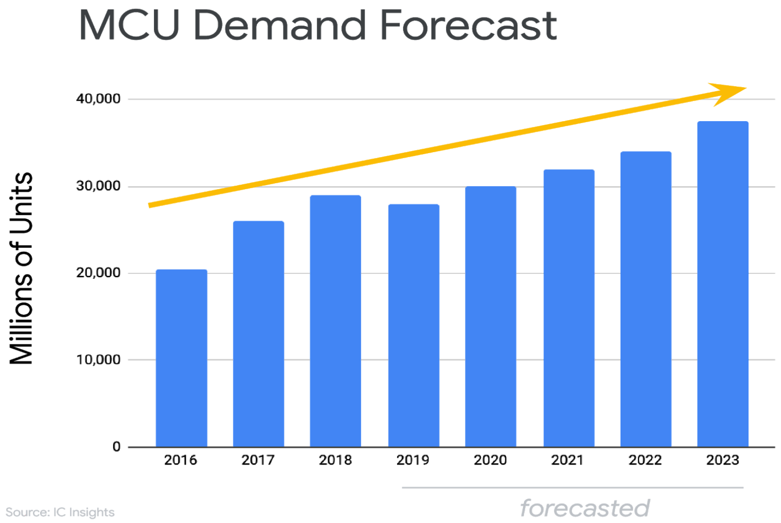
Why the Future of Machine Learning is Tiny and Bright

Tiny Computers Are Ubiquitous

It is estimated there are over 250 tiny microcontrollers in the world and that over 50 billion will be sold this year. Microcontrollers (or MCUs) are packages containing a small CPU with possibly just a few kilobytes of RAM and are embedded in consumer, medical, automotive, and industrial devices. To put MCUs into perspective, there are only about 10 million servers in use across the planet and 4 billion smartphone users. Microcontrollers typically do not get much attention because they are often used to replace functionality that older electro-mechanical systems could do, in cars, washing machines, or remote controls, but they are ubiquitous.

Driven by new applications in wearables, automotive use cases, smart appliances, etc., the demand for MCU is anticipated to increase steadily. The figure below shows the data from IC Insights. The number of units sold in 2023 is expected to exceed 35 Billion units per year!

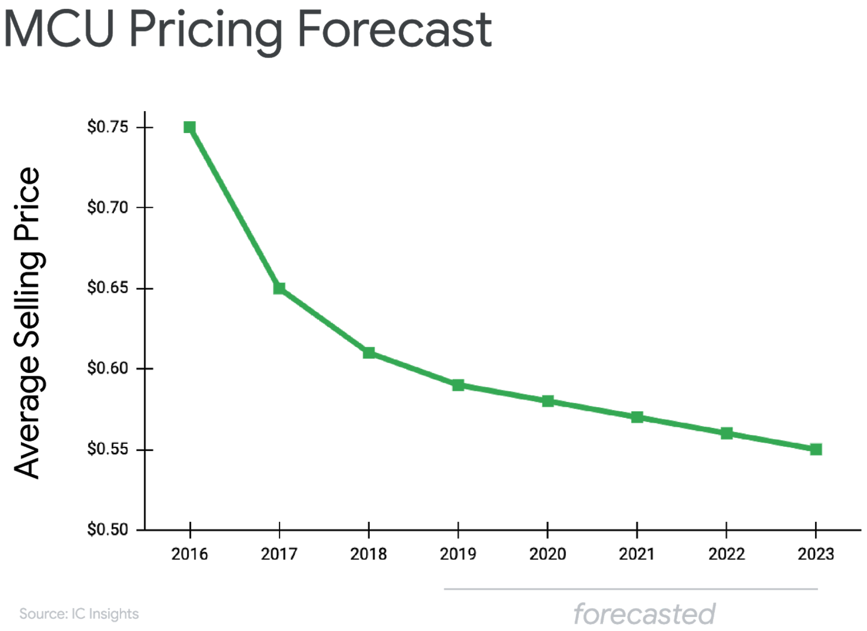


Source: IC Insights

Microcontrollers are Cheap

MCUs are designed to be cheap enough to include in almost any object that's sold. They are cheap because they aren't designed to be general-purpose computational workhorses that run complex workloads. Instead, they are often designed to perform specific tasks, slowly, albeit steadily. Therefore, they usually cost much less than a typical processor. The magnitude difference in price between an average MCU and an Intel or AMD processor found in your laptop or computer might be an order or two orders of magnitude difference in price.

The average price of an MCU is already less than USD $1. As seen in the figure below, the average sales price (ASP) for a microcontroller is hovering close to 55 cents, and as the demand surges, it is expected that this will fall below 50 cents in the future. On a global scale, the microcontroller market ($M) is poised to grow by USD $6.74 Billion between 2020-2024.



Source: IC Insights

MCUs are Resource-Constrained, Ultra-low Power Systems

The holy grail for almost any embedded device is for it to be deployable anywhere and require no maintenance like docking or battery replacement. Any device that requires tethered electricity faces a lot of deployment barriers. It can be restricted to only places with electrical wiring. Even where electrical wiring capability is available, it may be challenging for practical reasons to plug something new in, for example, on a factory floor or in an operating theatre. Putting something high up in the corner of a room means running a cord or figuring out alternatives like power-over-ethernet. The electronics required to step-down or convert the main-line high voltage to low voltage are expensive and waste energy.

