translates into a respectable energy efficiency for a 2.3-petaflop machine.

Local storage will be supplied by Apollo A4520 enclosures, providing 350 TB in the form of an all-flash Lustre appliance. Because of the relatively small capacity and high performance, it will primarily be used for operations needing extreme I/O bandwidth – things like burst buffering and file checkpointing.

Prior to the Astra announcement, most of the other action with regard to ARM-powered HPC was taking place in the United Kingdom. HPE had previously [announced](#) that three UK universities (Edinburgh, Leicester, and Bristol) had ordered Apollo 70 clusters, but each of these systems will be outfitted with just 64 nodes and will top out at a mere 74 teraflops. As far as computational capacity goes, the closest thing to Astra is Isambard, [a 10,000-core Cray XC50 supercomputer using these same ThunderX2 processors](#). It's set to be deployed at the Great Western 4 (GW4) Alliance, a research consortium of four UK universities (Bristol, Bath, Cardiff and Exeter).

Astra's delivery is the first production deployment of the of the Department of Energy's (DOE) National Nuclear Security Administration's (NNSA) Vanguard Project. The project's mission is to ensure a viable HPC ecosystem is established for ARM technology within the NNSA and the larger DOE community. Besides Sandia, a number of other national labs are involved in the project, including Lawrence Livermore, Oak Ridge, Argonne, and Los Alamos.

Over the next few years, these labs will help fill out the system software stack and perform application porting for various multi-physics codes, with the eventual goal of supporting ARM-based exascale systems at the agency. A number of ThunderX2-powered prototype clusters, based on pre-production Cavium silicon are already running at the labs, and are being used to develop operating system (OS) support, compilers, math libraries, file systems, and other elements of the toolchain. Lawrence Livermore, for example, has already ported the NNSA's Tri-Lab Operating System Stack (TOSS) to the ThunderX2 platform.

Given the size of Astra, it will be the first ARM system in the world that will be able to run HPC workloads at true supercomputing scale and demonstrate how much computational capacity can be extracted from the hardware. "One of the important questions Astra will help us answer is how well does the peak performance of this architecture translate into real performance for our mission applications," said Mark Anderson, program director for NNSA's Advanced Simulation and Computing program, which funds Astra.

Thanks to the local flash storage and the eight memory channels on each ThunderX2 socket, Astra is likely to be especially adept at analytics and other data-demanding codes. In particular, the eight-memory-design represents a 33 percent improvement on Intel's six-channel implementation of its Xeon Scalable processor. The better bandwidth is one of ThunderX2's most

important differentiating features and represents an attempt to provide a more balanced relationship between compute capacity and memory speed. (Note that you don't need an ARM processor to this; AMD has the same eight-memory-channel design with its x86 EPYC processor.)

This focus on optimizing data movement is due to the fact that this is where most of the system's energy is being consumed these days. "We can see clearly that the amount of power required to move data inside the system is an order of magnitude greater than the amount of power needed to compute that data," explained Vildibill.

That said, the DOE's main interest in ARM probably has more to do with the fact that it represents a third viable processor architecture for the datacenter and is poised to get much broader industry support. Maybe just as important though, the architecture is driven by an open licensing model than encourages innovation and diversity. And that model has already resulted in a partnership between ARM Ltd and Fujitsu to establish [an HPC implementation of ARM](#), known as the ARMv8-A Scalable Vector Extension (SVE). It's set to debut in the Post-K supercomputer, Japan's initial exascale system that is scheduled to be installed at RIKEN in 2021-2022. Future features, such as on-package high bandwidth memory and integrated high-performance interconnects are already being anticipated.

Astra is scheduled for deployment in late summer. It will be installed at Sandia in a part of a datacenter that originally housed the Red Storm supercomputer.

 **06/23/18--03:55: Thomas Sterling Talks Exascale, Chinese HPC, Machine Learning, and Non-von Neumann Architectures** 0  0   

On Wednesday evening at the ISC High Performance conference, HPC luminary Dr. Thomas Sterling will deliver his customary keynote address on the state of high performance computing. To get something of a preview of that talk, we caught up with Sterling and asked him about some of the more pressing topics in the space.

What follows is pretty much the unedited text of our email exchange.

#### **TOP500 News: What do you think the achievement of exascale computing will mean to the HPC user community?**

**Thomas Sterling:** As a particular point in capacity and capability, exascale is as arbitrary in the continuum of performance as any other. But symbolically it is a milestone in the advancement of one of mankind's most important technologies, marking unprecedented promise in modeling and information management.

Of a subtler nature, it is a beachhead on the forefront of nanoscale enabling technologies, marking the end of Moore's Law, the flatlining of clock rates due to power considerations, and the limitations of clock rate. The achievement of



exascale computing will serve as an inflection point at which change from conventional means is not only inevitable but essential. It also implies the need to replace the venerable von Neumann architecture of which almost all commercial computing

systems of the last seven decades are derivatives thereof.

Many will correctly argue the specific metrics by which this point is measured but at any dimension, it reflects progress, even if not as much as the community would like to think. This last consideration is a reflection of the Olympian heights at which almost all computing is excluded. The reality is that almost all systems operate at about two orders of magnitude lower capability. But then, most of us do not drive a Rolls Royce, while still admiring it.

#### **TOP500 News: Do you think it's important which nation reaches that milestone first?**

**Sterling:** It would be easy to dismiss the importance of the exact order in which nations realize exascale capability, in particular, based on High Performance Linpack (HPL). Perhaps a far more important metric is a nation's per capita number of systems deployed on the TOP500 list, suggesting the degree of access for high-end computing; this suggests that the number 500 system is the more important line on any such curve.

Further, thoughtful practitioners correctly observe that the accomplishment is the actual amount of science and engineering achieved as well as other important tasks, not an artificial test that has meaningful consequences. Finally, it's not even clear that we are looking at all of the world's big machines with industry deploying and operating enormous conglomerates of processing components and not even participating in the Linpack marathon.

Intellectually I agree with all of these cogent viewpoints. But there is an emotional aspect of this milestone and we are a species driven more by emotions than we are by predicate calculus. A nation is one delineation of a society and people – even HPC people – are atomic elements of societies. If a nation and therefore a societal identity is measured as competitive, then we as individuals inherit that property, sense of satisfaction – yes, even pride – and the tools of future achievement to which HPC contributes. If we fail significantly, then we accept our lesser stature. It does not so much matter who is at the front of the line around the race track at any instance. But it does matter if we are part of the race.

**TOP500 News: Potential exascale achievements aside, how much do you worry about the ascendance of China in HPC?**

**Sterling:** The recent dominance of China is important and of a concern, not that it is a Chinese accomplishment, but rather that it demonstrates a potential diminishing of US will, means, and ability of delivering the best that enabling technology can offer. I applaud the Chinese advances as well as those of Europe and Japan. The K machine has been at the top of the [Graph 500 list](#) and Europe is exploring alternative hardware structures for future HPC.

More, the Chinese have demonstrated significant innovation and are also competitive in terms of number of deployed HPC systems. They are learning a lot about how to apply these machines to real world applications rather than just paying for them. US funding has stalled and research in HPC has declined precipitously. Even with a recent increase in HPC budget by the US Senate, this has not been refactored into US HPC research but rather in system deployment. While [Summit](#) is a meaningful and long-awaited demonstration of American engagement in HPC progress, what does the [failure of the Aurora project](#) of similar scale portend for the future. It is not the Chinese success I worry about, it is the US stagnation I fear.

**TOP500 News: What's your perspective on the impact of machine learning and the broader area of data analytics on supercomputing – from both its effect on how its driving hardware – processors, memory, networking, etc. – and on how HPC practitioners are incorporating machine learning into their traditional workflows?**

**Sterling:** The techniques at the core of machine learning go back to the 1980's – neural-nets – and while creditable improvements have been inaugurated, the foundations are clearly similar. The important advance, even leap-frogging, is the explosive applications to unprecedented scale of data to which HPC is now being put to use. With little fear of contradiction, machine learning and data analytics is a major extension and market of HPC.

Further, it gives us a window into disorganized data sources in part made available through the internet and from many disparate origins from giant science experiments like the Large Hadron Collider, to existing and of criticality, for example, personalized medicine. I hope that these emerging data-intensive applications will stimulate innovative concepts in the memory side of supercomputer architecture with less focus on the FPU and more on the memory semantics, latency, bandwidth, and parallelism. It is long overdue.

But it should be noted that the term “machine learning” should not be misconstrued to mean machine intelligence. It is people who learn from the data processing of this paradigm, gaining human understanding and knowledge, not the machines. Machines do not know how to process the knowledge gained by human practitioners and they certainly don't achieve anything like “understanding”. For this, we need yet new breakthroughs to reach Machine Intelligence (MI).

**TOP500 News: There seem to be three computing technologies on the horizon that could potentially disrupt the market: quantum computing, neuromorphic computing, and optical computing. Can you give us your perspective on the potential of each of these for HPC kinds of applications?**

**Sterling:** The three identified technologies are certainly part of the exploratory road map “on the horizon” as you say. But they are neither alone nor necessarily the top three. The enormous funding being poured into quantum computing by industry and government is very positive and will ultimately lead to a new form of computing far different from conventional practices, and as distinct as [Vannevar Bush](#) machines were from the succeeding von Neumann generation. Quantum computing will be important but always serving domain-specific purposes where their advantages can be exploited.

