Tensor Processing Unit

By: Lucas Jodon
Yelman Khan

Overview

- History
- Neural Networks
- Architecture
- Performance
- Real-World Uses
- Future Development

History of TPUs

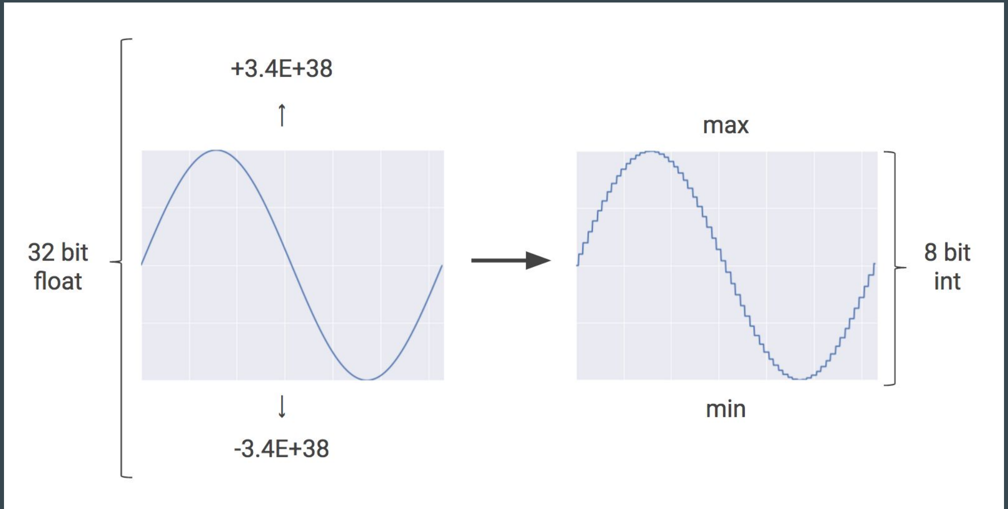
- Google began searching for a way to support neural networking for the development of their services such as voice recognition
    - Using existing hardware, they would require twice as many data centers
    - Development of a new architecture instead
- Norman Jouppi begins work on a new architecture to support TensorFlow
    - FPGA’s were not power-efficient enough
    - ASIC design was selected for power and performance benefits
    - Device would execute CISC instructions on many networks
    - Device was made to be programmable, but operate on matrices instead of vector/scalar
    - Resulting device was comparable to a GPU or Signal Processor

Neural Networks

- First proposed in 1944 by Warren McCullough and Walter Pitts
    - Modeled loosely on human learning
- Neural nets are a method of machine learning
    - Computer learns to perform a task by analyzing training examples
    - EX: pair several audio files with the text words they mean, the machine will then find patterns between the audio data and the labels
    - Each incoming pairing is given a weight, which is added to pre-existing node pairings
    - Once node weights pass a predefined threshold, the pairing is considered active
- Google began development on DistBelief in 2011
    - DistBelief became TensorFlow, which officially released version 1.0.0 in February 2017
    - TensorFlow is a software library with significant machine learning support
    - TensorFlow is intended to be a production grade library for dataflow implementation

