THE DOPPLER

VOLUME 2 · 2021

YOUR EDGE. YOUR FUTURE.

WHERE INSIGHT MEETS OPPORTUNITY

AN ENTERPRISE.NXT PUBLICATION FROM HEWLETT PACKARD ENTERPRISE

Editor-in-Chief
Paul Sloan

Editorial Team
David Chernicoff
David Crozier
Al Madden
Monica Sambataro
Larry Seltzer

Contributors
Alyson Behr
Glyn Bowden
Aaron Carman
Robert Christiansen
Jeff Enters
Tim Ferrell
Trent Fierro
Iveta Lohovska
Saadat Malik
Peter Moser
Lin Nease
Christopher Null
Garry Orsolini
David Rand
Yara-Yasemin Schuetz
Hartmut Schultze
Michael Tennefoss
Dan Tynan
James P. Woods Jr.

Creative
Tim Harms
Peter Sucheski
Darby West

Hewlett Packard Enterprise

THE DOPPLER is published by HPE and Enterprise.nxt.
Learn more at hpe.com/insights

IN THIS ISSUE

FOREWORD

ANTONIO NERI
President and CEO, HPE

It's not an overstatement to say that business growth and transformation—even societal progress at large—will be driven from the edge.

After all, the edge is where all of us live and work, whether that's a manufacturing plant, a sports stadium, or outer space. It's where transactions occur, where customer experiences happen, where businesses produce goods and people consume them.

Above all, the edge is about data. Eighty percent of all data will be processed at the source—primarily at the edge—by 2025. That makes understanding and analyzing all this data key to differentiating your business and succeeding. Those who can act on their data the fastest will be the winners.

The next wave of transformation—the one we're now in, which I refer to as the Age of Insight—demands a mindset shift. An edge-centric world by its nature requires technology landscapes that are more diverse and dispersed than ever before. The very definition of on-prem or off-prem will need to change. IT operations work best in controlled, standardized environments. Edge use cases are anything but. (To understand the evolution of all this, please read "Your edge, your future.")

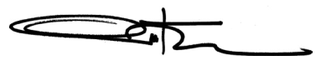
The winning approach to these challenges takes the agility, simplicity, and economic flexibility of the public cloud and applies it across this diverse edge landscape. This is not merely a technical challenge—it requires fundamental changes to organizational skills and ways of working.

In this edition of The Doppler, we explore the possibilities of this edge-centric world and provide some guideposts for success. A key to success in any domain is to recognize the nature of the opportunity and challenge in front of you. In "What is the edge, really?" we offer expert perspectives to help shape your thinking, and in "Boosting the bottom line," we look at some of the critical success factors in delivering enterprise-scale impact from your edge-to-cloud transformation.

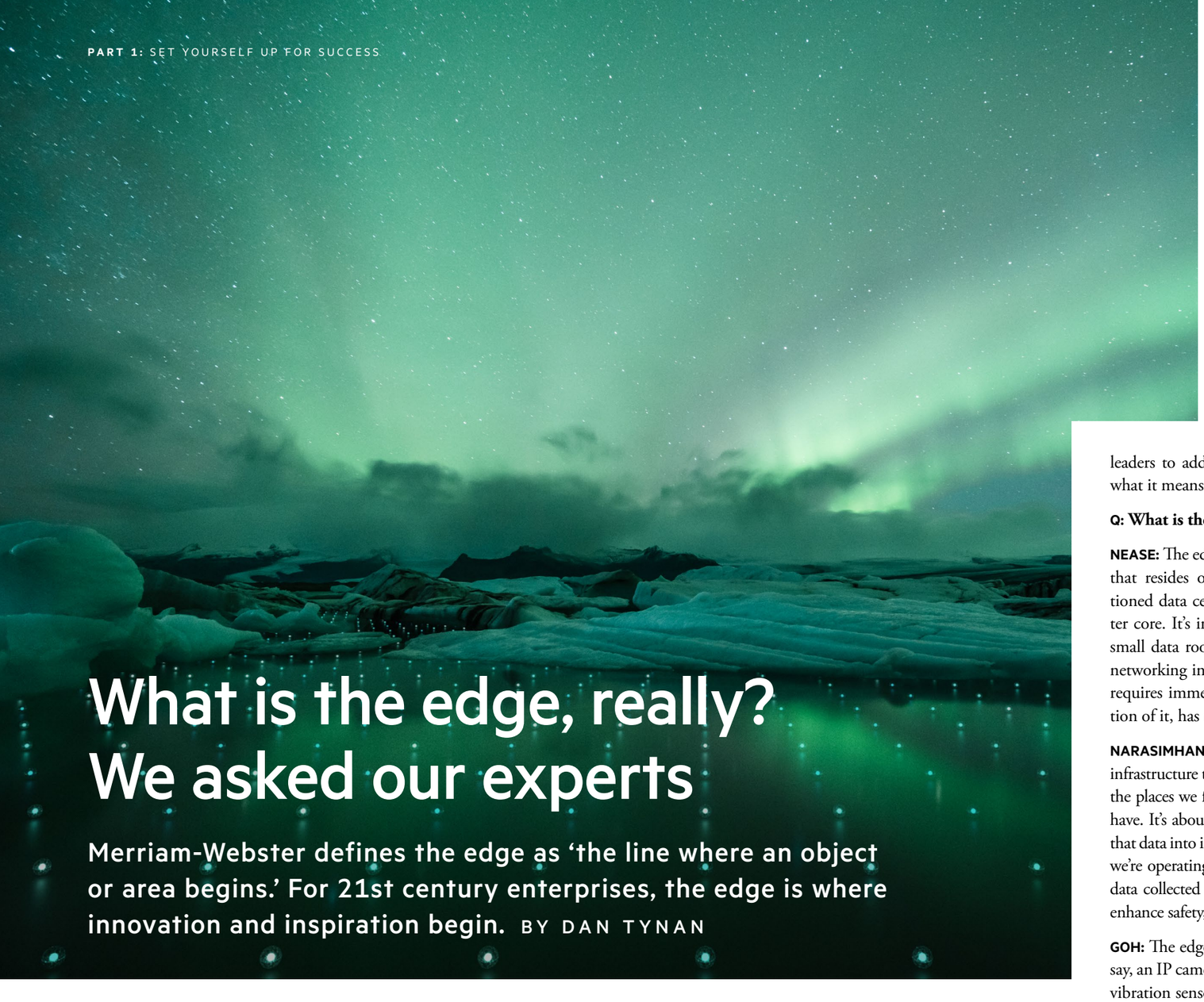
Edge initiatives present new challenges. We have devoted a section to addressing these challenges, with articles such as "Crushing complexity with automation" and "Meeting new security demands." Finally, it's important to understand that the edge is here and now. Leaders around the world are using insights from data at the edge to drive new experiences and transform the way they operate. We have shared a series of these stories, detailing the practical applications that are driving outcomes today.

No transformation worth doing is trivial. But with a clear view of your desired outcomes and a holistic approach to the journey, anything is possible.

We at HPE are delighted to be on this journey with you.



Set yourself up for success



What is the edge, really? We asked our experts

Merriam-Webster defines the edge as 'the line where an object or area begins.' For 21st century enterprises, the edge is where innovation and inspiration begin. BY DAN TYNAN

AN INCREASING AMOUNT OF computing power is moving away from data centers in the cloud and closer to the physical world. By the year 2025, IDC predicts there will be an estimated 56 billion connected IoT devices on the planet,[1] and 75 percent of all enterprise data will be generated at the edge.[2]

But what is edge, exactly? At HPE, we broadly define the edge as a place where people, places, things, and their data intersect. A more precise answer depends on how you apply the edge to your specific business needs.

Edge devices run the gamut from simple embedded sensors that keep your offices cool and energy efficient to complex industrial robots that are building the next Mars rover. They include the automated checkout machine you use at the grocery store, the intelligent cameras that keep a watchful eye on city streets, and the AI-powered MRI machine that helps your doctor interpret your latest scans.

To further refine what edge computing is and how organizations can use it to their advantage, we've assembled a dream team of subject matter experts. We asked these technology leaders to address some basic questions about the edge and what it means for enterprises.

1. "How You Contribute to Today's Growing DataSphere and Its Enterprise Impact," IDC, Nov. 4, 2019
2. "What Edge Computing Means for Infrastructure and Operations Leaders," Gartner, Oct. 3, 2018

MEET THE EXPERTS

LIN NEASE is chief technologist for IoT in HPE's advisory and professional services group, as well as an HPE Fellow.

PARTHA NARASIMHAN is CTO at Aruba, an HPE company and leading provider of next-gen network access technology.

DR. ENG LIM GOH is CTO for high-performance computing and artificial intelligence at HPE.

Q: What is the edge?

NEASE: The edge is the tentacles of the octopus. It's everything that resides outside those massive multipurpose air-conditioned data centers, which we now think of as the data center core. It's incredibly diverse and includes everything from small data rooms to smartphones to sensors, along with the networking infrastructure that connects them. Anything that requires immersion in the physical world, or a digital rendition of it, has to happen at the edge.

