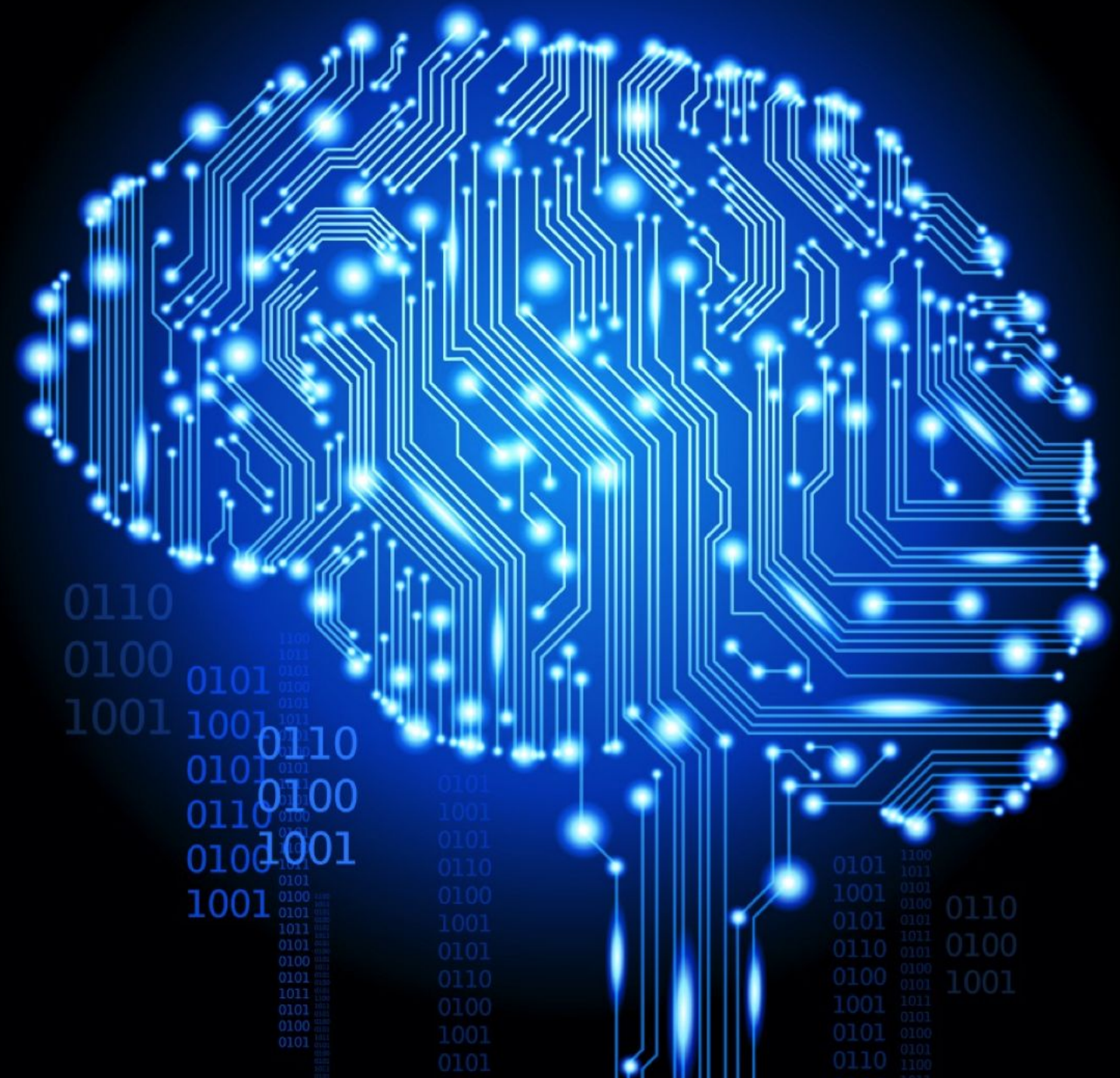
Neural Networks Deployment

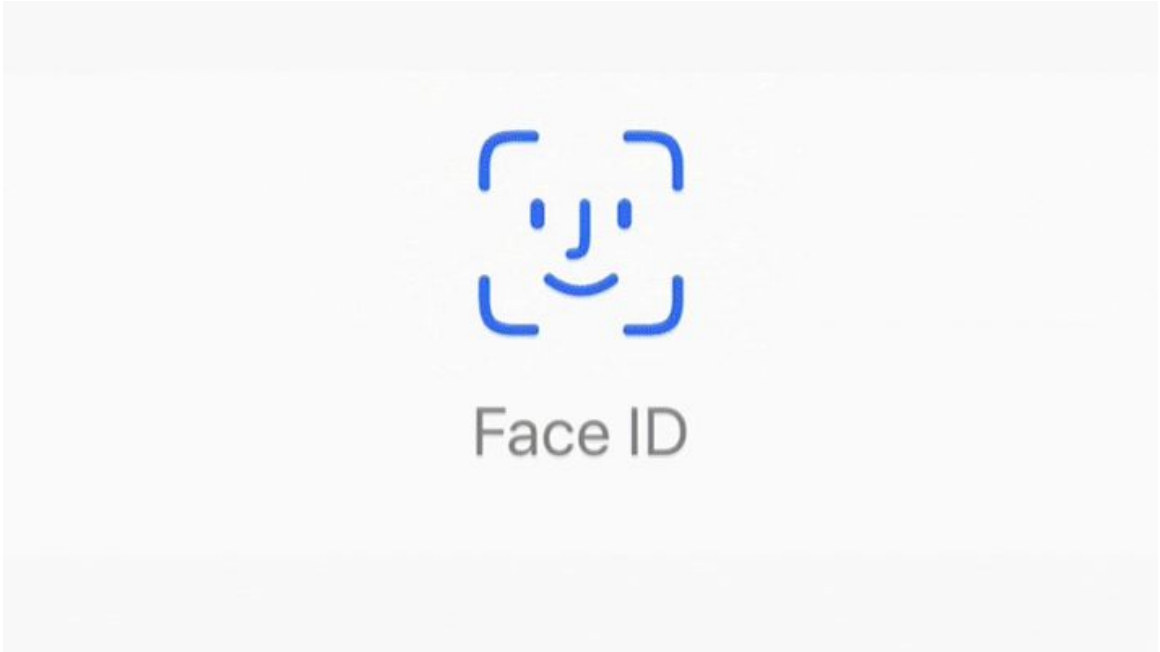
ESADE - MIBA (FALL 2017)