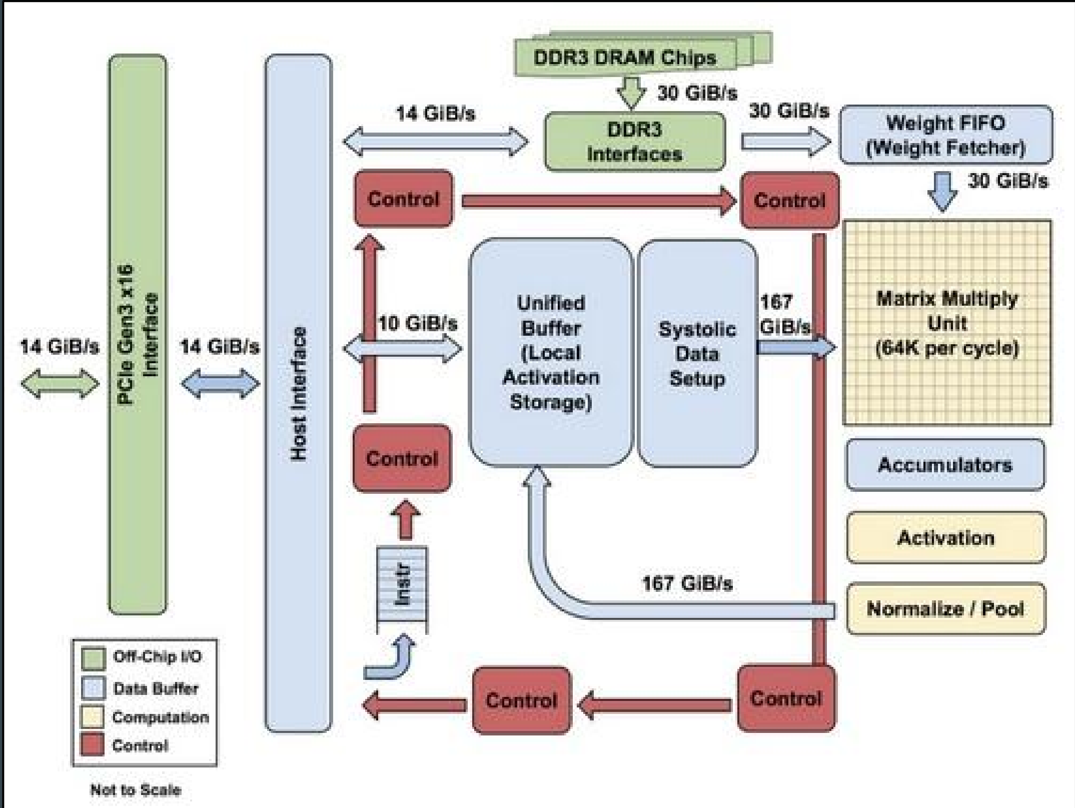
Quantization in Neural Networks

- Precision of 32-bit/16-bit floating points usually not required
- Accuracy can be maintained with 8-bit integers
- Energy consumption and hardware footprint is reduced