NARASIMHAN: At HPE, we begin by defining the edge as any infrastructure that's not inside the data center. But it's also about the places we find ourselves in and the experiences we want to have. It's about digitizing the physical world and transforming that data into insights that provide more visibility into the spaces we're operating in. In an industrial environment, for example, data collected at the edge can be used to improve operations, enhance safety, reduce costs, or increase profitability.

GOH: The edge is the point where the data is first collected—say, an IP camera connected to a server in a retail store, a smart vibration sensor on a machine inside a factory, or an electron microscope. A key characteristic of edge devices is that, on their network side, they have much more data flowing out than coming in. A third defining feature is that these devices are constrained in some way: by the amount of power you can provide to them, the kind of environment they must operate in, or the amount of bandwidth available.

Q: Why is the edge suddenly such a hot topic? What's driving this?

NEASE: We've reached the point with Moore's Law where we're able to bring compute and data processing into the physical world in the form of IoT devices. And that enables a wide range of new use cases. For example, we can build facial recognition logic into cameras, so they can identify people immediately. When we're driving and roaming between cell sites, our car's navigation system still has to work; that requires processing data at the edge. If you're operating a natural gas pipeline across a remote stretch of land, you might have at best a low-bandwidth satellite connection. The logic that's controlling the pumps has to happen at the edge.

NARASIMHAN: The number of personal and IoT devices is growing, which means data is continually increasing. Processing all of that data in a centralized location or the cloud is expensive. So organizations end up just dropping that data on the floor and not doing anything with it. But if you can process that data at the source, the costs become much lower and more use cases become viable. It's all about bringing a cloud-like experience to the edge and enabling new low-latency services like augmented and virtual reality.

GOH: Sensors are continually becoming more sensitive, and there are more and more of them. That means they're generating more data than ever. But the cost of bandwidth isn't decreasing at anywhere near the same rate, so it becomes very expensive to backhaul all that data. Collecting and processing it locally can be more cost efficient. At the same time, the number of applications requiring low latency is increasing. When you don't have time to send data up to the cloud and wait for an answer to come back, that's when you need processing power at the edge.

Q: How important are technologies like cloud-native computing and AI to making edge computing work?

NEASE: Extremely important. To take one example, HPE's own servers are built using AI at the edge. In our factory in the Czech Republic, the system performs 80 visual checks of a motherboard in less than 90 seconds. In the past, humans

> HPE built a high-performance computer for the International Space Station—the literal edge. It had to survive radiation levels 10 to 100 times higher than on the earth's surface."
>
> **DR. ENG LIM GOH**
> CTO, High-Performance Computing and Artificial intelligence, HPE

would do it, but they'd occasionally miss some things. With AI, the number of quality control issues has dropped dramatically. If this pattern recognition were performed in the cloud, it would probably take 10 minutes or more.

NARASIMHAN: Cloud-native technologies like microservices, containers, and orchestration are important for agility and scale. Nothing stays the same for very long, and you need the ability to keep pace as the technology changes. You may also want to run some applications on the edge and others in the cloud, then shift between them depending on the needs of the customer or the environment. Using cloud-native services at the edge gives us the flexibility to do that.

GOH: Nearly every edge computing scenario uses AI or rule-based machine learning models to some degree. Sensors upload data to the cloud, which is used to train a machine learning model. That model is pushed back down to edge devices using containers. These devices then use that model to make decisions independently. Today, edge devices are making inferences based on cloud-generated models; in the future, they may become smart enough to learn locally on their own.

Q: What are the barriers to wide-scale deployment on the edge?

NEASE: The biggest barrier is complexity. The number of systems we need to deploy, manage, and monitor multiplies dramatically at the edge. As a result, the attack surface becomes huge. How many times have we seen "the bad guys" round up a bunch of video cameras or HVAC systems and generate denial-of-service attacks on mundane infrastructure that still uses default admin passwords? Security at the edge is a big problem.

NARASIMHAN: It's really mindset. You can't think about the edge the same way. We're not shipping storage and infrastructure to customers and having them deploy and manage it. It needs to be centrally managed and orchestrated as a service. That's going to face a lot of resistance. The other big issues are security and privacy. How do we provide the right tools to manage data on behalf of the customer? The only way to gain their confidence is to be open about who has access to PII [personally identifiable information] and other data and how that process is managed.

GOH: The energy and environmental constraints I mentioned earlier: If you're building an edge device to go into a car, for example, it has to be able to run off a 12-volt battery. You can't use a fan for cooling because it will quickly become clogged by dust. If it's operating next to people, it can't be too loud. And so on. HPE built a high-performance computer for the International Space Station—the literal edge. One of the constraints was that it had to draw less than 500 watts. *[Editor's note: Supercomputers like the HPE-Cray Aurora require 60,000 times more than that.]* Another was that it had to survive the vibration of blast off and radiation levels 10 to 100 times higher than on the earth's surface. Those all had to be factored into the design.

Q: What else should enterprises be thinking about as they begin to deploy edge technologies?

NEASE: The biggest constraints of the edge aren't on the technology side. The constraints are in process and adoption patterns. Companies that embrace the edge faster than others will enjoy big advantages, and new businesses will emerge as a result.

NARASIMHAN: There's been such a mad rush to digitize and automate the physical world that we're underestimating the number of vulnerabilities it creates. I remember being at an Aruba user conference in 2017 when AWS went down and we couldn't run any of our demos. It was both amusing and alarming to see the number of people complaining that they couldn't open their garage doors or operate other devices in their homes. We just assume these things will be available when we need them. The level of attention cybersecurity requires is not keeping pace with the rate of edge adoption.

GOH: Customers need to be clear on why they're considering edge technologies. Are they thinking about it for technical reasons or strategic ones? If they have an unsolved technical issue that will solve business problems and where they think the intelligent edge will help, we approach it one way. But if they're saying, "We have this vision to pivot our company and we believe the edge will be a great way to start pivoting early," then that's a very different conversation. ■

Your edge, your future

Building the intelligent edge means making intelligent choices.

BY LIN NEASE

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.

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.

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

FIGURE 1: **REAL-TIME ACTION**

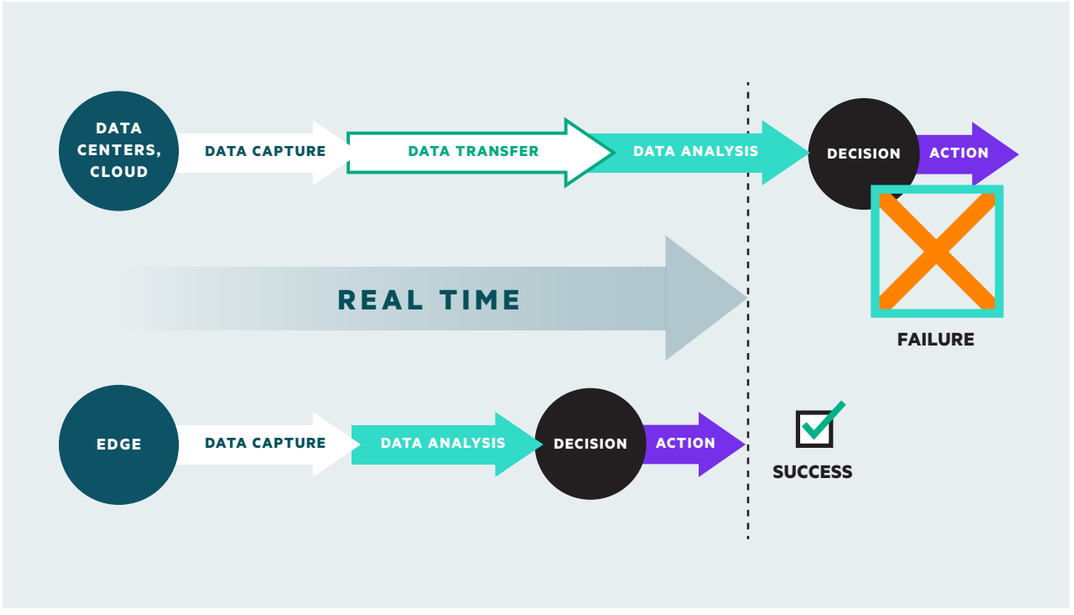


FIGURE 2: **KEY EDGE CAPABILITIES FOR BETTER DECISION-MAKING**

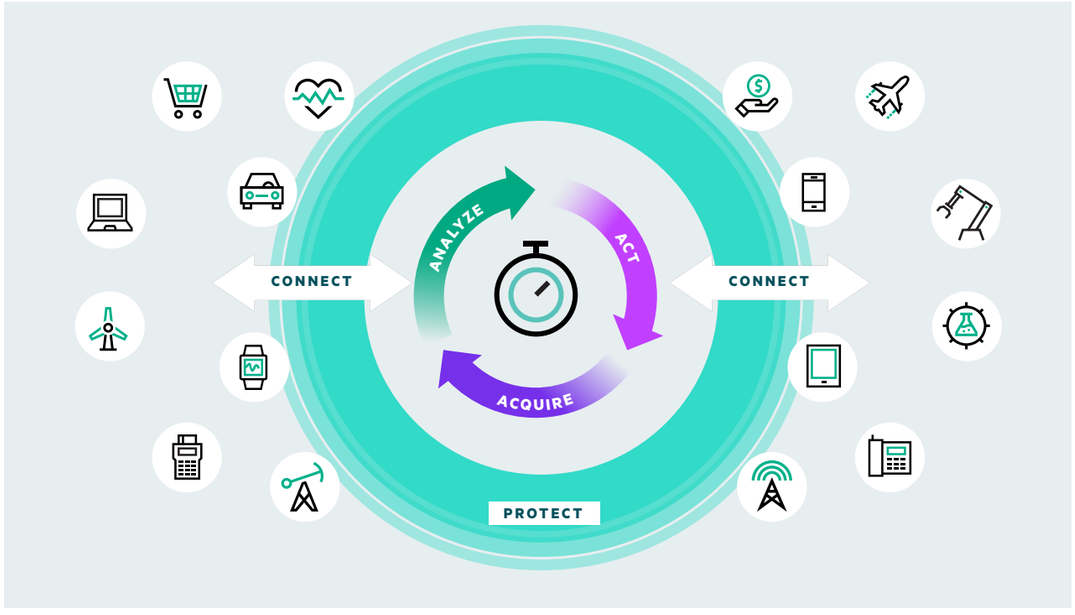


FIGURE 3: **WHAT TO CONSIDER WHEN DEPLOYING EDGE COMPUTING**

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.

