Google did its best to impress this week at its annual IO conference. While Google rolled out a bunch of benchmarks that were run on its current Cloud TPU instances, based on TPUv2 chips, the company divulged a few skimpy details about its next generation TPU chip and its systems architecture. The company changed from version notation (TPUv2) to revision notation (TPU 3.0) with the update, but ironically the detail we have assembled shows that the step from TPUv2 to what we will call TPUv3 probably isn't that big; it should probably be called TPU v2r5 or something like that.

You might want to take a look at the drill down into the TPUv2 that we did last year to update yourself on the architecture if you are not familiar with it. We use Google's definition of a Cloud TPU, which is a board containing four TPU chips. Google's current Cloud TPU beta program only allows users to access single Cloud TPUs. Cloud TPUs cannot yet be federated in any way, except by Google's in-house developers. We have learned over the past year that Google has abstracted its Cloud TPUs behind its TensorFlow deep learning (DL) framework. We don't expect that to change; no one outside of Google's in-house TensorFlow development team will have direct access to Cloud TPU hardware, probably ever.

We also believe that Google has funded a huge software engineering and optimization effort to get to its current beta Cloud TPU deployment. That gives Google incentive to retain as much of TPUv2's system interfaces and behavior – hardware abstraction layer and application programming interfaces (APIs) – as possible with the TPUv3. Google offered no information on when TPUv3 will be offered as a service, in Cloud TPUs or in a multi-rack pod configurations. It did show a photo of a TPUv3-based Cloud TPU board and pod photos. The company made the following assertions:

- The TPUv3 chip runs so hot that for first time Google has introduced liquid cooling in its datacenters
- Each TPUv3 pod will be eight times more powerful than a TPUv2 pod
- Each TPUv3 pod will perform at “well over a hundred petaflops”

However, Google also restated that its TPUv2 pod clocks in at 11.5 petaflops. An 8X improvement should land a TPUv3 pod at a baseline of 92.2 petaflops, but 100 petaflops is almost 9X. We can't believe Google's marketing folks didn't round up, so something is not quite right with the math. This might be a good place to insert a joke about floating point bugs, but we'll move on.

The Pods

It is obvious from the two photos of the full TPUv3 pod that Google scaled its next-generation way up:

- There are twice as many racks per pod
- There are twice as many Cloud TPUs per rack

This nets a 4X performance improvement over a TPUv2 pod if nothing else changes.
The Racks

The TPUv3 pod racks are spaced closer than the TPUv2 racks. But, like TPUv2 pods, there is still no storage evident in TPUv3 pods. TPUv3 racks are also taller, to accommodate adding water cooling.
Google moved the uninterruptable power supplies from the bottom of the TPUv2 rack to the top of the TPUv3 rack. We assume that the massive metal box now at the bottom of the rack contains a water pump or other water-cooling related gear.

Modern hyperscale datacenters don't use raised floors. Google's racks are heavy even before adding water, so they sit directly on concrete slabs. Water enters and leaves from the top of the rack. Google's datacenters have plenty of overhead space, shown in a photo of the TPUv3 pod. However, routing and hanging heavy water conduits must have been an additional operations challenge.

Note the twisted wire running in front of the rack on the floor, just in front of the big metal bottom-of-rack box. We suspect it is a moisture sensor.

Shelves And Boards

Google not only doubled the compute shelf density, it reduced the ratio of server boards to Cloud TPUs from one-to-one to one server board for every two Cloud TPU boards. This has an impact in power consumption estimates, as servers and Cloud TPUs in a TPUv3 pod will draw power from the same rack power supply.

Google bills the server board used with its current Cloud TPU beta instance as a Compute Engine `n1-standard-2` instance on its Cloud Platform public cloud, which has two virtual CPUs and 7.5 GB of memory. We think it is a safe bet to assume this is a mainstream dual-socket X86 server.

Recall that a TPUv2 pod contains 256 TPUv2 chips and 128 server processors. A TPUv3 pod will double the server processors and quadruple the TPU chip count.
We believe that Google over-provisioned the servers in its TPUv2 pod. This is understandable for a new chip and system architecture. After at least a year of tuning pod software and a minor revision of the silicon, it is likely that halving the number of servers had negligible effect on pod performance. This could be for many reasons, perhaps the servers were not compute or bandwidth bound, or Google could have deployed newer Intel Xeon or AMD Epyc processors with many more cores.

Integrating server boards into the Cloud TPU racks enabled Google to double the number of racks by using identical rack configurations. Standardizing on one rack configuration must help reduce the cost and complexity of hardware deployment.

Integrating server boards into the Cloud TPU racks enabled Google to double the number of racks by using identical rack configurations. Standardizing on one rack configuration must help reduce the cost and complexity of hardware deployment.

<table>
<thead>
<tr>
<th>Processor</th>
<th>TPU</th>
<th>CPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chips per Board</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Boards per Shelf</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Shelves per Rack</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Racks per Pod</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Chips per Pod</td>
<td>1,024</td>
<td>256</td>
</tr>
</tbody>
</table>

However, to achieve higher density, Google had to move from a 4U Cloud TPU form factor to a 2U high-density form factor. Google runs its datacenters warm (published figures are in the 80°F to 95°F range), so the TPUv2 air-cooled heat sinks had to be large. Google uses open racks, so moving enough air to cool a hot socket in a dense form factor becomes expensive — so expensive that water cooling becomes a viable alternative. Especially for a high value service like deep learning.