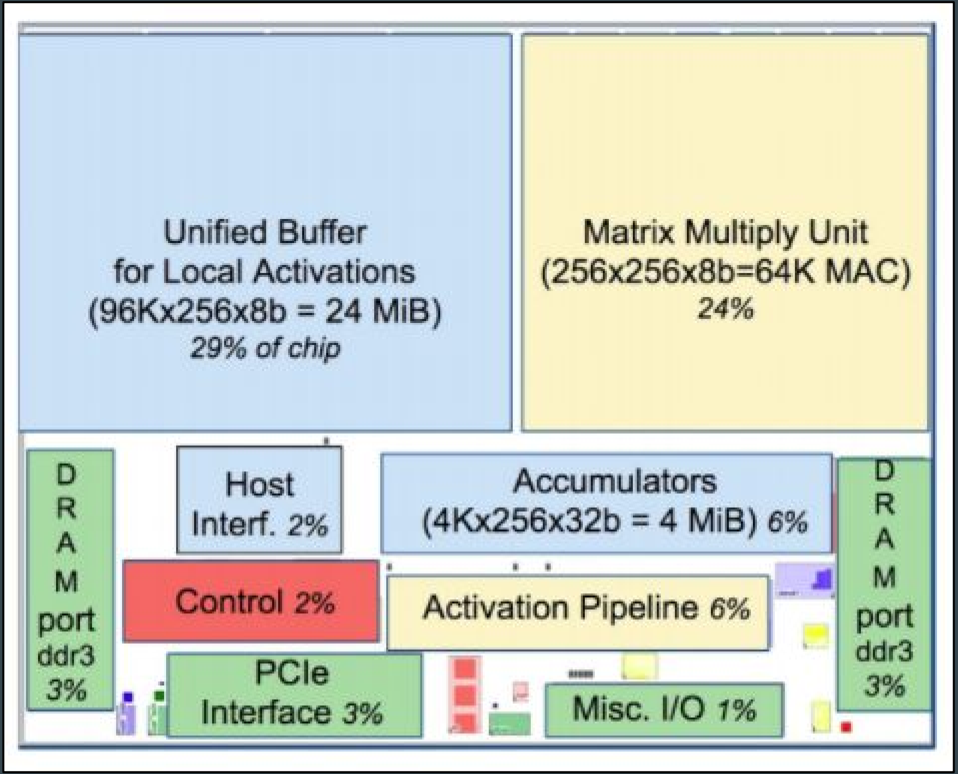
Architecture Overview

- Large, on-chip DRAM required for accessing pairing weight values.
- It is Possible to simultaneously store weights and load activations.
  - TPU can do 64,000 of these accumulates per cycle.
- First generation used 8-bit operands and quantization
  - Second generation uses 16-bit
- Matrix Multiplication Unit has 256 × 256 (65,536) ALUs



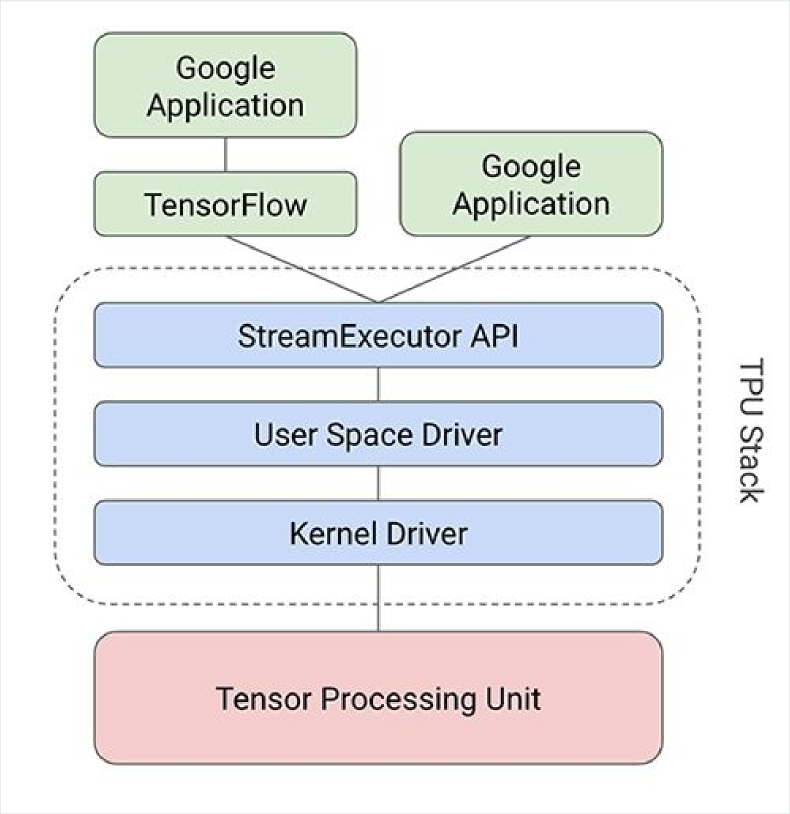
Architecture Overview Continued

- Minimalistic hardware design used to improve space and power consumption
  - No caches, branch prediction, out-of-order execution, multiprocessing, speculative prefetching, address coalescing, multithreading, context switching, etc.
  - Minimalism is beneficial here because TPU is required only to run neural network prediction
- TPU chip is half the size of the other chips
  - 28 nm process with a die size ≤ 331 mm
  - This is partially due to simplification of control logic



