

AI on Edge

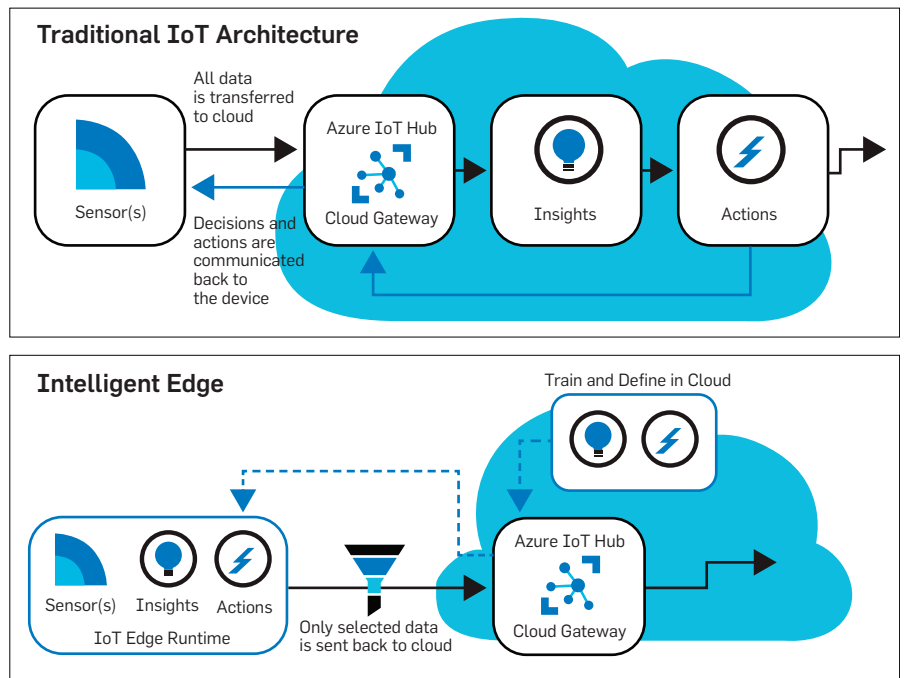
Shifting artificial intelligence to the “edge” of the network could transform computing ... and everyday life.

A REMARKABLE THING ABOUT artificial intelligence (AI) is how rapidly and dramatically it has crept into the mainstream of society. Automobiles, robots, smartphones, televisions, smart speakers, wearables, buildings, and industrial systems have all gained features and capabilities that would have once seemed futuristic. Today, they can see, they can listen, and they can sense. They can make decisions that approximate—and sometimes exceed—human thought, behavior, and actions.

Yet, for all the remarkable advancements, there’s a pesky reality: smart devices could still be a whole lot more intelligent—and tackle far more difficult tasks. What’s more, as the Internet of Things (IoT) takes shape, the need for low latency and ultra-low energy sensors with on-board processing is vital. Without this framework, “Systems must depend on distant clouds and data centers to process data. The full value of AI cannot be realized,” says Mahadev Satyanarayanan, Carnegie Group Professor of Computer Science at Carnegie Mellon University.

Edge AI takes direct aim at these issues. “To truly and pervasively engage AI in the processes within our lives, there’s a need to push AI computation away from the data center and toward the edge,” says Naveen Verma, a professor of electrical engineering at Princeton University. This approach reduces latency by minimizing—and sometimes complexly bypassing—the need for a distant datacenter. In many cases, computation takes place on the device itself. “Edge AI will enable new types of systems that can operate all around us at the beat of life and with data that is intimate and important to us,” Verma explains.

The power of this framework lies in processing data exactly when and where it is needed. “Edge AI introduces new computational layers between the cloud and the user devices. It distributes application computations be-



tween these layers,” says Lauri Lovén, a doctoral researcher and data scientist at the University of Oulu in Finland.

Pushing intelligence to the edge could also fundamentally alter data privacy. Specialized chips and cloudlets—essentially micro-clouds or ad hoc clouds that would function in a home, business, or vehicle—could control what information is sent from smart devices, such as TVs and digital speakers.

Beyond the Data Center

At the heart of edge AI is a simple but profound challenge: getting computing systems to make decisions at the pace of the human mind and real-time events. For artificial intelligence to fully blossom, any system incorporating AI must operate without any significant drop-off in speed and accuracy. This typically requires latency below 10 milliseconds. However, today’s clouds respond in the neighborhood of 70-plus milliseconds; connections that incorporate wireless connectivity are even slower, Satyanarayanan points out.

The current approach of forcing data

streams through a few large datacenters inhibits the capabilities of increasingly sophisticated digital technologies. Edge AI takes a different tack; it runs algorithms locally on chips and specialized hardware, rather than in distant clouds and remote datacenters. This means a device can operate without a persistent connection to a dedicated network, or the Internet, and it can access remote connections and transfer data on an “as needed” basis. Current frameworks, including “edge computing” and “fog” networks, offer only incremental benefits because chips are not optimized for AI and networks were not specifically designed for edge AI.

