



MENU



US

**MUST READ:** Ransomware attacks: Why and when it makes sense to pay the ransom

## AI is changing the entire nature of compute

Machine learning, especially deep learning, is forcing a re-evaluation of how chips and systems are designed that will change the direction of the industry for decades to come.



By [Tiernan Ray](#) | June 30, 2019 -- 13:26 GMT (06:26 PDT) | Topic: [Artificial Intelligence](#)

The world of computing, from chips to software to systems, is going to change dramatically in coming years as a result of the spread of machine learning. We may still refer to these computers as "Universal Turing Machines," as we have for eighty years or more. But in practice they will be different from the way they have been built and used up to now.

Such a change is of interest both to anyone who cares about what computers do, and to anyone who's interested in machine learning in all its forms.

In February, Facebook's head of A.I. research, Yann LeCun, gave a talk at the International Solid State Circuits Conference in San Francisco, one of the longest running computer chip conferences in the world. At ISSCC, LeCun made plain the importance of computer technology to A.I. research.

"Hardware capabilities and software tools both motivate and limit the type of ideas that AI researchers will imagine and will allow themselves to pursue," said LeCun. "The tools at our disposal fashion our thoughts more than we care to admit."

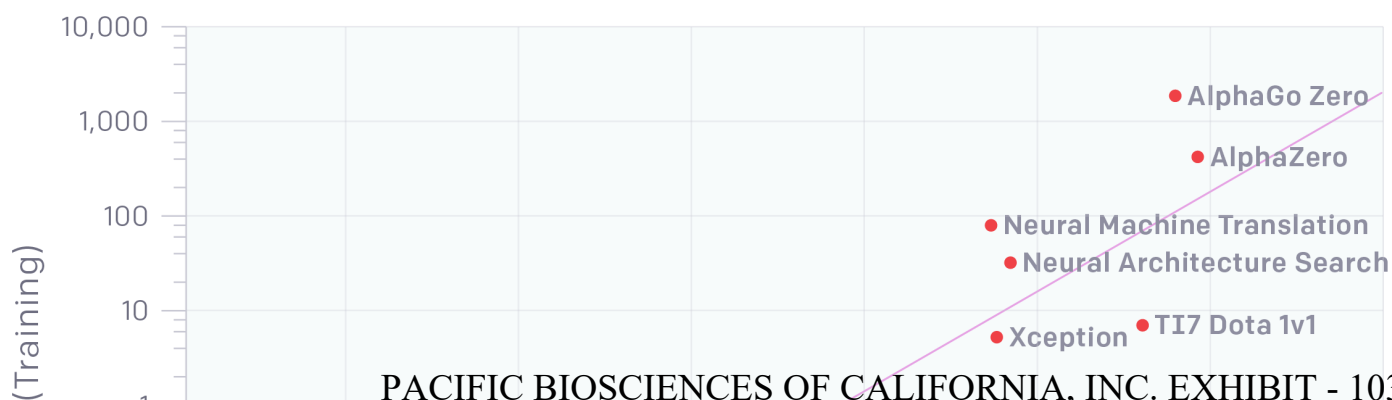
It's not hard to see how that's already been the case. The rise of deep learning, starting in 2006, came about not only because of tons of data, and new techniques in machine learning, such as "dropout," but also because of greater and greater compute power. In particular, the increasing use of graphics processing units, or "GPUs," from Nvidia, led to greater parallelization of compute. That made possible training of vastly larger networks than in past. The premise offered in the 1980s of "parallel distributed processing," where nodes of an artificial network are trained simultaneously, finally became a reality.

Machine learning is now poised to take over the majority of the world's computing activity, some believe. During that ISSCC in February, LeCun [spoke to ZDNet about the shifting landscape of computing](#) (<http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/facebook-yann-lecun-says-internal-activity-proceeds-on-ai-chips/>). Said LeCun, "If you go five, ten years into the future, and you look at what do computers spend their time doing, mostly, I think they will be doing things like deep learning — in terms of the amount of computation." Deep learning may not make up the bulk of computer sales by revenue, LeCun added, but, "in terms of how are we spending our milliwatts or our operations per second, they will be spent on neural nets."

