**TPUs**

**TENSOR PROCESSING UNITS**

**WHY TPUs?**

We have reached the end of Moore's Law. To match the increasing computational demands of high-end machine learning models, Google created special-purpose chips called TPUs.

**WHAT IS IT?**

TPUs are specialized hardware or "ASICs": Application-specific integrated circuits. They are engineered to accelerate machine learning workloads.

**TECH SPECS**

- Clock: 300 MHz
- Power consumption: 40W
- Compute: 92 Tflops

**APPLICATIONS**

TPUs were created to outperform GPUs and their inferior cousins, CPUs, on the inference phase of neural network applications.

- They are optimal for:
  - Large models
  - With large batch sizes
  - And workloads dominated by matrix-multiplication

- Real-world processes powered by TPUs:
  - Photo search
  - Text & speech translation/recognition
  - Search ranking

**UNDER THE HOOD**

TPUs have domain-specific architecture for deep learning:

- **Reduced Precision**
  - By quantization, TPUs map higher precision floating point numbers to 8 bit ints.

- **Matrix Processing**
  - Operations are narrow & hard-wired, not requiring memory access.

- **10x Performance**
  - * per watt *

**2ND GEN IMPROVEMENTS**

In 2017, Google released an improved TPU featuring 4 ASICs on a single motherboard, each with 2 cores.

- Compute: 180 Tflops
- TPU v2 excel at both training & inference.
- Uses a new datatype called “bf32float16”

  - Combines the range of a 32-bit float with the space of an 8-bit float.

**LIQUID COOLING**

Thanks to liquid cooling, the latest TPUs can be arranged into even larger pods.

- Compute: > 100 Pflops – 8 times more powerful than v2

**WHAT'S NEXT?**

v2 & v3 are now publicly available in beta.