Artificial Intelligence Chips: From Data Centers to Edge and IoT Computing

Version 1.0, 2018

Abderazek Ben Abdallah
Adaptive Systems Laboratory
benab@u-aizu.ac.jp
AI Hardware is … everywhere

Self-driving Car

Machine Translation

Smart Robots

Gaming

Bottom Image source: edition.cnn.com

Image source: roboticsbusinessreview.com

Bottom Image source: missqt.com

Bottom Image Source: newatlas.com
AI Hardware is ... everywhere

Self-driving Car

- Stereo vision camera
  1) Provides distance information
  2) Provides image information
- 3D LIDAR
  Generates a vehicle's local Environment Condition in 3D.
  The radius is around 100m.
- GPS/DGPS/RTK
  Tracking the location of the vehicle by radio signals from satellites.
- IMU
  Estimating the self-position by accelerometer, gyro, the magnetic sensor.
- In-wheel motors
  0.29[kw]×2
- Shaft encoder
  Provides input for the odometry component

Bottom Image source: edition.cnn.com

Smart Robots

Bottom Image source: roboticsbusinessreview.com

Machine Translation

Bottom Image source: missqt.com

Gaming

Bottom Image Source: newatlas.com
Brain implant allows paralysed monkey to walk

There really is a kind of intelligence inside the spinal cord. We are not just talking about reflexes that automatically activate muscles. In the spinal cord there are networks of neurons able to take their own decisions.

-Grégoire Courtine-
Neuroscientist, Federal Institute of Technology, Lausanne

PARALYSED PRIMATES WALK
A wireless implant bypasses spinal-cord injuries in monkeys, enabling them to move their legs.

Nature volume 539, pages 284–288 (10 November 2016)
AI Revenue & GDP Growth Rate in 2035 comparing Baseline Growth to AI scenario

Artificial Intelligence Revenue, World Markets: 2016-2025

Annual growth rates in 2035 of gross value added (a close approximation of GDP), comparing baseline growth in 2035 to an artificial intelligence scenario where AI has been absorbed into the economy.

Source: Accenture and Frontier Economics

AI Revenue & GDP Growth Rate in 2035 comparing Baseline Growth to AI scenario

Annual growth rates in 2035 of gross value added (a close approximation of GDP), comparing baseline growth in 2035 to an artificial intelligence scenario where AI has been absorbed into the economy

Source: Accenture and Frontier Economics

Governments are competing to establish advanced AI research, seeing AI as a way for greater economic power and influence.
Agenda

- Fundamental Trends

- AI – The 4th Industrial Revolution

- Survey of AI Hardware
  - Cloud AI Hardware, Chips
  - Mobile AI Chips
  - Edge and IoT AI Chips
  - Healthcare AI Chips

- Conclusions
Moore’s law is no longer providing more Compute

Source: Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Communications of the ACM, September 2018, Vol. 61 No. 9, Pages 50-59
Moore’s law is no longer providing more compute

**Dennard scaling:** As transistors get smaller their power density stays constant, so that the power consumption stays in proportion with area: both voltage and current scale (downward) with length (WP).

Source: Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Communications of the ACM, September 2018, Vol. 61 No. 9, Pages 50-59
Major improvements in cost-energy-performance must now come from domain-specific hardware.

**Dennard scaling**: As transistors get smaller their power density stays constant, so that the power consumption stays in proportion with area: both voltage and current scale (downward) with length (WP).

Source: Norman P. Jouppi, Cliff Young, Nishant Patil, David Patterson, Communications of the ACM, September 2018, Vol. 61 No. 9, Pages 50-19
Massive amounts of data is generated

A new style of IT emerging

Every 60 seconds

- 98,000+ tweets
- 695,000 status updates
- 11 million instant messages
- 698,445 Google searches
- 168 million+ emails sent
- 1,820 TB of data created
- 217 new mobile web users
# DNN Compute Requirements is Steadily Growing

<table>
<thead>
<tr>
<th>Metrics</th>
<th>LeNet-5</th>
<th>AlexNet</th>
<th>VGG-16</th>
<th>GoogLeNet (v1)</th>
<th>ResNet-50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top-5 error</td>
<td>n/a</td>
<td>16.4</td>
<td>7.4</td>
<td>6.7</td>
<td>5.3</td>
</tr>
<tr>
<td>Input Size</td>
<td>28x28</td>
<td>227x227</td>
<td>224x224</td>
<td>224x224</td>
<td>224x224</td>
</tr>
<tr>
<td># of CONV Layers</td>
<td>2</td>
<td>5</td>
<td>16</td>
<td>21 (depth)</td>
<td>49</td>
</tr>
<tr>
<td>Filter Sizes</td>
<td>5</td>
<td>3, 5, 11</td>
<td>3</td>
<td>1, 3, 5, 7</td>
<td>1, 3, 7</td>
</tr>
<tr>
<td># of Channels</td>
<td>1, 6</td>
<td>3 - 256</td>
<td>3 - 512</td>
<td>3 - 1024</td>
<td>3 - 2048</td>
</tr>
<tr>
<td># of Filters</td>
<td>6, 16</td>
<td>96 - 384</td>
<td>64 - 512</td>
<td>64 - 384</td>
<td>64 - 2048</td>
</tr>
<tr>
<td>Stride</td>
<td>1</td>
<td>1, 4</td>
<td>1</td>
<td>1, 2</td>
<td>1, 2</td>
</tr>
<tr>
<td># of Weights</td>
<td>2.6k</td>
<td>2.3M</td>
<td>14.7M</td>
<td>6.0M</td>
<td>23.5M</td>
</tr>
<tr>
<td># of MACs</td>
<td>283k</td>
<td>666M</td>
<td>15.3G</td>
<td>1.43G</td>
<td>3.86G</td>
</tr>
<tr>
<td># of FC layers</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td># of Weights</td>
<td>58k</td>
<td>58.6M</td>
<td>124M</td>
<td>1M</td>
<td>2M</td>
</tr>
<tr>
<td># of MACs</td>
<td>58k</td>
<td>58.6M</td>
<td>124M</td>
<td>1M</td>
<td>2M</td>
</tr>
<tr>
<td>Total Weights</td>
<td>60k</td>
<td>61M</td>
<td>138M</td>
<td>7M</td>
<td>25.5M</td>
</tr>
<tr>
<td>Total MACs</td>
<td>341k</td>
<td>724M</td>
<td>15.5G</td>
<td>1.43G</td>
<td>3.9G</td>
</tr>
</tbody>
</table>

Source: Joel Emer, ISCA Tutorial, 2017
What does it mean?

