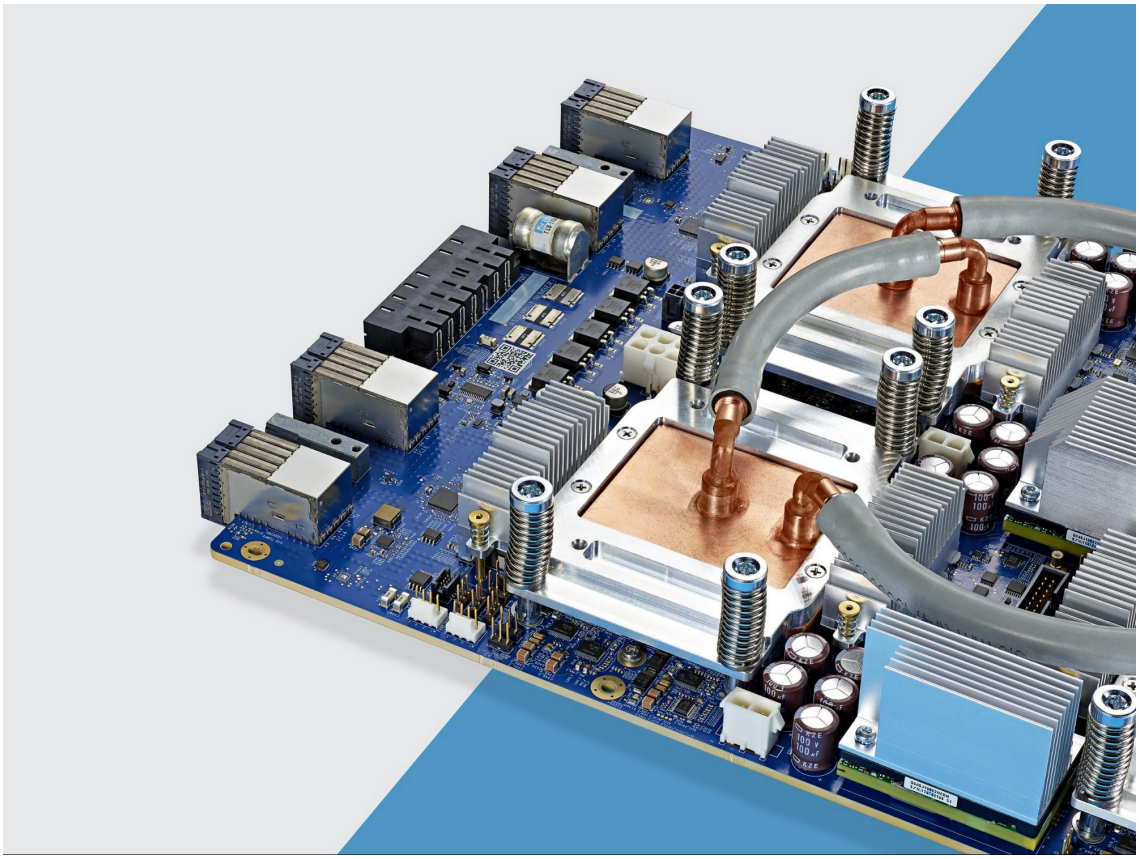
AI & MACHINE LEARNING

What makes TPUs fine-tuned for deep learning?

Kaz Sato

accounts, and the Cloud TPU v3 is available in alpha.



Third generation Cloud TPU

But many people ask me "what's the difference between a CPU, a GPU, and a TPU?" So we've created a demo site that is home to a presentation and animation that answer this question.