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. ■

Making the edge boost your bottom line

Expertise combined with an as-a-service model can help you overcome hurdles to achieve success at the edge.

BY SAADAT MALIK

EDGE COMPUTING PROMISES to deliver the next level of improved operations, unlock new opportunities, and increase business agility—not just in a single plant or location but across all locations. Every extra dollar that can be generated in one location is multiplied by the number of locations a given business operates. Making an extra $1,000 per location across a thousand locations delivers $1 million to the bottom line. It is this multiplication effect that makes edge computing so attractive and worth pursuing.

Moving computing closer to the data

By pushing compute power to the distributed edge locations of an enterprise, data at the location can be analyzed in real time. This is something cloud computing cannot do. Moving data to and from the cloud simply takes too long, and doing so pushes the time needed to analyze data and take corrective action outside the window of opportunity. Quick decisions are better made at the location where the activity takes place.

Edge projects can be complicated, however. A comprehensive edge computing solution may have hundreds of

TIME MATTERS

A factory producing 5,000 complex machines per month per production line has only 376 seconds to manufacture each unit. Burdening the production with a cloud-based QA process can add up to 79 seconds to the build time. Eliminating this cloud round-trip tax per unit increases the factory capacity by 21 percent.

The plant manager cannot ignore this ability to create "free capacity" and must therefore adopt systems that maximize the output of production equipment. To unlock the extra upside, QA data analytics must take place in the factory itself.

Edge computing is the answer, as it is much faster than cloud computing. And as an increasing amount of data is generated in the factory—at the edge—this strategy becomes increasingly attractive over time.

cameras and sensors as input devices. In addition, a given configuration may be replicated within a given site and across sites in different geographies.

Here are three components that can contribute to a successful edge computing strategy:

All as one

Edge computing cases vary widely in style and scope, ranging from a shoebox-size data center on an oil rig to thousands of Bluetooth low-energy tags at a retail site. So organizations need to choose the technology mix that best applies to the particular job. Edge computing installations can be stationary or highly mobile. And, as application needs change, enterprises have to be able to move workloads where they're needed most, so they can facilitate the business outcome they want to achieve.

Managing edge computing installations across the globe is a large undertaking, one more demanding in scope than managing a data center in a single location. Adopting a cloud-anywhere strategy, delivered in an as-a-service model, can help enterprises reduce complexity, retain control, and scale edge projects to the point where they can generate positive business outcomes.

Moving real-time computing, analytics, and control to each edge location does not mean complete autonomy for each location. While the location may operate autonomously, managerial oversight is still needed across locations. Orchestrating edge locations as a pool of distributed assets borrows from the trucking industry's concept of fleet management, whereby the organization inserts a control point above the locations that enables the entire distributed enterprise to move as one.

This is where the centralized nature of cloud computing provides a great opportunity to aggregate common data across sites and provide a single version of the truth across operations. Workloads can be published from a central location to each site as older workloads are retired, packaged in a container or a virtual machine, or offered as a service. What's important is moving fast and retaining flexibility.

Pay-as-you-go models

If a coffee shop chain wants to improve customer experience as well as profitability, adding compute power at its edge—at each of its 3,000 shops—will help. The profits can quickly add up by generating an additional 25 cents per cup in revenue. But rolling out new technologies across locations comes with cost and risk. The management challenge: How does the business gain the benefits of edge technologies while balancing investment capacity and risk? With an as-a-service approach, the company can structure a plan to pay a service organization to manage the rollout as profits roll in.

As-a-service models are maturing, and they're going to be important for edge computing. Edge projects need to scale quickly from proof of concept to production, and this requires economic flexibility. Outcomes can be projected during a pilot program, and project payouts can be structured based on money saved or revenue generated.

What do you do if something goes wrong? The answer is straightforward: Partner with a service provider with expertise, technology, and global reach to solve problems proactively.

Managerial expertise

Large scope and scale have a direct impact on the economic performance of an edge-driven company. But managing vast networks that are geographically dispersed and technologically diverse can be challenging. Just because you have a great idea for an edge project doesn't mean you have the expertise to envision the parameters, deploy the hardware, scale the operations, and optimize the entire system. It's far harder to fix a series of camera glitches across multiple locations than to fix a server malfunction in a data center down the hall. What do you do if something goes wrong? The answer is straightforward: Partner with a service provider with expertise, technology, and global reach to solve problems proactively.

The path forward

As edge strategies evolve, companies will face challenges. Taking advantage of flexible technologies, payment plans, and management models, they can clear initial hurdles and scale their projects to deliver the outcomes they desire.

By combining a global partnership with an as-a-service model, the benefits of edge computing are now within reach. Any business or government entity can improve operations by moving to an edge-driven strategy. And doing so is now more achievable than ever. ■

Getting your priorities right

A common language across stakeholder groups is critical to moving the right initiatives forward.

BY YARA-YASEMIN SCHUETZ

AS SOCIETY LOOKS BEYOND THE PANDEMIC, savvy enterprises are pulling out whiteboards and brainstorming ways they can seize emerging opportunities. They're thinking about what needs to change in their organizations to capture these opportunities and what role edge technologies can play.

There's plenty to consider. Teams can easily become paralyzed by choice. What's needed is a model that aids prioritization, balancing tradeoffs of the options available and their potential benefits and the opportunity costs of paths not taken.

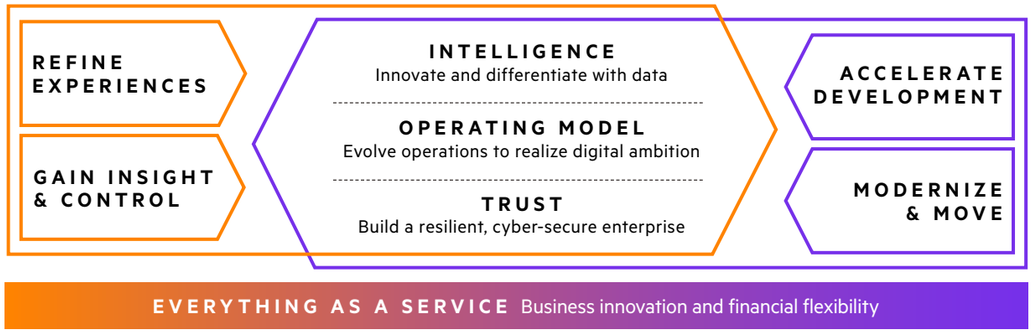
How do you architect technology platforms to digitize the physical world and project the experience into the digital world with customer-centric strategies? This requires anticipating the customer experience and an operating model that covers the customer journey all the way back to cloud ecosystems.

To best take advantage of the digital edge, start by doing three basic things: Understand your business goals, choose technologies that serve those goals, and move to an operating model that connects the two.

Where business happens

There are many definitions of the edge—so many, in fact, that we dedicated an article to it. From a technical perspective, it's the process of bringing data and computational power to the locations they're needed to deliver the best decision-making capacity for a given use case. But from a broader perspective, the edge is where business—and life—happens. It's the point of value exchange, where IT starts to meet the customer experience.

FIGURE 1: **COMMON ENABLERS FOR DIGITAL TRANSFORMATION**



Finding those points of value exchange requires a model that applies to all kinds of use cases. The model needs to address situations where digital engagement not only targets emerging growth opportunities but also helps a business recover from a series of bad quarters.

A common language for transformation

How does edge-based innovation or protecting the business by improving customer experience translate into a technology strategy? First and foremost, the entire organization needs to "get it." There is a moment when the value of the idea makes perfect sense to everyone, can be visualized, and can then be realized.

Using a model can help provide the language of digital transformation, a common framework on which to hang your organization's strategy and priorities. Figure 1 outlines a model that has helped a diverse range of digital transformation engagements globally. What the model does, at its basic, is help teams prioritize activities to orchestrate the teams' digital ambitions.