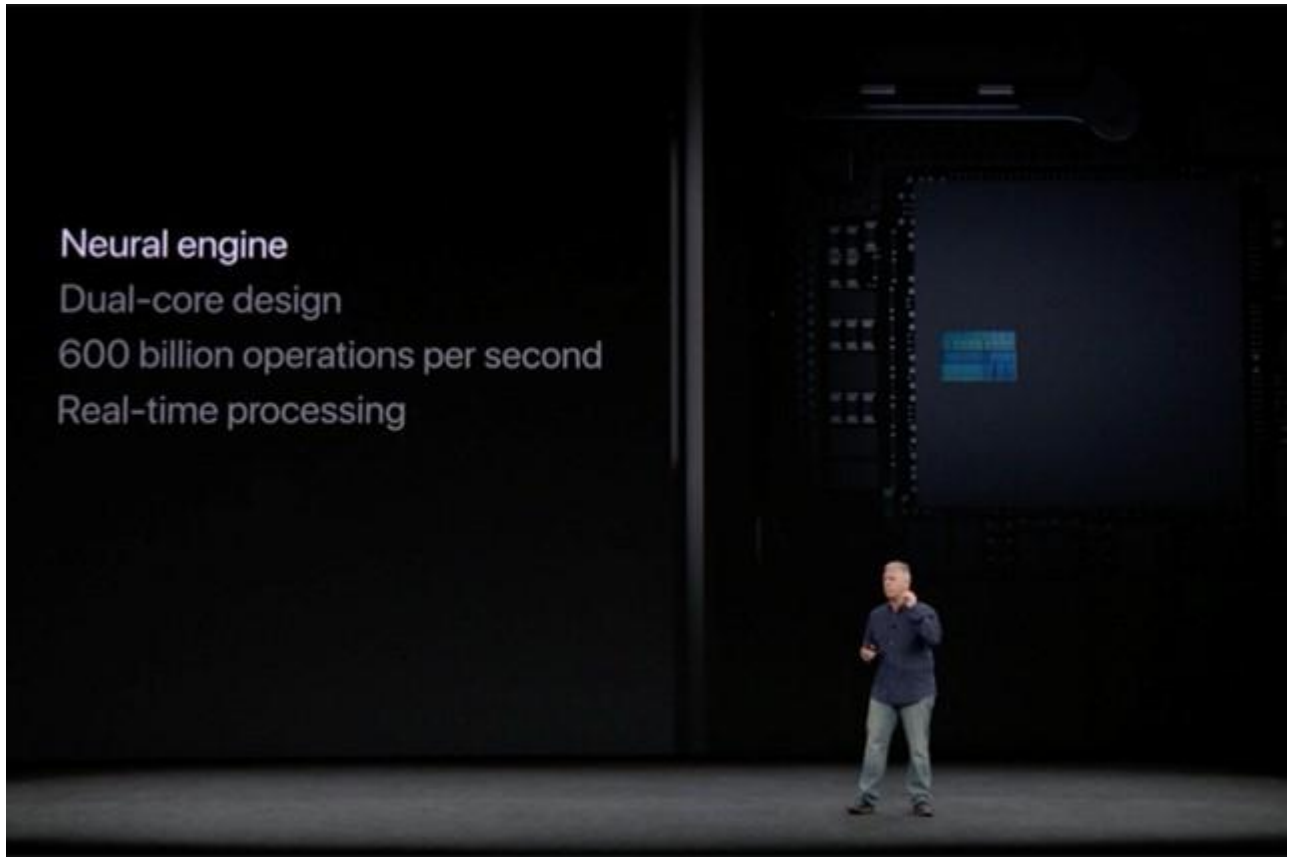
JORDI **TORRES** | FRANCESC **SASTRE**

Deploy Neural Network

- Research is not everything
- Need to apply the power of NN to commercial products