Tensor Processing Unit

Designed for fast and affordable AI

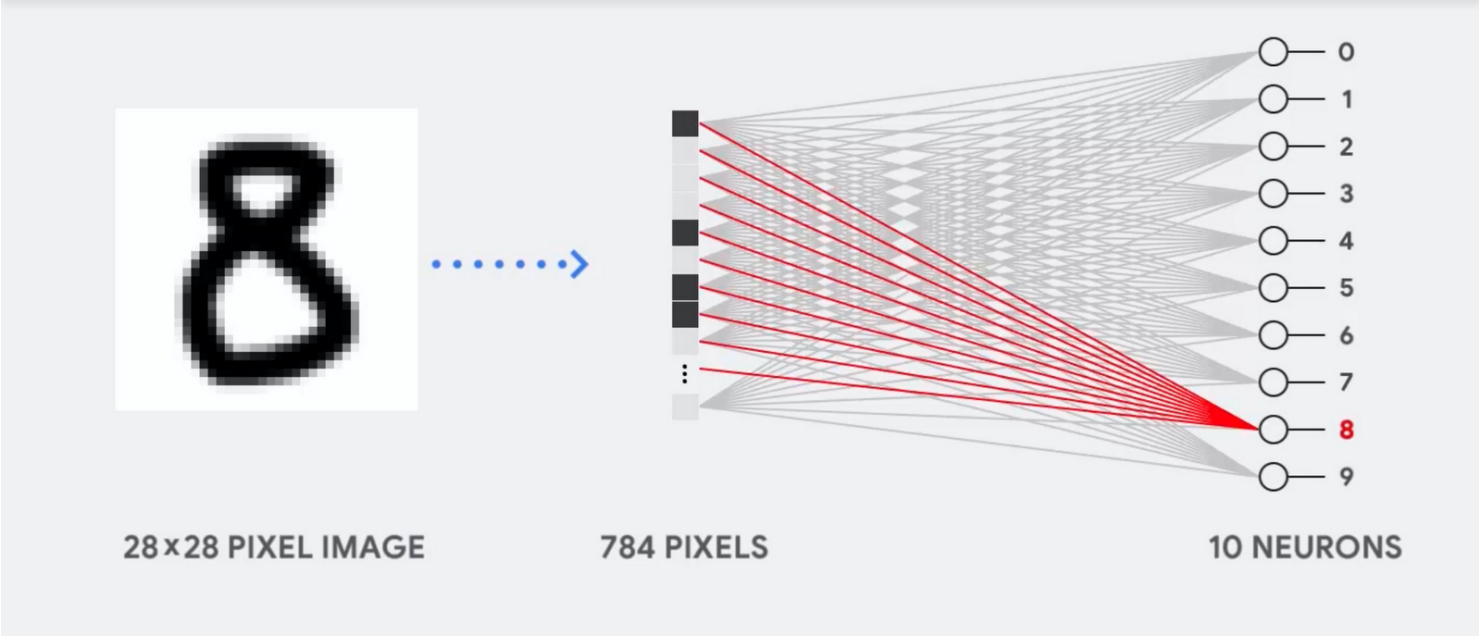
tpudemo.com, an explanatory presentation site for the Tensor Processing Unit

In this post, I'd like to highlight some specific parts of the site's content.

How neural networks work

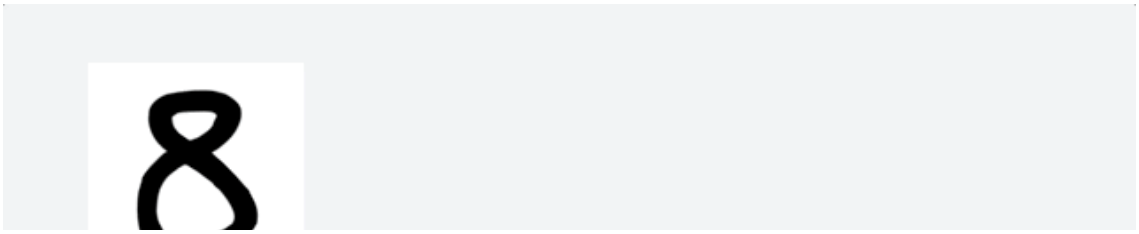
Before we start comparing CPU, GPU, and TPU, let's see what kind of calculation is required for machine learning—specifically, neural networks.