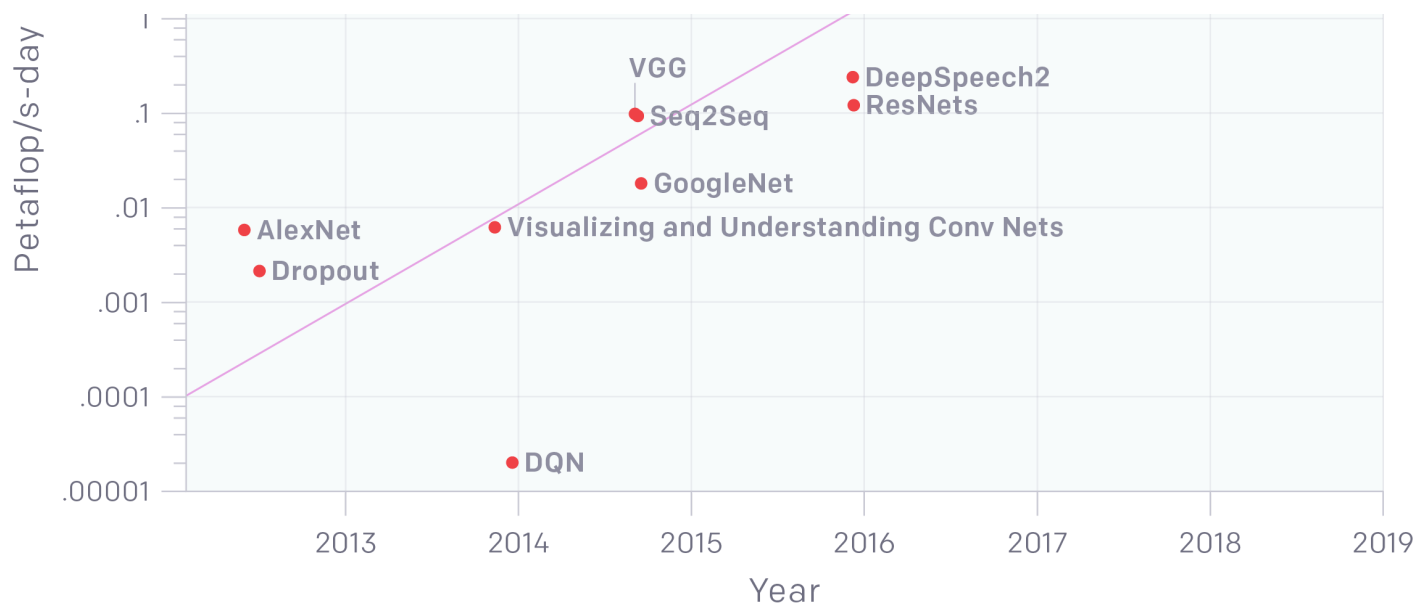
### Deep learning grows exponentially

As deep learning becomes the focus of computing, it is pushing at the boundaries of what today's computers can do, to some extent in the "inference task," where neural nets make predictions, but much more so for training a neural net, the more compute-intensive function.

## AlexNet to AlphaGo Zero: A 300,000x Increase in Compute



PACIFIC BIOSCIENCES OF CALIFORNIA, INC. EXHIBIT - 1031



According to OpenAI, the demand for compute by deep learning networks has been doubling every 3.5 months since 2012.

OpenAI

Modern neural networks such as OpenAI's GPT-2 are over a billion parameters, or network weights, that need to be trained in parallel. As Facebook's product manager for PyTorch, the popular machine learning training library, [told ZDNet in May](http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/facebooks-pytorch-1-1-does-the-heavy-lifting-for-increasingly-gigantic-neural-networks/) (<http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/facebooks-pytorch-1-1-does-the-heavy-lifting-for-increasingly-gigantic-neural-networks/>), "Models keep getting bigger and bigger, they are really, really big, and really expensive to train." The biggest models these days often cannot be stored entirely in the memory circuits that accompany a GPU.

**Also:** [Google says 'exponential' growth of AI is changing nature of compute](http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/google-says-exponential-growth-of-ai-is-changing-nature-of-compute/) (<http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/google-says-exponential-growth-of-ai-is-changing-nature-of-compute/>)

And the pace of demand for compute cycles is increasingly sharply. According to [data from OpenAI](http://web.archive.org/web/20190630155704/https://openai.com/blog/ai-and-compute/) (<http://web.archive.org/web/20190630155704/https://openai.com/blog/ai-and-compute/>), the venerable AlexNet image recognition system, created way back in 2012, consumed the equivalent of one thousand trillion floating point operations per second, a "petaflop," during total training time that amounted to a fraction of a day. But AlphaZero, the neural net built by Google's DeepMind in 2016 to beat the world champions of chess, go, and shogi, consumed more than one thousand days' worth of petaflops per second. That increase in compute cycles between AlexNet and AlphaZero constitutes a doubling of compute consumption every 3.5 months. And that was data gathered back in 2016. The pace will doubtless have increased by now.