Blending a "human edge" (digital engagement focusing on people) with a "physical edge" (controlling physical spaces through data and automation) creates a "digital edge" that channels experiences contextually distinct to each company.

Still, organizations won't get far without integrating a critical set of enablers: intelligence, an underlying foundation of trust that maintains integrity of data and transactions, and an operating model that delivers speed and adaptability.

Let's look at each of these in turn:

- Data is just data without intelligence, which is generated by implementing advanced techniques to capture, consume, and process the raw material into strategic insights.
- A successful trust model uses technology to secure physical assets, establish compliance, and govern brand perceptions inside and outside the company.
- A digitally enabled operating model rethinks how value is created by using digital technologies. It forces you to look at processes, the value chain, who's doing what, and how you're providing your value to your customer. It up-levels the perception of technology inside of organizations and enhances management systems to help drive better, faster, more reliable value creation.

Companies need to evaluate their digital transformation initiatives in terms of how well they help meet underlying business goals. These goals fall into four categories: driving revenues and productivity, improving business efficiency, speeding time to market, and optimizing costs and increasing agility.

Generating more data from this physical edge—from a set of sensors or a camera feed tracking consumers, for example—won't be beneficial if you can't derive digital insights. Can those insights help you apply new levels of control to your physical spaces or assets? Think about how fans in stadiums are encountering very different experiences than fans outside of stadiums. If we have all the data about where those fans are physically in relation to the services the venue offers, we can start to enrich that digital experience for everyone, even those outside the stadium.

And don't ignore the digital ecosystem as a whole. The ecosystem is a set of partners securing the digital initiative. There are actually two ecosystems of partners: the internal one that builds your value proposition and delivers it, and another that's external to you and part of a bigger system.

Organizations are as individual as the customer

While using a common framework to help focus and orchestrate strategy is a valuable starting point, the nature of ideation, and discussion sparked by the framework, is how each organization can bring the concepts to life in the context of its unique circumstances.

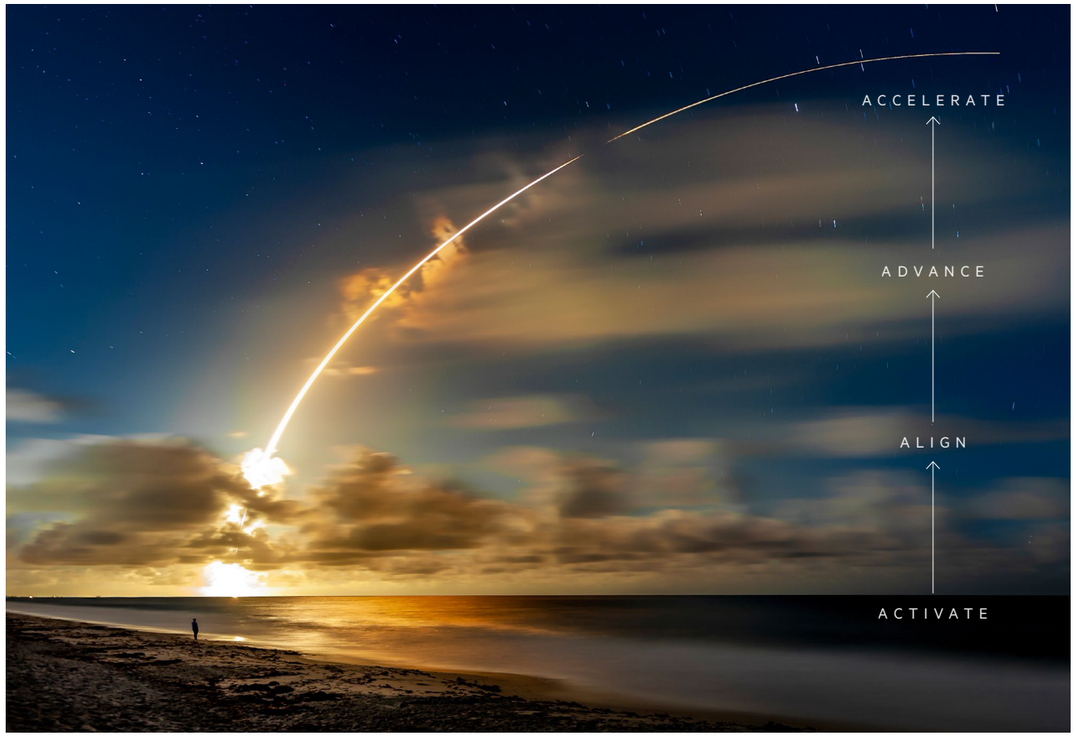
A number of organizational interactions—think of them as *moments*—can help set the right focus and tone for a digital transformation journey mapping exercise:

- **ACTIVATE:** Framing and communicating your digital transformation ambitions across your stakeholders. This is the ability to easily communicate and visualize what your digital agenda is all about. Aligning behind this common model to articulate your own journey is immensely powerful.
- **ALIGN:** Gaining an understanding of how to sequence and orchestrate your digital transformation initiatives. How do you see a clear path to value without dropping any of the many spinning plates you've been balancing during the transformation process? By isolating the common enablers for an initiative, it's easy to start to de-risk the path to value by connecting initiatives together and avoiding duplication.
- **ADVANCE:** Rapidly identifying a differentiated value proposition that can be delivered with technology. This is the time to create a minimum value proposition to attract investment in a uniquely new customer experience.
- **ACCELERATE:** Getting from idea to benefit realization as quickly as possible. Even if you can see the path forward and it's unique, how do you establish a minimum viable architecture to test the idea? What are the key building blocks? Stop spinning up lengthy and costly proofs of value—get to a minimum viable product with the minimal number of steps.

Takeaway

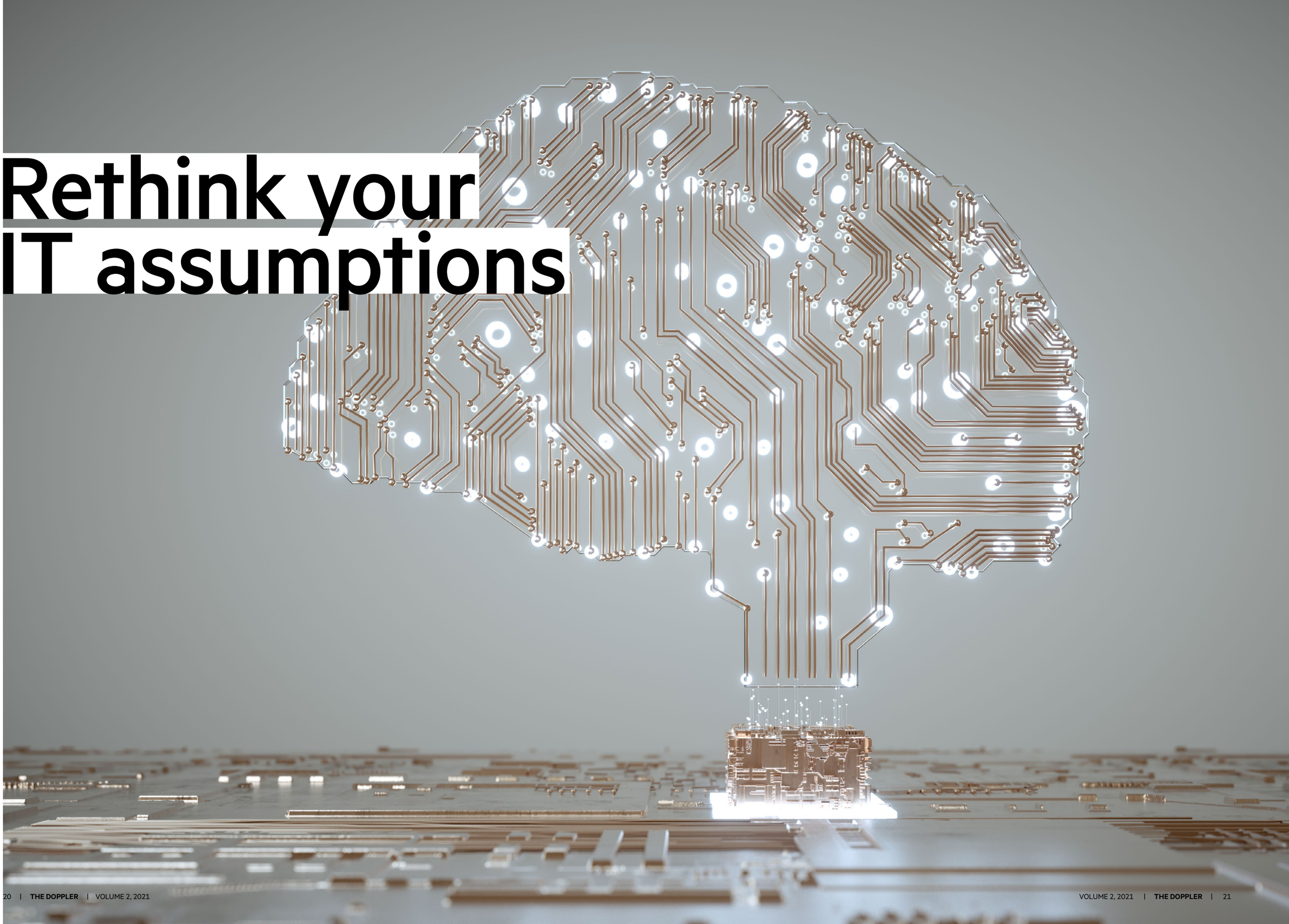
The edge is where you create new value. It is where digital ambition meets digital realization. Many enterprises that were undertaking digital transformation strategies pre-pandemic are now speeding them up in response. Each company will be looking at how to adjust its own digital edge to try to align projects that help it achieve its business goals.