End of Moore’s Law + Exponential Increase in Compute Requirements = Needs New Approach
Current State of the Art in Neural Algorithms HW Computing

Hardware

Domain-specific

- Programmable logic
  - FPGA
    - General; requires HDL
    - Moderate performance & efficiency
  - ASIC
    - Specific: executes STDP
    - HP & efficiency
    - Expensive, 40MB local memory
    - Example: IBM TrueNorth

General-purpose

- Fixed logic
  - CPU
    - Latency oriented
    - Most general; common programming languages
    - Lowest power efficiency and performance
    - Memory separate from chip
    - Example: Google deep learning study
  - GPU
    - Throughput oriented
Current State of the Art in Neural Algorithms HW Computing

- **Domain-specific**
  - Programmable logic
    - FPGA
  - Fixed logic
    - ASIC

- **General-purpose**
  - Latency oriented
    - CPU
  - Throughput oriented
    - GPU

**Hardware**

- **Current Industry Focus**
  - Most general; common programming languages
  - Lowest power efficiency and performance
  - Memory separate from chip
  - Example: Google deep learning study

- **Current State of the Art in Neural Algorithms**
  - General; requires HDL
  - Moderate performance efficiency

- **Specific**
  - Executes STDP
  - HP & efficiency
  - Expensive, 40MB local memory

**ANN**

**SNN**
The only tricky part is getting them to do AI computation quickly and efficiently.

Hardware: Flexibility vs Efficiency

Deployment alternatives for deep neural networks (DNNs) and examples of their implementations. (Image courtesy of Microsoft.)
Agenda

• Fundamental Trends

• **AI – The 4\textsuperscript{th} Industrial Revolution**

• Survey of AI Hardware
  > Cloud AI Hardware, Chips
  > Mobile AI Chips
  > Edge and IoT AI Chips
  > Healthcare AI Chips

• Conclusions
Four Main Factors in Promoting AI/Al HW

- **Algorithm**
  - AI algorithms are being applied to nearly everything we do.

- **Data**
  - Larger data sets and models lead to better accuracy but also increase the computation time.

- **Business**
  - Strong Gov. & Industry Engagements

- **Chip**
  - Growth of computational power

More compute means new solutions to previously intractable problems, i.e. G.
AI model performance scales with dataset size and the # of model parameters, thus necessitating more compute.

**IMAGE RECOGNITION**

- **16X Model**
  - 8 layers
  - 1.4 GFLOP
  - ~16% Error
- **152 layers**
  - 22.6 GFLOP
  - ~3.5% error

- **2012**
  - AlexNet
- **2015**
  - ResNet

**SPEECH RECOGNITION**

- **10X Training Ops**
  - 80 GFLOP
  - 7,000 hrs of Data
  - ~8% Error
- **465 GFLOP**
  - 12,000 hrs of Data
  - ~5% Error

- **2014**
  - Deep Speech 1
- **2015**
  - Deep Speech 2

*Microsoft*

*Dally, NIPS’2016 workshop on Efficient Methods for Deep Neural Networks*
AI Chips and systems are inspired by biology ➔ parallel computation

AI HW is inspired by Nature – Biological neuron

from pinterest.com
AI HW is inspired by Nature – Biological neuron

AI Chips and systems are inspired by biology ➔ parallel computation.

- # of neurons: \( \sim 10^{11} \)
- # of synapses: \( \sim 10^{15} \)
- Power consumption: \( \sim 20 \) W;
- Operating frequency: 10~100 Hz
- Works in parallel: \( 10^6 \) parallelism vs. \( <10^1 \) for PC (VN)
- Faster than current computers: i.e. simulation of a 5 s brain activity takes \( \sim 500 \) s on state-of-the-art supercomputer

Latest digital DL processors:

- \( \sim 10 \) TOPS/W
- Synapse op. in brain: 0.1~1 fJ/op
  - 1,000~10,000 TOPS/W
  - \( \approx 1 \)~10 POPS/W
...there are many topologies for mimicking the brain functions
# Different approaches to AI Chips

<table>
<thead>
<tr>
<th>Poor/Simple</th>
<th>Good/Complex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neuron</td>
<td></td>
</tr>
<tr>
<td>Digital, Analog, LIF</td>
<td>Izhikevich model</td>
</tr>
<tr>
<td></td>
<td>Huxley-Hodgkin model</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Synapse</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC (weighted sum)</td>
<td>Spiking STDP</td>
</tr>
<tr>
<td></td>
<td>Many nonlinear properties</td>
</tr>
</tbody>
</table>

Generally Used in DL algorithms

**Frequency**

10~100 Hz (brain)
Current AI Chip = Accelerator/Co-processor

Program Code

GPU

Compute-Intensive Functions

Use GPU to Parallelize

Rest of Sequential CPU Code

Acceleration with GPU

CPU
Accelerator Characteristics

### Memory subsystem
- **CPU**: L3, L2, L2, L1D, L1I, implicitly managed
- **GPU**: L2, SM, TX/L1, RF, RF, mixed
- **TPU**: Unified Buffer, FIFO, Acc, explicitly managed

### Compute primitives
- **CPU**: scalar
- **GPU**: vector
- **TPU**: tensor

### Data type
- **CPU**: fp32
- **GPU**: fp16
- **TPU**: int8

[Ref 3]
Deep Learning means using a neural network with several layers of nodes between input & output.

