The Deep Learning Revolution and Its Implications for Computer Architecture and Chip Design

DSIndy, 12/11/19

Andrew Hoblitzell

About the author

Jeffrey Adgate "Jeff" Dean is an American computer scientist and is currently the lead of Google AI, Google's AI division. He was elected to the National Academy of Engineering in 2009, which recognized his work on "the science and engineering of large-scale distributed computer systems.

Joined Google in 2009 and has worked on Spanner, BigTable, MapReduce, DistBelief, and Tensorflow.

Summary

https://arxiv.org/abs/1911.05289

“The past decade has seen a remarkable series of advances in machine learning, and in particular deep learning approaches based on artificial neural networks, to improve our abilities to build more accurate systems across a broad range of areas, including computer vision, speech recognition, language translation, and natural language understanding tasks. This paper is a companion paper to a keynote talk at the 2020 International Solid-State Circuits Conference (ISSCC) discussing some of the advances in machine learning, and their implications on the kinds of computational devices we need to build, especially in the post-Moore's Law-era…..

Summary

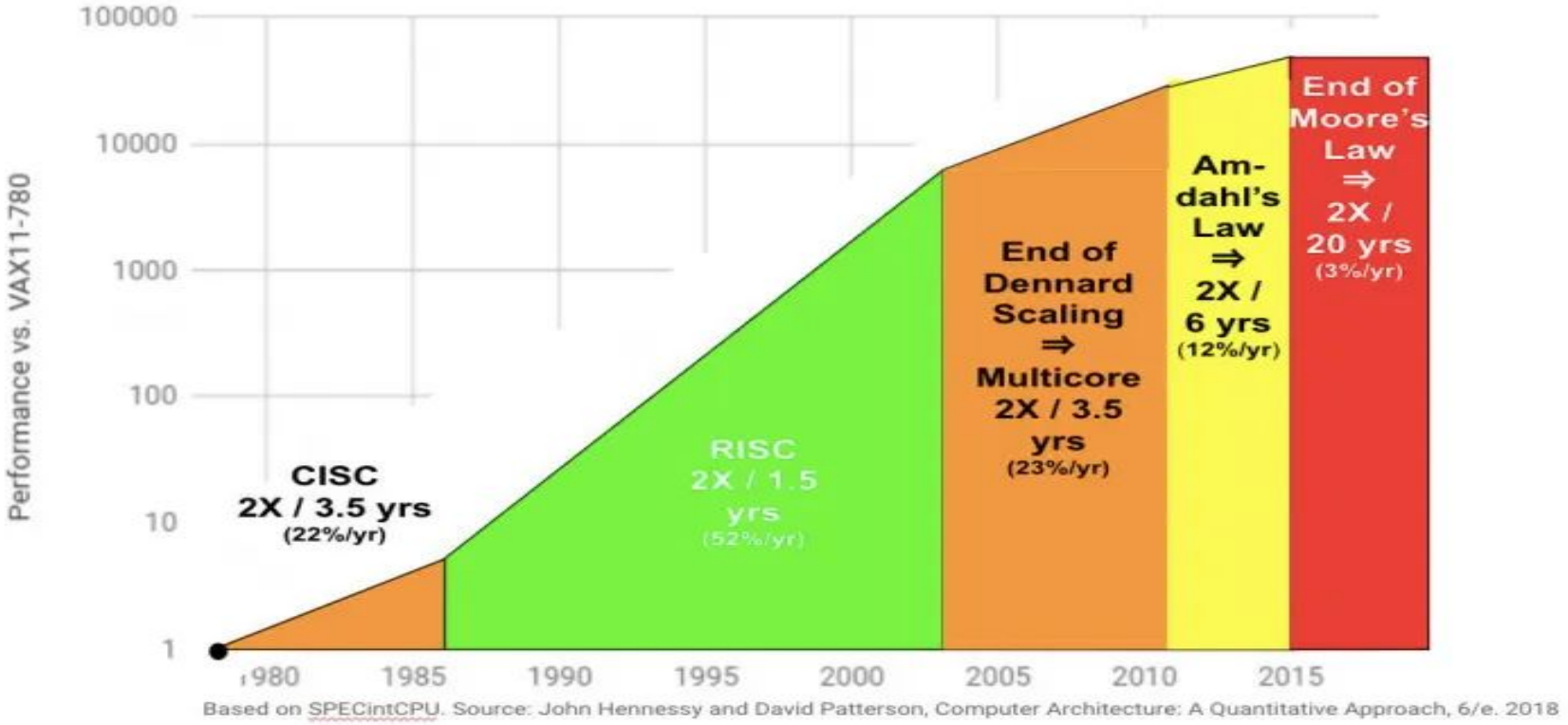
….it also discusses some of the ways that machine learning may also be able to help with some aspects of the circuit design process. Finally, it provides a sketch of at least one interesting direction towards much larger-scale multi-task models that are sparsely activated and employ much more dynamic, example- and task-based routing than the machine learning models of today.”