But perhaps the most significant barrier to achieving ubiquitous deployment computers everywhere is how much energy an electronic system uses. Here are some rough numbers for standard components based on figures from Smartphone Energy Consumption, and so it is no wonder that smartphones need to be tethered to the wall and recharged every night:

- A display might use 400 milliwatts.
- Active cell radio might use 800 milliwatts.
- Bluetooth might use 100 milliwatts.

- Accelerometer is 21 milliwatts.
- Gyroscope is 130 milliwatts.
- GPS is 176 milliwatts.

A key opportunity with using MCUs is they require a very minimal amount of energy. A microcontroller itself might only use a milliwatt or even less. Still, you can see that peripherals (accelerometer, gyroscope, GPU, etc.) require much more energy to stay powered on.

A coin battery might have 2,500 Joules of energy to offer, so even something drawing only one milliwatt will have severe consequences. Of course, most current products use “duty cycling” and power naps to avoid being always on, but we see how tight the budget is even then. The overall thing to take away from these figures is that while processors and sensors can scale their power usage down to microwatt ranges, displays and especially radios are constrained to much higher consumption, with even low-power wifi and bluetooth using tens of milliwatts when active. The physics of moving data around is generally well-known to require a lot of energy.

The general wisdom is that the energy an operation takes is proportional to how far you have to send the bits. CPUs and sensors send bits a few millimeters and are cheap. Radio sends them meters or more and is expensive. We don’t see this relationship fundamentally changing, even as technology improves overall. We expect the relative gap between computing and radio costs to get more expansive because we see more opportunities to reduce computing power usage.

Tiny Machine Learning (TinyML) on MCUs

Can you imagine a future where every one of the 250B tiny MCUs runs intelligent machine learning algorithms that can sense their surroundings, predict events in real-time, and even make nudges or recommendations based on the sensors’ activity? We can transform the world.

For instance, smart consumer devices like smart toothbrushes will have MCUs that are tightly coupled with sensors that dynamically adjust to your brushing intensity based on pressure. Medical devices in the future may use microcontrollers with biomedical sensors to control drug delivery as and when needed. Microcontrollers in automotive applications can aid functional safety on the road, detecting engine conditions, etc. to ensure our safety. Finally, MCUs will likely have widespread applications in industrial devices for continuous process monitoring and anomaly detection for applications such as predictive maintenance that can save millions of dollars in productivity loss and downtime by forewarning maintenance engineers.

TinyML is a relatively new machine learning paradigm, yet it is producing remarkably astounding results. Several recent examples include audio, visual, and sensor fusion applications, such as voice and facial recognition, voice commands, and natural language processing.

Looking more further into the future, we imagine a world where we have a tiny battery-powered image sensor that I could program to look out for things like particular crop pests or weeds and send an alert when one was spotted. These could be scattered around fields and guide interventions like weeding or pesticides in a much more environmentally friendly way.

TinyML at the Edge/Endpoint

Computing data locally, rather than streaming all the data to the cloud or edge servers, which is costly, is a feasible solution for intelligently processing the data available at the sensors. This will not only conserve energy, but it will also help unlock new applications and use cases. Consequently, given the existing widespread deployment of MCUs, and the minimal energy consumption of MCUs, there is growing interest in providing AI functionality at the edge/endpoint. This idea isn't particularly new. Both Apple and Google run always-on machine learning (ML) based neural networks for voice recognition on these kinds of power-efficient chips.

There are many more examples of AI capabilities at the endpoint, and people are just starting to realize that there is a unique match between deep learning and MCUs. If you are with us so far, then it's obvious there's a massive untapped market waiting to be unlocked with the right technology to enable AI on these cheap and ubiquitous devices. We need ML solutions that can work on cheap microcontrollers, which use very little energy, that rely on computing capability, not radio, and can turn all our continuously streaming sensor data into something useful. This is the gap that machine learning, and specifically deep learning, fills.

It has suddenly become possible to take noisy signals like images, audio, or accelerometers and extract meaning from them in the last few years by using neural networks. Because we can run these networks on microcontrollers and sensors themselves use little power, it becomes possible to interpret much more of the sensor data we're currently ignoring. For example, imagine we want to see that almost every device has a simple voice interface. By understanding a small vocabulary, and maybe using an image sensor to do gaze detection, we should control nearly anything in our environment without needing to reach it to press a button or use a phone app. We want to see a voice interface component that's less than 50 cents that runs on a coin battery for a year, and we believe it's possible with the technology we have right now.

The Future of TinyML is Bright

We can conjure up a thousand other products. But to get there, first and foremost, Pete, Laurence, and I are most passionate about ensuring that the technological imperative behind them is so compelling that we will all be able to build whole new applications that we can't even imagine today. For all of us, it feels a lot like being a kid in the Eighties when the first home computers emerged. None of us had any idea what they would become, and most people at the time used them for games or storing address books, but there were so many possibilities. A new world emerged out of those burgeoning systems. So we challenge you to invent the future. Come on and embrace the future with us by learning all we have to teach you about tinyML!