Electronic Design

WILLIAM WONG | Technology Editor

CPUs, GPUs and Now AI Chips

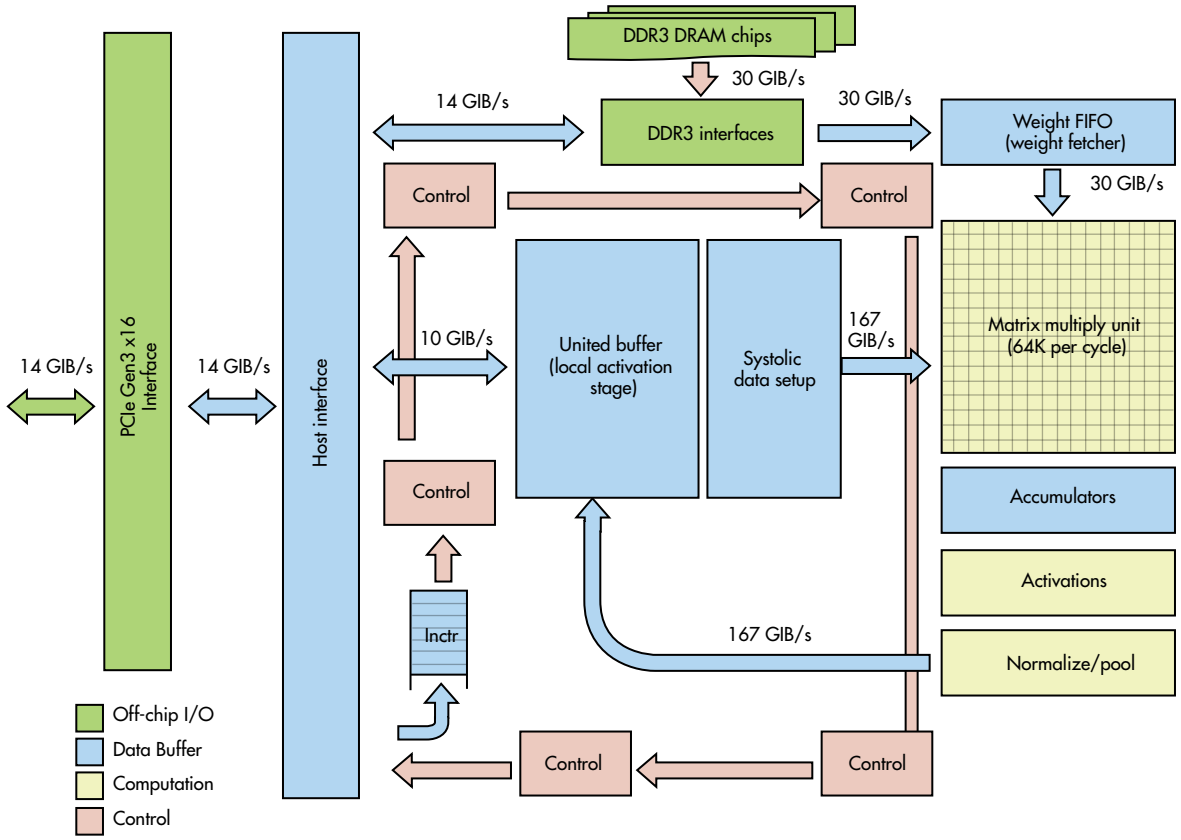
First it was CPUs. Next came GPUs. What's next? How about AI chips?

If you haven't heard about the artificial intelligence (AI) machine-learning (ML) craze that uses deep neural networks (DNN) and deep learning (DL) to tackle everything from voice recognition to making self-driving cars a reality, then you probably haven't heard about Google's new Tensor Processing Unit (TPU), Intel's Lake Crest, or Knupath's Hermosa. These are just a few of the vendors looking to deliver platforms targeting neural networks.

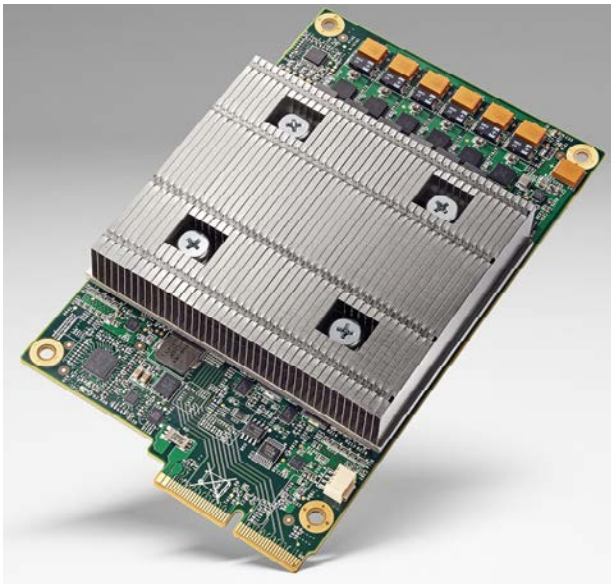
1. Google TPU

The TPU contains a large 8-bit matrix multiply unit *(Fig. 1)*. It essentially optimizes the number-crunching required by DNN; large floating-point number-crunchers need not apply.