CPU -> GPU rationale

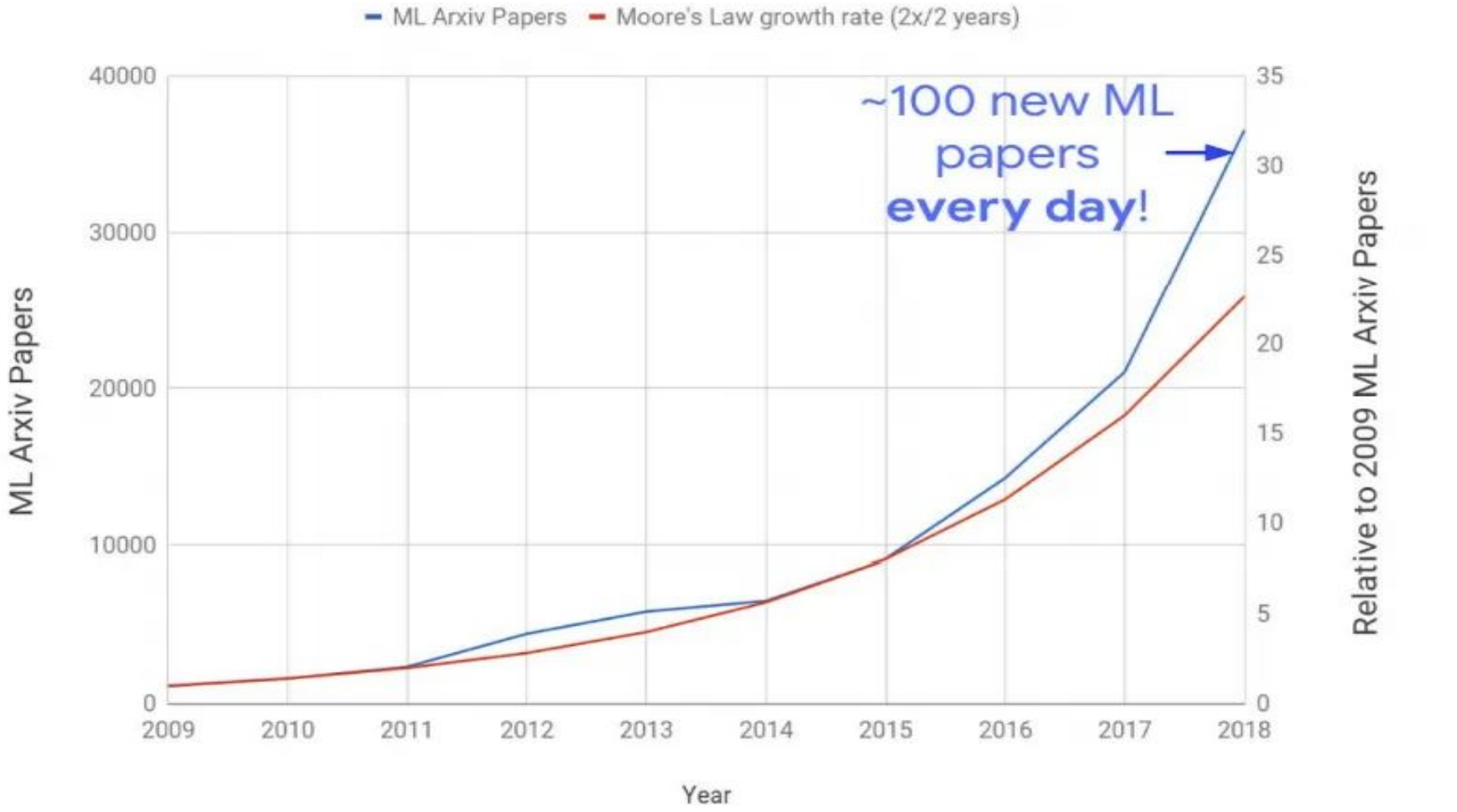
GPU cards' high floating point performance relative to CPUs, started to allow neural networks to show interesting results on difficult problems of real consequence.

