Your edge, your future

The edge is bursting with useful data. Architect and deploy your edge systems enabled by AI & ML that act in real time and help you make better business decisions.

HPE GreenLake

There's no turning back, and no reason anyone would want to: The edge, bursting with useful data, is the future. It's becoming the dominant source of all enterprise data. Its computer capabilities, combined with ever-advancing artificial intelligence, are growing at a fast clip. And add the economic forces at play—the expectation of computers doing more from afar, whether that's at a manufacturing site, a retail store, or inside your car—and we have an ever-evolving world of possibilities.

The upshot is that there doubtless will be far more aggregate computing happening at the several layers of the edge than exists in the data center today. There will also likely be as much networking, albeit in a different form. In fact, five years from now, we'll look back at this pivotal time in enterprise computing and it will seem intuitively obvious that IT organizations had to move their focus from the data centers that dominated corporate computing for five decades.

There is a parallel of sorts. The internet technologies commercialized during the dot-com boom in the late 1990s and early 2000s forever changed the scale of compute, storage, and networking in the data center as well as the technologies deployed to create that scale. Eventually, Big Data entered the picture and, combined with statistical and neural network software, allowed machine learning—envisioned three decades ago but impossible with the small datasets and puny parallel processors that existed until about a decade ago—to actually work.

With ML and AI, everything changes. Devices of all kinds gather up telemetry so they can help manage all aspects of themselves and deliver insights that compel either our action or that of another device. Put another way, all this alters the nature of the relationships that companies—and people—have with the world around them.

The edge is about scope. Instead of massive banks of compute and storage encapsulated in a data center, we have a swarm of orders of magnitude more computing elements out there, right where the real world is happening. Pushing this IT infrastructure to the edge is necessary because too much data is generated "out there," which makes relying on data centers costly and ineffective because things wouldn't happen fast enough.

The edge is about action

The edge will have precisely enough ML intelligence to turn live data streams, not large datasets, into some sort of work. The edge is about action, propelled by machine learning intelligence that was initially constructed in the data center and set free to roam the networks being stretched, quite literally, to cover the entire world. The edge devices may do their inferring and, perhaps someday soon, their own training right where the world is happening.

We are on the precipice of yet a new technological era. Today, of course, companies are not starting from scratch as they did during the initial computerization of the back office; there is no time for that. The people running companies need actionable insight, as the common phrase goes.

Everything outside of the data center—billions and perhaps someday trillions of devices—is being equipped with monitors of all kinds and networked to AI-enhanced compute. Data is generated in massive streams at speeds entirely too much for human beings to process.

But human capabilities aren't the point; the machines can run the show. These edge systems are about taking action, right there, right now. And that requires a new architecture, a new way of thinking, and expertise.

Intelligently build your intelligent edge

It's relatively easy to drop what amounts to a baby data center in a remote location. But any time you have more than a dozen applications running in a remote location and then many remote locations on top of that, IT organizations have to formalize the people, processes, and technologies.

Take the retail industry, which has had remote processing for running transactions and managing inventory for decades. Now, retailers are installing video cameras in their facilities to do pattern recognition and look for fraud within store aisles or at the self-checkout stand. The systems to do that are special servers driven by AI and supported by GPU processing. There are about a half dozen different pattern recognition applications, plus other sensing applications for asset management (including inventory and people), plus the point-of-sale applications already in place.