The series of layers between input & output do feature identification and processing in a series of stages, just as our brains seem to.

Deep Learning is considered as a sophisticated “rocket” of Machine Learning!!

Fuel = Data!
Example 1: Character Recognition on FPGA

Implementation of detecting 16 patterns from 16 inputs with BP.

Device: EP2C35F672C6
Family: Cyclone2
Synthesis: Quartus2 13.1

Table 1: ANN Performance Evaluation

<table>
<thead>
<tr>
<th>ALUs</th>
<th>Registers</th>
<th>Pins</th>
<th>Fmax</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,989 (33%)</td>
<td>5,814 (18%)</td>
<td>432 (89%)</td>
<td>76.02 MHz</td>
</tr>
</tbody>
</table>

Memory
DSP Block |
4,956 (1%) | 54 (77%) | 286.84 mW |

‘O’ letter
**Example 2: Handwriting Digit Recognition on FPGA**

**Input**

16 x 16 = 256
Ink → 1
No ink → 0

**Output**

Each dimension represents the confidence of a digit.

The image is “2”

is 1
is 2
is 0
Example of Neural Network

Sigmoid Function

\[ \sigma(z) = \frac{1}{1 + e^{-z}} \]
Example of Neural Network
Example of Neural Network

\[ f : \mathbb{R}^2 \to \mathbb{R}^2 \]

\[ f \left( \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right) = \begin{bmatrix} 0.62 \\ 0.83 \end{bmatrix} \quad f \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right) = \begin{bmatrix} 0.51 \\ 0.85 \end{bmatrix} \]

Different parameters define different function
Matrix Operation

\[
\begin{pmatrix}
1 & -2 \\
-1 & 1
\end{pmatrix}
\begin{pmatrix}
1 \\
-1
\end{pmatrix}
\begin{pmatrix}
1 \\
0
\end{pmatrix} =
\begin{pmatrix}
0.98 \\
0.12
\end{pmatrix}
\]
Neural Network

\[
\sigma(W^1 x + b_1) \rightarrow \sigma(W^2 a_1 + b_2) \rightarrow \sigma(W^L a^{L-1} + b_L)
\]
Parallel computing techniques are needed to speed up matrix operations.
The two phases of NN are called *training* (or learning) and *inference* (or prediction), and they refer to development versus production.

The Developer chooses the number of layers and the type of NN, and training determines the weights.

Virtually all training today is in floating point.

A step called *quantization* transforms floating-point numbers into narrow integers—often just 8 bits—which are usually good enough for inference.

8-bit integer multiplies can be 6X less energy and 6X less area than IEEE 754 16-bit FPMs, and the advantage for integer addition is 13X in energy and 38X in area [Dal16].
A more biological version: LIF/SRM Model

Spike Response Model

Reset of the membrane potential (action potential)

\[ \eta(t) = \eta(t - t_i^\wedge) + \sum_j \sum_{j} w_{ij} \varepsilon(t - t_j^f) \]

Spike reception: EPSP

\[ \varepsilon(t - t_j^f) \]

Spike emission: AP

\[ \eta(t - t_i^\wedge) \]

State of neuron i

\[ u_i(t) = \eta(t - t_i^\wedge) + \sum_j \sum_{j} w_{ij} \varepsilon(t - t_j^f) \]

Firing:

\[ u_i(t) = \emptyset \Rightarrow t_i^\wedge = t \]

[Ref. 18]
A more biological Model: Molecular Basis

-70 mV

Ions/proteins

K⁺

Ca²⁺

Na⁺
Electronic devise vs chemical device

- Deliver the concentration difference of K⁺, Na⁺
- Action potential ~ 70 mV
  - Extreme low voltage operation
  - Noise problem
  - Multiple signal input/ integration

- Spatial and temporal multiplexing ➔ Active sharing of the interconnect
- Chemical computing, extremely low operation voltage (<100mV) ➔ Low power
Hodgkin-Huxley Model

Inside the cell

\[ J_c = C_m \frac{\partial V_m}{\partial t} \]
\[ J_{K^+} = G_{K^+} \left( V_m - V_{K^+} \right) \]
\[ J_{Na^+} = G_{Na^+} \left( V_m - V_{Na^+} \right) \]
\[ J_L = G_L \left( V_m - V_L \right) \]

\[ J_m = J_c + J_{K^+} + J_{Na^+} + J_L \]

Outside the cell

Ion channels

Ion pump

\[ J_m = C_m \frac{\partial V_m}{\partial t} + G_{K^+} \left( V_m - V_{K^+} \right) + G_{Na^+} \left( V_m - V_{Na^+} \right) + G_L \left( V_m - V_L \right) \]

[Ref. 18]
Hodgkin-Huxley Model

\[ J_c = C_m \frac{\partial V_m}{\partial t} \]

\[ J_{Na^+} = G_{Na^+} (V_m - V_{Na^+}) \]

\[ J_{K^+} = G_{K^+} (V_m - V_{K^+}) \]

\[ J_L = G_L (V_m - V_L) \]

\[ J_m = J_c + J_{K^+} + J_{Na^+} + J_L \]

\[ J_m = C_m \frac{\partial V_m}{\partial t} + G_{K^+} (V_m - V_{K^+}) + G_{Na^+} (V_m - V_{Na^+}) + G_L (V_m - V_L) \]

[Ref. 18]
The electrical resistor is not constant but depends on the history of current that had previously flowed through the device.