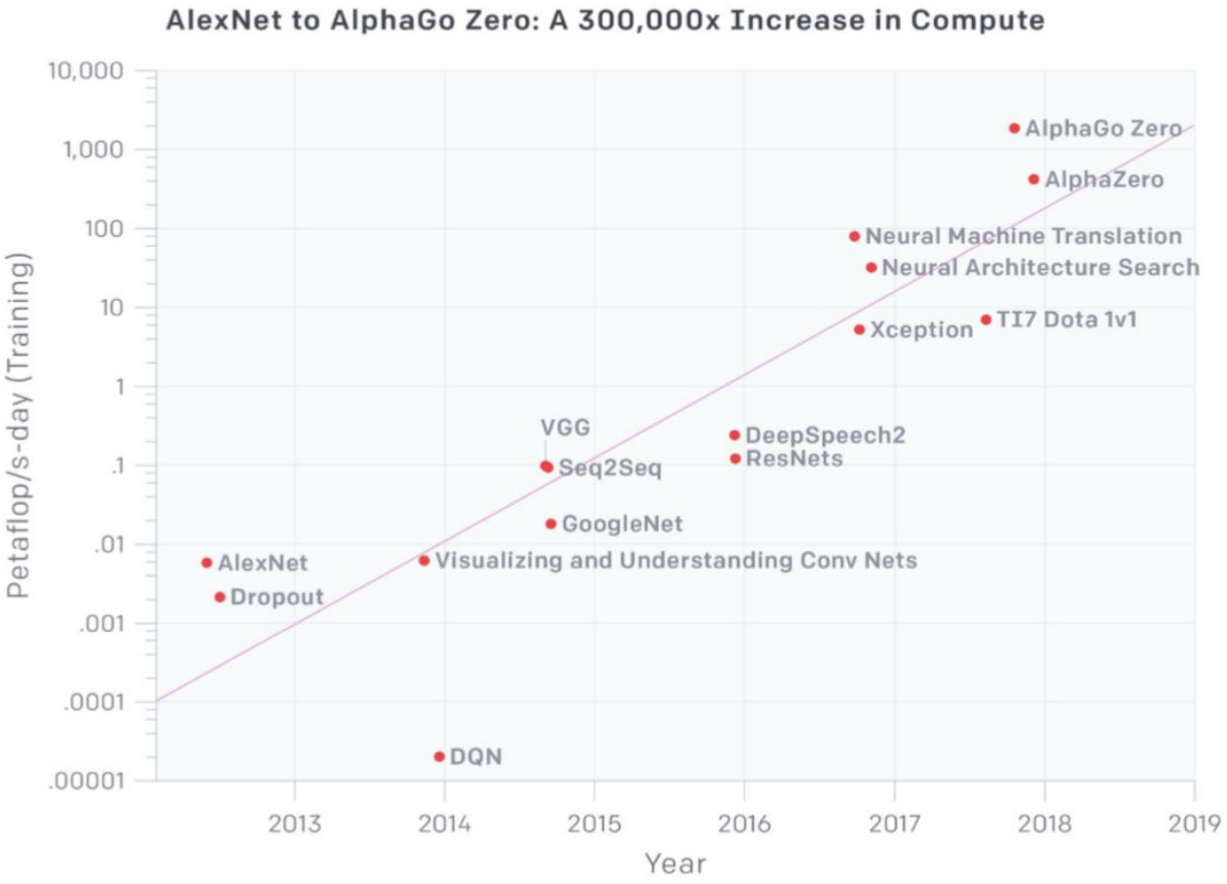
It is perhaps unfortunate that just as we started to have enough computational performance to start to tackle interesting real-world problems and the increased scale and applicability of machine learning has led to a dramatic thirst for additional computational resources to tackle larger problems, the computing industry as a whole has experienced a dramatic slowdown in the year-over-year improvement of general purpose CPU performance.

