azeem 10.5K Followers About Follow

Sign in Get started

azeem
Entrepreneur, inventor and creator — curator of The Exponential View
Follow

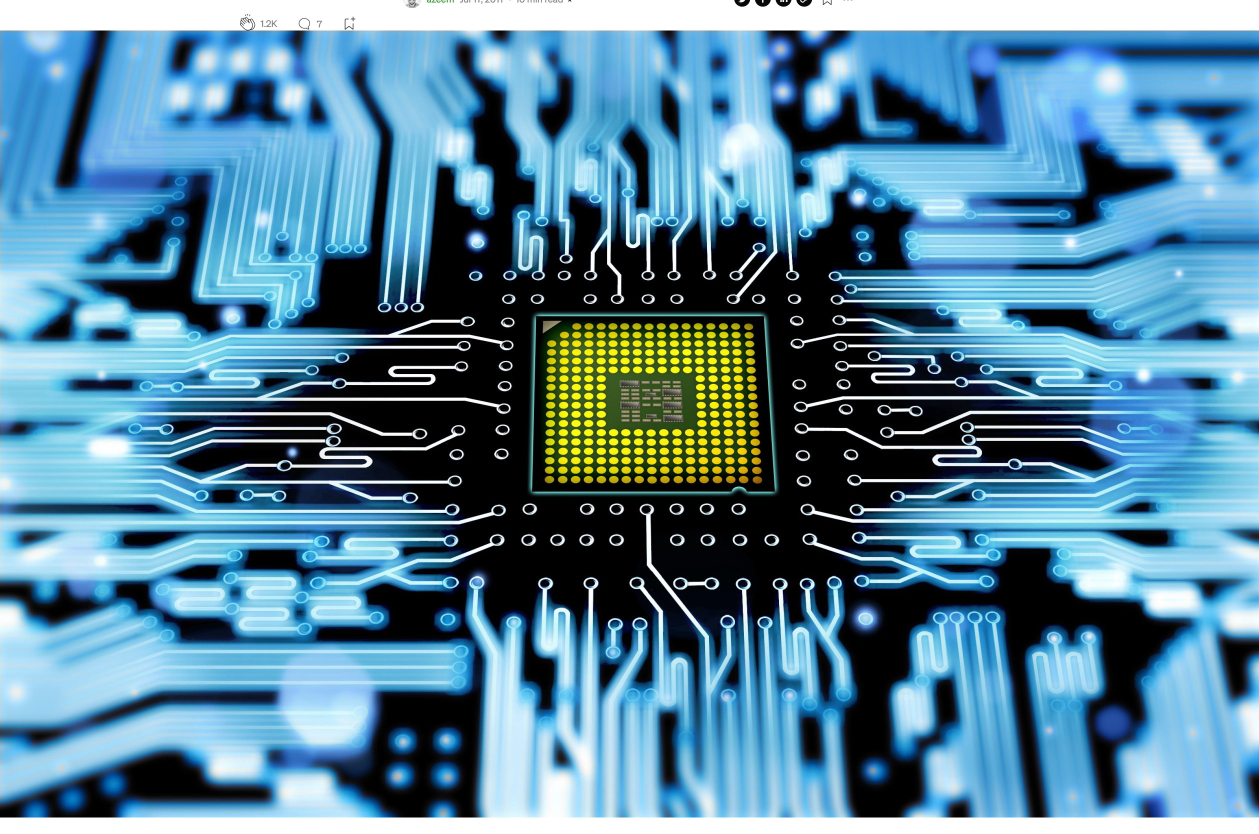
AZEEM FOLLOWS
Anthony Perkins
Ramez Naam
Stefan Wolpers
Enrique Dans
Clive Thompson
See all (1,797)

You have 2 free member-only stories left this month. Sign up for Medium and get an extra one

If We Want To Get To Real Time AI, We've Got To Build Another iPhone Industry. Five Times Over.

azeem Jul 17, 2017 · 10 min read

1.2K 7



Produced in partnership with NewCo Shift.

In October 2016, Tesla announced a significant change to its Advanced Driver Assistance System package. This is the combination of sensors and computer power that will enable Tesla to fulfill Elon Musk's promise to drive "all the way from a parking lot in California to a parking lot in New York with no controls touched in the entire journey" by the end of 2017.