Neuromorphic or “brain-inspired” computing is intriguing as it is uncertain. The diversity of approaches being explored is constructive as ideas and insights emerge through a human relaxation process. I personally don’t think it’s going to work the way many people think. For example, I don’t think we have to mimic the structural elements of the brain to achieve machine intelligence. But I do expect that the associative methods hardwired into the brain if duplicated in some analogous fashion will greatly enhance certain idioms of processing that are very slow with today’s conventional methods. Right now, the inspiration is catalyzing new ideas and resulting methods that are worth exploring. Who knows what we will find.

Optical computing in the sense of adopting optical technologies to digital computing have been heavily pursued but in active data storage and logical data transformation have not proved successful. I love the idea but do not have faith in its promise. However, optical in the analogous sense employing non-linear functionality in the analog versus digital means may be very promising in the long term – or not. A form of extreme heterogeneity mixing the strengths of optical in narrow operations with parallel digital systems may serve in a manner similar to structures integrating GPUs today.

But I’ll close here by mentioning two other possibilities that, while not widely considered currently, are nonetheless worthy of research. The first is superconducting supercomputing and the second is non-von Neumann architectures. Interestingly, the two at least in some forms can serve each other making both viable and highly competitive with respect to future post-exascale computing designs. Niobium Josephson Junction-based technologies cooled to four Kelvins can operate beyond 100 and 200 GHz and has slowly evolved over two or more decades. When once such cold temperatures were considered a show stopper, now quantum computing – or at least quantum annealing – typically is performed at 40 milli-Kelvins or lower, where four Kelvins would appear like a balmy day on the beach. But latencies measured in cycles grow proportionally with clock rate and superconducting supercomputing must take a very distinct form from typical von Neumann cores; this is a controversial view, by the way.

Possible alternative non-von Neumann architectures that would address this challenge are cellular automata and data flow, both with their own problems, of course – nothing is easy. I introduce this thought not to necessarily advocate for a pet project – it is a pet project of mine – but to suggest that the view of the future possibilities as we enter the post-exascale era is a wide and exciting field at a time where we may cross a singularity before relaxing once again on a path of incremental optimizations.

I [once said](#) in public and in writing that I predicted we would never get to zettaflops computing. Here, I retract this prediction and contribute a contradicting assertion: zettaflops can be achieved in less than 10 years if we

adopt innovations in non-von Neumann architecture. With a change to cryogenic technologies, we can reach yottaflops by 2030.

Thomas Sterling's ISC [keynote address](#) will take place at the Messe Frankfurt on Wednesday, June 27, at 5:30 - 6:15 pm CEST.

06/24/18--17:37: **US Regains TOP500 Crown with Summit Supercomputer, Sierra Grabs Number Three Spot**

0  0   

FRANKFURT, Germany; BERKELEY, Calif.; and KNOXVILLE, Tenn.—The TOP500 celebrates its 25<sup>th</sup> anniversary with a major shakeup at the top of the list. For the first time since November 2012, the US claims the most powerful supercomputer in the world, leading a significant turnover in which four of the five top systems were either new or substantially upgraded.



Summit supercomputer. Source: Oak Ridge National Laboratory

Summit, an IBM-built supercomputer now running at the Department of Energy's (DOE) Oak Ridge National Laboratory (ORNL), captured the number

one spot with a performance of 122.3 petaflops on High Performance Linpack (HPL), the benchmark used to rank the TOP500 list. Summit has 4,356 nodes, each one equipped with two 22-core Power9 CPUs, and six NVIDIA Tesla V100 GPUs. The nodes are linked together with a Mellanox dual-rail EDR InfiniBand network.

Sunway TaihuLight, a system developed by China's National Research Center of Parallel Computer Engineering & Technology (NRCPC) and installed at the National Supercomputing Center in Wuxi, drops to number two after leading the list for the past two years. Its HPL mark of 93 petaflops has remained unchanged since it came online in June 2016.

Sierra, a new system at the DOE's Lawrence Livermore National Laboratory took the number three spot, delivering 71.6 petaflops on HPL. Built by IBM, Sierra's architecture is quite similar to that of Summit, with each of its 4,320 nodes powered by two Power9 CPUs plus four NVIDIA Tesla V100 GPUs and using the same Mellanox EDR InfiniBand as the system interconnect.

Tianhe-2A, also known as Milky Way-2A, moved down two notches into the number four spot, despite receiving a major upgrade that replaced its five-year-old Xeon Phi accelerators with custom-built Matrix-2000 coprocessors. The new hardware increased the system's HPL performance from 33.9 petaflops to 61.4 petaflops, while bumping up its power consumption by less than four percent. Tianhe-2A was developed by China's National University of Defense Technology (NUDT) and is installed at the National Supercomputer Center in Guangzhou, China.

The new AI Bridging Cloud Infrastructure (ABCI) is the fifth-ranked system on the list, with an HPL mark of 19.9 petaflops. The Fujitsu-built supercomputer is powered by 20-core Xeon Gold processors along with NVIDIA Tesla V100 GPUs. It's installed in Japan at the National Institute of Advanced Industrial Science and Technology (AIST).

Piz Daint (19.6 petaflops), Titan (17.6 petaflops), Sequoia (17.2 petaflops), Trinity (14.1 petaflops), and Cori (14.0 petaflops) move down to the number six through 10 spots, respectively.

### General highlights

Despite the ascendance of the US at the top of the rankings, the country now claims only 124 systems on the list, a new low. Just six months ago, the US had 145 systems. Meanwhile, China improved its representation to 206 total systems, compared to 202 on the last list. However, thanks mainly to Summit and Sierra, the US did manage to take the lead back from China in the performance category. Systems installed in the US now contribute 38.2 percent of the aggregate installed performance, with China in second place with 29.1 percent. These numbers are a reversal compared to six months ago.

The next most prominent countries are Japan, with 36 systems, the United Kingdom, with 22 systems, Germany with 21 systems, and France, with 18 systems. These numbers are nearly the same as they were on the previous list.

For the first time, total performance of all 500 systems exceeds one exaflop, 1.22 exaflops to be exact. That's up from 845 petaflops in the November 2017



list. As impressive as that sounds, the increase in installed performance is well below the previous long-term trend we had seen until 2013.

The overall increase in installed capacity is also reflected in the fact that there are now 273 systems with HPL performance greater than one petaflop, up from 181 systems on the previous list. The entry level to the list is now 716 teraflops, an increase of 168 teraflops.

### Technology trends

Accelerators are used in 110 TOP500 systems, a slight increase from the 101 accelerated systems in the November 2017 lists. NVIDIA GPUs are present in 96 of these systems, including five of the top 10: Summit, Sierra, ABCI, Piz Daint, and Titan. Seven systems are equipped with Xeon Phi coprocessors, while PEZY accelerators are used in four systems. An additional 20 systems now use Xeon Phi as the main processing unit.

Almost all the supercomputers on the list (97.8 percent) are powered by main processors with eight or more cores and more than half (53.2 percent) have over 16 cores.

Ethernet, 10G or faster, is now used in 247 systems, up from 228 six months ago. InfiniBand is found on 139 systems, down from 163 on the previous list. Intel's Omni-Path technology is in 38 systems, slightly up from 35 six months ago.

### Vendor highlights

For the first time, the leading HPC manufacturer of supercomputers on the list is not from the US. Chinese-based Lenovo took the lead with 23.8 percent (122 systems) of all installed machines, followed by HPE with 15.8 percent (79 systems), Inspur with 13.6 percent (68 systems), Cray with 11.2 percent (56 systems), and Sugon with 11 percent (55 systems). Of these, only Lenovo, Inspur, and Sugon captured additional system share compared to half a year ago.

Even though IBM has two of the top three supercomputers in Summit and Sierra, it claims just 19 systems on the entire list. However, thanks to those two machines, the company now contributes 19.9 percent of all TOP500 performance. Trailing IBM is Cray, with 16.5 percent of performance, Lenovo with 12.0 percent, and HPE with 9.9 percent.

Intel processors are used in 476 systems, which is marginally higher than the 471 systems on the last list. IBM Power processors are now in 13 systems, down from 14 systems since November 2017.

### Green500 results

The top three positions in the Green500 are all taken by supercomputers installed in Japan that are based on the ZettaScaler-2.2 architecture using PEZY-SC2 accelerators, while all other system in the top 10 use NVIDIA GPUs.

The most energy-efficient supercomputer is once again the Shoubu system B, a ZettaScaler-2.2 system installed at the Advanced Center for Computing and Communication, RIKEN, Japan. It was remeasured and achieved 18.4

gigaflops/watt during its 858 teraflops Linpack performance run. It is ranked number 362 in the TOP500 list.

The second-most energy-efficient system is SuiRen2 system at the High Energy Accelerator Research Organization/KEK, Japan. This ZettaScaler-2.2 system achieved 16.8 gigaflops/watt and is listed at position 421 in the TOP500. Number three on the Green500 is the Sakura system, which is installed at PEZY Computing. It achieved 16.7 gigaflops/watt and occupies position 388 on the TOP500 list.

They are followed by the DGX SaturnV Volta system in the US; Summit in the US; the TSUBAME 3.0 system, AIST AI Cloud system, the AI Bridging Cloud Infrastructure (ABCI) system, all from Japan; the new IBM MareNostrum P9 cluster in Spain; the DOE's Summit system; and Wilkes-2, from the UK. All of these systems use various NVIDIA GPUs.