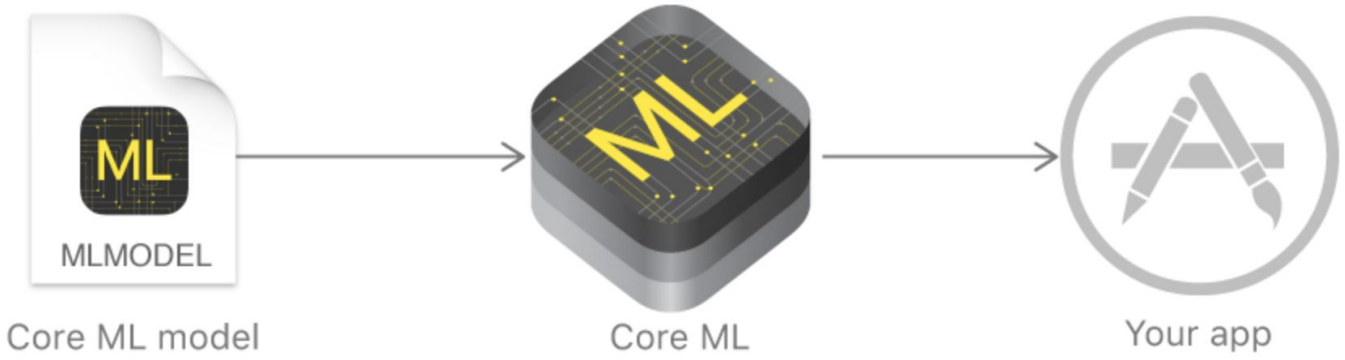
Case study: Apple - A11 Bionic

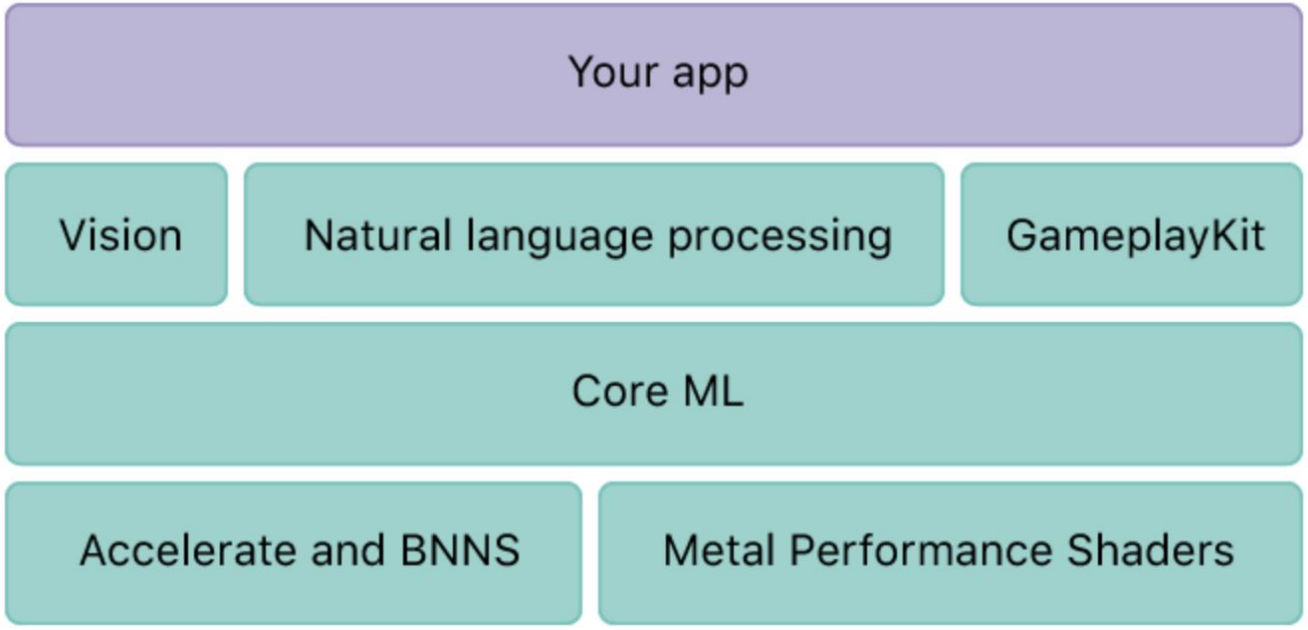


- FaceID
- Augmented Reality
- **CoreML**

Source: Apple

Case study: Apple - Core ML

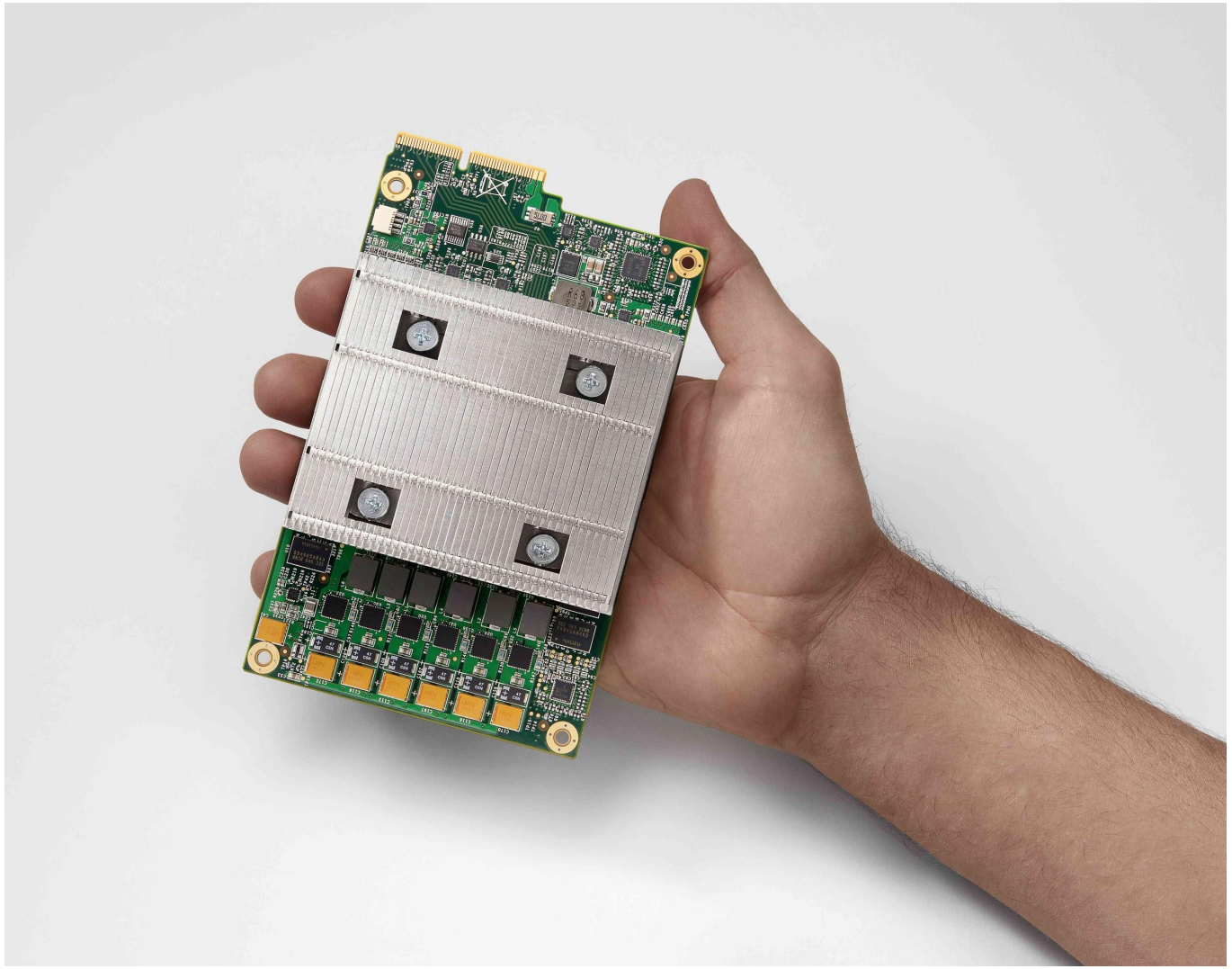




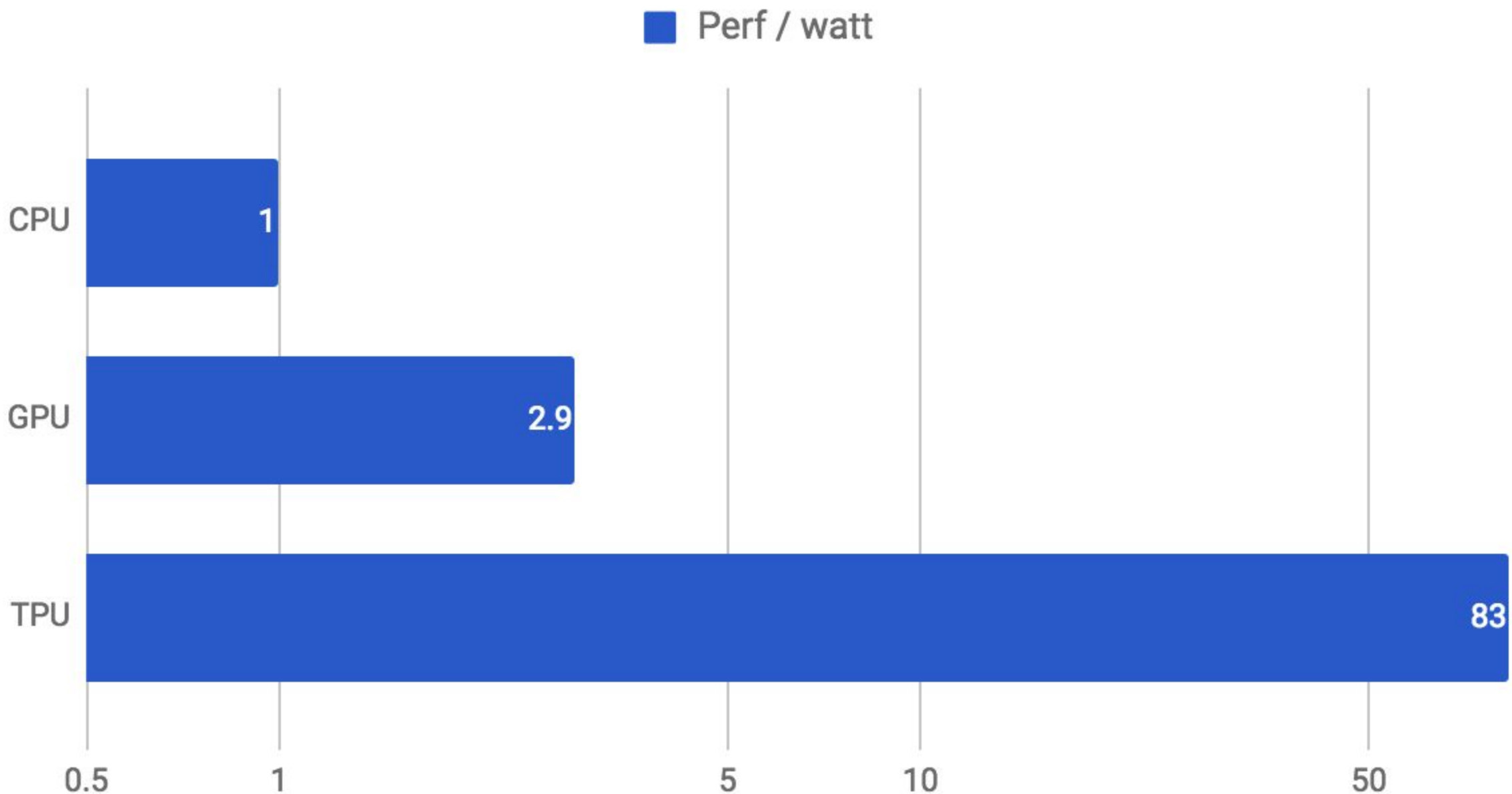
Source: Apple

Case study: Google - TPU v1

- Google made
- Inference
- Less consumption than GPU
- Same performance as GPU
- Less space than GPU
- Used on search queries, for neural machine translation, for speech, for image recognition, for AlphaGo match...