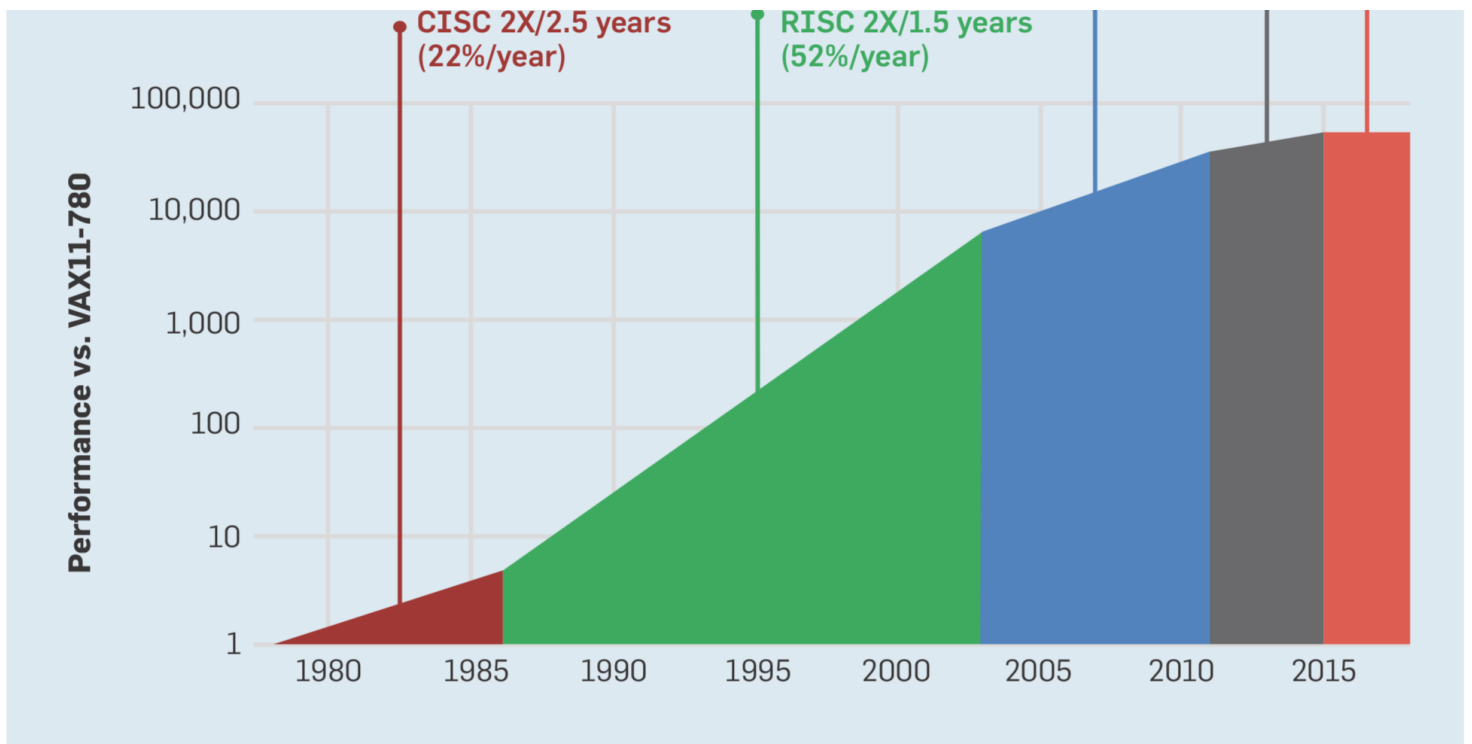
### A crisis in computer chips

The world doesn't even have petaflop chips to run on. A top-of-the line chip for deep learning training, such as Nvidia's Tesla V100, runs at 112 trillion operations per second. So, you would have to run eight of them for 1,000 days or else cluster many together into systems that expend more and more energy.

Worse, the pace of chip improvement in recent years has hit a wall. As UC Berkeley professor David Patterson and Alphabet chairman John Hennessy pointed out in an article earlier this year, Moore's Law, the rule of thumb that says chips double in power every twelve to eighteen months, has run out of gas. Intel has long denied the point, but the data is on the side of Patterson and Hennessy. As they mention in the report, chip performance is now only increasing by a measly 3% per year.

**Figure 6. Growth of computer performance using integer programs (SPECintCPU).**





Computer scientists David Patterson and John Hennessy have kept track of data on new chips that show the entire field is forming an asymptote, with the latest chips garnering no more than a 3% performance increase per year.

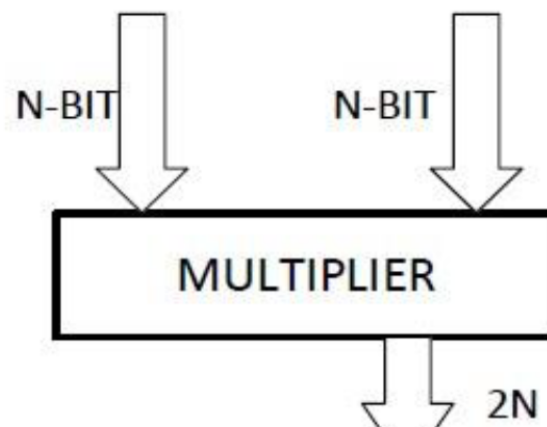
Association for Computer Machinery/John L. Hennessy, David A. Patterson

What that means, both authors believe, is that the design of chips, their architecture, as it's known, has to change drastically in order to get more performance out of transistors that are not of themselves producing performance benefits. (Patterson helped Google to create its "Tensor Processing Unit" chip, so he knows quite a bit about how hardware can affect machine learning, and vice versa.)