The most energy-efficient supercomputer that doesn't rely on accelerators of any kind is the Sunway TaihuLight, which is powered exclusively by ShenWei processors. Its 6.05 gigaflops/watt earned it 22nd place on the Green500 list.

### **HPCG Results**

The TOP500 list has incorporated the High-Performance Conjugate Gradient (HPCG) Benchmark results, which provided an alternative metric for assessing supercomputer performance and is meant to complement the HPL measurement.

The two new DOE systems, Summit at ORNL and Sierra at LLNL, captured the first two positions on the latest HPCG rankings. Summit achieved 2.93 HPCG-petaflops and Sierra delivered 1.80 HPCG-petaflops. They are followed by the previous leader, Fujitsu's K computer, which attained 0.60 HPCG-petaflops. Trinity, a Cray XC40 system installed at Los Alamos National Lab and Piz Daint, a Cray XC50 system installed at the Swiss National Supercomputing Centre (CSCS) round out the top five.

### **About the TOP500 List**

The first version of what became today's TOP500 list started as an exercise for a small conference in Germany in June 1993. Out of curiosity, the authors decided to revisit the list in November 1993 to see how things had changed. About that time, they realized they might be onto something and decided to continue compiling the list, which is now a much-anticipated, much-watched and much-debated twice-yearly event.

The TOP500 list is compiled by Erich Strohmaier and Horst Simon of Lawrence Berkeley National Laboratory; Jack Dongarra of the University of Tennessee, Knoxville; and Martin Meuer of ISC Group, Germany.

[More details on the current list](#)

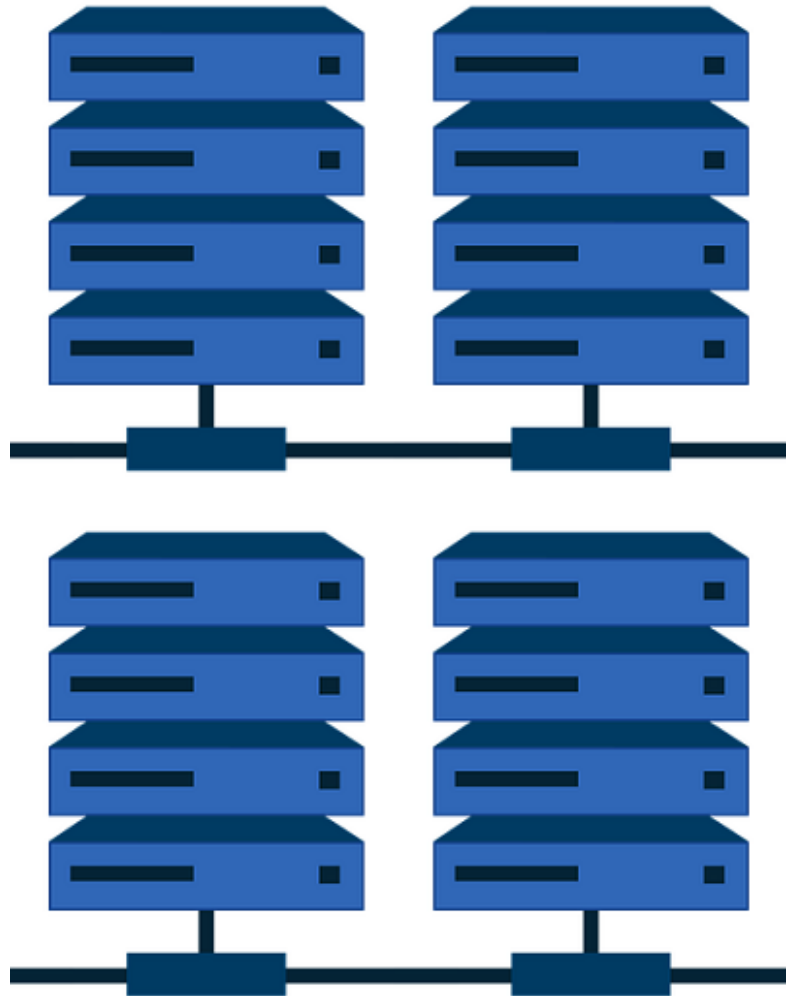
06/25/18--11:58: **Distortions, Trends and Quirks in the June 2018 TOP500 List**



What did the cynical and curious @hpcnotes spot in the TOP500 list?

### Stuffing the list

The TOP500 list is an intensely valuable tool for the HPC community, tracking aggregate trends over 25 years. However, a few observers have noted that recent publications of the TOP500 list have many duplicate entries, often at anonymous sites.



Let's park the debate on what is a legitimate HPC system or not for now and assume that any system that has completed a High Performance Linpack (HPL) run is a fair entry on the list.

But, there are 124 entries on the list that are identical copies of other entries. In other words, a single HPL run has been done on one system, and the vendor has said "since we have sold N copies of that machine, we can submit N entries on the list."

What happens to the list statistics if we delete all the duplicate systems?

The list of 500 reduces to 376 entries.

The biggest change is Lenovo, dropping from 117 entries to just sixty-one – yes, there are 56 duplicate entries from Lenovo! HPE drops from 79 to 62 entries and retakes the top spot for greatest share of the list with 16.5 percent. Lenovo drops to second place with 16.2 percent share.

Does this matter? Well, it probably does to Lenovo, who chose to submit many copies, and HPE, who probably sold many copies but chose not to submit the duplicates. And ultimately it matters to their market share PR.

For the rest of us, it comes down to what the list is about. If it supposed to list the 500 fastest supercomputers in the world, clearly it doesn't do that as many supercomputer owners choose not to acknowledge their systems. Is it the list of known supercomputers? No, because several known supercomputers are not listed, for example, Blue Waters. So, it can only be a list of acknowledged HPL runs, which would suggest that the "N copies" approach is wrong.

However, it isn't as simple as that. If the list is for tracking vendor/technology market share, then list stuffing is fine – even needed. If the list is for tracking adoption of technologies, vendor wins, the fortunes of HPC sites, and progress of nations, then I'd argue that stuffing in this way breaks the usefulness of the list.

The comparison of progress of nations is also affected. Do we measure how many systems deployed? Or who has the biggest system? I think the list stuffing is less critical here, as we can readily extract both trends.

I'm not sure there is a right or wrong answer to this but, as always, the headline statistics of market share only tell one side of the story, and users of the list must dig deeper for appropriate insight.

### **Anonymity rules**

Another value of the list over the years has been the ability to track who is buying supercomputers and what they are using them for.

In the June 2018 list, 97 percent of the systems do not specify an application area. This suggests this categorization is essentially meaningless. Either drop it from future lists or require future systems to identify an application area.

But, beyond that, the June 2018 List has 283 (!) anonymous entries. That is over half of the list where we do not know the company or site that has deployed the system nor, in most cases, what it is being used for.

The big question is: does that render meaningless the ability to track the who and what of HPC, or would we be in a worse position if we excluded anonymous systems?

There are maybe 238 systems that can be lumped into cloud / IT service / web hosting companies, and it is the sheer quantity of these that provides the potentially unhelpful distortion.

The remaining 45 arguably represent real HPC deployments and have enough categorization (e.g., "Energy Company") to be useful. Useful guesses can even be made for some of these anonymous systems. For example, one might

guess that “Energy Company (A)” located in Italy is actually Eni. The 45 systems are a mix of energy, manufacturing, government, or finance sectors. Most interestingly, a couple of university systems are listed anonymously too!

### Supercomputers in Industry

Of the 284 systems listed as “Industry,” only 16 systems actually have named owners that aren’t cloud providers (the other 268 are mostly the anonymous companies and “stuffing” discussed above). In fact, due to multiple systems per site, we actually only have a very small number of listed companies. These are Eni, Total, PGS, Saudi Aramco, BASF, EDF, Volvo and the vendors Intel, NVIDIA, and PEZY.

I assure you there are many more supercomputers out there in industry than this. I intend to explore the TOP500 trends, along lessons from my impartial HPC consulting work, of HPC systems in industry for a future article. Keep an eye out for it after ISC18.

### Hope

I have previously said that the HPC is lucky to have the TOP500 list. It is a data set gathered in a consistent manner twice per year for 25 years, with each entry recording many attributes – vendor, technology details, performance, location, year of entry, etc. This is a hugely rich resource for our community, and the TOP500 authors have done an amazing job over 25 years to keep the list useful as the world of HPC has changed enormously. I have high confidence in the authors successfully addressing the challenges listed here and keeping the list as an invaluable resource for years to come.

*Andrew Jones can be contacted via twitter (@hpcnotes), via LinkedIn (<https://www.linkedin.com/in/andrewjones/>), or via the TOP500 editor.*

📅 **06/25/18--21:43: CERN's Maria Girone: "We explain the magic!"**



The International Supercomputing Conference (ISC18) kicked off Monday in Frankfurt, Germany, with Maria Girone, CTO of CERN **openlab** delivering the opening keynote address. She explained how CERN’s needs will drive exascale computation and data science innovation in the future.



Founded in 1954, CERN straddles the Franco-Swiss border, and has an annual operating budget of one billion Swiss francs. With 22 member-states, its resources support 15,000 scientists around the world, and employs 2,500 at Swiss and French sites. CERN is home to the Large Hadron Collider (LHC), the world's largest and most powerful particle accelerator.

