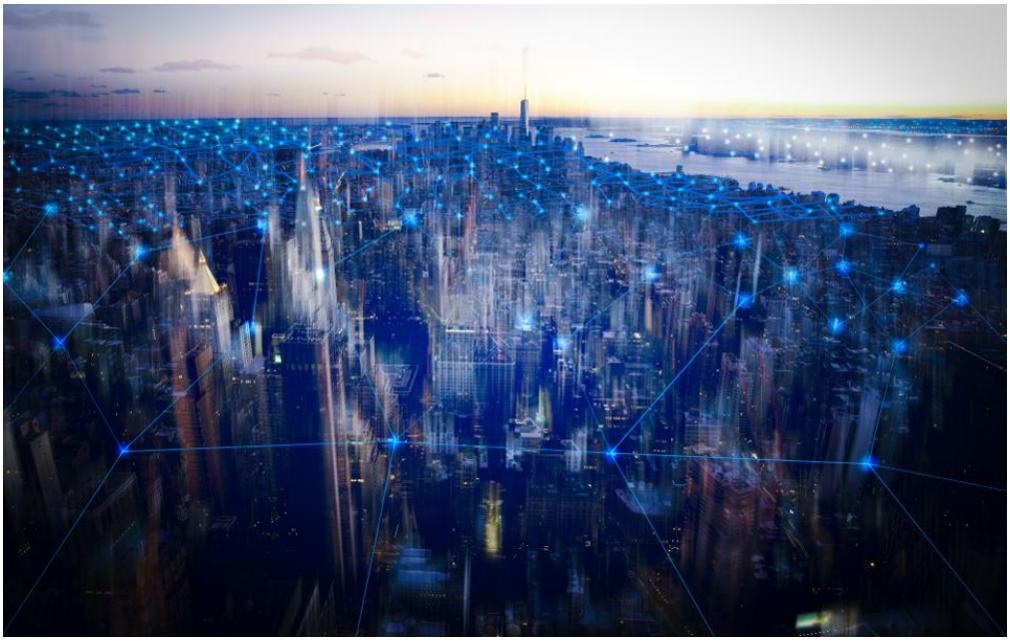
Why Apple And Microsoft Are Moving AI To The Edge

Mohanbir Sawhney Contributor
CMO Network
I analyze trends and current events in technology, marketing and AI.

-
-
-

Artificial intelligence (AI) has traditionally been deployed in the cloud, because AI algorithms crunch massive amounts of data and consume massive computing resources. But AI doesn't only live in the cloud. In many situations, AI-based data crunching and decisions need to be made locally, on devices that are close to the edge of the network.



GETTY

AI at the edge allows mission-critical and time-sensitive decisions to be made faster, more reliably and with greater security. The rush to push AI to the edge is being fueled by the rapid growth of smart devices at the edge of the network—smartphones, smart watches and sensors placed on machines and infrastructure. Earlier this month, Apple spent $200 million to acquire Xnor.ai, a Seattle-based AI startup focused on low-power machine learning software and

hardware. Microsoft offers a comprehensive toolkit called Azure IoT Edge that allows AI workloads to be moved to the edge of the network.

Will AI continue to move to the edge? What are the benefits and drawbacks of AI at the edge versus AI in the cloud? To understand what the future holds for AI at the edge, it is useful to look back at the history of computing and how the pendulum has swung from centralized intelligence to decentralized intelligence across four paradigms of computing.

Centralized vs. Decentralized

Since the earliest days of computing, one of the design challenges has always been where intelligence should live in a network. As I observed in an article in the *Harvard Business Review* in 2001, there has been an "intelligence migration" from centralized intelligence to decentralized intelligence—a cycle that's now repeating.

The first era of computing was the mainframe, with intelligence concentrated in a massive central computer that had all the computational power. At the other end of the network were terminals that consisted essentially of a green screen and a keyboard with little intelligence of their own—hence they were called "dumb terminals."