- Voltage **pulses** can be applied to a **memristor** to change its **resistance**, just as **spikes** can be applied to a **synapse** to change its **weight**.
Wiring via AER – address Event Representation

(Courtesy: iStock/Henrik5000)
Spike-timing-dependent plasticity (STDP)

- Adjusts the strength of connections between neurons in the brain.
  - Adjusts the connection strengths based on the relative timing of a particular neuron's output and input action potentials.
Agenda

• Fundamental Trends

• AI – The 4th Industrial Revolution

• Survey of AI Hardware
  - Cloud AI Hardware, Chips
  - Mobile AI Chips
  - Edge and IoT AI Chips
  - Healthcare AI Chips

• Conclusions
Big Corps AI Chips

**NVIDIA**
2017, 05
NVIDIA launches its Volta GPU computing architecture to boost AI inference, training and HPC.

**IBM**
2017, 07
IBM and the US AFRL announce a collaboration on a brain-inspired supercomputing system.

**Huawei**
2017, 09
Huawei introduces Kirin 970, its new flagship SoC with AI capabilities.

**Google**
2017, 06
Google introduces its TPU (Tensor Processing Units) which accelerate the TensorFlow framework in machine learning.

**Microsoft**
2017, 07
Microsoft announces that it is working on a processor for the second generation of HoloLens. The chip will enhance the AR headset's image recognition feature.

**Amazon**
2018, 02
Amazon is developing a new processor for its virtual assistant Echo to make Alexa faster and smarter.

[source: medium.com](https://medium.com)
There are two AI Chip Models: ANN and SNN

- The output of ANN Chip depends only on the current stimuli, the output of SNN depends on previous stimuli also.
- The SNN/Neuromorphic Chip operates on biology-inspired principles to improve performance and increase energy efficiency.

<table>
<thead>
<tr>
<th>Neuron</th>
<th>Digital, Analog, LIF</th>
<th>Izhikevich model</th>
<th>Huxley-Hodgkin model</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Synapse</th>
<th>MAC (weighted sum)</th>
<th>Spiking STDP</th>
<th>Many nonlinear properties</th>
</tr>
</thead>
</table>

Generally Used in DL algorithms

Frequency: 10~100 Hz (brain)
Training & Inference

TRAINING
Learning a new capability from existing data

INFEERENCE
Applying this capability to new data

Untrained Neural Network Model
Deep Learning Framework

TRAINING DATASET

Trained Model New Capability

NEW DATA
App or Service Featuring Capability

FPGA, GPU, Cloud
CPU, FPGA, GPU, ASIC
Neuromorphic Sensors - electronic models of retinas and cochleas.

Smart sensors — tracking chips, motion, pressor, auditory classifications and localization sensors.

Models of specific systems: e.g. lamprey spinal cord for swimming, electric fish lateral line.

Pattern generators — for locomotion or rhythmic behavior

Large-scale multi-core/chip systems — for investigating models of neuronal computation and synaptic plasticity.

Neurogrid (Stanford)

Brainscales/HBP (Heidelberg, Lausanne)

SpiNNaker (Manchester)

TrueNorth (IBM)
Example
Loihi AI-Chip - a 60-mm² chip fabricated in Intel’s 14-nm

Technology: 14nm
Die Area: 64 mm²
Core area: 0.41 mm²
NMC cores: 128 cores
x86 cores: 3 LMT cores
Max # neurons: 128K neurons
Max # synapses: 128M synapses
Transistors: 2.07 billion

Neuromorphic core
- LIF neuron model
- Programmable learning
- 128 KB synaptic memory
- Up to 1,024 neurons
- Asynchronous design

Parallel off-chip interfaces
- Two-phase asynchronous
- Single-ended signaling
- 100-200 MB/s BW

Embedded x86 processors
- Efficient spike-based communication with neuromorphic cores
- Data encoding/decoding
- Network configuration
- Synchronous design

Low-overhead NoC fabric
- 8x16-core 2D mesh
- Scalable to 1000's cores
- Dimension order routed
- Two physical fabrics
- 8 GB/s per hop

Cloud AI-Chips
Acceleration enterprise AI with DL Cloud
Custom ASIC: Tensor Processing Unit (TPU)

TPU is deployed in datacenters since 2015 that accelerates the **inference** phase of neural networks (NNs).

**Floor Plan of TPU die**

**TPU Printed Circuit Board**

**Source: TensorFlow.org**

TPU chip runs at only 700 MHz

**Source: In-datacenter Performance Analysis of a Tensor Processing Unit Jouppi et al, ISCA, 6/2017**
Custom ASIC: Tensor Processing Unit (TPU)

TPU is deployed in datacenters since 2015 that accelerates the inference phase of neural networks (NNs).

• The TPU board can perform 92 TeraOps/s (TOPS). It is 15 to 30 times faster than CPUs and GPUs tasked with the same work, with a 30- to 80-fold improvement in TOPS/W.
• The software used for comparison of systems was the TensorFlow framework.

Experience Cloud TPU: https://github.com/tensorflow/tpu
https://cloud.google.com/tpu/docs

Source: In-datacenter Performance Analysis of a Tensor Processing Unit Jouppi et al, ISCA, 6/2017
The portion of the application run on the TPU is typically written in TensorFlow and is compiled into an API that can run on GPUs or TPUs.

- The TPU has a CISC-like instructions set:
  - Read_Host
  - Read_Weights
  - MatrixMultiply/Convolve
  - Activate
  - Write_Host
TPU is based on the Systolic Array Idea

The matrix unit uses **systolic execution** to save energy by reducing reads and writes of the **Unified Buffer**.

**Benefit**: Maximizes computation done on a single piece of data element brought from memory.

**Systolic data flow of the Matrix Multiply Unit.**
SW has the illusion that each 256B input is read at once, and they instantly update one location of each of 256 accumulator RAMs.