40 years of Processor Performance



Based on SPECintCPU. Source: John Hennessy and David Patterson, Computer Architecture: A Quantitative Approach, 6/e. 2018


ML Arxiv Papers
Moore's Law growth rate (2x/2 years)
~100 new ML papers every day!
ML Arxiv Papers
Relative to 2009 ML Arxiv Papers
Year
40000
30000
20000
10000
0
35
30
25
20
15
10
5
0
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018


AlexNet to AlphaGo Zero: A 300,000x Increase in Compute
Petaflop/s-day (Training)
10,000
1,000
100
10
1
.1
.01
.001
.0001
.00001
2013
2014
2015
2016
2017
2018
2019
Year
AlphaGo Zero
AlphaZero
Neural Machine Translation
Neural Architecture Search
Xception
TI7 Dota 1v1
VGG
Seq2Seq
DeepSpeech2
ResNets
GoogleNet
AlexNet
Dropout
Visualizing and Understanding Conv Nets
DQN

GPU->ASIC

Deep learning models have three properties that make them different than many other kinds of more general purpose computations:

First, they are very tolerant of reduced-precision computations.

Second, the computations performed by most models are simply different compositions of a relatively small handful of operations like matrix multiplies, vector operations, application of convolutional kernels, and other dense linear algebra calculations [Vanhoucke *et al.* 2011].

Third, many of the mechanisms developed over the past 40 years to enable general-purpose programs to run with high performance on modern CPUs, such as branch predictors, speculative execution, hyperthreaded-execution processing cores, and deep cache memory hierarchies and TLB subsystems are unnecessary for machine learning computations.

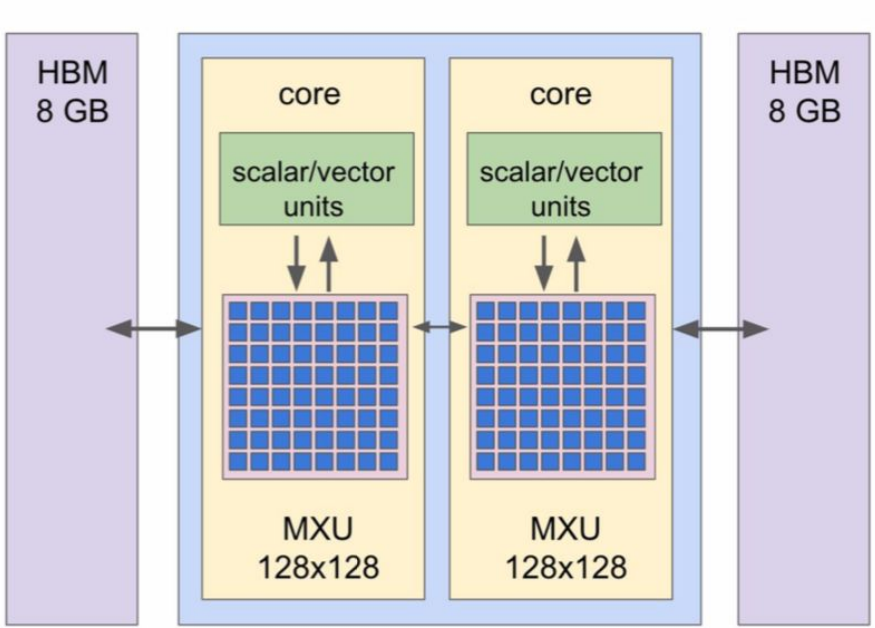


Figure 5: A block diagram of Google's Tensor Processing Unit v2 (TPUv2)

