How Important Are Transistor Sizes?

For years, the semiconductor industry has tracked progression of chip speeds by watching transistor sizes. As transistors got smaller, more of them could be packed onto a standard chip, and algorithms could run faster. At the time of writing of this book, Intel is currently operating on 10-nanometer transistors, and working on transitioning down to 7 nanometers. The rate of shrinkage of transistor sizes has slowed significantly in recent years, since formidable heat dissipation issues arise at these scales.

Nvidia's GPUs partially buck this trend. They tend to use transistor sizes a generation or two behind Intel's best, and focus on solving architectural and software bottlenecks instead of transistor engineering. So far, Nvidia's strategy has paid dividends and the company has achieved market domination in the machine learning chip space.

It's not yet clear how far architectural and software optimizations can go. Will GPU optimizations soon run into the same Moore's law roadblocks as CPUs? Or will clever architectural innovations enable years of faster GPUs? Only time can tell.

Tensor Processing Units

The tensor processing unit (TPU) is a custom ASIC (application specific integrated circuit) designed by Google to speed up deep learning workloads designed in TensorFlow. Unlike the GPU, the TPU is stripped down and implements only the bare minimum on-die needed to perform necessary matrix multiplications. Unlike the GPU, the TPU is dependent on an adjoining CPU to do much of its preprocessing work for it. This slimmed-down approach enables the TPU to achieve higher speeds than the GPU at lower energy costs.

The first version of the TPU only allowed for inference on trained models, but the most recent version (TPU2) allows for training of (certain) deep networks as well. However, Google has not released many details about the TPU, and access is limited to Google collaborators, with plans to enable TPU access via the Google cloud. Nvidia is taking notes from the TPU, and it's quite likely that future releases of Nvidia GPUs will come to resemble the TPU, so downstream users will likely benefit from Google's innovations regardless of whether Google or Nvidia wins the consumer deep learning market. Figure 9-3 illustrates the TPU architecture design.

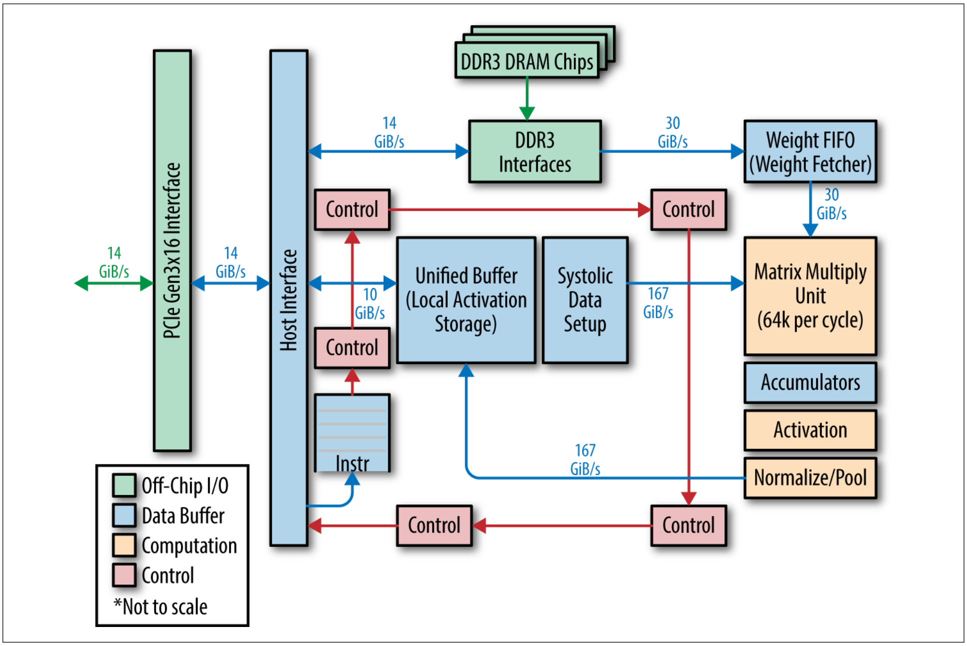

Figure 9-3. A tensor processing unit (TPU) architecture from Google. TPUs are specialized chips designed by Google to speed up deep learning workloads. The TPU is a coprocessor and not a standalone piece of hardware.

