EDGE AI: EDGE ARTIFICIAL INTELLIGENCE EXPLAINED

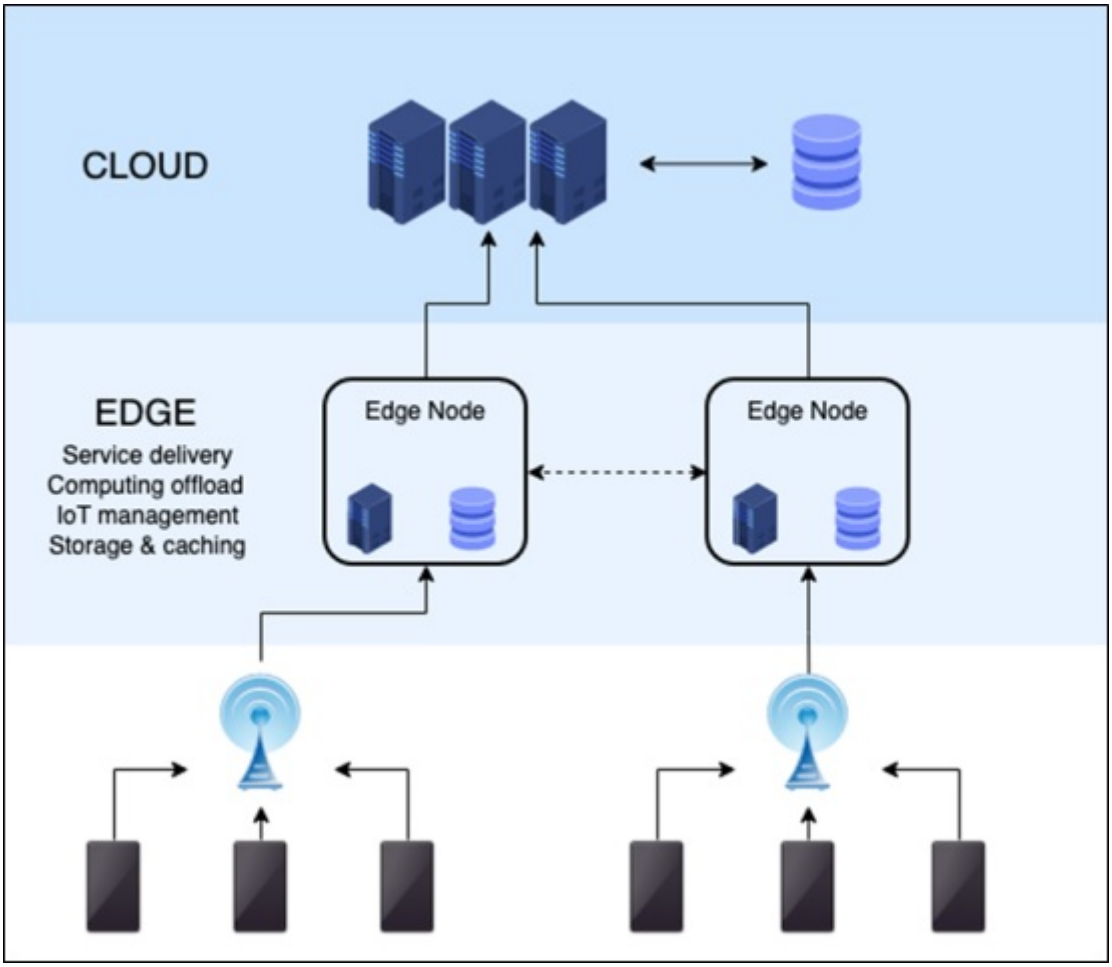
Edge AI utilizes the compute available on phones, sensors, raspberry pis, and other edge devices to train, load, and inference machine learning models.

Let's take a look at edge AI, including how it works and the pros and cons.

Computer dependencies

Computers depend on components to operate. In particular, there are:

- **Network dependencies**, like relying on a network connection to ping another computer on the network and get a return.
- **Resource dependencies**, like GPUs, memory, and CPUs.



Source

Computer tasks like rendering video, computing a function, or fetching data from the Twitter API can all be tagged as a <resource>-bound or <resource>-dependent.