In addition to probing the fundamental structure of our universe, understanding the very first moments of our universe after the big bang, and searching for dark matter, CERN technicians and scientists are always developing new technologies for accelerators and detectors. Its instrumentation advances medical diagnoses and therapies, trains the scientists and engineers of tomorrow, and unites people from different countries and cultures. "CERN is every bit as much of a feat of social engineering as it is a technical challenge," said Girone. Each participating lab builds their own parts that must work with everything else. There are 170 collaborative computing centers in 42 countries on most continents. She added with a smile, "we're even working on Antarctica!"

Two general-purpose detectors cross-confirm major discoveries, such as the



*Photo credit : CERN*

Higgs Boson on July 4, 2012. ALICE and LHCb (the “b” stands for beauty) are detectors that specialize in the study of specific high energy physics (HEP) phenomena. The LHCb experiment investigates the “slight differences between matter and antimatter by studying a type of particle called the beauty quark, or b quark.” There are 650 scientists from 48 institutions in 13 countries who participate in the LHCb experiment, alone.

CERN instrumentation has the capacity to generate 1 petabyte of data per second, and hundreds of petabytes per year. The Meyrin data center in Geneva is the heart of CERN’s computing infrastructure with 300,000 processor cores, 180 petabytes of disk and 230 petabytes of tape storage. A second data center in Wigner, Budapest (WLCG) features HPC systems with 100,000 cores, and 100 petabytes of disk storage. WLCG gives thousands HEP of scientists around the world near real-time access.

CERN is a leader in global data distribution and management. Massive amounts of data are moved between hemispheres each day via 340 Gbps transatlantic link, and identity management is handled by the eduGAIN federation. From a local management standpoint, eduGAIN saves managers time and effort because home credentials provide authentication and access to resources, instrumentation and data that are physically located at institutions in 48 member-countries that comprise an interfederated trust fabric. It’s

more secure and takes less time to manage since researchers must only remember one user name and password.

It's necessary for physicists to sift through 30 to 50 petabytes produced annually by the LHC experiments. "Searching for a single event is compared to finding a specific grain of sand in 20 volleyball courts," said Girone. There is so much data that scientists couldn't possibly access or manage raw data, so CERN exploits co-processors for software-based filtering and real-time construction, which prunes, packs and optimizes data for transfer and analysis.

CERN's road map for the future includes the construction of the High Luminosity (HL) LHC, which began a few weeks ago and is expected to be completed in 2035. It will be a challenge to address the exascale demands in the future. Considering current technologies, CERN will need 50 to 100 times the amount of current CPU computing capacity by 2028, as well as further advances in software development, advanced networks and storage solutions. The technology evolution expected to take place between now and 2028 will help meet this goal, along with close collaborations with industry partners.

CERN **openlab** is a broad science-industry partnership that fosters research and innovation to drive the innovation of ICT solutions for CERN and its stakeholders. To tackle the resource gap, **openlab** plans to fully exploit available hardware, expand dynamically to new computing environments, and introduce layered, virtualized services to provide flexibility and efficiency. CERN expects as much as 90 percent of resources will be delivered via the OpenStack private cloud platform, which will allow flexibility and dynamic deployment. Additionally, they have the option of elastically and dynamically expanding production to commercial clouds and currently have a joint procurement of R&D cloud services with several providers. While they employ accelerators, such as GPUs, FPGAs and others, they are exploring lower-performance, low-power alternatives, such as ARM. They have a focus on software optimization toward greater performance, and plan to explore the entire panorama of emerging innovations as they unfold.

CERN is well-suited for machine learning, and is expected to become a leader in the field of artificial intelligence (AI), specifically in the areas of monitoring automation and anomaly detection. Reconstruction and simulation—the most important applications for CERN—are data-intensive processes that will also benefit from machine learning and AI.

CERN has a relationship with the Square Kilometer Array (SKA) project, which has instrumentation in the Karoo region of South Africa, and in Australia. Phase One of that project is expected to start generating data in the mid 2020's and will function for the next 50 years. CERN has a joint exascale data-storage and processing challenge between HL-LHC & SKA. Additionally, CERN is accelerating innovation and knowledge transfer to medical applications via the MEDICIS-PROMED program.

Girone closed her presentation with a quote by actor Tom Hanks who had visited CERN in 2009 with actor-director Ron Howard. He said, "Magic is not happening here; magic is being explained here."

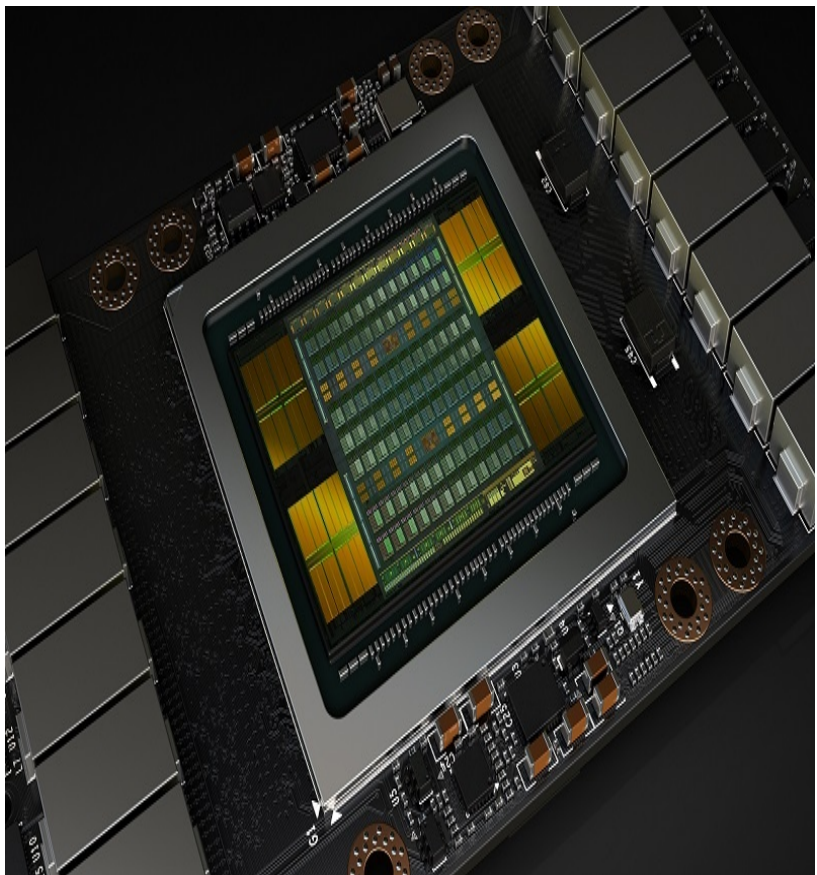
*For more information about CERN, the LHC and related projects, visit their [website](#). Follow @ISChpc and #ISC18 for more conference news. Elizabeth Leake, Founder and President of [STEM-Trek](#), can be followed at @STEMTrek.*



06/26/18--09:09: **New GPU-Accelerated Supercomputers Change the Balance of Power on the TOP500**

0  0   

For the first time in history, most of the flops added to the TOP500 list came from GPUs instead of CPUs. Is this the shape of things to come?



*Tesla V100 GPU. Source: NVIDIA*

In the latest TOP500 rankings announced this week, 56 percent of the additional flops were a result of NVIDIA Tesla GPUs running in new supercomputers – that according to the Nvidians, who enjoy keeping track of

such things. In this case, most of those additional flops came from three top systems new to the list: Summit, Sierra, and the AI Bridging Cloud Infrastructure (ABCI).

Summit, the new TOP500 champ, pushed the previous number one system, the 93-petaflop Sunway TaihuLight, into second place with a Linpack score of 122.3 petaflops. Summit is powered by IBM servers, each one equipped with two Power9 CPUs and six V100 GPUs. According to NVIDIA, 95 percent of the Summit's peak performance (187.7 petaflops) is derived from the system's 27,686 GPUs.

NVIDIA did a similar calculation for the less powerful, and somewhat less GPU-intense Sierra, which now ranks as the third fastest supercomputer in the world at 71.6 Linpack petaflops. And, although very similar to Summit, it has four V100 GPUs in each dual-socketed Power9 node, rather than six. However, the 17,280 GPUs in Sierra still represent the lion's share of that system's flops.

Likewise for the new ABCI machine in Japan, which is now that country's speediest supercomputer and is ranked fifth in the world. Each of its servers pairs two Intel Xeon Gold CPUs with four V100 GPUs. Its 4,352 V100s deliver the vast majority of the system's 19.9 Linpack petaflops.

As dramatic as that 56 percent number is for new TOP500 flops, the reality is probably even more impressive. According to Ian Buck, vice president of NVIDIA's Accelerated Computing business unit, more than half the Tesla GPUs they sell into the HPC/AI/data analytics space are bought by customers who never submit their systems for TOP500 consideration. Although many of these GPU-accelerated machines would qualify for a spot on the list, these particular customers either don't care about all the TOP500 fanfare or would rather not advertise their hardware-buying habits to their competitors.

It's also worth mentioning that the Tensor Cores in the V100 GPUs, with their specialized 16-bit matrix math capability, endow these three new systems with more deep learning potential than any previous supercomputer. Summit alone boasts over three peak exaflops of deep learning performance. Sierra's performance in this regard is more in the neighborhood of two peak exaflops, while the ABCI number is around half an exaflop. Taken together, these three supercomputers represent more deep learning capability than the other 497 systems on the TOP500 list combined, at least from the perspective of theoretical performance.