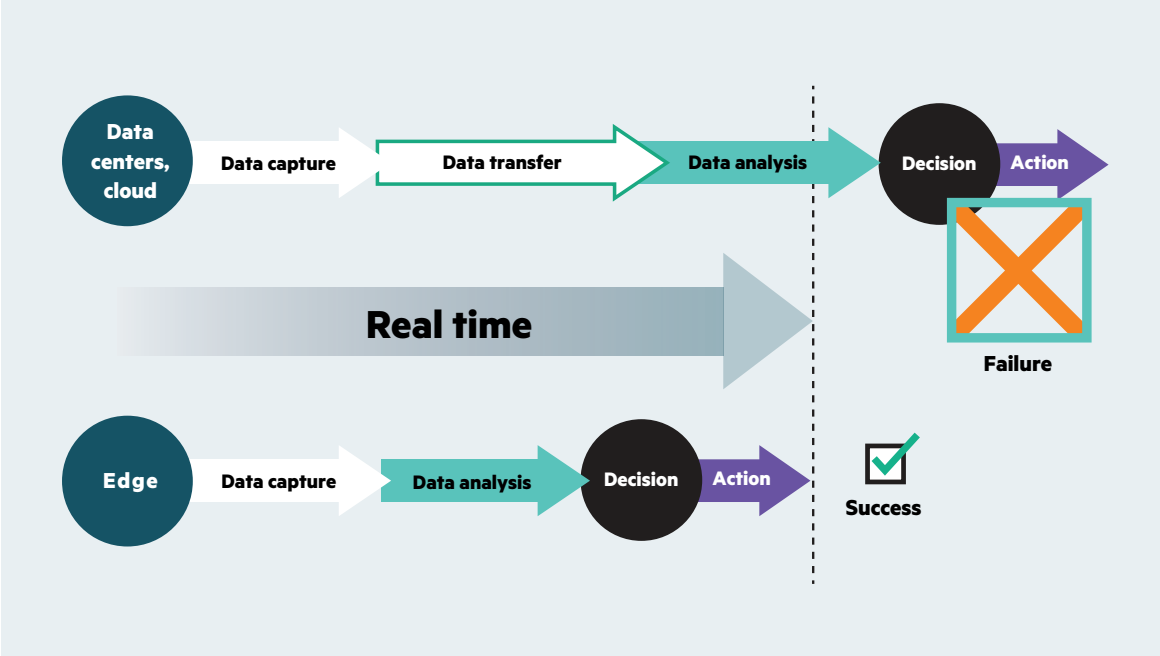

Figure 1. Real-time action

And all of the processes employed in the management of centralized IT—software release management, remote configuration and monitoring of hardware, change ticketing for tech support, remediation, and on and on—have to be applied to the edge.

That's why the edge is a more diffuse and potentially larger scale problem than the data center. The issues are more complex. And there will be many layers of edge on top of that, with infrastructure out there in the real world at the literal interface between people and things, but also points of presence and other kinds of aggregation edges that are not really part of the data center at all. All of these facts can catch IT organizations off guard if they just dive in.

The cultural shift will be just as jarring. Building data center infrastructure as we all know it—dense servers clustered to deliver scalable compute and storage over closed networks spanning a data center plus some legacy monolithic systems—will become an esoteric art as the focus of IT shifts from the data center to the edge.

The problem that IT organizations face will shift from "How do you stand up a room full of infrastructure to run 2,500 different applications?" to "How do you automate all kinds of interactions with people and machines and facilitate other kinds of infrastructure and deliver an experience or manage a process, end to end?"

Here is the other tricky bit. All edges are relative. Running some aspects of infrastructure in an intermediate place like a public cloud—perhaps in an aggregation layer with lots of analytics—makes it an aggregation edge relative to the further edge out there in the world. With the public clouds having hundreds of regions and then additional points of presence feeding into them, it is reasonable to expect that some of this capacity will be deployed as aggregation edges for expediency and data sovereignty reasons.

However, latency requirements for processing data in real time or near real time mean that compute should reside at the furthest edge. Also, there is not enough bandwidth in the world—not even in future 6G networks—to physically move all data back up to a centralized data center for processing. This demands that storage be local and some data be deleted after a period of time.