Amongst the many changes to the sensor package was a switch in the systems' brains. Previously powered by a processor from Mobileye (recently acquired by Intel) the package now sports a Nvidia Drive PX 2. Why?

It turns out that to be safe, self-driving cars need an extraordinary amount of data from sensor systems. And if it is to figure out what all those sensors are telling it, the car requires an unprecedented amount of processing. Once it knows what is going on in the environment, yet more processing is needed to help the car figure out what to do next.

The switch that Tesla made gives a clue to just how *much* processing. The Mobileye EyeQ3 processor was a significant chip. It was 42mm² in area (about a quarter of the size of a modern Intel i7 processor), packing transistors using a manufacturing process which arrays transistors 40nm apart.

The replacement chip from Nvidia is 610mm² in size and used a more advanced manufacturing technique, packing transistors at a 16nm node. This smaller node means that the transistors are packed 2.5 times more tightly than those in the EyeQ3 processor. In short, the replacement Nvidia chip had a **90x performance improvement** over the Mobileye one.

Even by the standards of Moore's Law, which represents an average 60% improvement of transistor packing or performance every year, this was a significant jump. In fact, the switch out represented the equivalent of *a decade* of Moore's Law processing.

AI is bigger than Moore's Law

In a nutshell, this shift by Tesla summarizes the kinds of demand machine learning-like applications are going to make on available processing. It isn't just autonomous vehicles. It will be our connected devices, on-device inferencing to support personal interfaces, voice interactions and augmented reality.

In addition, our programming modalities our changing. In the pre-machine learning world, a large amount of 'heavy lifting' was done by the brains of the software developer. These smart developers have the task of simplifying and representing the world mathematically (as software code), which then gets executed in a deterministic and dumb fashion.

In the new world of machine learning, the software developer needs to worry less about translating the detailed abstractions of the world into code. Instead, they build probabilistic models which need to crunch enormous datasets to recommend a best output. What the programmer saves in figuring out a mathematical abstraction they make up for by asking the computer to do many calculations (often billions at a time).

As machine learning creeps across the enterprise, the demand for processing in the firm will increasingly significantly. What kind of impact will this have on the IT industry, its hardware and software suppliers? How will practices change? What opportunities will this create?

Here are three of the key changes:

1. Massive increase in the amount of compute needed

2. The cloud will flourish, the edge will bloom

3. New species of chips will emerge

Massive compute requirements

Let's start with self-driving cars. At a recent event hosted by London venture capital firm Open Ocean, the product lead for Five.AI, an autonomous vehicle startup, summarized compute requirements needed for fully autonomous driving.

The key challenges are two fold. First, the car must map the actor-state space. The actor-state space is a representation of the external environment around the car, including all objects, vehicles, peoples and so on, and their current state (stationary, direction, velocity, acceleration, etc.)

Second, the car needs to figure out how to behave appropriately. What is its next best action given where the passenger is trying to go, given the hazards around the car right now?

To do this each car requires a bunch of data from Lidar, cameras and other sensors. Lidar, short for light detection and ranging, is a radar-like sensor which is commonplace on many self-driving car set-ups (although Tesla uses different technology).

A subsystem needs to handle sensor fusion and perception steps before additional subsystems plan the next action. All this processing happens on the device (the car); even with 5G networks the latency risk of sending signal data to the cloud for prediction is too great.

Taken together, to bring a car into full, safe self-driving mode, all of this processing and data ingestion will take an estimated 200 teraflops of processing, all of which would essentially need to be executed in a one second or smaller time window.

How much is 200 teraflops? It's a lot even by today's standards: many cycles of Moore's Law away. To complete 200 trillion floating point operations in one-second time window would roughly take ten current-model iPhones. That's significantly above the capacity of the Nvidia rig in current Tesla models. (To be fair, Mr. Musk may yet achieve his goal with the current Nvidia GPUs if algorithmic optimizations provide a step change in computational efficiency.)

The annual production volume of cars and vans runs to about 100 million cars globally. This means that to 'smarten' the entire output of vehicles under these assumptions will require the equivalent of one *billion* additional iPhones per year.

Current global iPhone output ran to around 200 million units in 2016. So smartening up global auto fleets is, in a very real sense, the equivalent of increasing the current size of the semiconductors supporting the iPhone industry by at least five-fold.