To carry out your organization's digital agenda, you need to not only prioritize the right initiatives but also scale those ideas to fit the constructs of the company's business plans. Moving more rapidly through proof of concept and proof of value to realization is better achieved through the use of a holistic model that steers companies to the right decisions, while extending the ecosystem from edge to cloud, across the transformation landscape. ■



PART 2

Rethink your IT assumptions



AI sharpens its edge

As artificial intelligence applications continue to shrink, they're increasingly appearing in a device near you.

BY CHRISTOPHER NULL

ARTIFICIAL INTELLIGENCE HAS TRADITIONALLY been confined to the data center, where powerful computers have been tasked with running complex algorithms that are managed by seasoned professionals. In many parts of the enterprise, that is changing, as the power of AI is rapidly making its way to devices on the edge.

To understand why this is happening and how it works, it's important to first understand what the edge is in today's enterprise. In the world of AI, there are two fundamental edges where AI is deployed, commonly known as the *near edge* (which is relatively near the central data center) and the *far edge* (which is close to where data is generated). The application of AI is different at each of them.

Glyn Bowden, CTO of the AI and data practice at Hewlett Packard Enterprise, explains that the near edge refers to areas that are located outside the data center but have capabilities similar to those in the data center. Near edge locations "could be something like a factory building or a hospital, where you can do computing in a reasonably robust environment," he says, though such locations may lack the traditional design of a classic server farm. As well, those locations are remotely managed and rarely have trained IT professionals on site—despite having ample computing, networking, and storage capabilities available.

The near edge is defined by fairly typical computing equipment and resources. Model training and predictive analysis can take place here, thanks to the availability of traditional computing power. Conversely, the far edge refers to a different class of devices, generally Internet of Things products such as cameras, industrial sensors, drones, and even personal devices like a user's smartphone. The far edge refers to the point where data is actually being captured or where a device is interacting with the end user directly. These devices can't train an AI model, but they can use one that's already been developed. AI at the far edge has to be designed to make faster decisions with less available data and is generally hyper-focused on a singular task.

Edge-based AI of both varieties is already here. Ian Hughes, an analyst at 451 Research, an S&P Global Market Intelligence company, says a 2020 Voice of the Enterprise survey found that 60 percent of respondents were analyzing and/or storing data at the edge.[1]

The case for the edge

Why does AI at the edge matter? It's not just becoming useful to have edge-based AI functions at the edge. In many cases, it's absolutely essential.

Take the near edge example of a hospital. AI has vast applicability in medical environments, from helping physicians make a proper diagnosis based on a large amount of data to research on treatments and vaccines for all manner of diseases. But while the modern hospital is filled with sensors, scanners, and mountains of patient data, making use of all that data isn't easy in a traditional computing environment, in part because privacy regulations like HIPAA and GDPR may prevent that data from leaving the confines of the building. "But if you can train your models on that data within the hospital boundaries," says Bowden, you can still effectively leverage AI tools.

60% **of respondents were analyzing and/or storing data at the edge**

The same is true for environments that are disconnected from the Internet or don't have reliable or fast enough network service to connect to a corporate data center. Many modern factories are located in developing regions, remote areas, or both, places where high-speed Internet access just isn't possible today—and if it is possible, it's extremely costly.

By moving AI functions directly into the factory environment, machine learning can be used to optimize operations, predict equipment failures, and uncover production errors quickly, minimizing the potential for financial loss. Relimetrics offers one example of this technology in action, developing industrial equipment that uses computer vision and machine learning to inspect components as they come off customers' production lines. Incorporating AI at the machine level in this way has reduced defects for the company's clients by an estimated 25 percent.[2]

Without an AI algorithm monitoring production, a machine will run for longer when it's in a less-than-ideal condition, and it can raise the risk of becoming damaged. "A machine can make decisions about whether to stop or not when things are flowing through, rather than going back to a data center and having an operator make that call," says Bowden. "Being able to make that decision in real time, at the machine, means that you can prevent a lot more loss and operate with a lot more efficiency."

At the far edge, the uses for AI are more focused. Here we're talking about individual devices that have nowhere near the processing power of the data center—such as the microchip inside an MRI machine or a robotic arm on the production line—and none of the stores of data that a broader, near edge environment would enjoy.

In these examples, AI has a simpler and more dedicated function, albeit one of equal importance. "Here, we have small models doing very specific things," says Bowden. That might include a medical imaging machine (which may not be connected to the data center at all) determining whether a patient is holding the correct pose before a photo is snapped or a single piece of industrial equipment that is constantly analyzing its operating temperature and making minor adjustments to keep production quality as high as possible.

With medical technology, there's a strong data privacy use case for edge-based AI as well. "The more decisions that can be handled at the edge where the data is stored, the better it is for the customer from a privacy standpoint," says Iveta Lohovska, senior data scientist at HPE Pointnext Services. "I think customers are becoming increasingly aware of those issues."

Ultimately, this is a market that's undergoing a massive amount of growth. "We're seeing lots of capabilities at the edge," says Bowden, "and it's changing all of the time."

AI hits the road

Edge-based AI isn't just improving things in hospitals and factories; it's also becoming a key tool for mobile applications—namely self-driving vehicles. "It makes autonomous driving possible," says Bowden. "You wouldn't want

1. Voice of the Enterprise survey on IoT technologies, 451 Research, 2020

2. "Making zero defects a reality," Relimetrics case study, Hewlett Packard Enterprise

your car having to communicate with a data center before deciding whether to stop in front of the wall or not. You want that decision being made there and then, as you're in the vehicle."

As well, consider drones that carry high-resolution cameras. If the drone can perform a real-time analysis on what it's observing to determine, say, the condition of crops in a field under observation or the legitimacy of a military target, that improves both performance and accuracy. "The drone might not have connectivity where it can push enough data to get a meaningful inference back," says Bowden. "So it has to be able to make its own inferences on the device."

For example, some drones are being used to examine tunnels where wireless signals won't penetrate, often a fleet of them flying at a time. "If an accident happens," which is common, "AI makes it easier for the rest of the fleet to fill in the gaps," he says.

AI is filling another key role in air travel. Luuk van Dijk, CEO of Daedalean, a maker of autonomous piloting systems, says, "Our applications give aircraft eyes and a visual cortex," and edge-based AI is a key technology on commercial airplanes. A plane can't carry a full data center onboard—it would be too large and too heavy, and consume too much power. Edge-based AI tools are now being used to provide navigation, air traffic monitoring, and landing guidance, much like their terrestrial counterparts.

"It's essential when you have bad connectivity, no remote dispatcher, or no GPS signal and have to rely on your eyes," van Dijk says. And the key is that all of this needs to happen virtually instantaneously: "A remote AI that receives signals from your cameras, processes them, and sends the situational information back to you is just not an option in flight."

Forging ahead

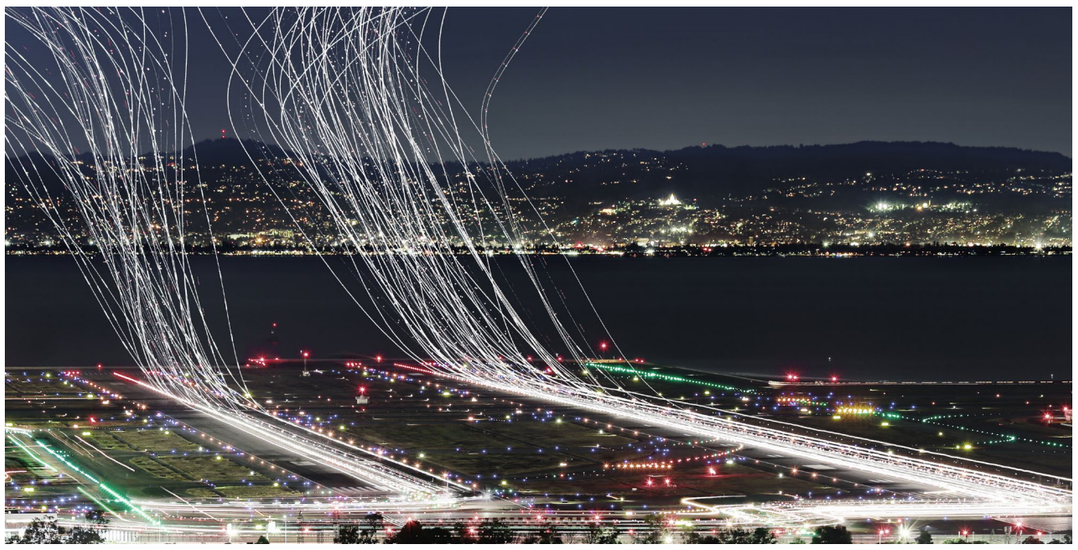
Edge-based AI is only now becoming possible because of a combination of factors all culminating at once. Naturally, miniaturization is allowing the creation of smaller and lower power microprocessors and denser storage, so more processing power can be wedged into a smaller space than ever. But perhaps more important is that the industry is getting better at streamlining AI models so they can become truly portable.