For example, imagine that we're using single layer neural network for recognizing a hand-written digit image, as shown in the following diagram:



If an image is a grid of 28 x 28 grayscale pixels, it could be converted to a vector with 784 values (dimensions). The neuron that recognizes a digit "8" takes those values and **multiply** by the parameter values (the red lines above).

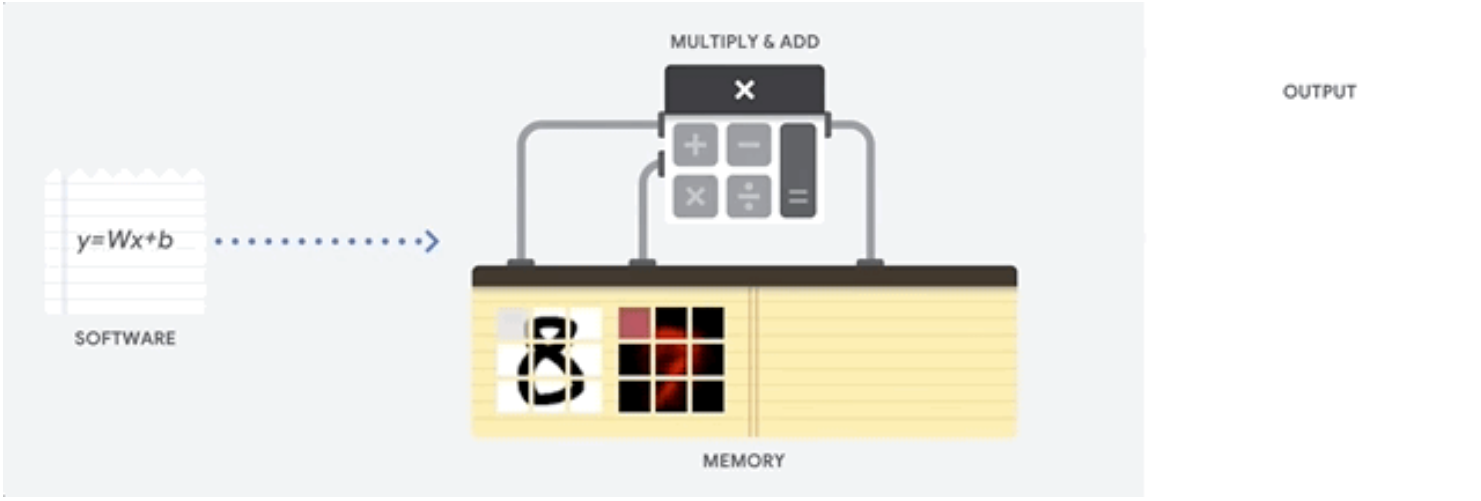
The parameter works as "a filter" to extract a feature from the data that tells the similarity between the image and shape of "8", just like this:



power consumption.