With processor improvement stalling, but machine learning demand doubling every few months, something's got to give. Happily, machine learning itself can be a boon for chip design, if looked at the right way. Because machine learning requires very little support for legacy code — it doesn't have to run Excel or Word or Oracle DB — and because of the highly repetitive nature of its most basic computations, machine learning is a kind of greenfield opportunity, as they say, for chip designers.

### Building a new machine

At the heart of convolutional neural networks, and long short-term memory networks, two of the mainstays of deep learning, and even in the more modern networks such as Google's Transformer, the majority of the computations are linear algebra computations known as tensor math. Most commonly, some input data is turned into a vector, and then that vector is multiplied by the columns of a matrix of neural network weights, and the products of all those multiplications are added together. Known as multiply-adds, these computations are rendered in the computer using what are called "multiply-accumulate" circuits, or "MACs." Thus, one can immediately improve machine learning just by improving the MAC and creating many more of them on a chip to increase parallelization.



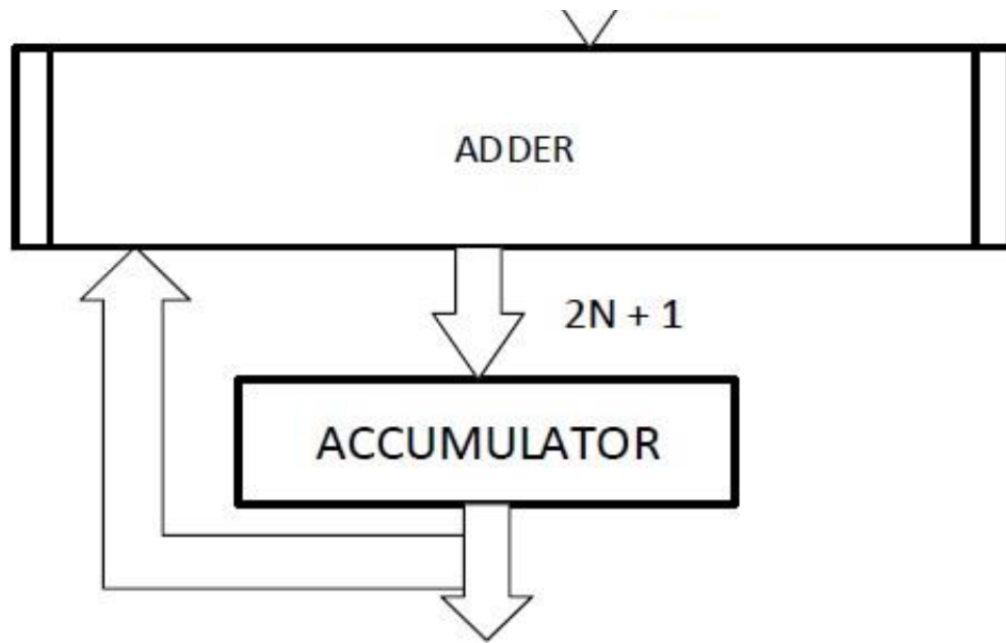


**Figure 13. Multiply and accumulate unit (MAC) basic structure**

*The multiply-accumulate circuit, or "MAC," one of the fundamental circuits involved in deep learning, from "A High Speed and Low Power 8 bit x 8 bit Multiplier Design Using Novel Two Transistor XOR Gates" ([http://web.archive.org/web/20190630155704/https://www.researchgate.net/publication/270898651\\_A\\_High\\_Speed\\_and\\_Low\\_Power\\_8\\_Bit\\_x\\_8\\_Bit\\_Multiplier\\_Design\\_using\\_Novel\\_Two\\_Transistor\\_2T\\_XOR\\_Gates](http://web.archive.org/web/20190630155704/https://www.researchgate.net/publication/270898651_A_High_Speed_and_Low_Power_8_Bit_x_8_Bit_Multiplier_Design_using_Novel_Two_Transistor_2T_XOR_Gates)), 2015, by Himani Upadhyay and Shubhajit Roy Chowdhury.*

*Himani Upadhyay and Shubhajit Roy Chowdhury*

Both Nvidia, which dominates A.I. training, and Intel, whose CPUs dominate machine learning inference, have tried to adapt their products to take advantage of those atomic linear algebra functions. Nvidia has added "tensor cores" to its Tesla GPUs, to optimize the matrix multiplications. Intel has spent \$30 billion buying companies that do stuff in machine learning, including Mobileye, Movidius, and Nervana Systems, the last of which is supposed to lead to a "Nervana Neural Network Processor" at some point, though there have been delays.

So far, these moves are not satisfying people in machine learning, such as Facebook's LeCun. During his chat with ZDNet in February, LeCun opined, "What we need are competitors, to the, you know, dominant supplier at the moment [Nvidia]." That's not because, he said, Nvidia don't make good chips, which they do. It's "because they make assumptions," he continued, "And it'd be nice to have a different set of hardware that makes different assumptions that can be used for complementary things that the current crop of GPUS are good at."

