Google

Deep Learning, Special-purpose hardware, and some hard problems

Cliff Young, cliffy@google.com

CCC Panel on AI and Amplifying Human Abilities

23 October 2017

Who is Cliff?

I think of myself as both a **researcher** and an **engineer**.

Exciting: build what has never been built before.

Harvard Ph.D. 1998, in Computer Science (compilers and computer architecture)

Office of Naval Research,

National Defense Science and Engineering Graduate Fellow (ONR-NDSEG)

My advisor, Michael D. Smith, had a Presidential Young Investigator award.

1997-2003: Bell Labs (Computing Sciences Research)

2003-2013: D. E. Shaw Research, special-purpose computers for molecular dynamics

2013-present: Google, Tensor Processing Units (TPUs) hardware for AI.

Google’s Mission

“Organize the world’s information and make it universally accessible and useful.”

The world’s information: started with the Web,
but the world is increasingly digital and digitized (books, maps, video,...).

Useful: **Search** is a form of human augmentation.
Many other Google products are, too: Assistant, Home, Translate.

Accessible: started with desktops.
Speech makes available on phones, and in the Internet of Things.

Google Services have Federated Structure

“Edge” devices are close to users

desktops and laptops; phones

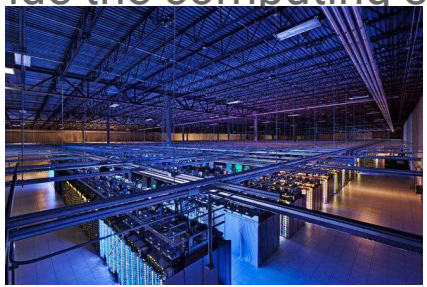
IoT: Nest products, Google Home, smart devices.

Only some of the smarts are in the edge.

“Datacenters” are huge, centralized, warehouse-scale computing.

DCs are some of the original specialization in Google’s computing infrastructure.

DCs provide the computing scale that lets us work with “Big Data”.



Two kinds of Scale required for Deep Learning

Scale in Data

- “Big Data”: huge datasets describe the world we live in.
- The web: over 1 billion pages.
- Google StreetView: 40 million miles of road on Earth; photo every 50 feet.
- YouTube: 5 billion videos watched per day; 300 hours uploaded per minute.

Scale in Compute

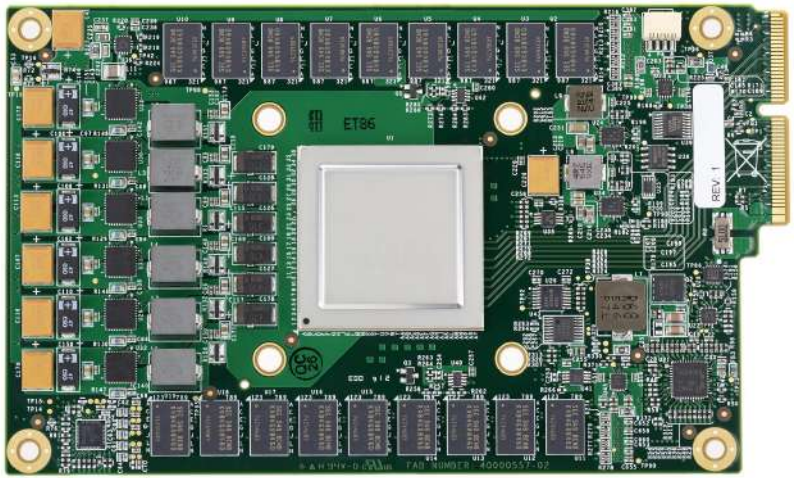
- Huge amounts of processing cycles to analyze these datasets.
- Millions of chips in our datacenters.
- Deep learning uses even more computation than before to do the analysis.