The addition of AI/machine learning/deep learning into the HPC application space is a relatively new phenomenon, but the V100 appears to be acting as a catalyst. "This year's TOP500 list represents a clear shift towards systems that support both HPC and AI computing," noted TOP500 author Jack Dongarra, Professor at University of Tennessee and Oak Ridge National Lab.

While company's like Intel, Google, Fujitsu, Wave Computing, Graphcore, and others are developing specialized deep learning accelerators for the datacenter, NVIDIA is sticking with an integrated AI-HPC design for its Tesla GPU line. And this certainly seems to be paying off, given the growing trend of using artificial intelligence to accelerate traditional HPC applications. Although the percentage of users integrating HPC and AI is still relatively small, this mixed-workflow model is slowly being extended to nearly every science and engineering domain, from weather forecasting and financial analytics, to genomics and oil & gas exploration.

Buck admits this interplay between traditional HPC modeling and machine learning is still in the earliest stages, but maintains “it’s only going to get more intertwined.” He says even though some customers will use only a subset of the Tesla GPU’s features, the benefits of supporting 64-bit HPC, machine learning, and visualization on the same chip far outweighs any advantages that could be realized by single-purpose accelerators.

And, thanks in large part to these deep-learning-enhanced V100 GPUs, mixed-workload machines are now popping up on a fairly regular basis. For example, although Summit was originally going to be just another humongous supercomputer, it is now being groomed as [a platform for cutting-edge AI](#) as well. By contrast, the [ABCI system was conceived from the beginning as an AI-capable supercomputer](#) that would serve users running both traditional simulations and analytics, as well as deep learning workloads. Earlier this month, the [MareNostrum supercomputer added three racks of Power9/V100 nodes](#), paving the way for serious deep learning work to commence at the Barcelona Supercomputing Centre. And even [the addition of just 12 V100 GPUs to the Nimbus cloud service](#) at the Pawsey Supercomputing Centre was enough to claim that AI would now be fair game on the Aussie system.

As Buck implied, you don’t have to take advantage of the Tensor Cores to get your money’s worth from the V100. At seven double-precision teraflops, the V100 is a very capable accelerator for conventional supercomputing. And according to NVIDIA, there are 554 codes ported to these graphics chips, including all of the top 15 HPC applications.

But as V100-powered systems make their way into research labs, universities, and commercial datacenters, more scientists and engineers will be tempted to inject AI into their 64-bit applications. And whether this turns out to be a case of the tail wagging the dog or the other way around, in the end, it doesn’t really matter. The HPC application landscape is going to be forever changed.

....

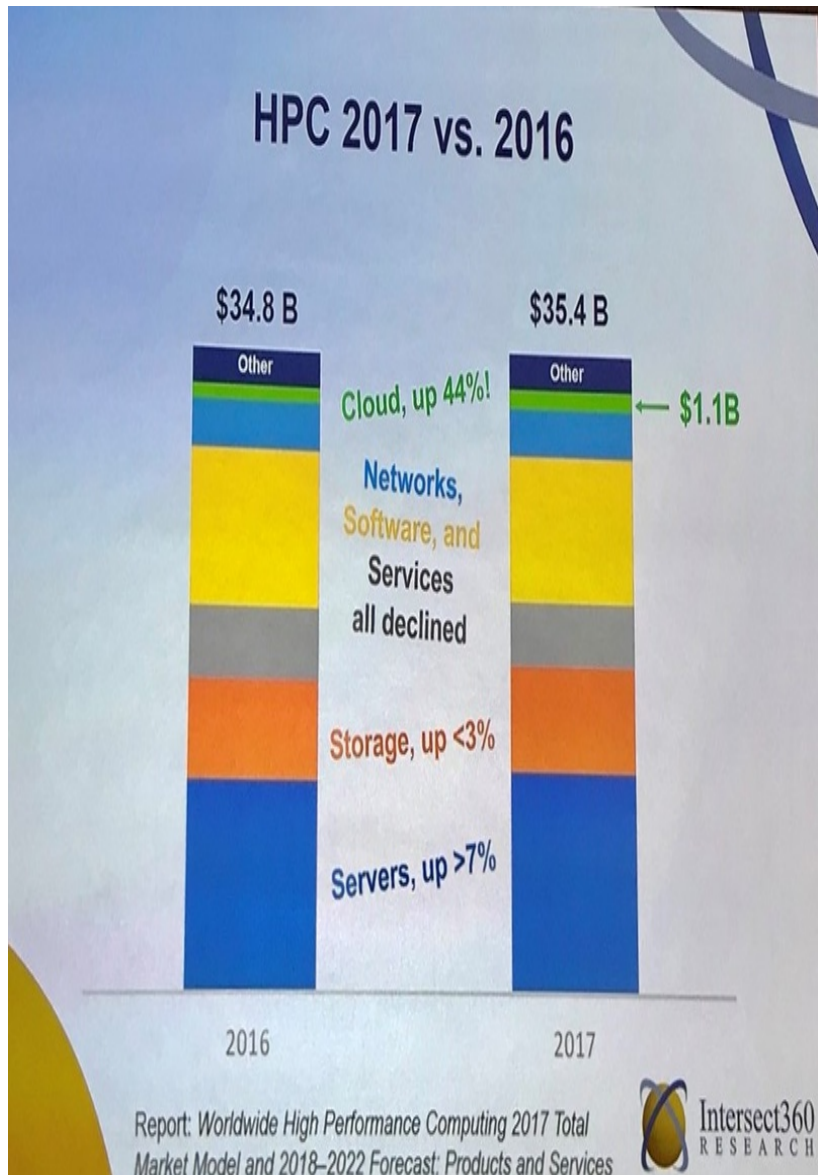
 **06/29/18--06:39: [Cloud Computing in HPC Surges](#)**  0  0  

According to the two leading analyst firms covering the high performance computing market, the use of the cloud for HPC workloads is looking a lot more attractive to users these days.

At the ISC High Performance conference this week, Intersect360 Research and Hyperion Research presented their respective HPC market forecasts, both of which included some rather encouraging news for the HPC-in-the-cloud crowd.

Intersect360 offered the most upbeat assessment in this regard, noting that cloud spending by HPC customers grew by a whopping 44 percent from 2016 to 2017, calling it a “breakout year” for this product category. According to the company’s market data, that put cloud-based spending at around \$1.1 billion

for 2017. And even though that represents only about three percent of total HPC revenue for the year, it's a high-water mark for cloud computing in this space.



Source: Intersect360 Research

To put that in perspective, Intersect360 had the entire HPC market during this period growing at a rather anemic 1.6 percent. While server spending was up by over seven percent and storage by less than three percent, overall growth was held back by decreased spending for networks, software, and non-cloud services.

The surge in cloud use must have been something of a surprise to the Intersect360 team, which as recently as last year was forecasting that spending wouldn't break \$1 billion until 2019. Now, with the new data in hand,


they are projecting cloud computing to be the fastest-growing product category for the next five years, reaching nearly \$3 billion by 2022.

Intersect360 noted big differences in the growth rates of the different cloud sub-categories. For example, spending on raw cycles grew only 5.6 percent from 2016 to 2017, while revenue for software-as-a-service (SaaS) increased by 125 percent.

The big jump in cloud spending was driven by a number of different factors, according to the Intersect360 folks, including “increasing facilities costs for hosting HPC, maturation of application licensing models, increased availability of high-performance cloud resources, and a spike in requirements for machine learning applications.”

The Hyperion Research team didn't offer specific cloud revenue spending or projections during their presentation at the ISC conference, but did note that 64 percent of HPC sites now run at least some of their work in public cloud. That's up from just 13 percent in 2011. However, according to Hyperion these same sites used cloud resources for only seven to eight percent of their jobs, which suggests a lot of these users were tapping into the cloud for bursting once their in-house capacity filled up. In fact, the need for extra capacity was the most cited reason by these sites for running some of their applications in the cloud.

According to Hyperion, another important factor driving HPC use in the cloud these days is the demand for special hardware or software features. (It was the number one reason given by government sites.) This is likely to be especially true for users running machine learning applications, where the most recently minted GPUs, like NVIDIA's V100, have a much higher capability for such workloads than previous models.

 **07/02/18--09:23: [Benchmarks in Hand, UK Academics See Promising Future for Arm Chips in HPC](#)**

0



0



Researchers at the Great Western 4 (GW4) Alliance have benchmarked the Cavium ThunderX2 processor that will soon power the Isambard supercomputer. But the most significant advantage of the Arm processor may have nothing to do with performance numbers.

GW4, which represents four universities in the South West of England and Wales (Bath, Bristol, Cardiff Exeter), will soon be installing Isambard, a 10,000-core machine that the alliance is calling the “world's first production Arm supercomputer.” When it comes online, it will be the most powerful Arm-powered supercomputer in the UK and second only to [the Astra system scheduled to be deployed at Sandia National Laboratories later this summer](#).

GW4 has been benchmarking pre-production versions of the ThunderX2 for at least a year, but with the [ThunderX2 chips now in production](#), more practical comparisons can now be made with the Intel Xeon processors already in the

field. In this case, a 32-core ThunderX2 was compared against a 22-core “Broadwell” Xeon and a 28-core “Skylake” Xeon. The University of Bristol’s Simon McIntosh-Smith summarized the results in a recent [blog post](#), which focused on floating point performance, as well as memory and cache bandwidth. The benchmark scores, illustrated below, were based on heavily used HPC codes that are run on ARCHER, the UK’s national supercomputer.