Similar to blood flow: **heart -> many cells -> heart**
Memory: heart
Data: blood
PEs: cells

Source: *In-datacenter Performance Analysis of a Tensor Processing Unit* Jouppi et al, ISCA, 6/2017
NN Training Works with Low-precision FP

**fp32: Single-precision IEEE Floating Point Format**

Exponent: 8 bits  
Mantissa (Significand): 23 bits

Range: \((10^{-45})\) to \((10^{38})\)

**fp16: Half-precision IEEE Floating Point Format**

Exponent: 5 bits  
Mantissa (Significand): 10 bits

Range: \(10^{-8}\) to 65504

- Represent the same range of numbers of fp32 just at a much lower position.
- It turns out that we don’t need all that precision for NN training, but we do actually need all the range.

**bfloat16: Brain Floating Point Format**

Exponent: 8 bits  
Mantissa (Significand): 7 bits

Range: \((10^{-45})\) to \((10^{38})\)
• One technique exploited by the new chips is using **low-precision**, often fixed-point data, **eight bits** or even fewer, especially for inference.

• One of the major open questions in all of this as far as hardware accelerators are concerned is **how far can you actually push this down** without losing classification accuracy?

• Results from **Google, Intel, and others** show that such low-precision computations can be very powerful when the data is prepared correctly, which also opens opportunities for novel electronics.
What are the differences between the three TPUs

FLOPS -> OPS (Fixed-point operations per sec.)
e.g. PC(Core i7) ~500GFLOPS

Operation performance: TOPS/GOPS (Tera/Giga Operations Per Second)
Energy efficiency: TOPS/W (Tera Ops. per sec. / Joule per sec = Tera ops. / Joule)
Energy consumption per op. (1/(TOPS/W) [pJ/op] = 1 [pJ/op])

Cloud TPU (v2, 2017)

Case Study: ResNet-50 and TF 1.11
Real data: ~4100 images/sec
Final accuracy: 93%
Training time: (90 epochs) 7h 47m
excluding startup overhead

Current training cost: $36
Current preemptible training cost: $11

Alpha

Synapse op. in brain: 0.1~1 fJ/op
(1fJ = 10^{-15} joules)
1,000~10,000 TOPS/W = 1~10 POPS/W
What are the differences between the three TPUs

ResNet-50 on Cloud TPU v2 Pod

- Real data: 219,000+ images/sec
- Final accuracy: 93%
- Training time: 8m 45s (90 epochs) excluding startup overhead

Cloud TPU Pod (v2, 2017)

TPU v3 Pod (2018)

https://cloud.google.com/tpu/
### TPU Performance on three Popular NNs

- Multi-Layer Perceptrons (MLP)
- Convolutional Neural Networks (CNN)
- Recurrent Neural Networks (RNN)

<table>
<thead>
<tr>
<th>Name</th>
<th>LOC</th>
<th>Layers</th>
<th>Nonlinear function</th>
<th>Weights</th>
<th>TPU Ops / Weight Byte</th>
<th>TPU Batch Size</th>
<th>% of Deployed TPUs in July 2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP0</td>
<td>100</td>
<td>FC 5</td>
<td>ReLU</td>
<td>20M</td>
<td>200</td>
<td>200</td>
<td>61%</td>
</tr>
<tr>
<td>MLP1</td>
<td>1000</td>
<td>Conv 4</td>
<td>ReLU</td>
<td>5M</td>
<td>168</td>
<td>168</td>
<td></td>
</tr>
<tr>
<td>LSTM0</td>
<td>1000</td>
<td>Vector 5</td>
<td>sigmoid, tanh</td>
<td>52M</td>
<td>64</td>
<td>64</td>
<td>29%</td>
</tr>
<tr>
<td>LSTM1</td>
<td>1500</td>
<td>Pool 5</td>
<td>sigmoid, tanh</td>
<td>34M</td>
<td>96</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>CNN0</td>
<td>1000</td>
<td>16</td>
<td>ReLU</td>
<td>8M</td>
<td>2888</td>
<td>8</td>
<td>5%</td>
</tr>
<tr>
<td>CNN1</td>
<td>1000</td>
<td>72</td>
<td>ReLU</td>
<td>100M</td>
<td>1750</td>
<td>32</td>
<td></td>
</tr>
</tbody>
</table>

Tensor Processing Unit (TPU) with MLP, CNN, RNN

<table>
<thead>
<tr>
<th>Model</th>
<th>mm²</th>
<th>mm</th>
<th>GHz</th>
<th>TDP</th>
<th>Measured</th>
<th>TOPS/s</th>
<th>GB/s</th>
<th>On-Chip Memory</th>
<th>Dies</th>
<th>DRAM Size</th>
<th>TDP</th>
<th>Measured</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Idle</td>
<td>Busy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Idle</td>
</tr>
<tr>
<td>Haswell E5-2699 v3</td>
<td>662</td>
<td>22</td>
<td>2300</td>
<td>145W</td>
<td>41W</td>
<td>145W</td>
<td>2.6</td>
<td>1.3</td>
<td>51</td>
<td>51 MiB</td>
<td>504</td>
<td>159W</td>
</tr>
<tr>
<td>NVIDIA K80 (2 dies/card)</td>
<td>561</td>
<td>28</td>
<td>560</td>
<td>150W</td>
<td>25W</td>
<td>98W</td>
<td>--</td>
<td>2.8</td>
<td>160</td>
<td>8 MiB</td>
<td>1838</td>
<td>357W</td>
</tr>
<tr>
<td>TPU</td>
<td>NA*</td>
<td>28</td>
<td>700</td>
<td>75W</td>
<td>28W</td>
<td>40W</td>
<td>92</td>
<td>--</td>
<td>34</td>
<td>28 MiB</td>
<td>861</td>
<td>290W</td>
</tr>
</tbody>
</table>

Benchmarked servers use Haswell CPUs, K80 GPUs, and TPUs. Haswell has 18 cores, and the K80 has 13 SMX processors.