The TPU is actually a coprocessor managed by a conventional host CPU via the TPU's PCI Express interface. The TPU chip runs at only 700 MHz, but can best CPU and GPU systems when it comes to DNN acceleration. Though not



1. Google's TPU has a large 8-bit matrix multiply unit to help it crunch numbers for deep neural networks.



2. Google's TPU module is designed to fill arrays of slots in its cloud data centers.

specifically a DNN processor, it handles the heavy lifting while consuming only 40 W of power. It has 28 Mbytes of on-chip RAM along with 4 Mbytes in the form of 32-bit accumulators used to compile the 16-bit results from the matrix multiply unit. The chip uses a 28-nm process and the die size is about 600 mm^2. The paper "In-Datacenter Performance Analysis of a Tensor Processing Unit" provides more details.

The TPU board *(Fig. 2)* can perform 92 TeraOps/s (TOPS). It is 15 to 30 times faster than CPUs and GPUs tasked with the same work, with a 30- to 80-fold improvement in TOPS/W. The software used for comparison of systems was the TensorFlow framework.

One thing to keep in mind is that TPU comparisons are done with respect to its limitations. Most CPUs are 64-bit platforms and GPUs can have wider word widths. They also tend to be optimized for larger data items, although most systems have support for smaller word sizes (including 8-bit vector operations). Likewise, different neural network applications benefit from different configurations, but the smaller 8-bit integers have found wide application in many DNN applications.

The TPU is has five primary instructions:

- Read_Host
- Read_Weights
- MatrixMultiply/Convolve
- Activate
- Write_Host

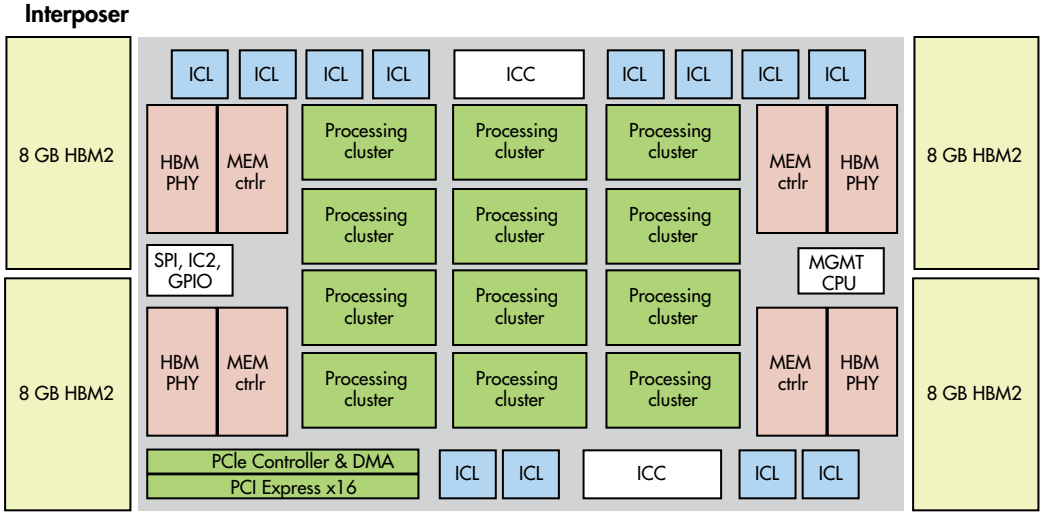
Weights are values within a neural network and are used by the matrix multiply unit. The activate function performs a nonlinear operation for an artificial neuron.

Google's TPU is expected to reduce the need for larger data centers that would otherwise need many more CPUs and GPUs to handle the AI applications addressing everything from voice recognition and analysis, to image and video processing, to providing services from search, to those little Google Home systems.