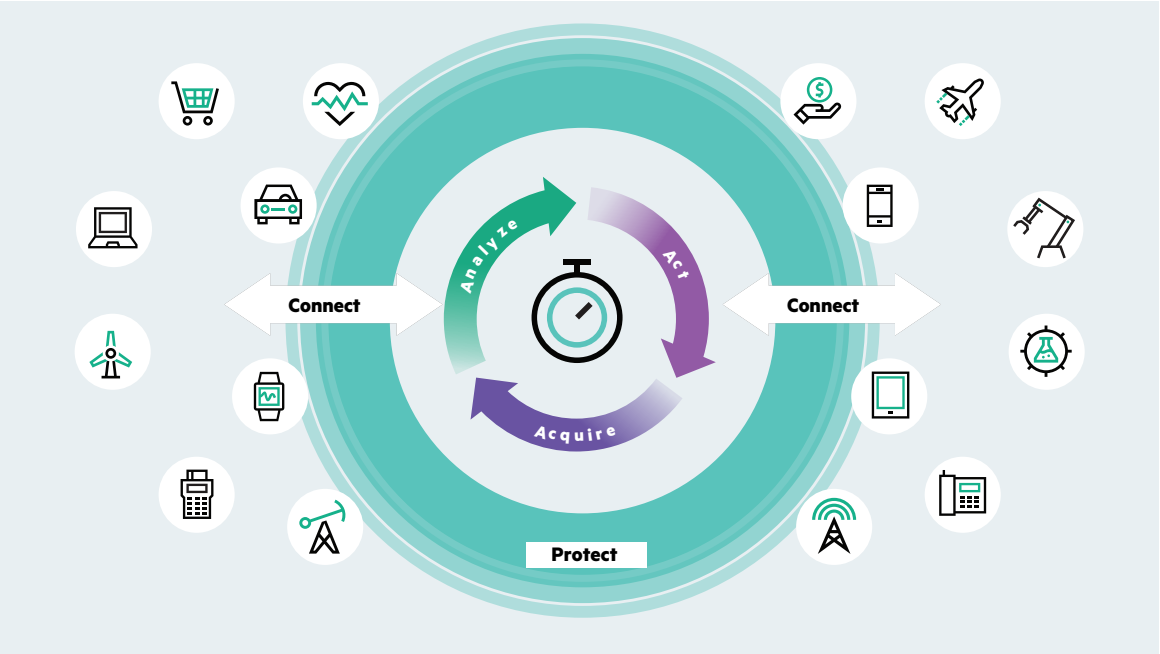

Figure 2. Key edge capabilities for better decision-making

Back to the video surveillance in the self-checkout line at a retailer. If someone scans the wrong barcode on an item, or forgets to scan one, the retailer would want to know immediately and send over an employee monitoring the queues to fix the issue. It has to happen in one second, maybe two seconds. Sending a video stream, even at 10 frames per second, up to a public cloud region will take tens of seconds, maybe longer, and while processing that could be fast given the scale of the infrastructure, the shopper will be driving home before the retailer figures it out.

A certain amount of compute, storage, and networking needs to be immersed in the physical environment because processing within that environment, instantly, is the whole value proposition of the edge. What is true of video surveillance in a self-checkout line at a retailer is equally true of a motherboard assembly line that has surveillance monitoring the quality of solder joints or a car manufacturer looking at its shop floor. Mistakes are costly everywhere in business, and catching them early, and often, means not losing time or money—or both.

Here is another subtle point to consider: Edges never exist by themselves. There are myriad applications that will use sensors and AI at the edge to control that physical environment in some way. There will be many different sensors and platforms that vendors will try to vertically integrate, but these will by necessity exist side by side in physical locations like stores, warehouses, factories, airports, and train stations. The edge is really a network of edges, and they will have to be managed in the aggregate in some fashion. Every edge device cannot have its own way of doing this because it is inefficient.

It remains to be seen who will do the installation and the management of these edge devices. It could be the companies using these devices. Or it could be the vendors that supply them and the companies merely provide a "data room" with a controlled environment where all of this stuff plugs in and the vendor manages it remotely.

System integrators will no doubt emerge to weave together different edges for specific situations and offer to manage the whole shebang. And customers will want as much of this infrastructure to look and feel like it is on-premises data center gear and yet be offered in a cloud-like subscription model. In the coming years, edge compute should comprise as much as 20 percent of server installations (using a much looser definition of what constitutes a server than is common today), and over time, it could grow to be two to three times the aggregate compute capacity across the data centers of the world.