The second era of computing was the desktop or personal computer (PC), which turned the mainframe paradigm upside down. PCs contained all the intelligence for storage and computation locally and did not even need to be connected to a network. This decentralized intelligence ushered in the democratization of computing and led to the rise of Microsoft and Intel, with the vision of putting a PC in every home and on every desk.

The third era of computing, called client-server computing, offered a compromise between the two extremes of intelligence. Large servers performed the heavy lifting at the back-end, and "front-end intelligence" was gathered and stored on networked client hardware and software.

The fourth era of computing is the cloud computing paradigm, pioneered by companies like Amazon with its Amazon Web Services, Salesforce.com with its SaaS (Software as a Service) offerings, and Microsoft with its Azure cloud platform. The cloud provides massively scaled computational power and very cheap memory and storage. It only makes sense that AI applications would be housed in the cloud, since the computation power of AI algorithms has increased 300,000 times between 2012 and 2019—doubling every three-and-a-half months.

The Pendulum Swings Again

Cloud-based AI, however, has its issues. For one, cloud-based AI suffers from latency—the delay as data moves to the cloud for processing and the results are transmitted back over the network to a local device. In many situations, latency can have serious consequences. For instance, when a sensor in a chemical plant predicts an imminent explosion, the plant needs to be shut down immediately. A security camera at an airport or a factory must recognize intruders and react immediately. An autonomous vehicle cannot wait even for a tenth of a second to activate emergency braking when the AI algorithm predicts an imminent collision. In these situations, AI

must be located at the edge, where decisions can be made faster without relying on network connectivity and without moving massive amounts of data back and forth over a network.

The pendulum swings again, from centralization to decentralization of intelligence— just as we saw 40 years ago with the shift from mainframe computing to desktop computing.

However, as we found out with PCs, life is not easy at the edge. There is a limit to the amount of computation power that can be put into a camera, sensor, or a smartphone. In addition, many of the devices at the edge of the network are not connected to a power source, which raises issues of battery life and heat dissipation. These challenges are being dealt with by companies such as Tesla, ARM, and Intel as they develop more efficient processors and leaner algorithms that don't use as much power.

But there are still times when AI is better off in the cloud. When decisions require massive computational power *and* do not need to be made in real time, AI should stay in the cloud. For example, when AI is used to interpret an MRI scan or analyze geospatial data collected by a drone over a farm, we can harness the full power of the cloud even if we have to wait a few minutes or a few hours for the decision.

Training vs. Inference

One way to determine where AI should live is to understand the difference between training and inference in AI algorithms. When AI algorithms are built and trained, the process requires massive amounts of data and computational power. To teach an autonomous vehicle to recognize pedestrians or stop lights, you need to feed the algorithm millions of images. However, once the algorithm is trained, it can perform "inference" locally—looking at one object to determine if it is a pedestrian. In inference mode, the algorithm leverages its training to make less computation-intensive decisions at the edge of the network.

AI in the cloud can work synergistically with AI at the edge. Consider an AI-powered vehicle like Tesla. AI at the edge powers countless decisions in real time such as braking, steering, and lane changes. At night, when the car is parked and connected to a Wi-Fi network, data is uploaded to the cloud to further train the algorithm. The smarter algorithm can then be downloaded to the vehicle over the cloud—a virtuous cycle that Tesla has repeated hundreds of time through cloud-based software updates.

Embracing the Wisdom of the "And"

There will be a need for AI in the cloud, just as there will be more reasons to put AI at the edge. It isn't an either/or answer, it's an "and." AI will be where it needs to be, just as intelligence will live where it needs to live. I see AI evolving into "ambient intelligence"—distributed, ubiquitous, and connected. In this vision of the future, intelligence at the edge will complement intelligence in the cloud, for better balance between the demands of centralized computing and localized decision making.

Follow me on Twitter.

Mohanbir Sawhney

I am a professor of marketing and technology at the Kellogg School of Management, where I also direct the Center for Research in Technology & Innovation. I have co-a...

Read More

https://www.forbes.com/sites/mohanbirsawhney/2020/01/27/why-apple-and-microsoft-are-moving-ai-to-the-edge/#41c8e5b32570