Source: In-datacenter Performance Analysis of a Tensor Processing Unit, Jouppi et al, ISCA, 6/2017
Quantifying the performance of the TPU, our first machine learning chip:
NVIDIA’s Volta GPU is specially designed for AI

- NVIDIA’s Volta GPU is specially designed for ML, and it offers 100 TFLOPS of DL performance, according to the company.
- GPUs were built for graphics workloads and evolved for high performance computing and AI workloads.
- While GPUs are used extensively for training, they’re not really needed for inference.
The HGX-2, announced at NVIDIA GTC May 2018

Multi-precision computing platform for scientific computing (high precision) and AI workloads (low precision).
NVIDIA’s GPU Performance

30x Higher Throughput than CPU Server on Deep Learning Inference

Source: NVIDIA
At Facebook, for example, primary use case of GPUs is offline training rather than serving real-time data to users.

**Offline training uses a mix of GPUs and CPUs**

<table>
<thead>
<tr>
<th>Service</th>
<th>Resource</th>
<th>Training Frequency</th>
<th>Training Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>News Feed</td>
<td>Dual-Socket CPUs</td>
<td>Daily</td>
<td>Many Hours</td>
</tr>
<tr>
<td>Facer</td>
<td>GPUs + Single-Socket CPUs</td>
<td>Every N Photos</td>
<td>Few Seconds</td>
</tr>
<tr>
<td>Lumos</td>
<td>GPUs</td>
<td>Multi-Monthly</td>
<td>Many Hours</td>
</tr>
<tr>
<td>Search</td>
<td>Vertical Dependent</td>
<td>Hourly</td>
<td>Few Hours</td>
</tr>
<tr>
<td>Language Translation</td>
<td>GPUs</td>
<td>Weekly</td>
<td>Days</td>
</tr>
<tr>
<td>Sigma</td>
<td>Dual-Socket CPUs</td>
<td>Sub-Daily</td>
<td>Few Hours</td>
</tr>
<tr>
<td>Speech Recognition</td>
<td>GPUs</td>
<td>Weekly</td>
<td>Many Hours</td>
</tr>
</tbody>
</table>

**TABLE II**

Frequency, duration, and resources used by offline training for various workloads.

**However, online training is CPU-heavy**

<table>
<thead>
<tr>
<th>Services</th>
<th>Relative Capacity</th>
<th>Compute</th>
<th>Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>News Feed</td>
<td>100X</td>
<td>Dual-Socket CPU</td>
<td>High</td>
</tr>
<tr>
<td>Facer</td>
<td>10X</td>
<td>Single-Socket CPU</td>
<td>Low</td>
</tr>
<tr>
<td>Lumos</td>
<td>10X</td>
<td>Single-Socket CPU</td>
<td>Low</td>
</tr>
<tr>
<td>Search</td>
<td>10X</td>
<td>Dual-Socket CPU</td>
<td>High</td>
</tr>
<tr>
<td>Language Translation</td>
<td>1X</td>
<td>Dual-Socket CPU</td>
<td>High</td>
</tr>
<tr>
<td>Sigma</td>
<td>1X</td>
<td>Dual-Socket CPU</td>
<td>High</td>
</tr>
<tr>
<td>Speech Recognition</td>
<td>1X</td>
<td>Dual-Socket CPU</td>
<td>High</td>
</tr>
</tbody>
</table>

**TABLE III**

Resource requirements of online inference workloads.
Exhibit 11: Offerings for AI command significantly higher prices
Google Compute Engine price/hour/single compute instance (i.e. per 1CPU, GPU, TPU, etc)

Source: Google, Goldman Sachs Global Investment Research.
Mobile AI-Chips
Mobile AI-Chips

- Much of the data captured by the smartphone, including images, video, and sound, is unstructured.
- **Training and Inference are Two Vital Components of AI on Smartphones.**
- Unlike structured data — information with a degree of organization — unstructured data makes compilation a time- and energy-consuming task.
- Huawei’s Kirin 970 chipset comes with its own neural processing unit (NPU).
- Huawei has its own APIs that developers need to use to tap the power of the “neural” hardware.
- Google has its mobile AI framework - TensorFlow Lite.

---

**On-Device AI**

- Security
- Power Efficiency
- Low Latency
- Connectivity Independent

**Cloud AI**

- Big Data
- Natural Language Processing
- Image Recognition
- Speech Recognition
- Microsoft Cortana

**Deep Learning**

- Machine Learning

---

**Hi-Silicon AI**

<table>
<thead>
<tr>
<th>8-Core CPU</th>
<th>12-Core GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>up to 2.4GHz</td>
<td>Mali G72MP12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Kirin NPU</th>
<th>Image DSP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.92T FP16 OPS</td>
<td>512bit SIMD</td>
</tr>
</tbody>
</table>