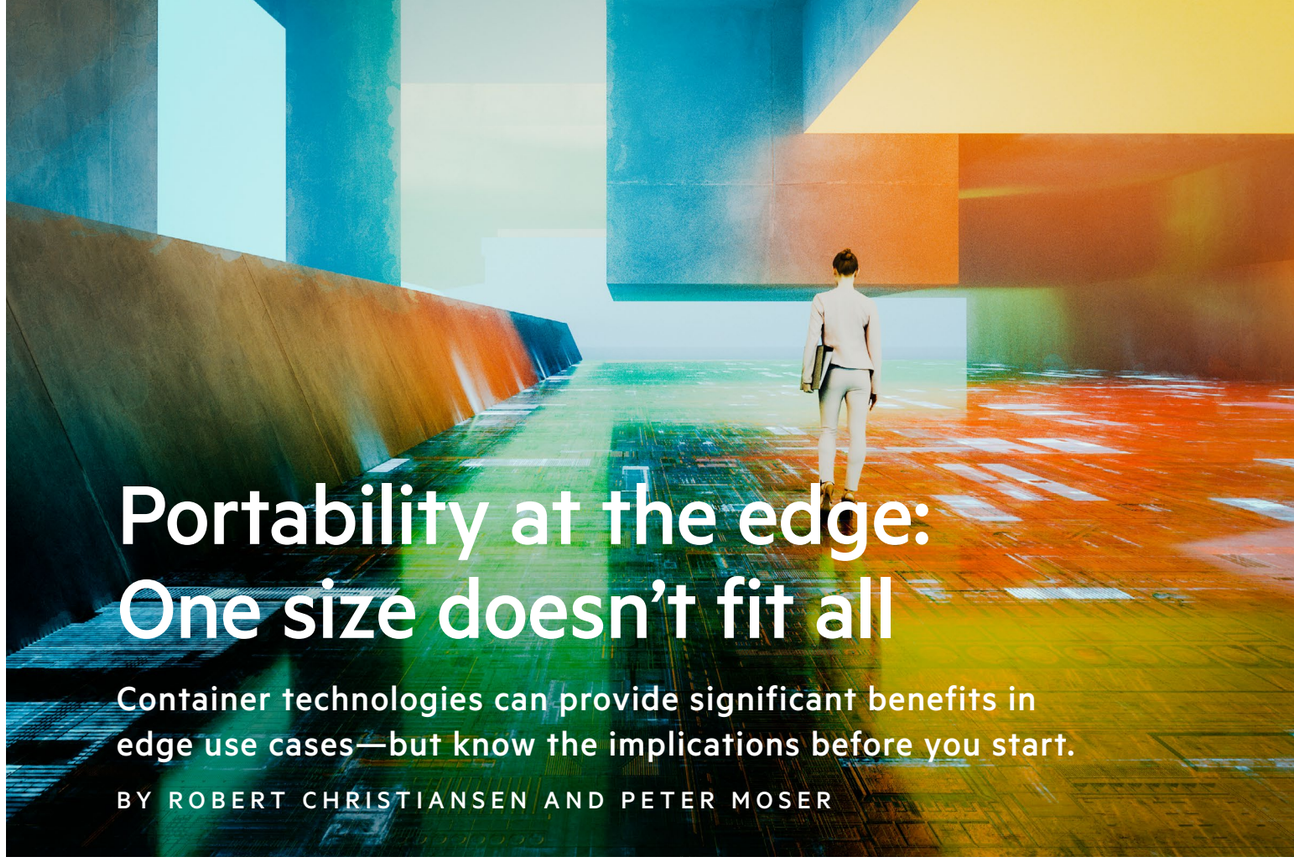
"Models are being compressed and made more lightweight," says Lohovska. Techniques like alpha-beta pruning are being used in edge scenarios to decrease the total number of nodes in an AI algorithm, essentially trimming away the least likely possibilities to improve speed and decrease the size of the model.

Smaller AI doesn't just mean a more efficient AI; it means more AI. "The fact that we're shrinking them also means that we're able to use many more of them," says Lohovska. That proliferation, in turn, is creating a virtuous cycle where more and more processing is able to be pushed to the edge.

"The more you can do at the edge—the more you distribute that workload—the less you have to worry about congestion on the network and available computing power at the data center," says Bowden. "Edge-based AI is really about the immediacy of data and the ability to do things in a disconnected environment."

It's also, of course, about making technology more capable at the point of usage. "Applications that understand us and what we need to do should over time remove some of the complexity of the devices we use," says Hughes. "Most computing technology has evolved around the needs of the machine, not the people using them… and AI has the potential to help in that area." ■





Portability at the edge: One size doesn't fit all

Container technologies can provide significant benefits in edge use cases—but know the implications before you start.

BY ROBERT CHRISTIANSEN AND PETER MOSER

IN RECENT YEARS, IT DEPARTMENTS HAVE SCRAMBLED to equip their infrastructures to accommodate the arrival of two disruptive forces: containers and edge technologies. These two sets of projects were often executed in parallel. Now, as edge use cases become more common, more sophisticated, and more impactful on the business's bottom line, containers and the edge are becoming more intertwined.

What does it mean to extend container deployment out to the edge? There's a perception that edge technologies such as the Internet of Things, machine learning, and 5G-connected stations are ubiquitous resource consumers that can be served by a one-size-fits-all container strategy. But it's more complicated than that.

Edge use cases differ widely, and a strategy for one might not work as well for another. A hospital's edge might be hundreds of patient rooms, while a manufacturer's edge might be on production lines. Other examples might be a transportation company that generates data throughout its fleets of disconnected trucks or a bank's edge with tens of thousands of ATMs.

Where's your edge? And how do you devise deployment and portability strategies that serve your particular edge computing requirements? Answers to these questions will go a long way in determining whether your edge initiatives will generate the information, insights, and overall business value you're expecting.

What to consider

The first factor to consider is what kind of role virtual machines will play in portability strategies for the edge. VMs are often used at the edge because many traditional, legacy workloads can't be containerized. Additionally, many edge workloads are Windows-based and, unlike their Linux-based counterparts, aren't ideal for containerization.

But, for most applications, containers' lightweight nature presents an advantage when moving workloads to and from edge applications. Moving a workload from the core or cloud to the edge is typically done over lower bandwidth WANs. This can be a challenge when moving a large workload over the network. With container-based workloads, the size of the programs are much smaller and thus easier to move over lower bandwidth circuits.

Edge use cases ranging from autonomous vehicles to augmented reality to predictive maintenance to blockchain all have one principal thing in common: They consume mountains of data. The question is where and how to process all that data. For some initiatives, the business case will support data processing at the edge, but not always. And you can't send it all back to the core. So you'll need to develop a strategy that determines what data will stay at the edge and what will be sent back.

That strategy will need to incorporate storage for older, brownfield data sources packaged up with VMs and newer,

more data-intensive greenfield sources that will likely be deployed through containers. They can't interfere with each other. You have to think about edge data solutions in a holistic manner.

You'll also need to ensure you have a consistent data fabric from your edge locations to your core—whether you're using a data center, a cloud, or a mix of the two. The process of moving, replicating and accessing data from edge locations to your core has to be consistent, and it has to meet compliance requirements. If government rules regarding data privacy require your data to stay in a particular location, your data strategy needs to take that into account.

Security's a priority, too. How do you maintain a security posture across thousands of edge locations without collapsing your existing security model today? Customers are using different forms of connectivity—LTE, Wi-Fi, Bluetooth, and 5G, for starters. The attack surface is becoming more complex, so you have to rethink your security strategy at the edge for your apps and the data itself. You can't use the technologies of the past because the scale of the problem demands automated responses. You have to have something much more automated and self-healing.

The priority is to get these worlds—VM, legacy, container—to coexist at the edge.

Finally, how are you pushing out your container-based applications to the edge and the rest of the world? Application lifecycle management becomes critical as the number of edge nodes increases. You'll need a platform that deploys the applications and their supporting components as well as a system that tracks the application version at each node and applies appropriate governance. There will be a blend of legacy edge applications deploying workloads on VMs and newer ones in Linux or containers. If a customer can get a platform that orchestrates both, that can create economies of scale. The priority is to get these worlds—VM, legacy, container—to coexist at the edge. A successful container strategy includes the vehicle of consistent application deployment anywhere, anytime.

What needs to change?

Companies have adopted Kubernetes as the industry's de facto container orchestration platform, providing the foundation for app portability and common platform-as-a-service functions, no matter which private or public cloud platform they use. Kubernetes streamlines workflows, automating conventional tasks like networking, storage, logs, and alerting. But scaling a Kubernetes platform to manage tens of thousands of nodes at edge deployments is hard.