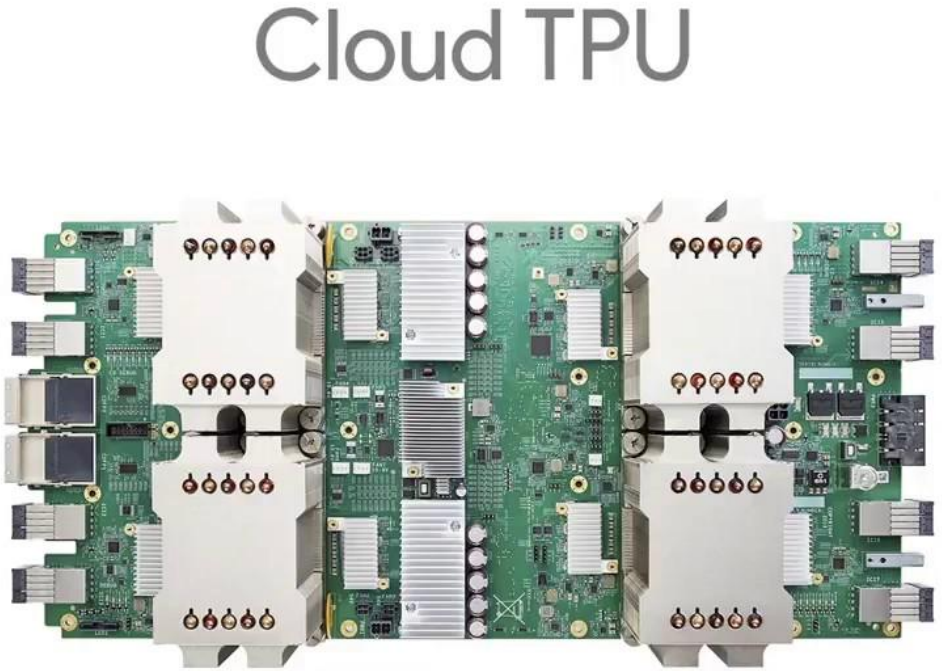


Source: Google Brain


Perf / watt
CPU
1
GPU
2.9
TPU
83
0.5
1
5
10
50

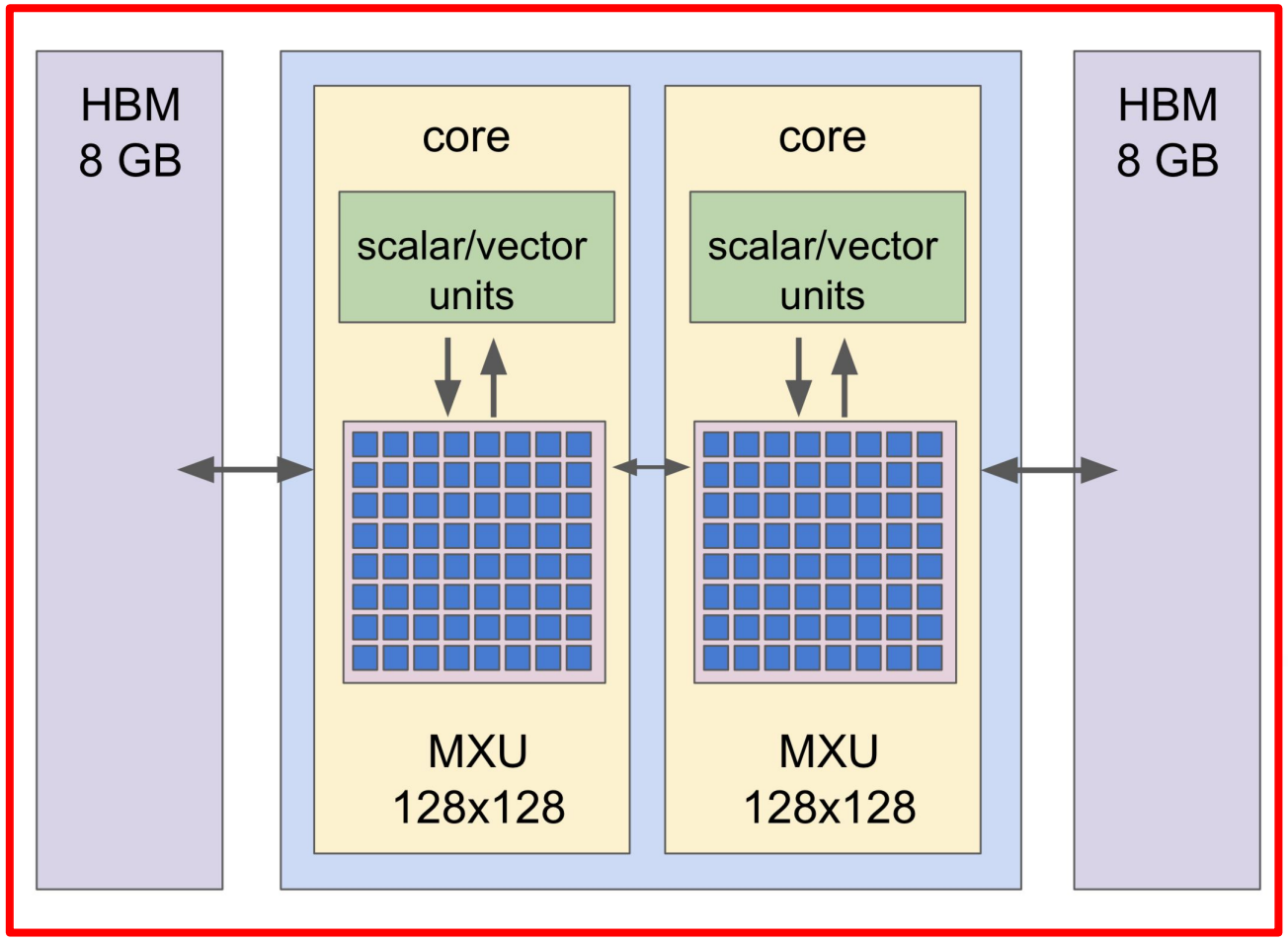
Case study: Google - TPU v2

- Google made
- Training and inference
- Less consumption than GPU
- Same performance as GPU
- Less space than GPU
- Available on Google Cloud
- Keras | TensorFlow code works out of the box