**Global-Mode Modem**

- 1.2Gbps@LTE Cat18
- HiFi Audio 32bit / 384k

**Dual Camera ISP**

- with face & motion detection
- UFS 2.1

**i7 Sensor Processor**

- Security Engine
- inSE & TEE

---

Source: IDC, Huawei, Qualcomm, Nvidia, 2017
# Summary of Mobile AI Chips

<table>
<thead>
<tr>
<th>Design</th>
<th>System-on-chip (SoC)</th>
<th>A11 Bionic</th>
<th>A12 Bionic</th>
<th>Kirin 970</th>
<th>Kirin 980</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplier</td>
<td></td>
<td>Apple</td>
<td></td>
<td>Hisilicon</td>
<td></td>
</tr>
<tr>
<td>Released date</td>
<td></td>
<td>9.12.2018</td>
<td></td>
<td>8.31.2018</td>
<td></td>
</tr>
<tr>
<td>64 Bit</td>
<td></td>
<td></td>
<td></td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Manufacturing process</td>
<td></td>
<td>10 nm TSMC</td>
<td>7nm TSMC</td>
<td>10nm TSMC</td>
<td>7nm TSMC</td>
</tr>
<tr>
<td>Transistors</td>
<td></td>
<td>4.3 billion</td>
<td>6.9 billion</td>
<td>5.5 billion</td>
<td>6.9 billion</td>
</tr>
<tr>
<td>CPU Cores</td>
<td></td>
<td>2+4</td>
<td>2+4</td>
<td>4+4</td>
<td>2+2+4</td>
</tr>
<tr>
<td>Performance CPU</td>
<td></td>
<td>Monsoon</td>
<td>New CPU x 2 + 15% performance</td>
<td>Cortex-A73 x 2</td>
<td>Cortex-A76 (2.6GHz) x 2 + Cortex-A76 (1.92GHz) x 2</td>
</tr>
<tr>
<td>Efficiency CPU</td>
<td></td>
<td>Mistral x 4</td>
<td>New CPU x 4 + 50% efficiency</td>
<td>Cortex-A53 x 4</td>
<td>Cortex-A55 x 4</td>
</tr>
<tr>
<td>Max Clock (GHz)</td>
<td></td>
<td>2.4</td>
<td>N/A</td>
<td>2.4</td>
<td>2.6</td>
</tr>
<tr>
<td>GPU</td>
<td></td>
<td>Internally-designed GPU</td>
<td>Internally-designed GPU</td>
<td>Mali-G72 MP12</td>
<td>Mali-G76</td>
</tr>
<tr>
<td>GPU Cores</td>
<td></td>
<td>3</td>
<td>4</td>
<td>12</td>
<td>10</td>
</tr>
<tr>
<td>AI Accelerator</td>
<td></td>
<td>2-core Neural Engine</td>
<td>8-core Neural Engine</td>
<td>NPU</td>
<td>Dual NPU</td>
</tr>
<tr>
<td>Performance</td>
<td></td>
<td>600 billion operations per second</td>
<td>5 trillion operations per second</td>
<td>2005 pictures per minute</td>
<td>4500 pictures per minute</td>
</tr>
<tr>
<td>Memory</td>
<td></td>
<td>LPDDR4X</td>
<td>LPDDR4X</td>
<td>LPDDR4x</td>
<td>LPDDR4X</td>
</tr>
<tr>
<td>Ram Interface</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>1833</td>
<td>2133</td>
</tr>
<tr>
<td>Max Bandwidth</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
<td>29.9</td>
<td>34.1</td>
</tr>
</tbody>
</table>

Source: [medium.com](https://medium.com)
# Summary of Mobile AI Chips

<table>
<thead>
<tr>
<th>Design</th>
<th>Supplier</th>
<th>A11 Bionic</th>
<th>A12 Bionic</th>
<th>Kirin 970</th>
<th>Kirin 980</th>
</tr>
</thead>
<tbody>
<tr>
<td>64 Bit</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Transistor Size</td>
<td>10 nm TSMC</td>
<td>7nm TSMC</td>
<td>10nm TSMC</td>
<td>7nm TSMC</td>
<td>7nm TSMC</td>
</tr>
<tr>
<td>CPU Cores</td>
<td>2+4</td>
<td>2+4</td>
<td>4+4</td>
<td>2+2+4</td>
<td></td>
</tr>
<tr>
<td>Performance CPU</td>
<td>Monsoon</td>
<td>New CPU× 2 + 15% performance</td>
<td>Cortex-A73 × 2</td>
<td>Cortex-A76 (2.6GHz) × 2 + Cortex-A76 (1.92GHz) × 2</td>
<td></td>
</tr>
<tr>
<td>Efficiency CPU</td>
<td>Mistral × 4</td>
<td>New CPU× 4 + 50% efficiency</td>
<td>Cortex-A53 × 4</td>
<td>Cortex-A55 × 4</td>
<td></td>
</tr>
<tr>
<td>Max Clock (GHz)</td>
<td>2.4</td>
<td>N/A</td>
<td>2.4</td>
<td>2.6</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GPU</th>
<th>Supplier</th>
<th>A11 Bionic</th>
<th>A12 Bionic</th>
<th>Kirin 970</th>
<th>Kirin 980</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU Core</td>
<td>Internally-designed GPU</td>
<td>Internally-designed GPU</td>
<td>Mali-G72 MP12</td>
<td>Mali-G76</td>
<td></td>
</tr>
<tr>
<td>AI Accelerator</td>
<td>2-core Neural Engine</td>
<td>8-core Neural Engine</td>
<td>NPU</td>
<td>Dual NPU</td>
<td></td>
</tr>
<tr>
<td>Performance</td>
<td>600 billion operations per second</td>
<td>5 trillion operations per second</td>
<td>2005 pictures per minute</td>
<td>4500 pictures per minute</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Memory</th>
<th>Supplier</th>
<th>A11 Bionic</th>
<th>A12 Bionic</th>
<th>Kirin 970</th>
<th>Kirin 980</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ram Interface</td>
<td>LPDDR4X</td>
<td>LPDDR4X</td>
<td>LPDDR4x</td>
<td>LPDDR4X</td>
<td></td>
</tr>
<tr>
<td>Ram Frequency</td>
<td>N/A</td>
<td>N/A</td>
<td>1833</td>
<td>2133</td>
<td></td>
</tr>
<tr>
<td>Max Bandwidth</td>
<td>N/A</td>
<td>N/A</td>
<td>29.9</td>
<td>34.1</td>
<td></td>
</tr>
</tbody>
</table>

Source: medium.com
Edge and IoT AI-Chips

~Processing Real-Time Data~
Edge Computing: Edge AI Chip

• The need for no latency, higher security, faster computing, and less dependence on connectivity will drive the adoption of devices that can offer AI at the edge.