A second consideration is the cost and power-load of the processing. Self-driving cars need to be affordable. And as they are likely to be electrically operated, the brains will need to be power efficient. A Tesla burns about 200 watt-hours per kilometer. A high-performing but high-power output GPU rig might add an extra 1.2% power consumption to the load (Nvidia PX2 consumes 250W). This increase in the power-load will decrease the range concomitantly.

Shifting bits is almost as expensive as shifting atoms

Self-driving cars are just the sexier end of our increased compute demands. Modern deep learning approaches also have meaningful compute requirements. A modern neural net might have dozens of connected layers and billions of parameters, requiring a step-function increase in compute power from where we are today.

As AI analyst Libby Kinsey outlines in a recent essay, most breakthrough approaches in deep learning use enormous amounts of compute. In any machine learning system, teaching the model how to predict effectively is the most expensive computational step. The second step, applying to produce useful output (also known as inferencing), is much cheaper. But is still not free.

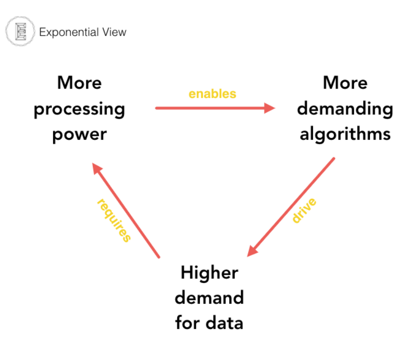
It's only recently, for this reason, that object recognition has started to move from the cloud onto mobile handsets itself. Local inferencing will allow machine learning on the device — meaning it will be faster and more reliable when data networks are patchy. A good example would be the face recognition biometric lock used by recent models of Samsung phone.

TensorFlow, the most popular framework for developing deep learning models, has recently become available on Android devices. However, this framework currently only allows the cheaper inferencing step. It won't be until the end of 2017 that TensorFlow will allow the building of leaner deep learning models on devices. Of course, the hit show *Silicon Valley* already put this capability to a noble use with its "Not Hot Dog" app (which is apparently a real thing now).

The reinforcing cycle

Algorithms and processing are just two parts of the cycle. **The third part is data.** As processing power increases, we can use more demanding algorithms, which can apply more data (so the demand for sensor data to train from or infer upon will increase). This in turn will increase the demand for efficient processing, which will allow us to increase algorithmic complexity.

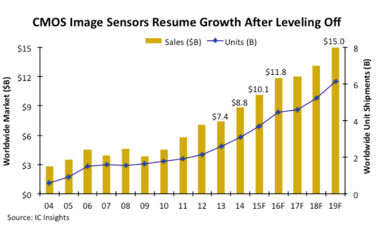
This cycle is reminiscent of the relationship between Microsoft and Intel during the establishment of the Wintel duopoly. With Intel's development of processing, Microsoft could write bloated code and create features that absorbed all the capability. However, with Microsoft's new features, Intel was urged to improve. The incremental headroom created by new chips allowed Microsoft (and its ecosystem of independent software vendors) to use that headroom for new things.



What this reinforcing loop suggests is that the combination of increasing processing power and more demanding algorithms should drive a great demand for data.

And we are seeing just that. One clear example is machine vision, which is now getting sufficiently good to be used as a primary data source for software (rather than only log data, database entries, or user input). Self-driving cars are a great example of this, but also biometric identity systems or the Amazon Go shop, which relies heavily on machine vision as a primary input.

If you want to see all this in action, take a look at the camera sensor industry. Between 2009 and 2019 the number of CMOS sensors shipped is expected to rise three-fold.



The sensors shipped in 2008/9 had a primary job of taking photos for human eyes to process. But increasingly, sensors will capture images that will be processed by machine vision algorithms. Many will never be seen by humans. Those pixels may now be programmatically accessed only by software.

Assuming a five-year lifespan for a typical CMOS sensor, we would expect by 2019 that a total of about 45bn digital camera sensors will be in operation. The resolution power of these sensors will increase as well. Hendy's Law loosely describes a Moore's Law-like relationship with pixel density, a 59% annual improvement year on year. This means that the typical sensor shipped in 2019 has about 100-times the pixels of the sensors shipped in 2009. And with higher volumes of sensors shipping as well, we'll see a 100% increase in unit shipment from 2009 to 2019.