By creating “smarter” devices and curbing reliance on conventional datacenters, it is also possible to dramatically lower energy consumption. Chip maker Qualcomm claims its edge AI-optimized chips produce energy savings as great as 25x compared to conventional chips and standard computing approaches. Low-power wireless connectivity also reduces reliance on batteries that must be swapped

out or recharged constantly. Yet another benefit is that edge AI introduces stronger controls for sensitive and private information because data stays on the device or chip, or in a local cloud the user controls.

A low-latency framework requires new chips, storage devices, and algorithms, however. It significantly alters conventional computing models. “Modern mainstream data-driven AI methods, particularly in decision-making and machine learning (ML), are designed to be run in a cloud environment, with all data items always available for learning or inference on abundant and homogeneous computing resources,” Lovén observes. “The cloud-native paradigm is a poor fit for the opportunistic, distributed, and heterogeneous edge computing environment, where devices appear and disappear, connections fail, and device batteries run out—and where user and edge devices have widely varying computational resources.”

Smarter Devices

Pushing decision-making and other functions to the edge of the network produces dramatic changes, Verma says. For example, an autonomous vehicle could use onboard machine learning to adapt to different conditions and drivers dynamically. A collection of sensors in a home or hospital could better track patients, including the elderly, and detect potential problems, such as a patient’s inability to get out of bed or failing to take medications. Edge AI also could monitor the condition of underground pipes without any need to change a hard-to-reach sensor battery for decades. “Right now, what we do at the edge is fairly basic, but within a few years we will likely see robust functionality,” says Kurt Busch, CEO and co-founder of Syntiant Corp., a company developing Edge AI chips.

While many of these things already take place today without edge AI, eliminating the round trip to the cloud would significantly alter functionality. For example, it’s a safe bet that a language translation app today will function reasonably well in Barcelona or Beijing, but things get trickier in, say, the Gobi Desert of Mongolia, where there is no cellular connection. Yet, even when a strong signal exists, the process of bouncing phrases to the

cloud and back takes time, and it creates awkward, and often unacceptable, lags. Edge AI could solve the problem by storing all the needed data on the device and hitting the Internet only when it is necessary and desirable.

Another particularly appealing feature of edge AI is wake-on-command functions. These systems can dial down power consumption to near zero when a device isn’t in use. This allows some devices to operate for years or decades without a recharge or a new battery. Remote video cameras, medical implants, and embedded sensors would benefit from this feature. What’s more, many appliances—microwave ovens or coffee makers, for example—don’t require vast processing capabilities, or a Siri or Alexa, to operate; a couple of hundred hard-wired words will do. “The device becomes more responsive and delivers better privacy because you don’t have to deal with the roundtrip of the cloud,” Busch explains.

Edge AI could add new, more advanced features to smartphones, watches, smart glasses, smart TVs, Bluetooth ear buds, hearing aids, remote control devices, smart speakers, medical devices, and various IoT devices. However, Amit Lal, professor of electrical engineering at Cornell University, believes edge AI could have an impact far beyond microwave ovens that let people bark out cooking instructions, or a hearing aid that automatically adjusts to the user and the surrounding environment. As part of a team that oversaw the NZERO program for the U.S. Defense Advanced Research Projects Agency (DARPA) between 2017 and 2019, Lal and others explored ultra-low-power or zero-power nanomechanical learning chips that could harness acoustical signals or other forms of ambient energy and wake as needed. At some point, this research could lead to vehicles and other machines that can be detected by a unique acoustical signature. “You would verify the identity of the vehicle or other device before it gets close and poses a threat,” he says.

Rethinking and Rewiring AI

Realizing the full potential of edge AI requires a focus on things both practical and technical. There is a need for new devices and network models that bypass virtual assistants, smart speakers, and the cloud. A starting point for

addressing this task is engineering microprocessors designed specifically for deep learning and on-chip AI functions, including speech processing and wake-on-demand features. “Edge AI requires an entirely different framework for data collection, modeling, validation, and the production of a deep learning model,” Syntiant’s Busch says.

Syntiant is one of several companies developing chips specifically engineered for edge AI. Others include Ambient, BrainChip, Coral, GreenWaves, Flex Logix, and Mythic. Such chips typically run machine learning algorithms as 8-bit or 16-bit computations, which optimizes local performance but also reduces energy consumption, in some cases by orders of magnitude. Unlike traditional Von Neumann or stored-program chips such as central processing units (CPUs) and digital signal processors (DSPs), edge AI chips don’t need to swap data between the memory and the processor; instead, they typically rely on in-memory or near-memory data flow designs that place the logic and the memory data closer together. Busch says Syntiant’s Neural Decision Processor produces a 100x efficiency improvement over stored program architectures such as CPUs and DSPs.