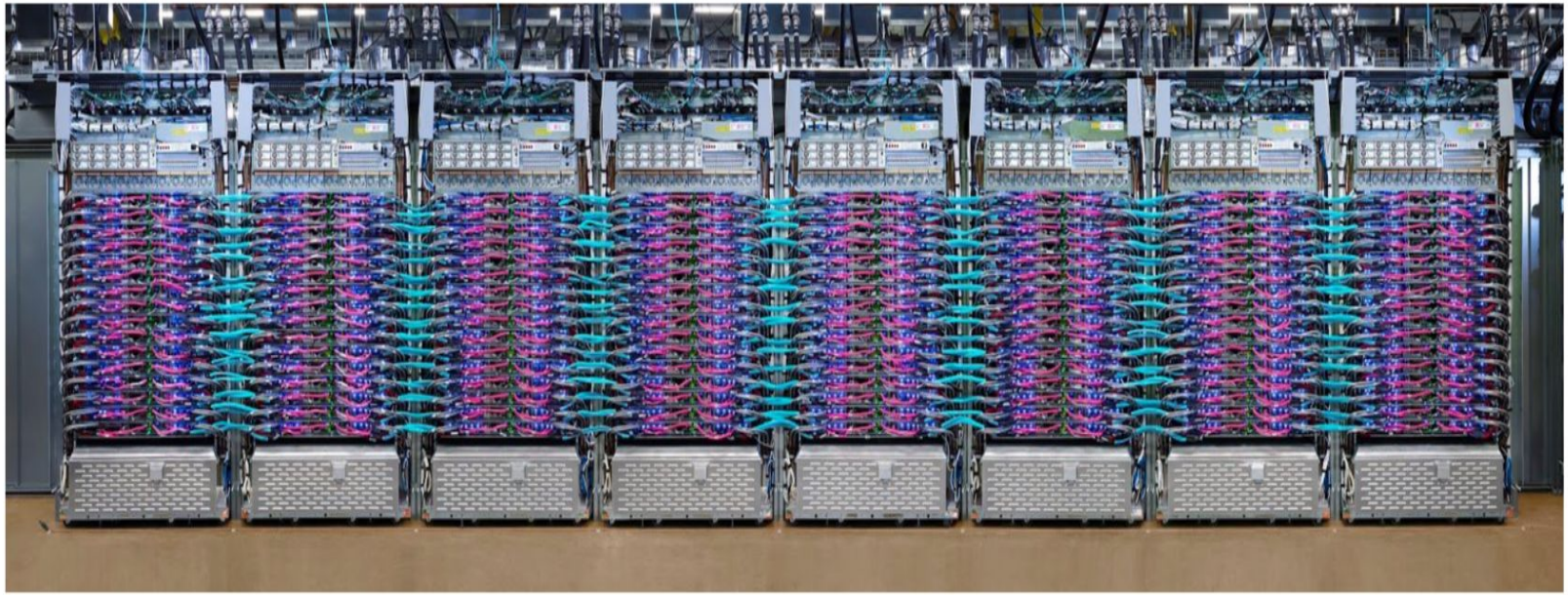


Figure 6: Google's TPUv3 Pod, consisting of 1024 TPUv3 chips w/peak performance of >100 petaflop/s

Future Directions

Work on sparsely-activated models, such as the sparsely-gated mixture of experts model [Shazeer *et al.* 2017], shows how to build very large capacity models where just a portion of the model is “activated” for any given example.

Work on automated machine learning (AutoML), where techniques such as neural architecture search [Zoph and Le 2016, Pham *et al.* 2018] or evolutionary architectural search [Real *et al.* 2017, Gaier and Ha 2019] can automatically learn effective structures and other aspects of machine learning models or components in order to optimize accuracy for a given task.