What Are ASICs?

Both CPUs and GPUs are general-purpose chips. CPUs generally support instruction sets in assembly and are designed to be universal. Care is taken to enable a wide range of applications. GPUs are less universal, but still allow for a wide range of algorithms to be implemented via languages such as CUDA.

Application specific integrated circuits (ASICs) attempt to do away with the generality in favor of focusing on the needs of a particular application. Historically, ASICs have only achieved limited market penetration. The drumbeat of Moore's law meant that general-purpose CPUs stayed only a breath or two behind custom ASICs, so the hardware design overhead was often not worth the effort.

This state of affairs has started shifting in the last few years. The slowdown of transistor shrinkage has expanded ASIC usage. For example, Bitcoin mining depends entirely on custom ASICs that implement specialized cryptography operations.

Field Programmable Gate Arrays

Field programmable gate arrays (FPGAs) are a type of "field programmable" ASIC. Standard FPGAs can often be reconfigured via hardware description languages such as Verilog to implement new ASIC designs dynamically. While FPGAs are generally less efficient than custom ASICs, they can offer significant speed improvements over CPU implementations. Microsoft in particular has used FPGAs to perform deep learning inference and claims to have achieved significant speedups with their deployment. However, the approach has not yet caught on widely outside Microsoft.

Neuromorphic Chips

The "neurons" in deep networks mathematically model the 1940s understanding of neuronal biology. Needless to say, biological understanding of neuronal behavior has progressed dramatically since then. For one, it's now known that the nonlinear activations used in deep networks aren't accurate models of neuronal nonlinearity. The "spike trains" is a better model (see Figure 9-4), where neurons activate in short-lived bursts (spikes) but fall to background most of the time.

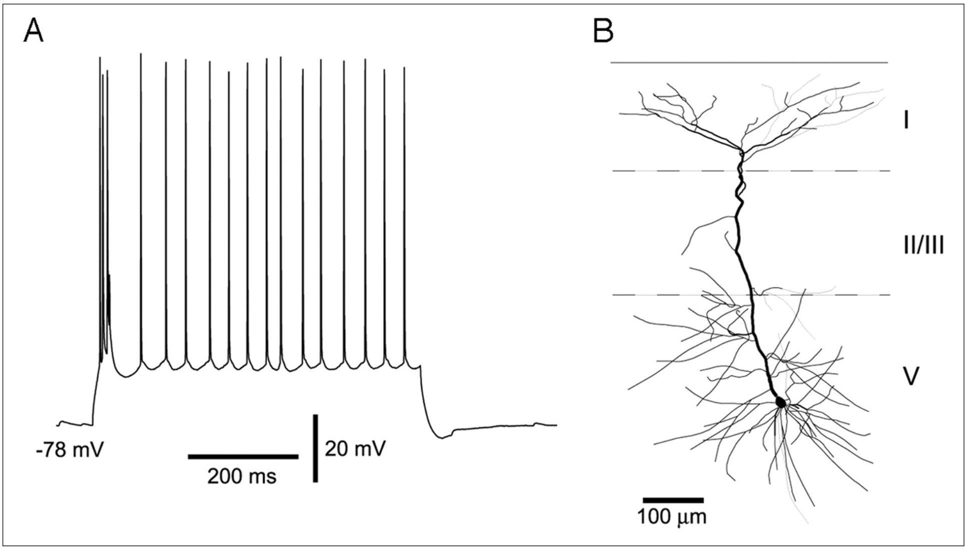

Figure 9-4. Neurons often activate in short-lived bursts called spike trains (A). Neuromorphic chips attempt to model spiking behavior in computing hardware. Biological neurons are complex entities (B), so these models are still only approximate.

Hardware engineers have spent significant effort exploring whether it's possible to create chip designs based on spike trains rather than on existing circuit technologies (CPUs, GPUs, ASICs). These designers argue that today's chip designs suffer from fundamental power limitations; the brain consumes many orders of magnitude less