Google Data Center

Google now builds its own chips, “TPUs”

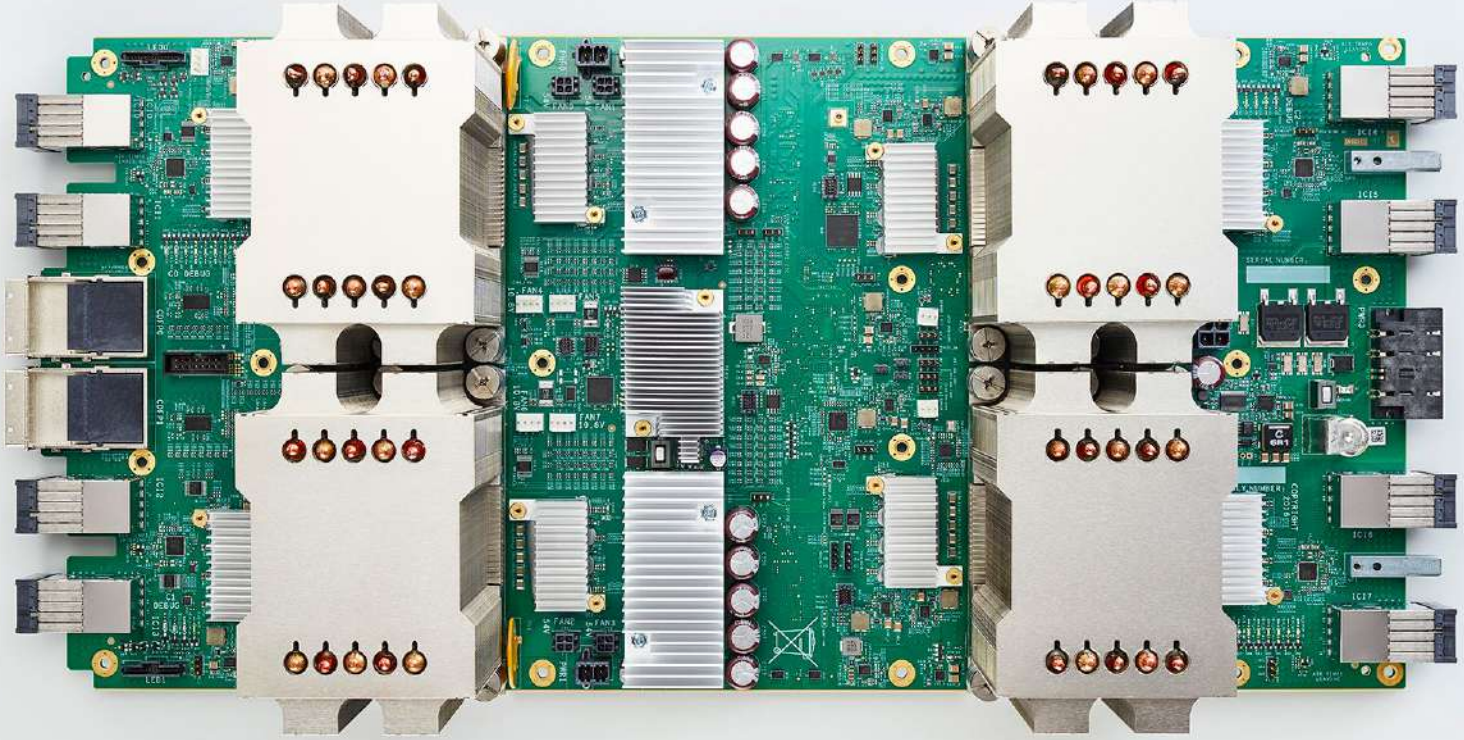
TPU = “Tensor Processing Unit”
Special-purpose hardware, focused on deep-learning calculations only.

First-generation TPU focuses on serving:
Search
Speech
StreetView (Maps) Imagery
Photos
Translation
AlphaGo



“Success disaster” motivation: both development and research

Tensor Processing Unit v2



Google-designed device for neural net **training** and **inference**

TPU Pod
64 2nd-gen TPUs
11.5 petaflops
4 terabytes of HBM memory

TensorFlow
RESEARCH CLOUD

Making 1000 **Cloud TPUs** available **for free** to top researchers who are committed to **open machine learning research**

We’re excited to see what researchers will do with much more computation!
g.co/tpusignup

Industry, Academia, and Government: Complementary Roles

Most of Industry does immediate development.

- 6-month timeframe. Even the first TPU had a 15-month timeframe.
- Very few industrial efforts look even 5 years out (some parts of Google do).

Academia can and should look farther out.

- This is inherently risky: the AI field has a term, “AI winter”.
- Geoffrey Hinton and other deep learning researchers spent **decades**.
- How do you pick the next field like deep learning? You can’t.

Academia trains the next generation of engineers and scientists.

Government is in the business of **portfolio** management.

- Diversify, and make a mixture of bets with different time-frames and risks.