On-device approach helps reduce latency for critical applications, lower dependence on the cloud, and better manage the massive data being generated by the IoT device.
Examples of Edge AI Applications

<table>
<thead>
<tr>
<th>In-home smart cameras can recognize that a person(s) has entered an area</th>
<th>On-device facial recognition and object recognition, where user data doesn’t leave the device</th>
<th>On-board AI making instantaneous driving decisions</th>
<th>Vision for baby monitors, drones, robots, and other devices that can respond to situations without internet connection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eg: nest IQ cameras, aws DeepLens</td>
<td>Eg: apple neural engine</td>
<td>Eg: TESLA autopilot</td>
<td>Eg: intel Myriad X</td>
</tr>
</tbody>
</table>

Cloud stores large datasets, trains algorithms, collects edge data, pushes AI model updates

Source: CBINSIGHTS 2018
Examples of Edge AI Applications

Combining a 4K sensor with HDR and Intelligent Imaging
Uses on-device vision processing to watch for motion, distinguish family members, and send alerts only if someone is not recognized or doesn’t fit pre-defined parameters.
Apple, Intel, and Google Edge AI-Chips

- **Apple** released its **A11 chip** with a “neural engine” for iPhone 8 and X. Apple claims it can perform machine learning tasks at up to 600B operations per second.
  - It powers new iPhone features like FaceID, which scans a user’s face with an invisible spray of light, without uploading or storing any user data (or their face) in the cloud.

- **Intel** released an on-device vision processing chip called **Myriad X** (initially developed by Movidius, which Intel acquired in 2016).
  - Myriad X promises to take on-device deep learning beyond smartphones to devices like baby monitors and drones.

- **Google** proposed a similar concept with its “federated learning” approach, where some of the machine learning “training” can happen on your device. It’s testing out the feature in **Gboard**, the Google keyboard.

- **AI on the edge reduces latency. But unlike the cloud, edge has storage and processing constraints.**
Healthcare AI-Chips
Healthcare AI-Chips

Applications/Research Areas

- **Neuroscience**: neuroinformatics; brain simulation
- **Medicine**: medical informatics; early diagnosis; personalized treatment
- **Future computing**: interactive supercomputing; neuromorphic computing

### SpiNNaker-1 machine
- Many-core system
- 0.5 (1.0) Million ARM cores
- Real-time simulator

### BrainScaleS-1 machine
- Physical model system
- 4M neurons, 1B plastic syn.
- Accelerated emulator

### SpiNNaker-2 prototype
- 144 Cortex M4F per chip
- 36 GIPS/Watt per chip
- x10 with constant power

### BrainScaleS-2 prototype
- On-chip plasticity processor
- Flexible hybrid plasticity
- Active dendritic spatial structure

Healthcare AI-Chips

SpiNNaker-1 machine
- Many-core system
- 0.5 (1.0) Million ARM cores
- Real-time simulator

BrainScaleS-1 machine
- Physical model system
- 4M neurons, 1B plastic syn.
- Accelerated emulator

SpiNNaker-2 prototype
- 144 Cortex M4F per chip
- 36 GIPS/Watt per chip
- x10 with constant power

BrainScaleS-2 prototype
- On-chip plasticity processor
- Flexible hybrid plasticity
- Active dendritic spatial structure

https://www.humanbrainproject.eu/en/
Healthcare AI-Chips

The Human Brain Project
An EU ICT Flagship project (€1B budget)
80 partner institutes, led by Henry Markram, EPFL

The basic idea of the Human Brain Project
From Science to Infrastructures to Science and Innovation

The Joint Platform - Unified access through SW Collaboratory
- Neuroinformatics
- Brain Simulation
- HPAC
- Medical Informatics
- Neuromorphic
- Neurorobotics

Co-Design
- Mouse
- Human
- Cognition
- Theory
- HBP Neuroscience

Knowledge About the brain
Basic Science

Application in brain technology
Innovation

What USERS get from the platforms

https://www.humanbrainproject.eu/en/
Agenda

• Fundamental Trends

• AI – The 4th Industrial Revolution

• Survey of AI Hardware
  ➢ Cloud AI Hardware, Chips
  ➢ Mobile AI Chips
  ➢ Edge and IoT AI Chips
  ➢ Healthcare AI Chips

• Conclusions
Conclusions

• DNNs are a key component in the AI revolution.

• Efficient processing of DNNs is an important area of research with many promising opportunities for innovation at various levels of hardware design, including algorithm co-design

• It’s important to consider a comprehensive set of metrics when evaluating different DNN solutions: accuracy, speed, energy, and cost
Conclusions

Memory access in AI-Chip is the bottleneck
- **Worst case**: ALL memory R/W are DRAM accesses
Ex. AlexNet [NIPS 2012] has 724M MACs → 2896M DRAM accesses required

Possible HW/SW techniques to cope with the memory access problem:

- **Advanced Storage Technology**
  - Embedded DRAM (eDRAM) → Increase on-chip storage capacity
  - 3D Stacked DRAM → Increase memory bandwidth
  - Use memristors as programmable weights (resistance)

- **Reduce size of operands for storage/compute**
  - Floating point → Fixed point
  - Bit-width reduction

- **Reduce number of operations for storage/compute**
  - Network Pruning; Compact Network Architectures
Conclusions

- For SNN Chips, the physical properties of silicon make it hard to control ions (the current between artificial neurons).

- A new design from an MIT team used different materials – single-crystalline silicon and silicon germanium. The hope is to get well controlled flow of ions.

- A team in Korea is also investigating other materials called Tantalum Oxide.

- A team in Colorado is implementing magnet to precisely control the flow of ions/current.
References

1. Dally, W. February 9, 2016. High Performance Hardware for Machine Learning, Cadence ENN Summit
5. Dally, NIPS’2016 workshop on Efficient Methods for Deep Neural Networks
7. Top AT Trends to watch in 2018, CBINSGHTS, 2018
https://www.youtube.com/watch?v=E8n_k6HNAgs