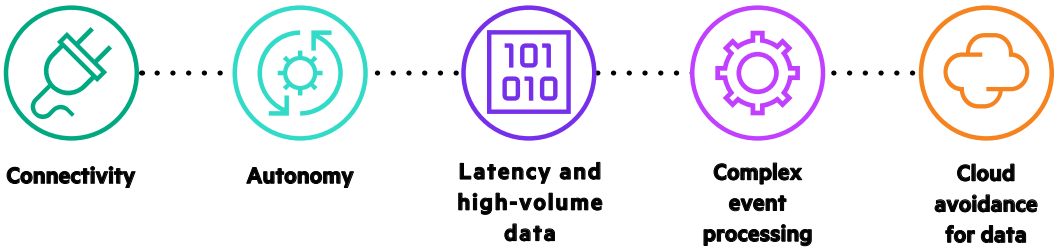

Figure 3. What to consider when deploying edge computing

Pushed to the edge

Here are five things companies looking at deploying edge computing need to think about as they architect their platforms (see Figure 3):

Connectivity: There is no such thing as pure hardware anymore. Everything is connected, and at the edge, everything starts with the sensors in touch with the physical world, no matter what they are, and works its way backward from there. And things have to work locally even when the connectivity up the chain of data and processing doesn't.

Autonomy: Most factories today already have some form of automated guided vehicles—what some might call robot forklifts—but AGVs are not limited to that form factor. They used to need network connectivity to function, but with machine learning techniques, they can continue operating safely without that connectivity. Or take mining operations. Sensors on motors in drilling equipment are used to show that a drill bit is about to break or has broken. This could be even more useful. The same telemetry coming off the drill bit can be monitored and run through AI algorithms to warn miners that they are hitting a particularly hard kind of rock before the bit breaks. This saves the bit and the motor and the downtime associated with the bit breaking.

Latency and high-volume data: The ability to make split-second decisions is necessary out there on the edge, and like a lunar or Mars rover, decisions have to be made locally, in real time, because it can't take minutes to make a decision because of network latency. Similarly, if it takes high volumes of data to make a decision, and that data can't be moved efficiently or in a cost-effective manner, that also forces decisions to be made and actions to be taken at the edge.

Complex event processing: In the drill bit example above, what is really at work is complex event processing across multiple sensors, all in real time. There are vibration sensors in the drill as well as gauges that look at the energy input to and current draw by the motors. Or in augmented agriculture, farmers will take the historical data from the national weather centers and mash it up with sensors at the edge to get real-time conditions in the actual fields and figure out when to sow specific plants. The more sensors, the better the assessment of the situation and the better the action that is taken out on the edge.

Cloud avoidance for data: No one is avoiding putting data onto the cloud because the storage is not good storage. Rather, there's just going to be far too much telemetry to keep, and it's going to be far too expensive to store and even more expensive to use—even at the relatively cheap, and declining, prices of cloud storage. The data volumes at the edge will grow faster than the price of data storage on the cloud comes down, so storing edge data and processing it in the cloud is a losing proposition from the get-go. Here is a great example: autonomous vehicles. If you are training models for self-driving cars, you need to gather lots of information from real-world conditions to train those cars—typically dozens of cameras and hours of driving, which is petabytes of data. Even if you could upload that data to the cloud for storage, moving to an AI cluster to do training routines would cost a fortune. The easiest thing to do is to build a replaceable storage array and drive it in the car to the data center and swap it out.

The edge is competitive, and companies need a competitive edge. That means building an edge strategy that is focused on achieving very specific results, not installing technology for its own sake. As a starting point, it means doing projects that take data and telemetry from many sources in the physical world—and doing so in a secure fashion—and then applying machine learning models to them to take automated action in the physical world.

Learn more at

greenlake.hpe.com/edge

Explore **HPE GreenLake**

Make the right purchase decision.
Contact our presales specialists.

Chat now (sales)

Call now

Get updates

HPE GreenLake

a00121148ENW, Rev. 1