Also: [Why is AI reporting so bad?](http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/why-is-a-i-reporting-so-bad/) (<http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/why-is-a-i-reporting-so-bad/>)

One of those assumptions that are faulty, he said, is the assumption that training a neural network will be a matter of "a neat array" that can be operated on. Instead, future networks will probably make use of a lot of network graphs, where elements of the compute graph of a neural network are streamed to the processor as pointers. Chips have to do plenty of multiply-adds, LeCun said, but with different expectations of how those multiply-adds will be presented to the processor.

Cliff Young, a Google software engineer who was one of the contributors to the TPU chip, put matters more bluntly [when he gave a keynote last October](http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/google-says-exponential-growth-of-ai-is-changing-nature-of-compute/) (<http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/google-says-exponential-growth-of-ai-is-changing-nature-of-compute/>) at a chip event in Silicon Valley. "For a very long time, we held back and said Intel and Nvidia are really great at building high-performance systems," said Young. "We crossed that threshold five years ago."

### Rise of the startups

Into the breach, new chips are arriving from both the A.I. titans themselves, such as Google, but also a raft of venture-backed startups.

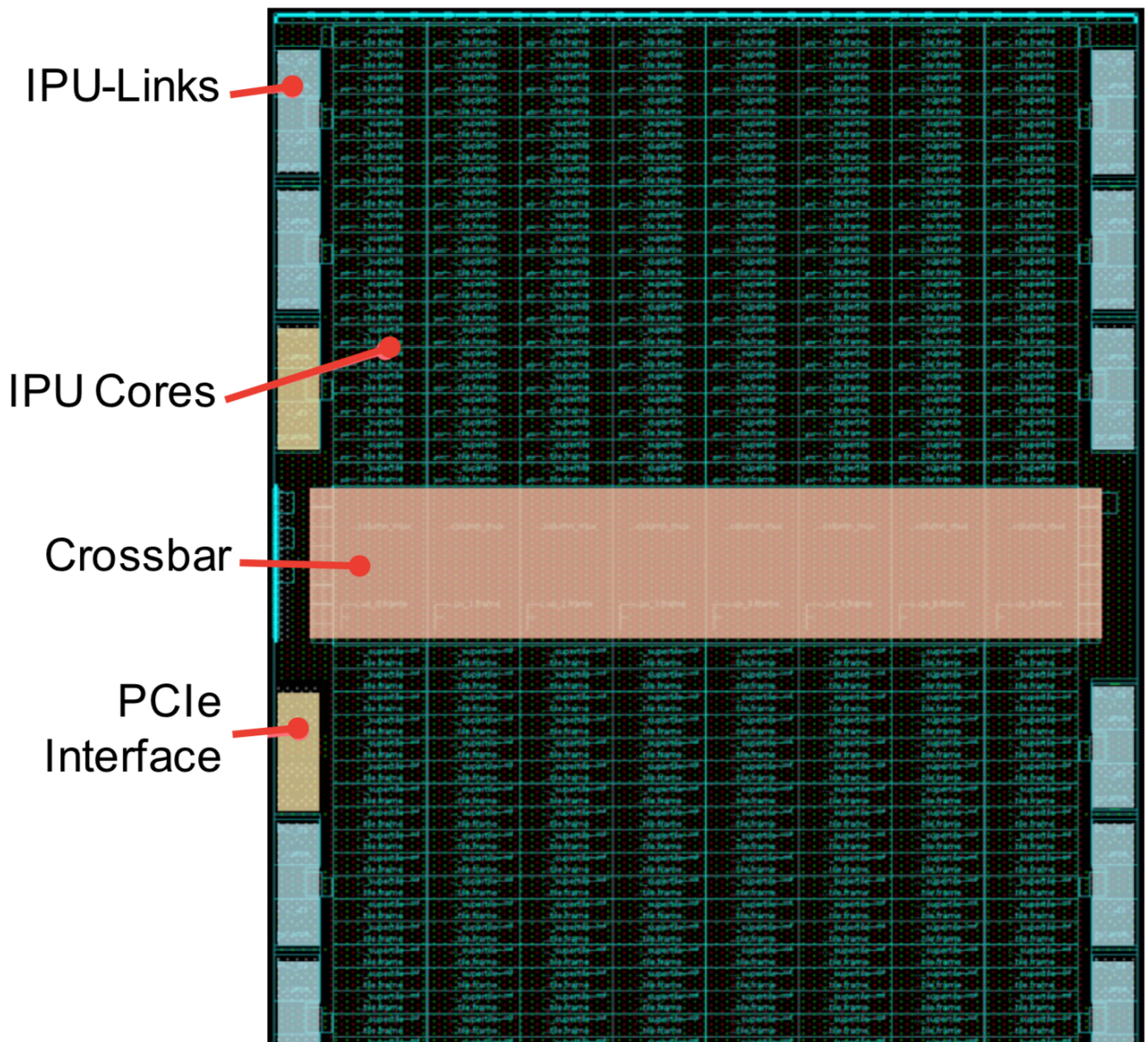
In addition to Google's TPU, now on its third iteration, Microsoft has a programmable processor, an "FPGA," called Project Brainwave, which customers can rent through its Azure cloud service. Amazon has said it will have its own custom chip later this year, called "Inferentia." When LeCun talked to