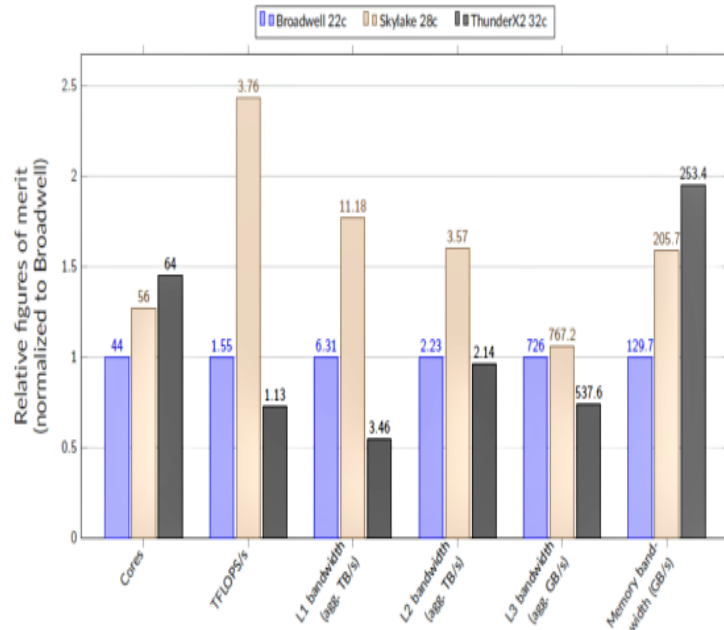


FIGURE 2 Comparison of properties of Broadwell 22c, Skylake 28c and ThunderX2 32c. Results are normalized to Broadwell.

Source: Simon McIntosh-Smith, University of Bristol

If you’ve been following our coverage of ThunderX2 and Arm, the results are pretty much as expected. The floating-point-heavy Xeon processors were substantially better in the flops department than the ThunderX2. This is especially true for the Skylake Xeon, which incorporates 512-bit Advance Vector Extensions (AVX). By contrast, ThunderX2’s vector width is a measly 128 bits.

On the other hand, thanks to Cavium’s 8-channel memory design, the ThunderX2 has about 23 percent more memory bandwidth than the Skylake Xeon and about 95 percent more than the Broadwell chip. The cache performance results were somewhat mixed, although the Skylake Xeon came out on top across all three cache levels.

The obvious conclusion is that the Xeons are preferable for floating point-intensive codes, while the ThunderX2 are the better choice for applications that are memory bound and less reliant on floating point performance. For codes that can perform the majority of their work out of cache, the Xeon would also be the chip of choice.

One advantage the Cavium processors have is lower price. McIntosh-Smith says Arm-based processors are “significantly cheaper than those shipping from the incumbent vendors, by a factor of 2-3X, depending on which SKUs you compare.” As a consequence, from a price-performance perspective, the flops-lite ThunderX2 could look a lot more attractive.

McIntosh-Smith also makes the point that having a diverse set of architectures from which to choose is a good thing, independent of any specific performance advantages one particular chip may have over another. “For users, this means that we have a new set of processor vendors becoming relevant, giving us much more choice today than at any point in the last decade,” he writes.

One might be tempted ask, why can't we have lots of flops, cache performance, and memory bandwidth on the same die? Unfortunately, with a limited amount of transistor real estate, chip designers are forced to make trade-off and hope the selected capabilities deliver the right balance for the largest number of applications.

In a way, this is easier to navigate for upstarts like Cavium, who just need to find a way to edge out the competition in one or two areas in order to grab some initial market share, rather than having to worry about a legacy customer base and the all the applications that go with them. It's interesting that AMD, with a similar mandate to snatch some of Intel's market share with its EPYC processor, also focused on memory performance rather than big fat vectors.

Of course, despite all the talk in HPC circles these days of analytics, AI, and other data-intensive applications, the bread and butter of high performance computing is still floating point. Which suggests that for the HPC space, either the Arm community needs to start forging tighter alliances with accelerator providers like NVIDIA, AMD, and Xilinx, or find ways to spread its home-grown acceleration technology, such as the ARMv8-A SVE (Scalable Vector Extension) architecture. The SVE design is getting its first test in Japan, where Fujitsu is developing an implementation for RIKEN's Post-K supercomputer. That system is expected to debut in 2021.

Along similar lines, McIntosh-Smith also points to what he believes is Arm's biggest advantage, namely that the licensing of its shrink-wrapped IP enables people to build customized processors at lower cost than would be possible with a commodity chip business model. The implication is that some enterprising startup or startups could construct specialized Arm processors for the HPC market, incorporating custom circuitry for vector processing, AI, or other interesting types of acceleration.

“These processors will be highly differentiated from high-volume mainstream datacenter parts, and should bring significant steps forward in performance for scientists around the world who have become increasingly frustrated with the relatively small improvements in performance and performance per dollar we've seen in recent years,” writes McIntosh-Smith. “As such, Arm's entry into the HPC market, and the injection of new ideas, innovation and competition this brings, could trigger a revolution in scientific computing of the kind not seen since the commodity CPU revolution of the late 1990's. Exciting times are ahead.”

We'll see.

07/05/18--13:27: **European Program to Develop Supercomputing Chips Begins to Take Shape**



The European Processor Initiative (EPI), an ambitious program to develop a pair of chips for domestic supercomputers, is poised to change the way Europe does HPC. And although the work is still very much in its early stages, it looks like the Europeans have selected their preferred processor architectures: Arm and RISC-V.



[Launched in March 2018](#) by the European Commission, the EPI's overall aim is to develop domestically produced low-power microprocessors for the European market. Even though the work is focused on delivering chips for HPC, and in particular for exascale supercomputers, the technology will also be applied to the broader datacenter market, as well as the automotive industry. The rationale for this more expansive strategy is provide higher volume markets that can economically sustain the considerable effort involved in chip R&D and support..

The first generation of these HPC processors are expected to be delivered toward the end of the decade, in time to form the basis for pre-exascale supercomputers scheduled to be deployed across the EU in the 2020 to 2021 timeframe. The second-generation chips will power Europe's first exascale systems in 2023 and 2024. The system work is being led by EuroHPC, a group formed to bring Europe on par with the US, China, and Japan in high performance computing technology. Part of the mission involves developing home-grown componentry so that EU members have more control over what goes into their supercomputers.

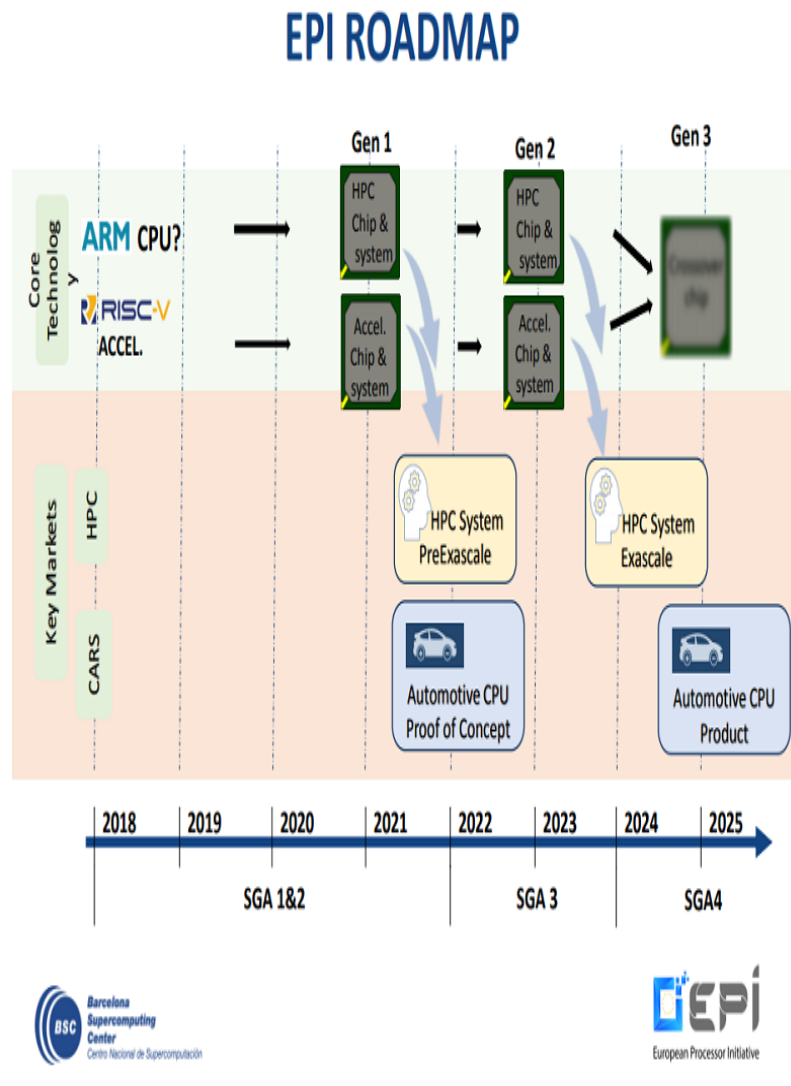
The most central element of these systems is the processor, which puts EPI in the critical path for the EuroHPC work. In a statement delivered at the launch of EPI in March, Vice-President Andrus Ansip, who heads the Digital Single Market, and Mariya Gabriel, the Commissioner for Digital Economy and Society, summed up the strategy as follows:

"The European Processor Initiative is an important step of a strategic plan to develop an independent and innovative European supercomputing and data ecosystem and will ensure that the key competence of high-end chip design remains in Europe, a critical point for many application areas. Thanks to such new European technologies, European scientists and industry will be able to access exceptional levels of energy-efficient computing performance. This will



benefit Europe's scientific leadership, industrial competitiveness, engineering skills and know-how and the society as whole.”