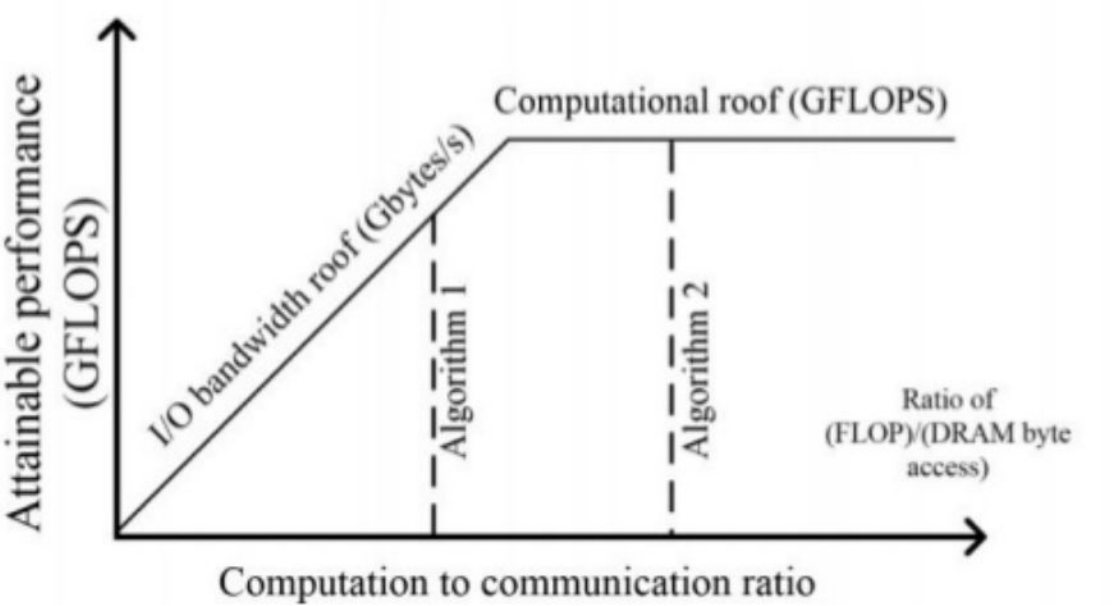
Multi-task training at modest scales of a few to a few dozen related tasks, or transfer learning from a model trained on a large amount of data for a related task and then fine-tuned on a small amount of data for a new task, has been shown to be very effective in a wide variety of problems.

Discussion

1) Each company still manages to publish benchmarks showing their processing unit outperforming at some specific task.
2) Pre-trained models have actually shown the same performance as from scratch models in some domains.
3) There is a growing focus on cost to train, and with these and even more growing focus on cost to infer (think edge computing, or on a mobile phone)

Breadcrumbs for discussion or study

Not only FLOPS: Roofline Performance Model



https://blog.inten.to/hardware-for-deep-learning-part-3-gpu-8906c1644664

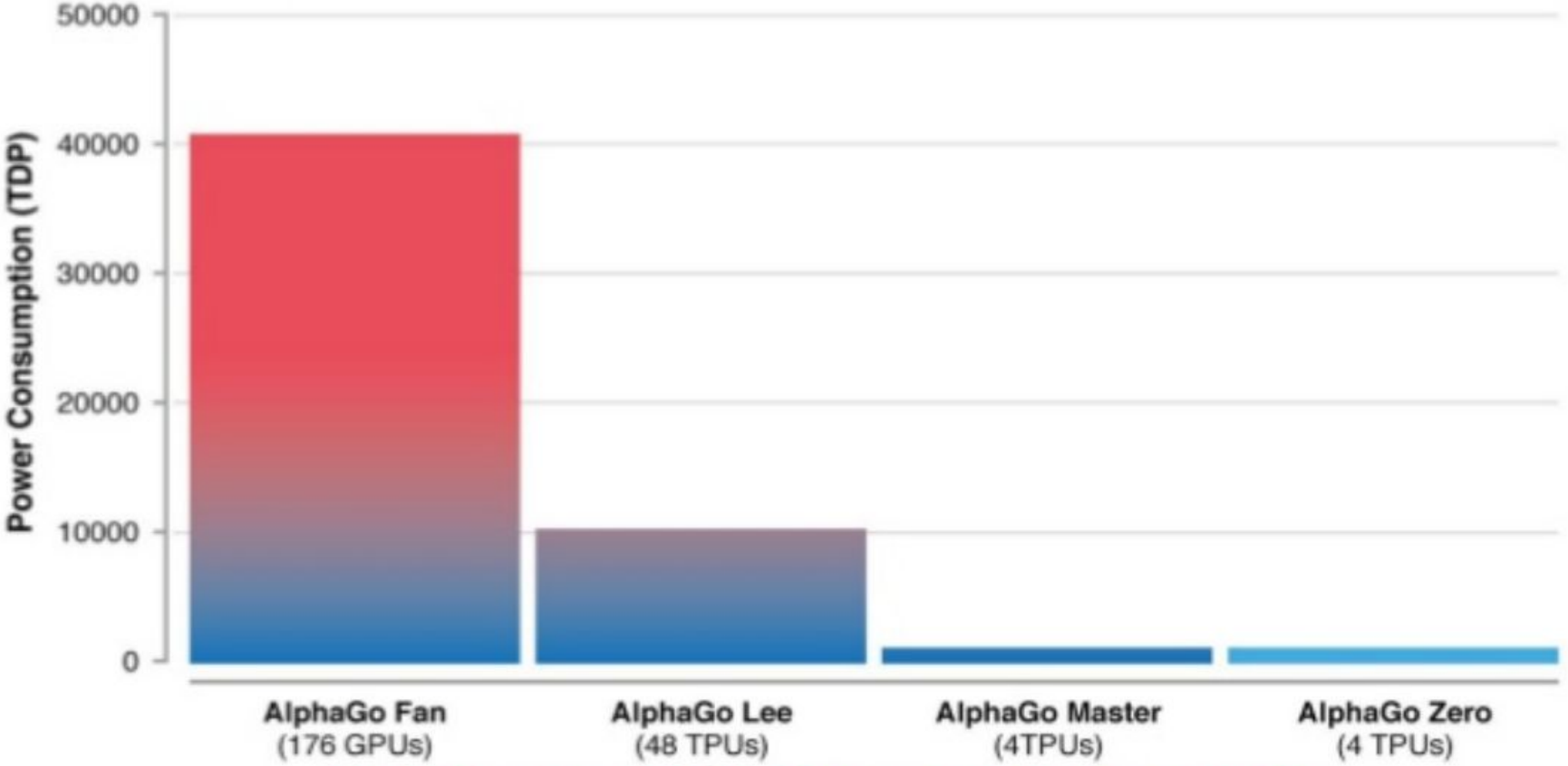
Serious problems with the current processors (CPU/GPU) are:

- **Energy efficiency**:
  - The version of AlphaGo playing against Lee Sedol used 1,920 CPUs and 280 GPUs (https://en.wikipedia.org/wiki/AlphaGo)
  - The estimated power consumption of approximately **1 MW** (200 W per CPU and 200 W per GPU) compared to only **20 watts** used by the human brain (https://jacquesmattheij.com/another-way-of-looking-at-lee-sedol-vs-alphago/)
- **Architecture**:
  - good for matrix multiplication (still the essence of DL)
  - but not well-suitable for brain-like computations

There is a lot of movement to ASIC right now:

- **Google** has Tensor Processing Units (TPU) in the cloud.
- **Intel** just demonstrated their Nervana processors for training and inference.
- **Mobileye** (Intel) chips with specially developed ASIC cores are used in BMW, Tesla, Volvo, etc.
- **Movidius** (acquired by Intel) Myriad X VPU - a dedicated hardware accelerator for deep neural network inferences. https://www.movidius.com/myriadx
- **Alibaba** Hanguang 800
- **Huawei** Ascend 310, 910
- **Bitmain** Sophon
- ...

Case: AlphaGo Zero



https://deepmind.com/blog/alphago-zero-learning-scratch/

Problems

Even with FPGA/ASIC and edge devices:

- **Energy efficiency**:
  - Better than CPU/GPU, but still far from **20 watts** used by the human brain
- **Architecture**:
  - Even more specialized for ML/DL computations, but...
  - Still far from brain-like computations

Neuromorphic chips

- Neuromorphic computing - brain-inspired computing - has emerged as a new technology to enable information processing at very low energy cost using electronic devices that emulate the electrical behaviour of (biological) neural networks.
- Neuromorphic chips attempt to model in silicon the massively parallel way the brain processes information as billions of neurons and trillions of synapses respond to sensory inputs such as visual and auditory stimuli.
- DARPA SyNAPSE program (Systems of Neuromorphic Adaptive Plastic Scalable Electronics)
- IBM TrueNorth; Stanford Neurogrid; HRL neuromorphic chip; Human Brain Project SpiNNaker and HICANN.
  https://www.technologyreview.com/s/526506/neuromorphic-chips/