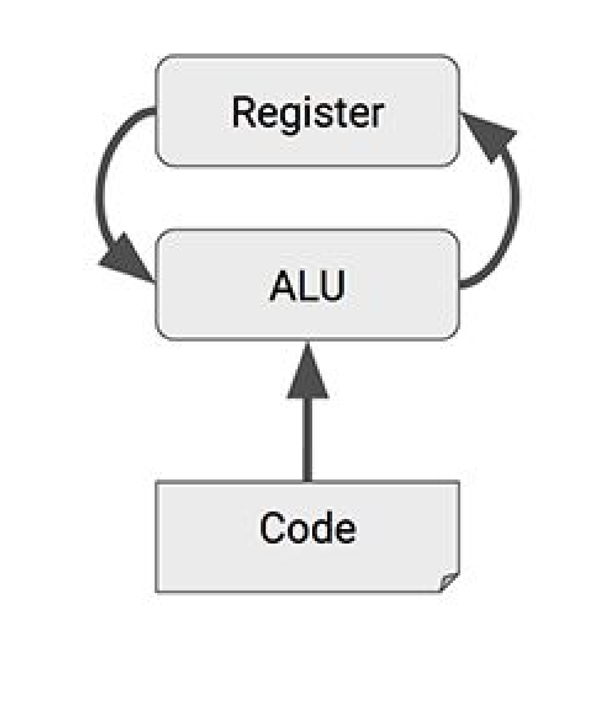
TPU Stack

- TPU performs the actual neural network calculation
- Wide range of neural network models
- TPU stack translates the API calls into TPU instructions



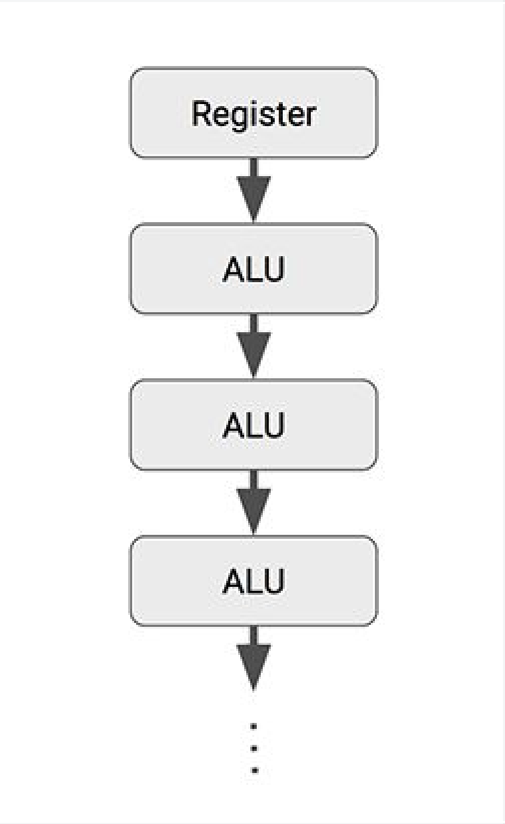
CPUs & GPUs

- CPUs and GPUs store values in registers
- A program tracks the read/operate/write operations
- A program tells ALUs :
  - Which Register to read from
  - What operation to perform
  - Which Register to write to