Network-bound	CPU-bound	GPU-bound	Memory-bound
Web scraping	Processing files	3D Rendering	Storing data for computations like 3+4 = 7
API calls	Moving data	Video editing	Storing data to be processed
	Making computations	Video games	
		Machine Learning	

A new breed of processing units

Human ability and the increase in hardware technologies have aligned to create the demand for AI to perform on edge devices.

Machine learning models benefit from training on lots of data and reconfiguring its model's weights. It needs to run this task many, many times. CPUs are limited by their one-at-a-time processing capabilities. Even a quad-core or 16-core CPU gets 4 or 16 processes running simultaneously, but that pales to the processing power of a GPU whose design already allows for parallel processing, a multi-lane bridge crossing for processing large blocks of data.

The past decade has seen a different kind of chip emerge, specifically designed to handle tasks for A.I. Some of these new AI chips include:

- AMD's Accelerated Processing Unit (APU)

- iPhone's AI Chip
- Google's Tensor Processing Unit (TPU)
- Intel's Nervana

These chips are already available in most computers and in every smartphone. That means phones, one type of edge device, can begin to train and inference machine learning models.

Edge AI performs on your device

Today, we use cloud computing and an API to train and serve an ML model. Edge AI, then, performs ML tasks close to the user. Let's compare.

Without edge AI

Cloud computing has handled most of the heavy-lifting for models such as face detection and language generation.

Using that model (that is, no edge AI), the user's phone would merely transmit the data—an image or a piece of text—over a network to the service and let the service do its calculations. Then the results are sent back to the user.

With edge AI

With edge AI, data does not need to be sent over the network for another machine to do the processing. Instead, data can remain on location, and the device itself can handle the computations.

Benefits of edge AI

Eliminating the cloud service portion results in two benefits:

- Privacy is increased because data isn't passed over a network.
- Cloud computers are less strained.

Increased privacy

Like a vault getting robbed by an outlaw gang when it is transported on a train from the East coast to the West coast, data gets stolen when it is in transit. Even if it's not stolen, third parties can still know that something was transmitted between one party and another. With a little bit of digging, they can figure out what kind of information is being sent across the network.

Making ML inferences on location means that the data, and the predictions made on that data, never risk being seen while in transit. Your data doesn't get compromised, and the relationship between you and the AI service provider can remain unknown.

This is great for people. With edge AI, we have more control over who knows what about us—an important feature for society (and the dance people play with one another...controlling their image). Edge AI prevents third parties from knowing a person is seeing their therapist once a week and holds that conversation between patient and therapist private; it permits a person to disclose that information themselves.

Less cloud compute strain

Secondly, edge AI relieves workloads from cloud computers. Networks aren't strained. CPU, GPU, and memory usage drop significantly as their workloads get distributed across edge devices.

When cloud computing performs all the calculations for a service, some central location is doing a lot of work. Networks see a lot of traffic in order to get data to the source. Machines get started to perform their tasks, and the networks get busy again, sending data back to the user. Edge devices eliminate this back-and-forth transfer.

Like a busybody learning to delegate tasks to others, setting boundaries, the networks and machines are a lot less stressed when they're not doing everything.

Drawbacks of edge AI

Edge AI results in a couple drawbacks:

- Less compute power than cloud computing
- Much more machine variation

Less compute power

Naysayers might point out that edge computing is good, but it still lacks the computing power available in a cloud-computing system.

So, only select AI tasks can be performed on an edge device. Cloud-computing will still handle creating and serving large models, but edge devices can perform on-device inference with smaller models. Edge devices can also handle small transfer learning tasks.

Machine variations

Relying on edge devices means there's significantly more variation in machine types. So, failures are more common.

In the event of failure, orchestrators can help move jobs to other clusters, ensuring system resiliency. Still, there will be more failures to deal with overall.

Edge AI examples

These interactive tools are great examples of Edge AI:

- Text generation
- Toxic Comment Classifier

Additional resources

For more on this topic, explore the BMC Machine Learning & Big Data Blog or read these articles:

- What is the "Intelligent Edge"?
- The Empowered Edge: An Introduction
- BYOD Policies: Best Practices for BYOD in the Enterprise

- Shadow AI: A 2020 Trend
- A Primer: Machine Learning, Data Science, Artificial Intelligence, Deep Learning & Statistics