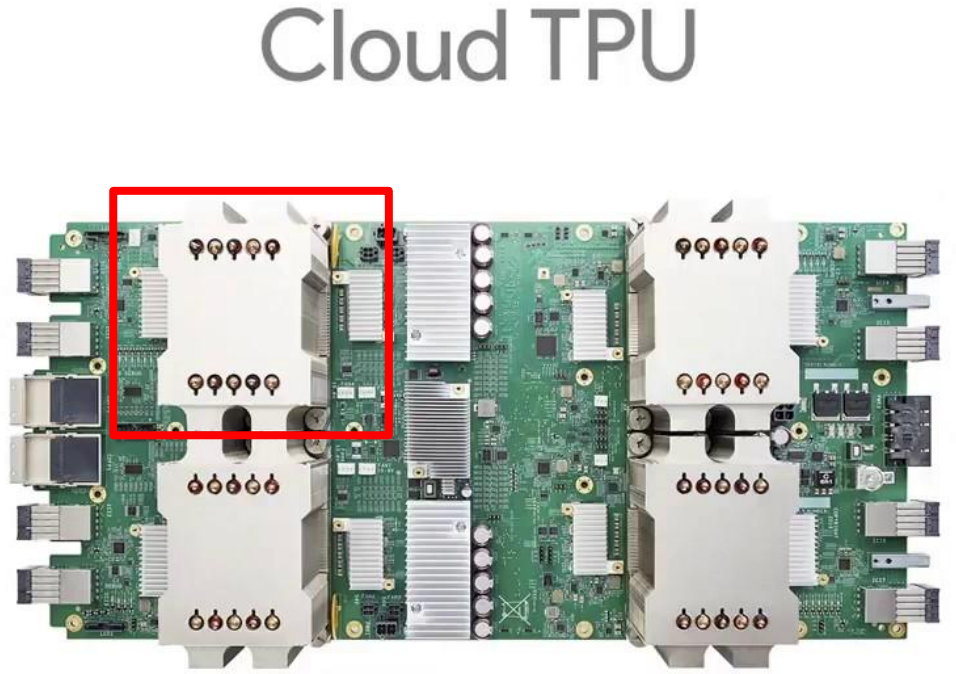


Source: Google Brain

Case study: Google - TPU v2



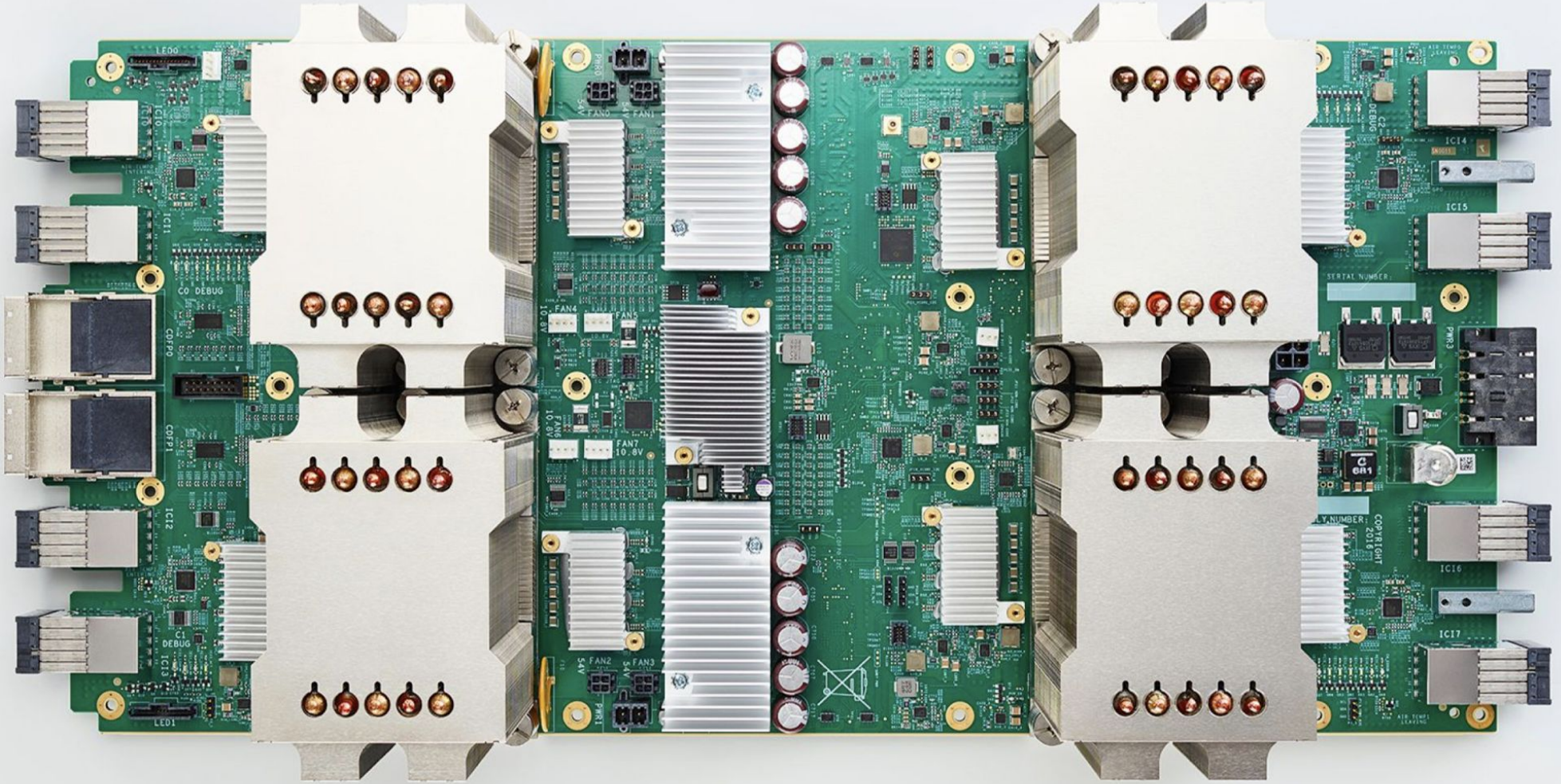
Source: Google Brain



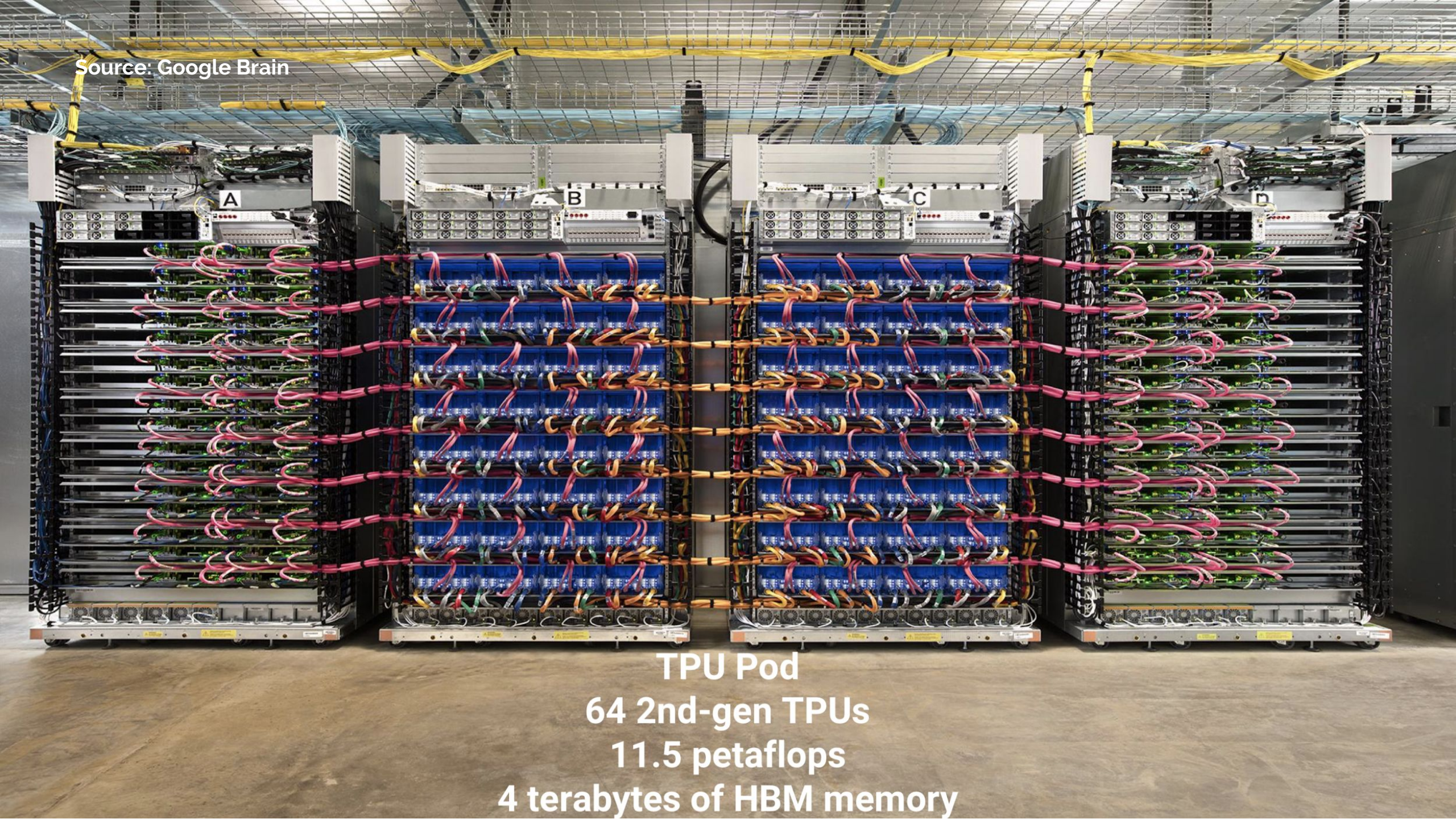
- 45 TFLOPS
- 600 GB/s BW
- FLOAT 32 except multiplications