How a CPU works

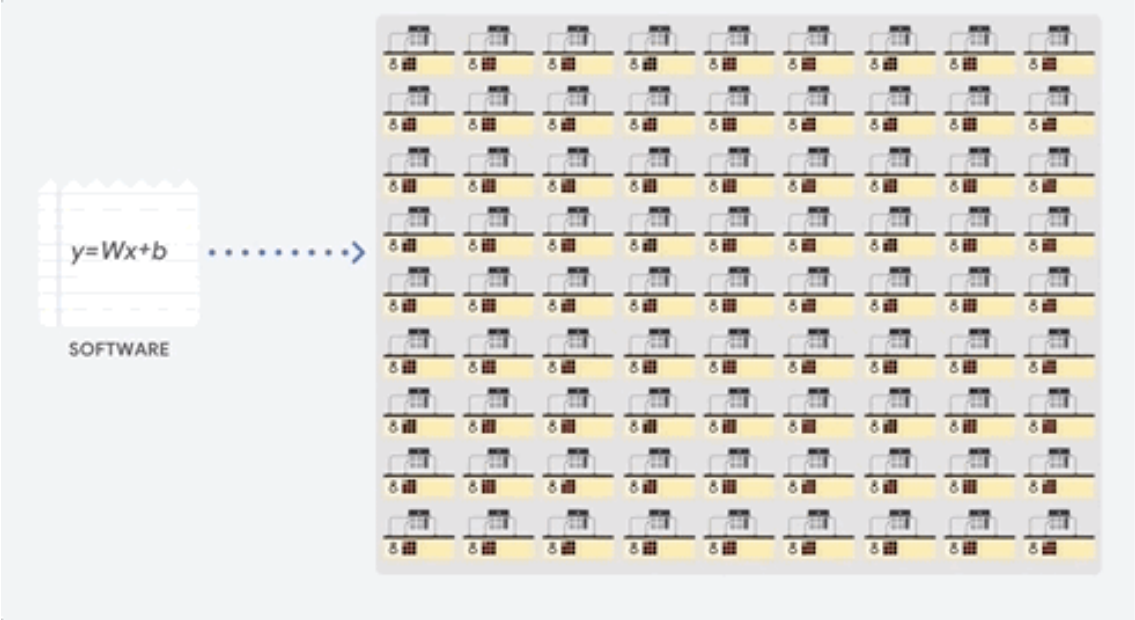
So, how does a CPU approach this task? The CPU is a general purpose processor based on the von Neumann architecture. That means a CPU works with software and memory, like this:



How a CPU works
(This animation is designed for conceptual presentation purpose only, and does not reflect the actual behavior of real processors.)

The greatest benefit of CPU is its **flexibility**. With its Von Neumann architecture, you can load any kind of software for millions of different applications. You could use a CPU for word processing in a PC, controlling rocket engines, executing bank transactions, or

thousands of ALUs in a processor? The modern GPU usually has 2,500–5,000 ALUs in a single processor that means you could execute thousands of multiplications and additions simultaneously.



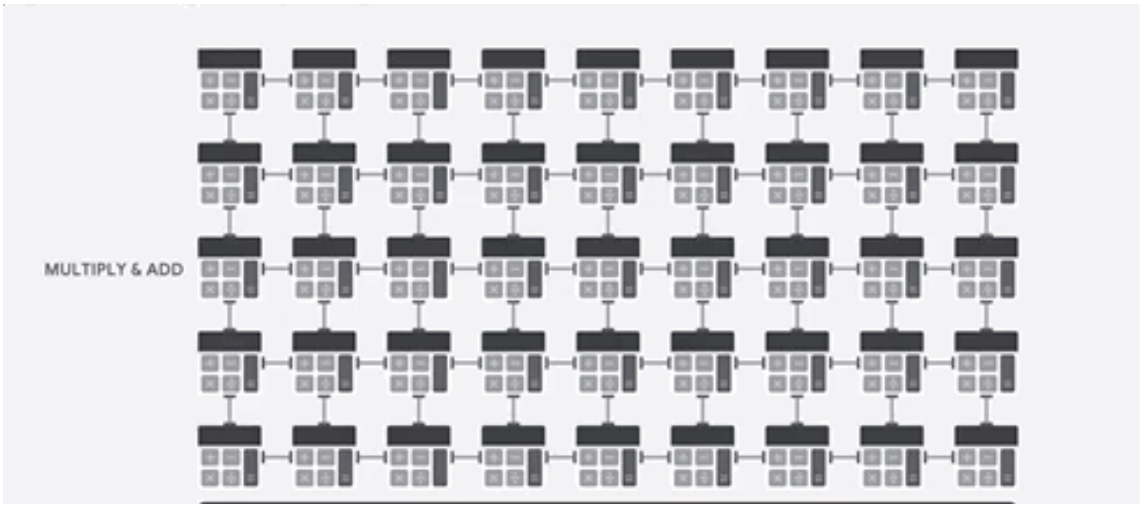
How a GPU works
(This animation is designed for conceptual presentation purpose only, and does not reflect the actual behavior of real processors.)