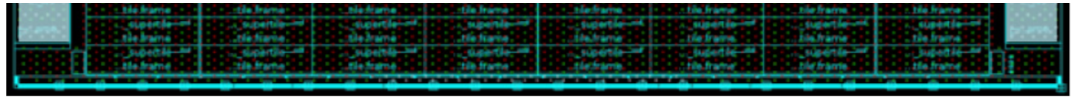
ZDNet in February, he mentioned that Facebook has its own chip efforts.

"Certainly it makes sense for companies like Google and Facebook who have high volume to, you know, work on their own engines," said LeCun. "There is internal activity on this."

Startups include companies such as Graphcore, a five-year-old startup in Bristol, a port city an hour and a half southwest of London; [Cornami](http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/ai-startup-cornami-reveals-details-of-neural-net-chip/) (<http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/ai-startup-cornami-reveals-details-of-neural-net-chip/>), [Efinix](http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/chip-startup-efinix-hopes-to-bootstrap-ai-efforts-in-iot/) (<http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/chip-startup-efinix-hopes-to-bootstrap-ai-efforts-in-iot/>), and [Flex Logix](http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/ai-startup-flex-logix-targets-nvidias-market-for-inference-chips/) (<http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/ai-startup-flex-logix-targets-nvidias-market-for-inference-chips/>), all of which have been profiled by ZDNet; and Cerebras Systems of Los Altos, in Silicon Valley, a company still in stealth mode.

There's a common thread with many of these startups, which is to greatly increase the amount of the area of a computer chip devoted to matrix multiplications, the MAC units, to squeeze the most parallelization out of each clock cycle. Graphcore is the farthest along of any of the startups, being the first to actually ship production chips to customers. One of the things that most stands out about its first chip is the huge amount of memory. Colossus, as the chip is called (<http://web.archive.org/web/20190630155704/https://www.graphcore.ai/posts/colossus-poplar-and-tommy-flowers>), in honor of the world's first digital computer, is gigantic, measuring 806 square millimeters. Chief technology officer Simon Knowles boasts it is "the most complex processor chip that's ever been built."





Graphcore's "Colossus" chip, named for the first digital computer, has over one thousand identical vector processor cores that let it achieve high parallelism, aided by an unprecedented 304 megabytes of on-chip SRAM. At 806 square millimeters, it is one of the largest chips ever made.

Graphcore

The Colossus is made up of 1,024 individual cores dubbed "intelligence processing units," each of which can independently process matrix math. And each IPU, as they're known, has its own dedicated memory, 256 kilobytes of fast SRAM memory. In total, the 304 megabytes of memory is the most ever built into the chip.

No one knows how the presence of so much memory on chip will alter the kinds of neural networks that are built. It might be that with access to increasing amounts of memory, with very low-latency access, more neural networks will focus on reusing values stored in memory in new and interesting ways.