The program will develop two types of processors: one for general-purpose HPC processing, the other for throughput acceleration. By the looks of it, an Arm design will be the basis for the former, while RISC-V will provide the architecture for the latter. A [presentation](#) delivered by Barcelona Supercomputing Centre (BSC) director Mateo Valero in May, reflected some of the current thinking on this strategy, including the roadmap for chip development, as illustrated below.



Source: [RISC-V Workshop presentation from Prof. Mateo Valero](#), May 9, 2018,

The choice of Arm certainly makes sense, given its European roots in the UK, and its commercial licensing scheme. Anyone can buy an Arm license and develop a processor of their own design, something not possible with Intel,

AMD, or NVIDIA technology. In some ways, RISC-V is even more attractive, since it's offered as an open standard architecture that can be had at no cost under a BSD license, either for developing free implementations or proprietary designs.

It's worth noting that the selection of Arm and RISC-V has not been officially announced. However, given inclination of some of the major European players — the long history of the Mont-Blanc exascale project with Arm ([including its latest project to build an Arm-based SoC for exascale machines](#)), BSC's enthusiasm for RISC-V ([it recently hosted the RISC-V Workshop](#)), and [Atos's interest in both architectures](#) — it's hard to fathom any other choice. Of course, OpenPower and even MIPS are possibilities, but neither one has been the focus of any European HPC research. Coming up with a completely new processor architecture is the least likely option, given the timeframes for the pre-exascale and exascale deployments.

There's a good chance the EPI Arm implementation for HPC will be based on the SVE (Scalable Vector Extensions) variant of the architecture, which is the same one that Fujitsu is using to develop the processor that will power RIKEN's Post-K exascale supercomputer. If some of the Japanese work, especially the system software and tools development, could be reused for the EPI project, a lot of effort could be saved.

The development of the RISC-V accelerator is going to entail a good deal more work, if only for fact that there are no examples of high-end designs to draw upon. RISC-V is of recent vintage, having been introduced at the University of California, Berkeley in 2010. Its vector capability is somewhat lacking for an HPC architecture, although at the previously mentioned RISC-V workshop, a 128-bit vector extension was discussed. The fact that RISC-V is being used as high-throughput accelerator could simplify the design effort to some extent since there it wouldn't have to incorporate all the control flow logic expected in a general-purpose processor.

Up until this year, the development of these domestic chips could be considered something of an exercise in isolationism or perhaps even vanity. The multinational nature of the semiconductor industry doesn't limit European access to the latest technology from US or anywhere else. But with a trade war now being instigated by the current administration in Washington, even computer chips could soon run into tariffs on their journey through the global supply chain. If that turns out to be the case, the EU decision to develop a home-grown chip capability would have been prescient.

The EU has initially invested €120 million toward the EPI program, but the 23 industry and research partners are expected to kick in a certain amount as well. The key players include Atos, BSC, CEA, Jülich Supercomputing Centre, and STMicroelectronics, to name a few. BSC has taken the lead for the accelerator work, while Atos has taken on the role of system/chip integrator for the general-purpose processor. Work on both designs was supposed to begin mid-year, so we can assume that the development effort is now underway.

📅 **07/16/18--09:18: [A Tightening Exascale Race Reveals Underlying Forces Shaping Supercomputing](#)**



The competition between the US, China, and Japan to field the first exascale supercomputer looks a lot closer than it did a couple of years ago. But the real significance of the narrowing schedules reflects a shift in technology preferences and a trend toward domestic control of HPC hardware.



The three latest developments in the exascale sweepstakes that point to a tightening race include Fujitsu's production of the prototype of the ARM processor that will power Japan's Post-K supercomputer, the recapture of the number one spot on the TOP500 list by the US, and the report that China's exascale effort could be delayed by up to a year.

The latter development, [reported in MIT Technology Review](#), could move the deployment of China's initial exascale supercomputer into 2021, which happens to be the same year that Japan and the US are planning power up their first machines. The article quotes Depei Qian, a Beihang University professor who helps direct China's exascale effort and who admitted: "I don't know if we can still make it by the end of 2020. There may be a year or half a year's delay."

The problem apparently stems from the fact the Chinese are having trouble deciding between the different types of systems under consideration. Three different exascale processor architectures are being pursued in parallel, all of which rely on domestic development: one based on some version of the ShenWei chip, another using a Chinese-designed Arm CPU, and the third using licensed x86 technology. According to Qian, the evaluation process for these approaches has dragged out and a call for proposals to build the exascale systems has been pushed back.

Of the three approaches, the only one the Chinese have any experience with in the HPC realm is their native ShenWei processor, which is currently being used to power the nation's most powerful supercomputer, the Sunway TaihuLight. Because ShenWei is a non-standard processor architecture, it relies exclusively on custom software tools for application development. Any effort to establish the processor as a more widely accepted architecture for servers, HPC or otherwise, would probably take a decade or more.

The indigenous Arm and x86 efforts in China don't have that problem, but as far as we know, these approaches have yet to produce a working prototype. As we [reported](#) earlier this month, Chinese chipmaker Hygon just recently began manufacturing Zen-based x86 CPUs based on a licensing agreement with AMD. That chip would almost certainly need to be deployed in tandem with some sort of accelerator to deliver an exascale-capable machine based a reasonable number of servers. The Chinese have such an accelerator in the Matrix general-purpose DSP, but the current Matrix-2000 implementation delivers only about three teraflops per chip. The Arm effort, which appears to be based on [a Phytium Technology design](#), would likewise require an accelerator coprocessor to achieve a practical exascale system, unless a much more performant version that incorporated Arm's Scalable Vector Extension (SVE) was developed .

The dilemma of having too many choices may indeed be delaying China's exascale plans, but the more obvious explanation is that it's time-consuming to develop new processors, not to mention systems based on those processors. That's true even if the architects involved have some experience with the underlying technology. And when those systems have to operate at the exascale level, those challenges are magnified by the additional demands of energy efficiency, scalability, and reliability.

Japan ran up against such challenges early on. In 2016, Fujitsu and RIKEN had committed to developing an Arm SVE-powered supercomputer for the country's first exascale system, known as Post-K. The original schedule had RIKEN installing the system in 2020. But a few months after the plans were announced, Dr. Yutaka Ishikawa, who was the project lead at the time, admitted [the Post-K deployment could be delayed by as much as two years](#). Last month, [Fujitsu revealed it had built a prototype of the Arm chip](#), which is now being tested and benchmarked. Currently, Post-K appears to be track for a 2021 deployment.

At this point, the US appears to be closest to reaching the exascale milestone, inasmuch as IBM's newly-deployed Summit system at Oak Ridge National Lab is, from a Linpack perspective, is within 12 percent of that goal. That's a bit of mirage though. America's first exascale supercomputer will be Aurora, an Intel-based system [whose architecture and even processor design are still largely unknown](#). If this was just a matter of developing a manycore Xeon processor with enough computational horsepower (at least 20 teraflops per chip) to power Aurora's 50,000 nodes and integrating the company's second-generation Omni-Path fabric, silicon photonics, and Optane NVDIMMs, that all seems pretty doable. But considering Intel's latest problem in moving to its 10nm process node, the chipmaker's curious abandonment of its Xeon Phi roadmap, and the chipmaker's general lack of specifics regarding its HPC plans, this is no slam dunk.

Even though Summit's Power/GPU hybrid design will not be the architecture for the first exascale system in the US, it will almost certainly be the model for

subsequent machines. In one respect, it is the most mature architecture for these future supercomputers, since it will be based on processors and other componentry with long-established product roadmaps, namely IBM Power processors, NVIDIA Tesla GPUs, and Mellanox InfiniBand.

The EU countries have conceded that their first exascale supercomputer will be a couple of years behind the initial systems deployed in the US, China, and Japan. Like their counterparts, the Europeans are also developing domestic processors for these next-generation machines, in this case [based on Arm and RISC-V](#). The driving effort for this work, known as the European Processor Initiative (EPI), is part of a larger push to develop an indigenous HPC capability on the continent and free itself from its dependency on the North American chipmakers, specifically, Intel, AMD, NVIDIA, and IBM. EPI recently got underway and is supposed to produce pre-exascale versions of these chips by the end of the decade.

An additional complication in all these efforts is the realization that machine learning is emerging as a new application requirement for HPC work. This is forcing all the players to demand additional mixed precision support in the processor or coprocessor and ensure memory capacity and performance will be adequate for these workloads. Large-scale machine learning also places extra requirements on the system interconnect. It's doubtful if any exascale supercomputers will be built without these additional capabilities.