Source: Google Brain

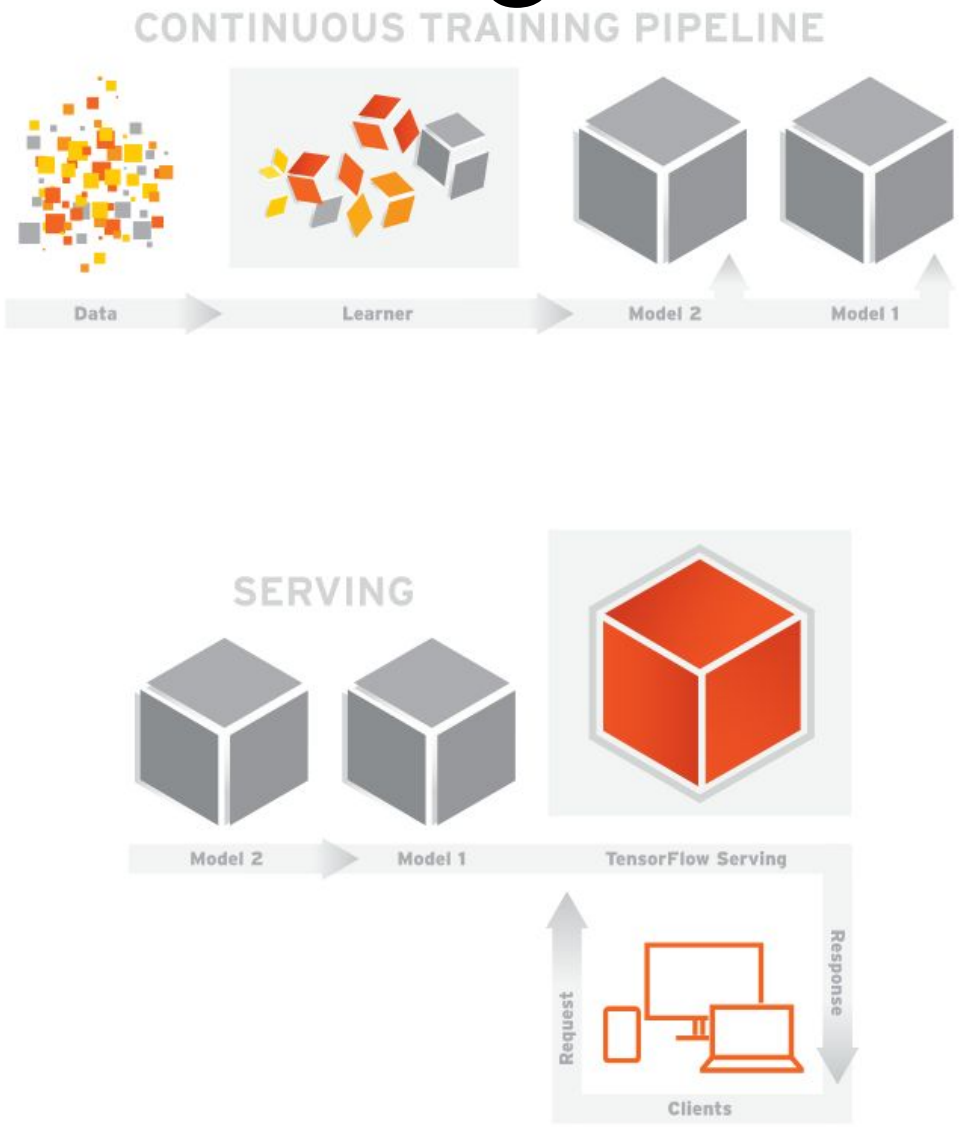
Tensor Processing Unit v2



- 180 teraflops of computation, 64 GB of HBM memory, 2400 GB/s mem BW
- Designed to be connected together into larger configurations


Source: Google Brain
A
B
C
D
TPU Pod
64 2nd-gen TPUs
11.5 petaflops
4 terabytes of HBM memory