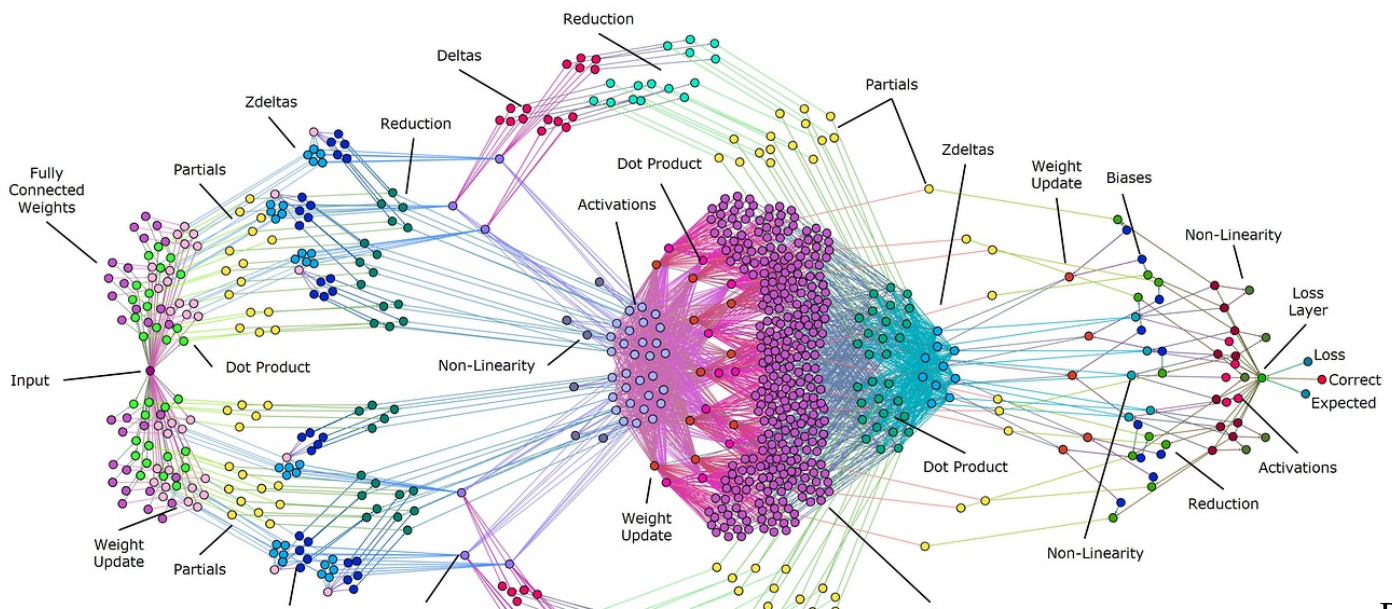
- [What is AI? Everything you need to know](http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/what-is-ai-everything-you-need-to-know-about-artificial-intelligence/) (<http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/what-is-ai-everything-you-need-to-know-about-artificial-intelligence/>)
- [What is deep learning? Everything you need to know](http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/what-is-deep-learning-everything-you-need-to-know/) (<http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/what-is-deep-learning-everything-you-need-to-know/>)
- [What is machine learning? Everything you need to know](http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/what-is-machine-learning-everything-you-need-to-know/) (<http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/what-is-machine-learning-everything-you-need-to-know/>)
- [What is cloud computing? Everything you need to know](http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/what-is-cloud-computing-everything-you-need-to-know-from-public-and-private-cloud-to-software-as-a/) (<http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/what-is-cloud-computing-everything-you-need-to-know-from-public-and-private-cloud-to-software-as-a/>)

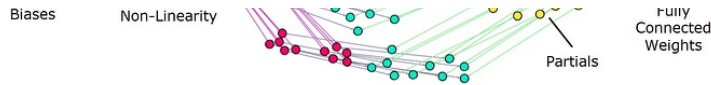
## The software conundrum

For all these chip efforts, the problem, of course, is that they don't have the years of software built up for Nvidia thanks to the company's "CUDA" programming technology. The answer by Graphcore and others will be two-fold. One is that the various programming frameworks for machine learning, such as TensorFlow and Pytorch, provide a way to avoid the details of the chip itself and focus on the program structure. All the chips coming to market support these frameworks, which their creators believes levels the playing field with Nvidia.

The second point is that Graphcore and others are building their own programming technologies. They can make the case that their proprietary software both translates the frameworks, but also intelligently assigns parallel computations to the numerous MAC units and vector units on a chip. That's the argument Graphcore makes for its "Poplar" software. Poplar breaks up the compute graph of a neural network into "codelets" and distributes each codelet to a different core of Colossus to optimize parallel processing.

### Poplar Graph : 2 Layer Fully Connected MNIST Training





Graphcore®

Graphcore's "Poplar" compiler takes a neural network and distributes its various functional elements efficiently throughout the Colossus processor as independent "codelets."

Graphcore

In the past twenty years, big data and fast parallel computation became the norm and propelled machine learning, bringing about deep learning. The next wave of computer hardware and software will probably be about vast amounts of memory and neural networks that are dynamically constricted to take advantage of highly parallel chip architectures. The future looks very interesting.

**The tech that changed us: 50 years of br...** (</web/20190630155704/https://www.zdnet.com/pictures/1968-2018-50-years-of-technology-breakthroughs/>)

**SEE FULL GALLERY** (</web/20190630155704/https://www.zdnet.com/pictures/1968-2018-50-years-of-technology-breakthroughs/>)



(</web/20190630155704/https://www.zdnet.com/pictures/1968-2018-50-years-of-technology-breakthroughs/>)

1 - 5 of 52

**NEXT** 0

## ARTIFICIAL INTELLIGENCE

**Use-case selection will make or break your AI strategy in healthcare** (<http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/use-case-selection-will-make-or-break-your-ai-strategy-in-healthcare/>)

**Salesforce adds AI bias module to Trailhead** (<http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/salesforce-adds-ai-bias-module-to-trailhead/>)

**Using machine learning to solve your dark data nightmare** (<http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/using-machine-learning-to-solve-your-dark-data-nightmare/>)

**How do neural networks see depth?** (<http://web.archive.org/web/20190630155704/https://www.zdnet.com/article/how-do-neural-networks-see-depth/>)

**AI in Healthcare: Saving lives at population scale (CNET)** (<http://web.archive.org/web/20190630155704/https://www.cnet.com/paid-content/news/ai-in-healthcare-saving-lives-at-population-scale/?ftag=CMG-01-10aaa1b>)