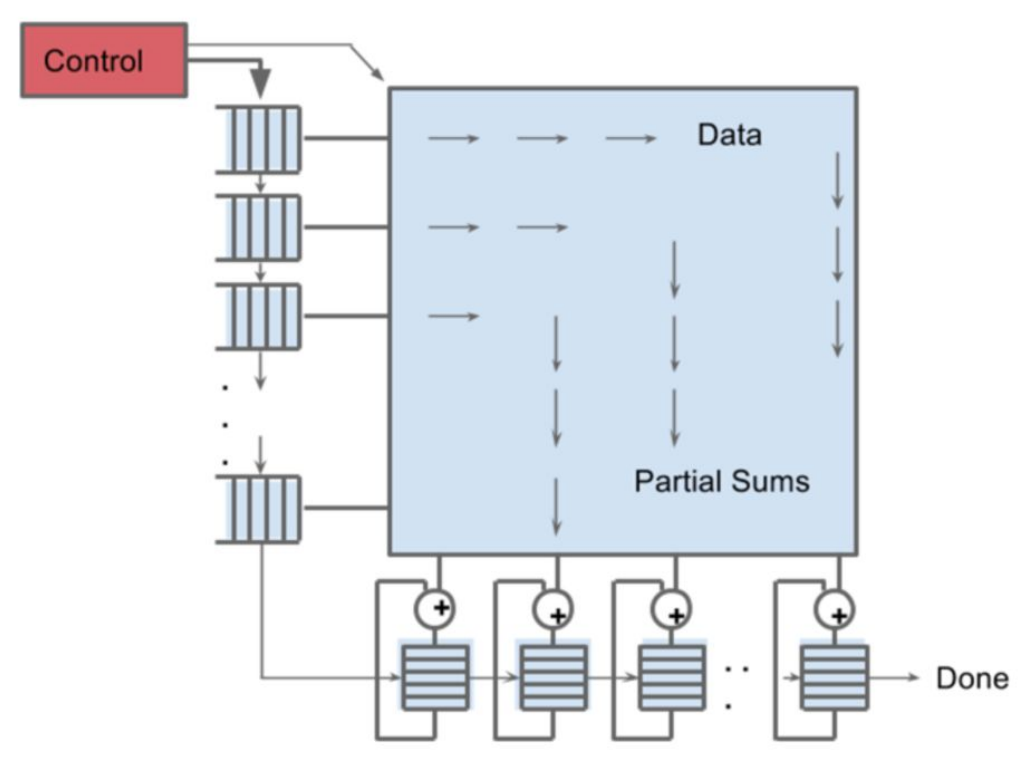
Performance

- TPU consists of Matrix Multiplier Unit (MXU)
- MXU performs hundreds of thousands of operations per clock cycle
- Reads an input value only once
- Inputs are used many times without storing back to register
- Wires connect adjacent ALUs
- Multiplication and addition are performed in specific order
- Short and energy efficient
- Design is known as systolic array



Matrix Multiplication Unit

- Contains 256 x 256 = 65,536 ALUs
- TPU runs at 700 MHz
- Able to compute 46 x 1012 multiply-and-add operations per second
- Equivalent to 92 Teraops per second in matrix unit

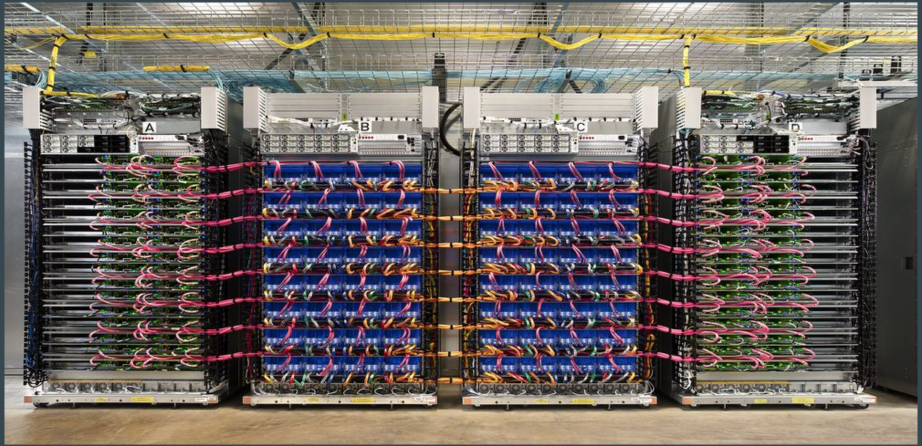


USES

- RankBrain algorithm used by Google search
- Google Photos
- Google Translate
- Google Cloud Platform

Future Development

- Google Cloud TPUs
  - Uses TPU version 2
  - Each TPU include a high-speed network
  - Allows to build machine learning supercomputers called “TPU Pods”
  - Improvement in training times
  - Allows mixing and matching with other hardware which includes Skylake CPUs and NVIDIA GPUs



Works Cited

- First In-Depth Look at Google's TPU Architecture
  https://www.nextplatform.com/2017/04/05/first-depth-look-googles-tpu-architecture/
- An in-depth look at Google's first Tensor Processing Unit (TPU) | Google Cloud Big Data and Machine Learning Blog | Google Cloud Platform
  https://cloud.google.com/blog/big-data/2017/05/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu
- Google Cloud TPU Details Revealed
  Servethehome - https://www.servethehome.com/google-cloud-tpu-details-revealed/
- TensorFlow - Google
- https://research.googleblog.com/2015/11/tensorflow-googles-latest-machine_9.html
- Explained: Neural networks
- Larry Hardesty | MIT News Office -
  http://news.mit.edu/2017/explained-neural-networks-deep-learning-0414
- https://www.nextplatform.com/2017/04/05/first-depth-look-googles-tpu-architecture/
- Google cloud TPU -
  https://www.blog.google/topics/google-cloud/google-cloud-offer-tpus-machine-learning/

QUESTIONS?