TensorFlow Serving

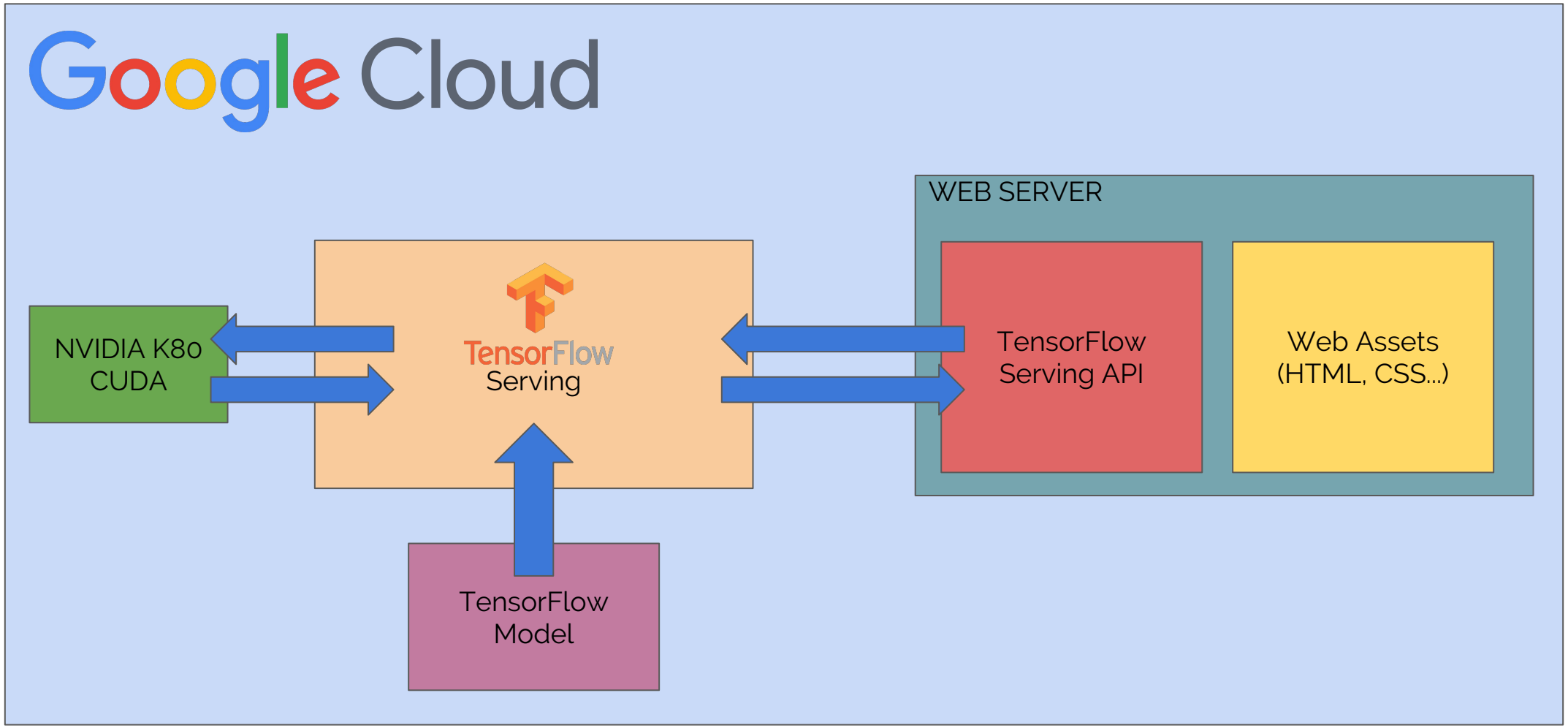


Source: TensorFlow docs

Example: Web Classifier

- Webpage that uses a NN to classify images
- Use TensorFlow serving
- Inception V3
- Google Cloud

Example: Web Classifier



Example: Web Classifier

Create a Google Cloud Server

http://jorditorres.org/dl-with-keras-on-gpu-cloud/

Example: Web Classifier

- Webserver
  - Python
  - Flask
  - No BD needed
  - Manage image upload
  - TensorFlow Serving API
    - Connect to TensorFlow Serving

Example: Web Classifier

Webserver code:

https://github.com/jorditorresBCN/ESADE-MIBA-2017/tree/master/webserver

Example: Web Classifier

- TensorFlow Serving
  - https://www.tensorflow.org/serving/
  - Import model
    - https://github.com/tensorflow/serving/blob/master/tensorflow_serving/example/inception_saved_model.py
  - Configure run parameters

ESADE MIBA 2017

Image classifier

Upload an image

Select a file

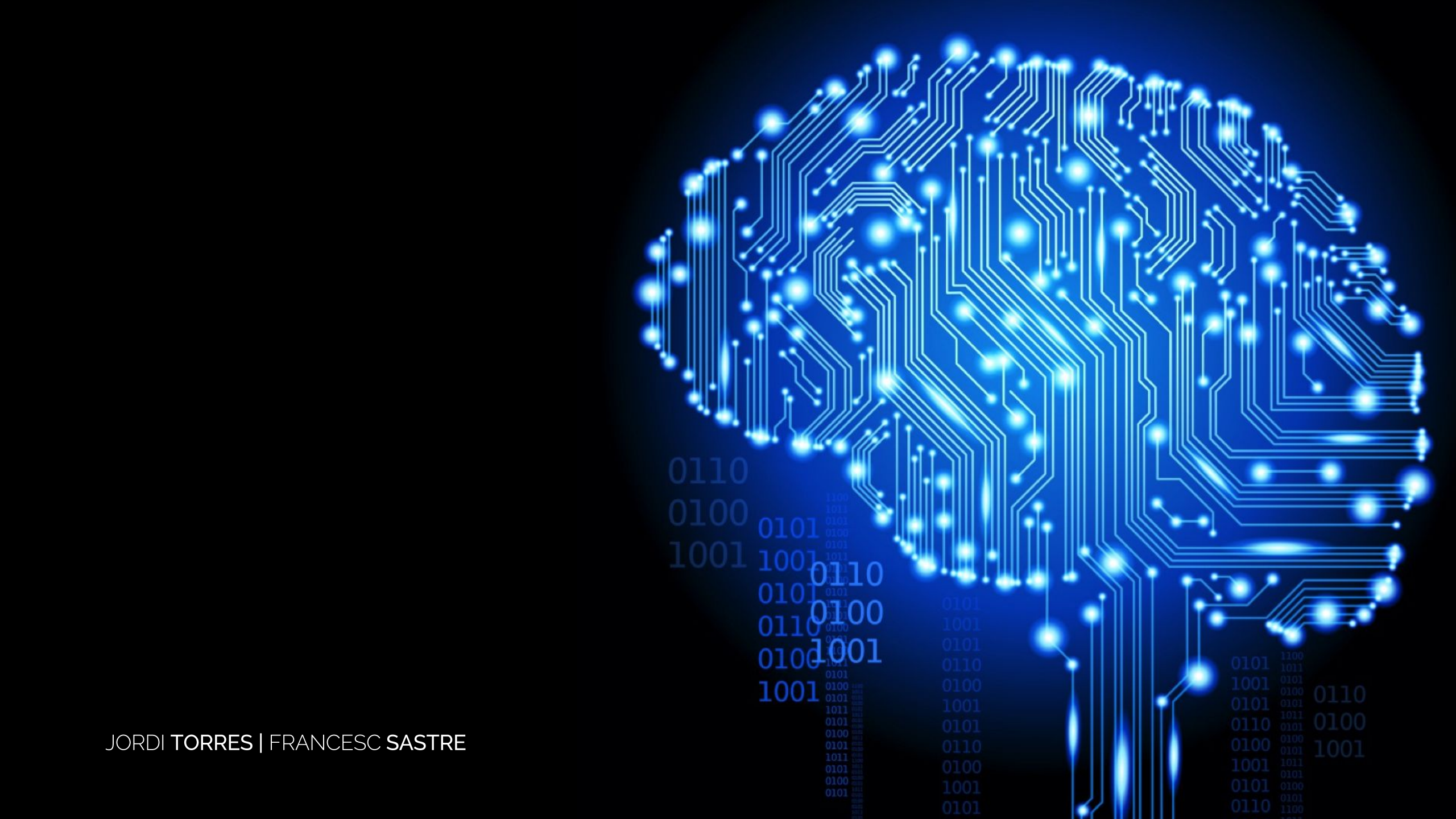
Browse... 04126_theoverlookoflakelucern_1920x1080.jpg

Upload

Image

Results

- alp
- valley, vale
- cliff, drop, drop-off
- lakeside, lakeshore
- mountain tent



JORDI **TORRES** | FRANCESC **SASTRE**