All of these sensors are capable of generating massive amounts of data. Some, like the third camera on my iPhone 7+, will not be used very often. Others, like CCTV or IoT sensors, will be online 24/7 streaming images that need processing. The self-driving car team reckons that a typical self-driving car will be have 120–150 megapixels of camera sensors streaming at all times to assess its environment. (For comparison, an iPhone 7 camera has a 12 megapixel resolution. A Canon EOS5DS professional digital camera has a 50 megapixel sensor. A megapixel is a unit of graphic resolution equivalent to 1,048,576.)

Lidar will increase that very significantly. As Intel CEO Brian Krzanich points out in a recent article:

> *In an autonomous car, we have to factor in cameras, radar, sonar, GPS and LIDAR — components as essential to this new way of driving as pistons, rings and engine blocks. Cameras will generate 20–60 MB/s, radar upwards of 10 kB/s, sonar 10–100 kB/s, GPS will run at 50 kB/s, and LIDAR will range between 10–70 MB/s. Run those numbers, and each autonomous vehicle will be generating approximately 4,000 GB — or 4 terabytes — of data a day.*

Top highlight

For comparison, the most data hungry mobile internet users in the world, the Finns, used on average 2.9 Gb of data per month back in 2015. (In other words, a self-driving car's daily data demands are the equivalent of about 40,000 eager Finns surfing the internet.)

It is expected that anywhere from 20–30 billion more IoT devices are coming online by 2020, streaming data that helps build smarter objects, homes, inform consumer lifestyle, ensure security monitoring, and energy consumption.

These are just a handful of the new applications coming online. There are at least four others which will create heavy demands, which we don't have room to dig into deeply here.

- **Augmented reality & virtual reality.** AR & VR, which rely heavily on machine vision and 3D modeling and simulation will create incredible demands for processing. (An example of how the AR/VR space is part of a wider *keiretsu* is how Softbank has invested in virtual world pioneer, Improbable; acquired ARM, the semi-conductor design firm and taken a reasonable stake in Nvidia.)
- **Bioinformatics** relies on heavy-weight computing for billions of pieces of data that require processing, analysis, and storage. According to a paper published by four German computer scientists in 2009, the size of DNA sequence databases doubles twice a year, scaling at a rate that computing performance can hardly keep up with.
- **Computational biology** where computational simulations and modeling are used to study biological systems will also increase the demand for cycles. (For a simple introduction to computational biology read this.)
- **Cryptocurrencies** have their own computational demands at the mining level.

This flywheel of computing begetting new applications begetting new algorithmic solutions begetting more demand for data begetting more demand for compute will not just continue. It will accelerate and make previous virtuous cycles of computing and software pale into comparison.

The demands of self-driving cars alone represent an opportunity approaching five iPhone industries annually. Many other new applications in the machine learning domain will further increase demand for compute which the semiconductor industry will rise to meet.

This is the end of part one of this series. In part two we will discuss the shifting location of computing and the balance between cloud and edge. We'll also discuss the shifts in the architecture of CPUs and novel modes of computing. In part three, we'll discuss implications for the industry and opportunities for investors.

Research assistance for this was provided by Marija Gavrilov. Thanks to Kenn Cukier, Sandy Mackinnon, Libby Kinsey, Gerald Huff and Pete Warden for commenting on drafts of this essay.

I curate a weekly newsletter about the impact of exponential technologies on businesses, society and polity. Join more than 18k happy readers.

Thanks to Gerald Huff and John Battelle.

1.2K 7

More from azeem
Entrepreneur, inventor and creator — curator of The Exponential View
Follow

More From Medium

Artificial Intelligence and Music offshoot World Music Hub
Safaque Kagdi

Is AI Just as Biased as the Humans Creating It?
ReadWrite in ReadWrite

AI for Everyone: Important Trends and Business Aspects of AI
Chuck

FlexiGAN: An End-to-End Solution for FPGA Acceleration of Generative Adversarial Networks
Sarang Patil in Accelerated Image Processing using FPGAs.

Automating an AI to find the shortest route using reinforcement learning
Pragati Baheti in Heartbeat

Mudley Project#2 Timebase Rebalance AI —Payoff#10
Mudley Token in MudleyToken

Artificial Intelligence is Taking Over Android App Development…
Aspirantedge

Can AI Mitigate Human Bias?
Katharine Stevenson

Medium
About Write Help Legal