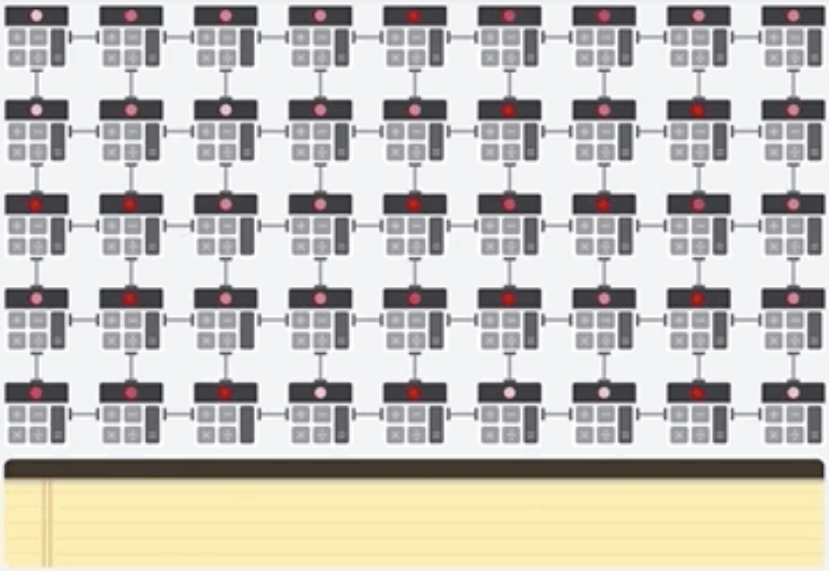
This GPU architecture works well on applications with massive parallelism, such as matrix multiplication in a neural network. Actually, you would see order of magnitude

instead of designing a general purpose processor, we designed it as a **matrix processor** specialized for neural network work loads. TPUs can't run word processors, control rocket engines, or execute bank transactions, but they can handle the massive multiplications and additions for neural networks, at blazingly fast speeds while consuming much less power and inside a smaller physical footprint.

The key enabler is a major reduction of the von Neumann bottleneck. Because the primary task for this processor is matrix processing, hardware designer of the TPU knew every calculation step to perform that operation. So they were able to place thousands of multipliers and adders and connect them to each other directly to form a large physical matrix of those operators. This is called systolic array architecture. In case of Cloud TPU v2, there are two systolic arrays of 128 x 128, aggregating **32,768 ALUs** for 16 bit floating point values in a single processor.

Let's see how a systolic array executes the neural network calculations. At first, TPU loads the parameters from memory into the matrix of multipliers and adders.





This is why the TPU can achieve a high computational throughput on neural network calculations with much less power consumption and smaller footprint.

The benefit: the cost reduces to one fifth

So what's the benefit you could get with this TPU architecture? The answer is cost. The following is the pricing of Cloud TPU v2 in August 2018, at the time of writing:

Google Cloud Region	Normal (hourly)	Preemptible (hourly)
us-central1	$4.50 / hour	$1.35 / hour
europe-west4	$4.95 / hour	$1.485 / hour

Learn more

Interested in Cloud TPU? Please go to cloud.google.com/tpu to try it today.

Acknowledgements

Special thanks to BIRDMAN who authored the awesome animations. Also, thanks to Zak Stone and Cliff Young for valuable feedback on this content.

POSTED IN: AI & MACHINE LEARNING — GOOGLE CLOUD PLATFORM — TPUS

RELATED ARTICLES

How 20th Century Fox uses ML to predict a movie audience

AI in Motion: designing a simple system to see, understand, and react in the real world (Part III)

Follow Us