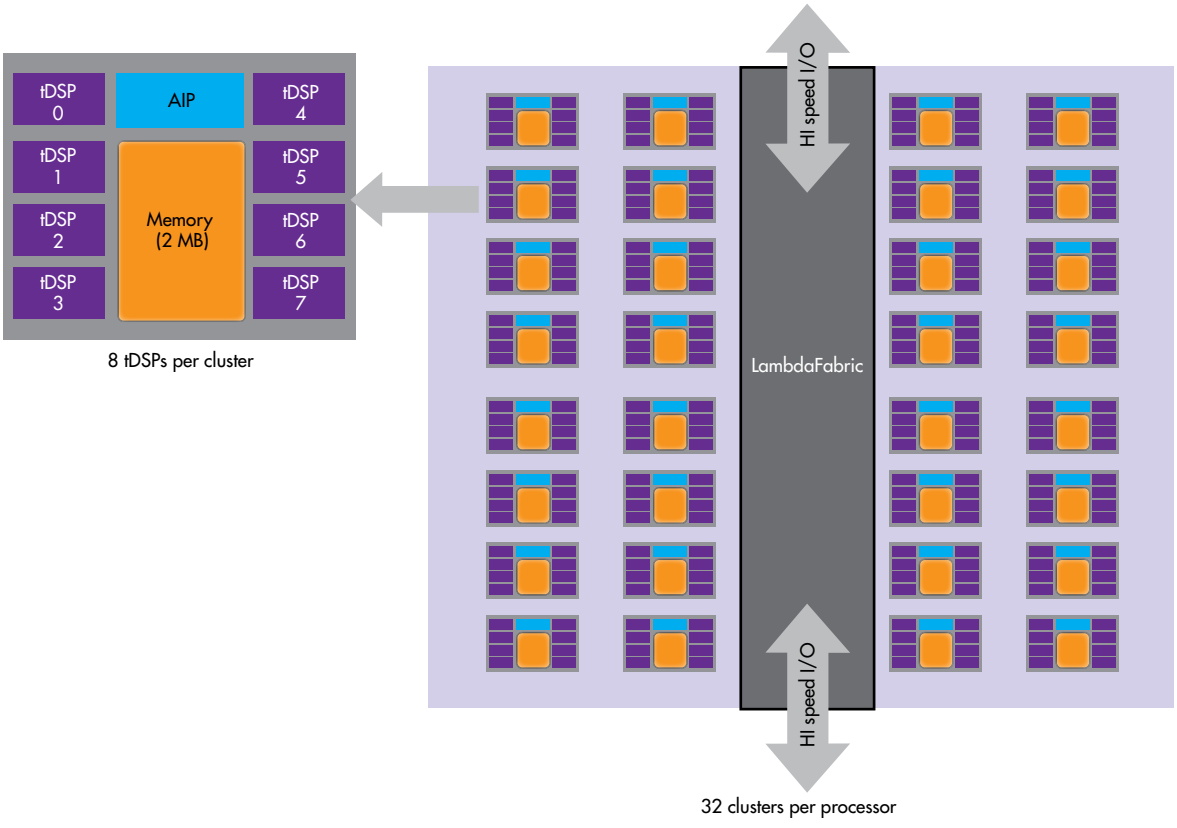
2. Intel Lake Crest

Lake Crest *(Fig. 3)* is the codename for an Intel platform designed to complement the many-core Xeon Phi *(see "Integrated Fabric is Key to Many Core Platforms" on electronicdesign.com)*. The Xeon Phi has been tasked with many AI chores, but it can be challenged by applications that Google's TPU or Intel's Lake Crest will readily handle more efficiently. Lake Crest technology was originally developed by Nervana, which is not part of Intel.

The new chip will employ a range of advanced features from MCM (Multi Chip Module) design to the "Flexpoint" architecture with a dozen specialized, multicore processing nodes like the TPU's matrix multiply unit. The chips will have 32 Gbytes of High Bandwidth Memory 2 (HBM2) with an aggregate bandwidth of 8 Tbytes/s attached via an interposer. HBM2 has become common in high-performance SoCs and



3. Intel's Lake Crest uses processing clusters optimized for AI applications.



4. Knupath's Hermosa many core processor has 256 DSP cores organized in eight clusters of eight cores connected by its Lambda Fabric.

GPUs. Lake Crest does not have any caches. Software will be used to optimize memory management.

Lake Crest is expected in the 2017 timeframe.

3. Knupath Hermosa

Knupath's Hermosa *(Fig. 4)* has 64 DMA engines and 256 DSP cores organized in eight clusters of eight cores connected by its Lambda Fabric. The Lambda Fabric is also designed to link thousands of Hermosa processors in a low latency and high throughput mesh.

The Hermosa has an integrated L1 router with 32 ports and a 1 Tbit/s bandwidth. Links to the outside world include 16 10 Gbit/s bidirectional ports. The chip has 72 Mbytes of data RAM organized in 32 banks and 2 Mbytes of program RAM.

Although Hermosa targets AI applications, it may be more akin to the many-core Xeon Phi than the more specialized Lake Crest or TPU platforms. Hermosa only uses 34 W to deliver 384 GFLOPS of computing power, making it very interesting for a wide range of applications—not just AI ones.

GPGPUs Continue to Reign (for Now)

NVidia and AMD have a vested interest in their GPU platforms, which have been the backbone for most high-end neural network work. This could change as specialized AI chips become available. The question is how tailored these chips will be to a particular application, how available they will be, and how well they can be applied to different applications.

Right now GPU platforms like NVidia's Jetson TX2 *(see "DNN Popularity Drives NVidia's Jetson TX2" on electronicdesign.com)* are being used in everything from drones to medical devices. It is actually possible to be used in an AI accelerator in Intel's tiny Curie module, as well *(see "What Is Inside an IoT Chip?" on electronicdesign.dom)*. One size does not fit all, but AI will only continue grow in importance for computer applications.