The move toward greater processor diversity and more specialization in high performance computing is a good thing, and probably necessary considering the slowdown in Moore's Law. The emerging importance of machine learning is also a welcome development since it promises to advance the breadth and usefulness of HPC applications. But these developments come at a time when the top tier of supercomputing users seem overly focused on attaining the artificial milestone of exascale and national policies are favoring domestic designs. Would we really be building the most powerful systems in the world with such new technology were it not for the constraints imposed by exascale computing and the political winds of economic nationalism?

Consider that the Arm architecture features more prominently in current exascale plans than either x86 or Power, but has no representation on the TOP500 list ([although it will soon](#)). Even if these systems arrive on schedule, there are bound to be growing pains due to immature software tools, application porting challenges, and a general lack of support for the technology in the datacenter. A broader market for such systems is far from assured.

One plausible scenario is that these first supercomputers will be one-off machines – not stunt systems, per se, but not widely adopted for general use. Depending upon what Intel, AMD, and NVIDIA do, the more traditional combo of an x86 CPU married to a GPU accelerator could turn out to be the most practical architecture for a good chunk of exascale systems and most of the non-exascale systems.

Regardless of how this all shakes out, the next decade of supercomputing is bound to be an interesting time for everyone involved.

**07/20/18--10:16: AMD May Be About to Beat Intel at Its Own Game**

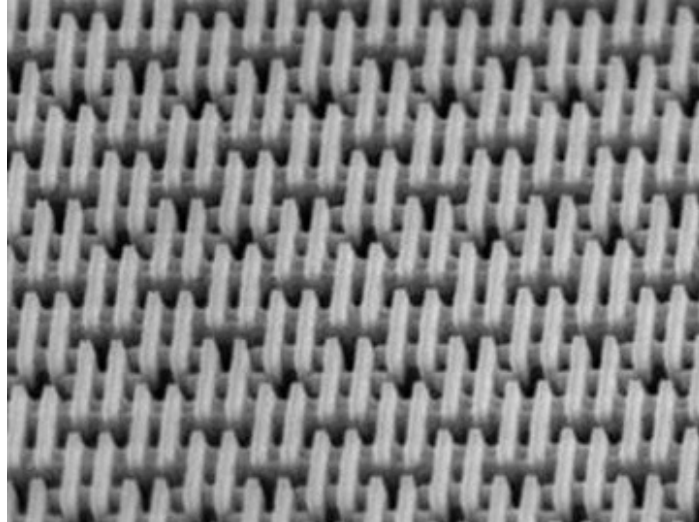
0



0



The upcoming battle between AMD's "Rome" server CPU and Intel's "Ice Lake" Xeon scalable processor promises to be an interesting matchup. But this time around, AMD could have an advantage it has never had before.



Rome, which will be based on AMD's second-generation Zen architecture (Zen 2), is scheduled to start sampling in the second half of 2018, with full production in 2019. It will be manufactured by either TSMC or GlobalFoundries, or perhaps both, using their respective 7nm process technologies. According to the [latest rumors](#), the chip will deliver a 10 to 15 percent improvement in instructions per clock (IPC) compared to the first-generation EPYC processor and will be equipped with up to 64 cores. That's a doubling of the core count of the EPYC CPUs, which are being manufactured by GlobalFoundries on its 14nm process.

Very little is known about Ice Lake, Intel's next-generation Xeon microarchitecture that will purportedly be built on the company's 10nm semiconductor process. The Power Stamp Alliance has provided a few details revealing the product will include up to eight memory channels and draw as much as 230 watts of power. That suggests a higher count than the current "Skylake" Xeon-SP processors, which top out at 28 cores. It could also mean special IP blocks – graphics processing, extra-wide vector units, integrated FPGAs, etc. – will be offered as part of the package.

What is better known is that Intel's 10nm technology has suffered multiple delays and is currently not expected to be ready for mass production until sometime in 2019. Originally, the chipmaker was planning to deliver its "Cascade Lake" Xeon-SP on 10nm as a process shrink of "Skylake," but the delays forced them to move this product to their latest 14nm node (14nm++). The Cascade Lake Xeon-SP is slated for release later this year.

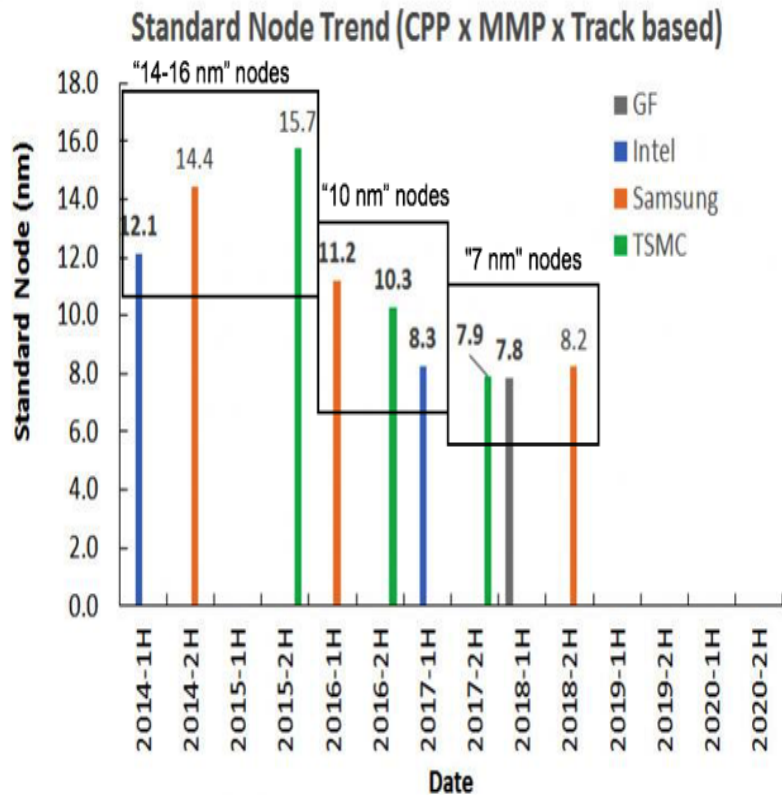
As a consequence, Ice Lake's debut on 10nm will be one of those rare instances where Intel will advance the microarchitecture and process node of a Xeon product in tandem. The state of affairs was summed up by Intel CEO

(and now ex-CEO) Brian Krzanich, during the first quarter earnings report back in April:

*"We continue to make progress on our 10-nanometer process. We are shipping in low volume and yields are improving, though the rate of improvement is slower than we anticipated. As a result, volume production is moving from the second half of 2018 into 2019. We understand the yield issues and have defined improvements for them, but they will take time to implement and qualify. We have leadership products on the roadmap that continue to take advantage of 14-nanometer, with Whiskey Lake for clients and Cascade Lake for the data center coming later this year."*

Anyone who has been following the semiconductor tech space knows that transistor feature sizes advertised by different manufacturers are not comparable. In other words, a 10nm Intel process is not the same size as a 10nm process from TSMC. That's mostly because there are different ways to measure transistors geometries in the era of 3D FinFET technology. In general, Intel is more conservative with its measurements, such that its nodes correspond to smaller nodes from Samsung, TSMC, and GlobalFoundries.

So how will Intel's 10nm technology match up against the 7nm process nodes from TSMC and GlobalFoundries? If you can believe Scotten Jones' [analysis](#) at SemiWiki.com, the three are pretty much on par, although the 7nm nodes from both TSMC and GlobalFoundries have a slight edge when compared Intel's 10nm node. Jones uses a "standard node" designation that normalizes feature sizes across manufacturers based on transistor density. His overall conclusion is that "Intel has lost their multiyear density lead over the foundries." In a [writeup](#) by Mark Hibben at Seeking Alpha, he overlays the process nodes on top of Jones's data to offer a clearer picture of what's going on:



Source: <https://seekingalpha.com/article/4151376-tsmc-intel-lead-semiconductor-processes>

The bottom line is that for the first time, AMD server chips will likely be manufactured with better transistor densities than competing silicon from Intel. As we've [reported](#) previously, the EPYC CPUs already have a number of architectural advantages over the Skylake Xeon-SP products. A process node advantage could enable the Rome processors to be more competitive in basic areas like clock speed and energy efficiency.

That said, the actual differences in transistor density between Intel's 10nm node and the 7nm nodes from TSMC and GlobalFoundries appear to be relatively small. Plus, architectural design tends to be much more important than transistor size, especially when the differences are not all that significant. On the other hand, even reaching parity with Intel in process technology is a huge accomplishment for AMD.

There is, however, one other development in AMD's favor. The Rome processors seem to be pretty much on schedule for a 2019 release date. In June, AMD CEO Lisa Su [held up a same chip](#) at the Computex computer tradeshow in Taipei. Meanwhile, Intel hasn't offered much in way of a timeline for the Ice Lake rollout. If the company releases the 14nm Cascade Lake Xeon-SP later this year, there would be little incentive to follow it with a new CPU that would make its predecessor immediately obsolete. Plus, Intel's 10nm problems may prevent good yields of the chips until much later in 2019. The gang at Wccftech are reporting that the [Ice Lake-SP processors aren't expected to arrive until 2020](#), which would give Rome a year head-start over its targeted competition.



Either way, AMD prospects for growing its datacenter business is probably the best it's been since the early days of the Opteron CPU. And while Intel has plenty of experience in leap-frogging its competition, in the short-term, it may not have very many options to exercise. As always, we'll have to wait and see how it plays out.