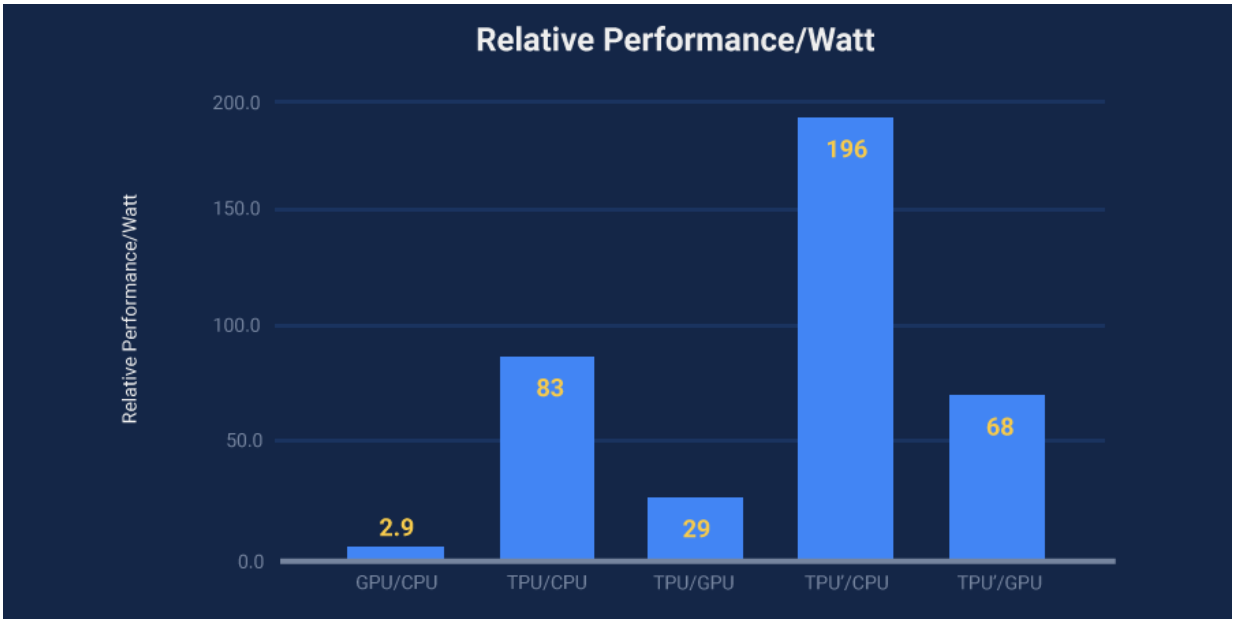
Machine learning chip earns another feather in Google thinking cap

April 7 2017, by Nancy Owano



Credit: Google

(Tech Xplore)—Google has said the TPU beat Nvidia and Intel. Let's explain that. There is so much to explain. TPU stands for Tensor Processing Unit. This is described by a Google engineer as "an entirely new class of custom machine learning accelerator."

OK, but what exactly is a TPU? *IEEE Spectrum*: "Google's Tensor Processing Unit is a printed-circuit card, which inserts into existing

servers and acts as a co-processor, one tailored for neural-network calculations."

It beat Xeon (Intel) and Nvidia GPU in machine learning tests—not only that but beat them by an order of magnitude, reports said—that is the TPU is claimed as an order of magnitude faster than contemporary CPUs and GPUs with relative performance per watt even larger.

As important, Google revealed details about how its custom TTPU speeds up machine learning, said *InfoWorld*.

Serdar Yegulalp, Senior Writer, *InfoWorld*, said the company "is providing details of exactly how much juice a TPU can provide for machine learning, courtesy of a paper that delves into the technical aspects.

Rick Merritt, *EE Times*, said the paper "gives a deep dive into the TPU and benchmarks showing that it is at least 15 times faster and delivers 30 times more performance/watt than the merchant chips."

One of the hardware engineers, Norm Jouppi, on Wednesday wrote a blog about the chip, "Quantifying the performance of the TPU, our first machine learning chip."

He said, "Today, [Wednesday] in conjunction with a TPU talk for a National Academy of Engineering meeting at the Computer History Museum in Silicon Valley, we're releasing a study that shares new details on these custom chips."

The blog is useful reading because, beyond benchmarks, Jouppi is able to deliver a scenario in which TPUs would play a role.

"The need for TPUs really emerged about six years ago, when we started

using computationally expensive deep learning models in more and more places throughout our products. The computational expense of using these models had us worried. If we considered a scenario where people use Google voice search for just three minutes a day and we ran deep neural nets for our speech recognition system on the processing units we were using, we would have had to double the number of Google data centers!"

And TPU faces that challenge. "TPUs allow us to make predictions very quickly, and enable products that respond in fractions of a second. TPUs are behind every search query; they power accurate vision models that underlie products like Google Image Search, Google Photos and the Google Cloud Vision API..."

Senior Editor David Schneider, *IEEE Spectrum*: "The TPU is built for doing inference, having hardware that operates on 8-bit integers rather than higher-precision floating-point numbers."

This is what Thomas Claburn in *The Register* had to say about its inference role:

"The internet king assembled a team to produce a custom chip capable of handling part of its neural network workflow known as inference, which is where the software makes predictions based on data developed through the time-consuming and computationally intensive training phase. The processor sits on the PCIe bus and accepts commands from the host CPU: it is akin to a yesteryear discrete FPU or math coprocessor, but obviously souped up to today's standards."

Google has deployed TPU in its data centers since early in 2015.The benchmark tests used to reach these conclusions, said Schneider, were based on "a half dozen of the actual kinds of neural-network programs that people are running at Google data centers."

A team of more than 70 engineers contributed to the TPU, said Merritt.("It really does take a village to design, verify, implement and deploy the hardware and software of a system like this," remarked Jouppi.)

The paper is titled "In-Datacenter Performance Analysis of a Tensor Processing Unit."

Why it matters: "Google's approach will influence future development of machine learning powered by custom silicon," said *InfoWorld*.

"It ought to be clear by now that custom silicon for machine learning will drive the development of both the hardware and software sides of the equation. It's also clear Google and others have barely begun exploring what's possible."

More information: — cloudplatform.googleblog.com/2 … e-learning-chip.html
— drive.google.com/file/d/0Bx4ha … MzRNcy1vSUxtcEk/view

Citation: Machine learning chip earns another feather in Google thinking cap (2017, April 7) retrieved 27 April 2024 from https://techxplore.com/news/2017-04-machine-chip-feather-google-cap.html