To serve edge requirements, you need to change your context on several fronts. Regarding Kubernetes, edge node size will determine which deployment model to choose. Kubernetes tends to be bulky and not easily deployed on small compute nodes. Consider K3S (lightweight Kubernetes) as an alternative for thinner compute nodes.

Also, non-x86-based chips will become more and more prominent in edge nodes. These simply will not run a bulky K8S (standard Kubernetes) distribution. Therefore, consider deploying containers without the orchestrator. The challenge here is trying to perform orchestration without an orchestrator! Your operations practices—and skills—may need an upgrade to handle container deployments to bare-metal nodes. Again, this is harder than initially thought. Application lifecycle management is the main vehicle to simplify this process.

You have to change your context around data fabric and data motions. Today, you're probably working with disparate and disconnected data services that have specific connectors to applications. If you want to share data from one network-attached storage device to a whole new set of applications, you need to make sure the replication strategy doesn't cause complications. Investigate data fabric software platforms that support a global namespace for consistent data access, such as HPE Ezmeral Data Fabric.

Most use cases today include intelligence for machine learning capability. Your strategy around data curation and developing systems that learn will require a data fabric that is significantly more robust and more flexible than what is typically deployed today.

Questions to ask

As you extend your portability strategy out to the edge, here are key questions you should ask yourself:

- *What is the network connectivity to our edge?* Is it highly connected, where you're always talking to it, or less highly connected, like in an autonomous vehicle where you're only connecting once every day or two?
- *Does our edge unit need to act autonomously?* Does it need the intelligence to self-heal, fix security problems, and manage data? And when it does connect to the broader network, can it rapidly receive updates and push data back and forth in a timely fashion?
- *What does the service model look like for our edge location?* Do you have 25 locations that are highly managed by people, or do you have 2,500 with a team of three? This will dictate the level of automation your edge deployment will need to manage for the quality of service you desire.
- *How sophisticated are the connected devices at our edge?* Do your IoT devices have embedded intelligence? If so, what are the capabilities that can remain inside the edge vs. capabilities that need to be connected back to the core for an answer? Your quality of service requirements will dictate what intelligence needs to be at your last mile vs. what the device needs for a phone-home system.
- *How much resiliency do our edge applications need?* If you have an AI use case at a pump station that samples data three times a day, a service failure wouldn't be costly. If you're running data analytics for quality control on a plant floor and the server goes down, that can throw off production timelines.
- *What are the application lifecycle management requirements?* The higher the number of edge nodes, the greater the challenges with application management in those nodes.

Adapting your container strategy

Edge uses are here, and they're not going away. That means container strategies need to adapt. By considering the specific opportunities and challenges edge use cases present to container deployments, enterprises can position themselves for innovation and success no matter what applications they deploy or where they decide to deploy them. ■

Crushing complexity with automation

New automation technologies provide a way to effectively manage the complexities of evolving edge infrastructures.

BY AARON CARMAN AND TRENT FIERRO

JUST A FEW YEARS AGO, A POWER SUPPLY ISSUE at a major airline data center led to a cascading series of problems that crashed its IT and business operations worldwide, canceling thousands of flights and stranding tens of thousands of passengers. It was a public relations disaster that in large part resulted from a relatively minor issue—re-plugging a power outlet—that could not be identified and managed quickly enough before its consequences spiraled out of control.

And despite the notoriety of the event and its obvious business repercussions, something just like it can still happen at any moment to any organization.

Traditional enterprise technology infrastructures are just too complex to manage. Despite decades of investment in tools and systems, the current IT management landscape is siloed and more complex than ever. In an Enterprise Strategy Group survey, more than 60 percent of IT decision-makers said their infrastructures were more complex than they were just two years earlier.[1] And in another study, more than three-fourths of survey respondents said the complexity of their environments presented the biggest barrier to the productivity of their IT operations.[2]

The enterprise technology environment cannot be managed effectively or efficiently without help. Across the board, the decentralization of network systems and compute resources has upended the game. The requirements of big data, artificial intelligence (AI) platforms, and edge architectures exceed the capabilities of data systems that were never designed to handle them. And all these trends are dramatically accelerating in the wake of the COVID-19 pandemic.

Overnight, enterprises challenged with managing a fragmented IT landscape find themselves in a world in which their entire workforce now works from home. That same phenomenon is driving a renewed urgency for accelerating digital transformation. But perhaps most challenging is the proliferation of new edge devices and infrastructures: Nearly 9 billion Internet-connected devices are in use today—and that number is projected to triple to nearly 56 billion by 2025.[3]

Deploying IT workloads and supporting infrastructure across such a decentralized environment is a challenge that requires software management capabilities outside the typical enterprise data center. There are just too many devices to manage and too many issues that could occur, many of which cannot be understood. So complex and challenging is the situation that nearly 90 percent of business leaders surveyed said their organizations would need intelligent automation just to keep pace with changing business requirements.[4]

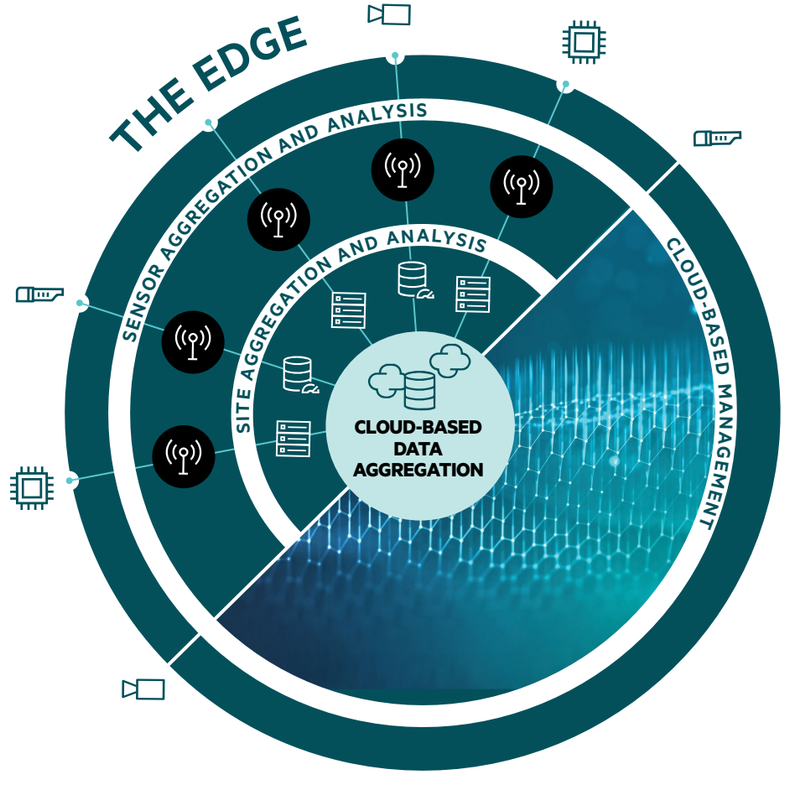
The revolution will be automated

Automation offers businesses a way to gain control of digital transformation and the tsunami of new edge infrastructures. When automation combines with AI, the edge is transformed. It allows enterprises to take control and manage their edge infrastructures in much the same way they manage their on-prem data center and cloud operations.

So, what needs to be done to create the intelligent automated edge? There are three core components:

- Holistic infrastructure intelligence and control
- Integrated software, including third-party workloads
- Infrastructure sensors and closed-loop remediation

FIGURE 1: **INTELLIGENT EDGE AUTONOMY WITH CENTRALIZED MANAGEMENT**



1. "ESG Research Report: 2019 Technology Spending Intentions Survey," Enterprise Strategy Group, Feb. 15, 2019
2. "The Impact of Automation on IT Operations: Are You Ready for the Software-Defined Data Center?" Freeform Dynamics, July 2017
3. "How You Contribute to Today's Growing DataSphere and Its Enterprise Impact," IDC, Nov. 4, 2019
4. "Today's State of Work: At the Breaking Point," ServiceNow, 2017

The first core capability requires adding AI and machine learning (ML) techniques into the software that runs edge data applications and processing. Infusing the edge with intelligence allows it to operate on its own. It understands and reacts appropriately to unusual behavior or other user or operational deviations from the norm.

The second core component is management software that orchestrates seamlessly with third-party tools and workloads. Very few, if any, enterprise environments are greenfields, meaning any intelligent infrastructure requirements need to augment, not replace, existing infrastructure investments.

The third core component is providing the sensors that will connect, communicate, alert, control, and solve problems on the edge. This last piece is especially critical for edge environments in environmentally complicated locations.

These three components provide the critical capabilities for an intelligent automated management system that can improve performance from the network to edge computing infrastructure and software.

Automated edge management systems can see and anticipate problems on the edge before they begin; reallocate workloads in the event of a problem; and, using closed-loop remediation, repair and reestablish themselves independently. This frees enterprise IT staff to focus on new value-creating tools or services for the organization rather than monitoring screens for alerts.

The cost of data center downtime is

$9,000

per minute

Major benefits

So, what are some of the benefits of automating network management operations?