Yet, the current class of edge AI chips is only a starting point. Busch says future edge chips likely will take on different designs and features, depending on the use case. Emerging memory technologies like Magnetoresistive Random-access Memory (MRAM) and Resistive Random-Access memory (ReRAM) could further optimize performance and power for specific uses cases, including ultra-low-power applications running independent of a data center. Other chipmakers are studying nonvolatile flash memory (NOR) as a way to store code on devices for more advanced machine learning functionality.

It will take more than new and better chips to push edge AI into the mainstream, however. Satyanarayanan says there’s a need to deploy cloud computing in entirely new ways. A decade ago, he introduced the idea of cloudlets—essentially a datacenter in a box—that could operate in planes, trains, automobiles, houses, and offices. “The same Xeon hardware that occupies a football-sized building would be adapted to a small box or rack to fit the envi-

acm

Advertise with ACM!

Reach the innovators and thought leaders working at the cutting edge of computing and information technology through ACM's magazines, websites and newsletters.



Request a media kit with specifications and pricing:

Ilia Rodriguez

+1 212-626-0686

acmm mediasales@acm.org

acm

media

“Widely deployed cloudlets would fundamentally change the way data flows, processes take place, and machines handle decisions.”

ronment. These hyperconverged clouds bring compute closer to the user. You wind up with high bandwidth and low latency,” he says. Such systems, and edge AI, could be further enhanced with the introduction of 5G, which better supports IoT frameworks.


Over the last couple of years, the idea of cloudlets and edge AI has begun to take shape. Amazon Web Services has introduced Wavelength, and Google has introduced Edge TPU, hardware and software solutions that accommodate edge functionality. Although edge AI technology poses questions, including how to approach physical protection and cybersecurity optimally, the model is garnering attention and gaining momentum. “Widely deployed cloudlets would fundamentally change the way data flows, processes take place, and machines handle decisions,” Satyanarayanan says.

On the Leading Edge

Moving edge AI off the drawing board and into everyday life will require a few other things. One particularly important requirement is distributed learning and inference algorithms that function in a dispersed, opportunistic, and heterogeneous edge environment with non-IID data (data that is dependent or unidentically distributed), Lovén says. How well these systems accomplish the task will determine how effectively they work and how much value they provide—particularly in highly connected IoT ecosystems.

In addition, there’s a need for libraries and frameworks that implement new and more efficient algorithms. Edge AI application developers and on-chip or on-device machine learning

tasks will require ready-made tools and resources. Moreover, these libraries must operate in different edge environments, including ad hoc clouds or cloudlets from different manufacturers. Lacking this framework, compatibility and data quality issues will emerge, and edge AI could stumble. “Existing frameworks such as Spark, Tensorflow, or Ray are essentially cloud-native, and their computational models are a poor fit to the edge environment,” Lovén says.

Despite technical challenges and new security concerns, edge AI will almost certainly gain momentum over the next few years. Not only will edge chips and other components appear in appliances, devices, and sensors, they will introduce entirely new ways to tap AI, neural nets, and machine learning—while perhaps recapturing a sense of privacy that has been largely lost in the digital era. Says Lal, “There are an incredible number of applications and possibilities for edge AI. If you make machines and sensors smarter and lower their power requirements, you open up a world of possibilities.” 

Further Reading

Satyanarayanan, M. and Davies, N. **Augmenting Cognition Through Edge Computing**, IEEE Computer Society, Volume: 52, Issue: 7, July 2019, Pages 37-49. <https://ieeexplore.ieee.org/document/8747287>

Lovén, L., Leppänen, T., Peltonen, E., Partala, J., Harjula, E., Pörömböge, P., Ylianttila, M., and Riekk, J. **Edge AI: A Vision for Distributed, Edge-native Artificial Intelligence in Future 6G Networks**, 6G Wireless Summit, March 24-26, Levi, Finland. <http://jultika.oulu.fi/files/nbnfi-fe2019050314180.pdf>

Rausch, T. and Dustdar, S. **Edge Intelligence: The Convergence of Humans, Things, and AI**, 2019 IEEE International Conference on Cloud Engineering (IC2E), 24-27 June 2019. <https://ieeexplore.ieee.org/abstract/document/8789967>

Murshed, M.G.S., Murphy, C., Hou, D., Khan, N., Ananthanarayanan, G., and Hussain, F. **Machine Learning at the Network Edge: A Survey**, <https://deeplearn.org/arxiv/113246/machine-learning-at-the-network-edge-a-survey>

Samuel Greengard is an author and journalist based in West Linn, OR, USA.

© 2020 ACM 0001-0782/20/9 \$15.00