Moving the server board into the TPUv3 rack also shortens connecting cables, so in general we believe Google saved significant cable costs and eliminated dead space in the TPUv2 Pod’s server racks.

Google did not show photos of the board to rack water interconnect.

Cloud TPUs

However, it did show two views of the TPUv3 Cloud TPU. The TPUv3 Cloud TPU has a similar layout to the TPUv2 Cloud TPU. The addition of water cooling is the obvious change. The back of board power connector looks the same. However, there are four additional connectors on the front of the board. The two large silver squares on the front (left) side of the photo are clusters of four connectors each.
Last year we speculated on several types of interconnects and called it wrong – Google connects a server to a Cloud TPU using 32 lanes of cabled PCI-Express 3.0 (28 GB/s for each link). We believe that it is unlikely that Google increased bandwidth between the server boards and the Cloud TPUs, because PCI-Express bandwidth and latency are probably not big performance limiters.

While interconnect topology will help deep learning tasks scale better in a pod, it doesn't contribute to raw theoretical petaflops of performance.

TPU Chips

Now that we are down to the chip level to answer the question: “Where does the remaining 2X performance improvement come from?” Google described its TPUv2 core in general terms:

- There are two Matrix Units (MXUs)
- Each MXU has 8 GB of dedicated high bandwidth memory (HBM)
- Each MXU has a raw peak throughput of 22.5 teraflops
- But the MXU does not use a standard floating point format to achieve its floating point throughput

Google invented its own internal floating point format called “bfloat” for “brain floating point” (after Google Brain). The Bfloat format uses an 8-bit exponent and 7-bit mantissa, instead of the IEEE standard FP16’s 5-bit exponent and 10-bit mantissa. Bfloat expresses values from \(-10^{-38}\) to \(-10^{38}\), which is orders of magnitude wider dynamic range than IEEE FP16. Google invented its bfloat format because it found that it needed data science experts while training with IEEE FP16 to make sure that numbers stayed within FP16’s more limited range.
We believe that Google has implemented hardware format conversion inside the MXU itself to virtually eliminate conversion latencies and software development headaches. Format conversion from FP16 to bfloat looks like a straightforward precision truncation to a smaller mantissa. Converting FP16 to FP32 and then FP32 to FP16 is known practice; the same techniques can be used to convert from FP32 to bfloat and then bfloat to FP16 or FP32.

Google claimed “great” reuse of intermediate results as data flows through the MXU’s systolic array. That last sentence will take more words to unpack than we can devote to this post.

Given how well the MXU appears to be performing for Google, we believe it is unlikely that Google will change the MXU significantly from TPUv2 to TPUv3. The far more likely scenario is that Google will simply double the number of MXUs for TPUv3.

One year between chip announcements is a very short chip design cadence. It does not leave time for significant architectural development. But it is enough time to shrink an existing MXU core to a new manufacturing process, tune power consumption and speed paths while doing so, and then stamp more MXU cores onto a die with a little additional floor-planning work. The following table contains what little hard information we have and our best estimates of where Google is headed with its TPUv3 chips.
Last year, we estimated that TPUv2 consumed 200 watts to 250 watts per chip. We now know that includes 16 GB of HBM in each package, with 2.4 TB/sec bandwidth between the MXU and HBM.

We will stick with last year's estimate of 36 kilowatts of rack power supply (288 kilowatts total for a TPUv3 pod). If we assume 400 watts for each dual-socket server, we work backwards to about 200 watts per TPUv3 chip, including 32 GB of HBM. If these chips were not packed so densely onto boards and into racks, or they were deployed in cooler datacenters, they might not need water cooling. Another alternative might be that Google is deploying single-socket servers in their new TPUv3 clusters. Lowering server power to under 250 watts might give TPUv3 enough headroom to burn up to 225 watts.

Given a conservative initial TPUv2 MXU design, followed by TPUv3 process shrink, wider and faster HBM, and speed path tuning, it is reasonable to expect that performance per core can stay the same across the two generations without a radical MXU redesign.

### Market Recap

Google is still deploying TPUv1 add-in cards for inferencing tasks, four to a server. Google has deployed TPUv1 to accelerate web searches and other large-scale inferencing tasks – if you've used Google's search engine lately, you have probably used a TPUv1.

Google is only offering TPUv2 access via its beta Cloud TPU instances and gave no predictions for when it would start production availability with a service level agreement. Google did state this week that it would make TPUv2 pods available to customers “later this year,” but it is unclear whether that would be a production service. Our best guess is that Google will wait until it has validated and debugged TPUv3 pods to deploy TPU pods at worldwide scale. Google is using TPUv2 pods internally for some training tasks. This week, Google made no statements at all about when it would deploy any capability or service based on TPUv3 chips. We believe the TPUv3 announcement was intended to highlight Google's long-term commitment to controlling its own destiny for accelerating its TensorFlow deep learning framework.

However, we look at TPUv3 as more of a TPUv2.5 than a new generation of chips. Most of the new hardware development appears to be happening at a system level around the TPUv3 chip.

Paul Teich is a technologist and a principal analyst at TIRIAS Research, covering clouds, data analysis, the Internet of Things and infrastructure-scale user experience. He is also a contributor to Forbes/Tech, and focuses on how people interact with technology-based products and services. For three decades, Teich immersed himself in IT design, development, and strategy, including two decades at AMD in product management roles. Teich holds 12 US patents and earned a BSCS from Texas A&M University and an MS in Technology Commercialization from the University of Texas' McCombs School of Business.