**Only 29% of companies regularly use AI (TechRepublic)** (<http://web.archive.org/web/20190630155704/https://www.techrepublic.com/article/only-29-of-companies-regularly-use-ai/?ftag=CMG-01-10aaa1b>)

### RELATED TOPICS:

CLOUD

DIGITAL TRANSFORMATION

CXO

INTERNET OF THINGS

INNOVATION

ENTERPRISE SOFTWARE

SHOW COMMENTS

### MORE RESOURCES

**Tech Pro Research: Using Tech to Make Shopping Easier and More Enjoyable**

Downloads from [TechRepublic.com](http://TechRepublic.com)

DOWNLOAD NOW

**Tech Pro Research: IT Leader's Guide to Deep Learning**

White Papers from [TechRepublic.com](http://TechRepublic.com)

READ NOW

**Tech Pro Research: IT Budget Research Report 2019**

White Papers from [TechRepublic.com](http://TechRepublic.com)



## MICROSOFT STIRS SUSPICION

Windows security-only update contains telemetry files

---

### JUST IN

**US mayors group adopts resolution not to pay any more ransoms to hackers**

7 minutes ago

**Samsung Galaxy Note 10: Latest rumors, features, and release date**

2 hours ago

**You can't ignore GDPR in customer service**

3 hours ago

**Raspberry Pi 4: Hardware first impressions**

3 hours ago

**Corel Painter 2020, hands on: Performance and interface tweaks enhance user experience**

4 hours ago

**Python is eating the world: How one developer's side project became the hottest programming language on the planet**

4 hours ago

**AWS intros new EventBridge, CDK services as CTO Vogels urges customers to 'encrypt everything'**

4 hours ago

**What happened to mass adoption of virtual and augmented reality?**

4 hours ago

---

### TODAY ON ZDNET



## SPECIAL FEATURE

Managing the Multicloud



## **Microsoft defends planned partner program changes, but many aren't buying it**

4 hours ago by Mary Jo Foley in Cloud

## **One issue is probably to blame for your IT outages and data losses - and it's not hackers**

4 hours ago by Steve Ranger in Security

## **Hacked surveillance firm pitches NYC with invasive camera tech to track driver journeys**

4 hours ago by Charlie Osborne in Security

## **Apple disables Walkie-Talkie app due to snooping vulnerability**

5 hours ago by Adrian Kingsley-Hughes in Security

## **Amazon CTO Vogels battles through protesters in AWS keynote as company pledges \$700 million to upskill workers**

5 hours ago by Larry Dignan in Cloud



VIDEO

Huawei ban: Winners, losers, and what's at stake (a whole lot)



## After Trump Twitter ruling, Ocasio-Cortez is sued for blocking critics

5 hours ago by Liam Tung in Tech Industry



## Cisco Live 2019: Updates and news

from Tech Pro Research

## Big-eyed robots in supermarkets are really annoying people

6 hours ago by Chris Matyszczyk in Robotics

## Nvidia expands colocation program around the globe

6 hours ago by Stephanie Condon in Data Centers

## Juniper updates its multi-cloud container platform Juke

6 hours ago by Stephanie Condon in Cloud

## Retailers' main weapon to counter Amazon Prime Day will be email, says Adobe

6 hours ago by Larry Dignan in E-Commerce



### GALLERY

10 super sweet laptops that come with Linux pre-installed

LOAD MORE



Article

Windows 10 update: The complete guide for businesses of every size

What happened to mass adoption of virtual and augmented reality?

UTS says RoboCup is about building robots to improve lives, not winning

Reverse engineering of BlueKeep patch reveals how dangerous it is

Think before you click: Watch out for these malicious Word documents

Amazon Prime Day 2019: The business strategy behind the sale

AI and China's 21st century strategy

Article

Top cloud providers 2019: AWS, Microsoft Azure, Google Cloud; IBM makes hybrid move

**Wyndham Hotels & Resorts tackled technical debt, cloud, hybrid cloud in a hurry [Cloud TV]**

**HSBC charts out its move to the cloud [Cloud TV]**

**How Brinker International thinks through cloud, data, Apple iPads [Cloud TV]**

**Why security is the top barrier in enterprise cloud adoption [Hybrid Cloud TV]**

**How New Belgium Brewing evaluated managed vs. private cloud [Hybrid Cloud TV]**

**With Red Hat, IBM to become the leading hybrid cloud provider**

---

**MORE RESOURCES**

---

Tech Pro Research: IT Leader's Guide to Deep Learning

White Papers from [TechRepublic.com](#)

[READ NOW](#)

Don't Risk It. Higher Education Cybersecurity 101

White Papers from [Cisco](#)

[READ NOW](#)

The Zero Impact Database Migration

White Papers from [Quest Software](#)

[READ NOW](#)

Lears Journey to Digital Procurement Excellence

Webcasts from [Coupa x KPMG](#)

[WATCH NOW](#)