- **REDUCED COMPLEXITY:** The typical enterprise has a large investment in legacy tools and systems. Managing such tools and systems was a complex process even before digital transformation and the rise of edge infrastructures. By unifying and automating these systems, organizations can be centrally managed from a single pane of glass. Multiple data streams can now be aggregated using AI and ML techniques, minimizing touch points and the amount of data needed to be humanly evaluated.
- **AUTOMATED WORKLOAD PROTECTION:** Automating workload protection on the edge provides the eyes and support needed to identify issues before they become problems and automatically troubleshoot and remediate problems before they become crises. Automated workloads can reallocate workloads geographically so they continue to function without interruption.
- **LOWER DOWNTIME AND COSTS:** Automating edge infrastructure reduces network downtime and its associated costs. A Ponemon Institute study estimates the cost of data center downtime is $9,000 per minute, or $540,000 per hour.[5] But just as valuable is the fact that edge systems are traditionally the biggest revenue producers for most companies. One survey projects that by the year 2025, automated edge systems could produce as much as $5 billion per year in new revenue.[6]
- **DIGITAL TRANSFORMATION SUPPORT:** Automating edge network management can demonstrate the success of the organization's digital transformation initiatives and encourage buy-in at all levels of the enterprise. Successfully rolling out an automation edge management initiative is a powerful proof of concept for validating a digital transformation project and how to handle edge use case deployment at scale.
- **IMPROVED SECURITY:** The FBI has reported that since the start of the coronavirus pandemic, cyberattacks have increased 300 percent,[7] with 95 percent of all enterprise security breaches due to human error.[8] Especially now, in the new normal of work from home, remote workers continue to be the principal target of cybercriminals. Automating security from the network to the edge enables the enterprise to automatically enforce identity, device, and access security policies across all enterprise end points.

Given these benefits, here are some best practices leading enterprise adopters suggest for automating edge infrastructures faster and more efficiently:

- **ARCHITECT FOR THE FUTURE.** Enterprises should begin with an overall strategy that looks beyond the initial use case to how automation will be implemented across the entire organization. Future-proof the infrastructure so it can adapt to new technologies and scale easily in response to changes in the business.
- **BUILD IN AI.** Preventive maintenance capabilities in AI infrastructure are critical to creating an automated intelligent edge. Such functionality is key to successfully managing the complexity of the technology environment and especially for building in proactive and eventually self-healing capabilities across the entire IT environment. One expert recommends that organizations make this element a priority in planning for three-year technology infrastructure refreshes. Equally important is the value of AI and data. Collecting and analyzing data with AI is a key capability that will continue to grow in importance in the years ahead.
- **CREATE A BASELINE.** Many strategies are siloed, as successful automation planning requires many different organizations to be involved. It helps to view the enterprise holistically: What are the business units doing? What workflows are third party? Design for an end state that integrates all of these aspects of the organization.
- **FOCUS ON LOW-HANGING FRUIT FIRST.** For many global organizations, the problem is often with their Wi-Fi systems or older switches. Go after that low-hanging fruit first. Demonstrating success with those initial use cases will result in multiple benefits, which is critical to continuing the momentum of digital transformation. It shows leadership that digital transformation is a net gain for the organization and deserves continued funding for the next use case. But just as important, the improved user experience generates momentum from the bottom up.
- **TAKE A CONSULTATIVE APPROACH.** Enlist a partner with domain expertise. Many use cases cross multiple domains. Expertise in this area can greatly reduce the pain and speed the gain of the automation strategy and overall progress of an enterprise digital transformation.

The bottom line

Ultimately, the goal of an organization's automation strategy—the metric on which it will be judged a success—is how it changes the user experience for the better, or for a service provider, how it improves the customer experience. In almost all cases, that means making the edge experience the same as that in the data center and the cloud.

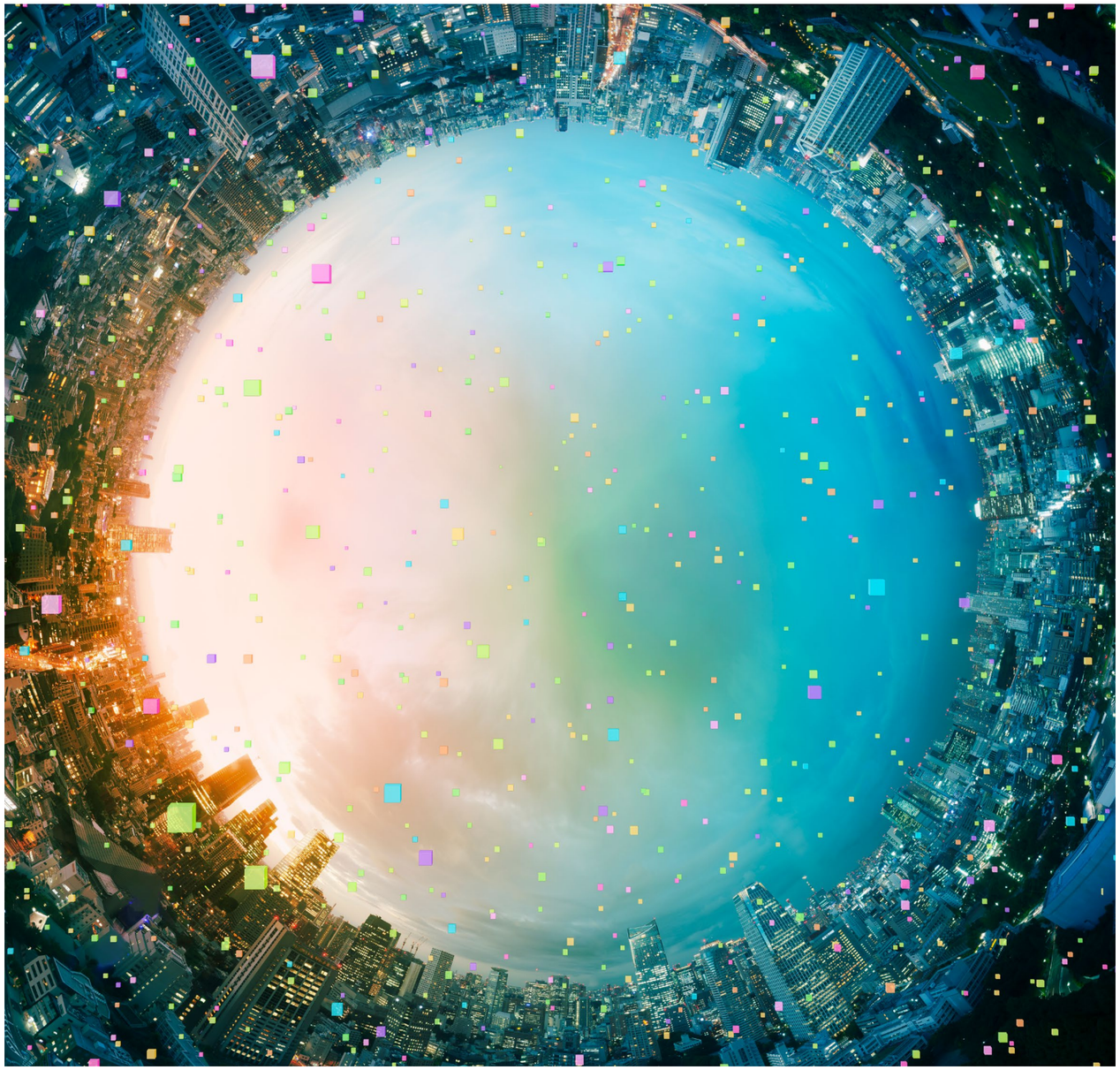
Enterprises need help to manage the complexity of their evolving edge to cloud environments. Automation is the answer. ■

5. "Emerson Network Power Study Says Unplanned Data Center Outages Cost Companies Nearly $9,000 Per Minute," Emerson Network Power and Ponemon Institute, Jan. 19, 2016
6. "The Edge Computing Infrastructure Market: A Technology and Market Forecast 2020-2024," Research and Markets, March 2021
7. "COVID-19 News: FBI Reports 300% Increase in Reported Cybercrimes," IMC Grupo, May 2, 2020
8. "15 Alarming Cyber Security Facts and Stats," Cybint, Dec. 23, 2020

5G and Wi-Fi 6 are up to the task of edge networking

The latest generations of Wi-Fi and cellular have the speed and capacity to handle busy and complex edge networks.

BY JEFF ENTERS



WITH THE NUMBER OF EDGE DEVICES already in the tens of billions and growing rapidly, the need for pervasive and fast bandwidth is obvious. The industry and standards bodies saw this coming, and the current wireless standards are built to support large numbers of devices with different kinds of network requirements.

Wi-Fi and cellular, like 5G, are often portrayed as competitive technologies. There's something to this, but it's all to the customer's benefit, and the two technologies can complement and work together.

The latest and greatest

Wi-Fi 5 and earlier versions were powerful and compelling enough to make it a dominant presence on edge networks. At the same time, Wi-Fi is mature technology and people are familiar with it. But older versions do not have the capability to support mobile communications over a long range, and older Wi-Fi deployments may not have the bandwidth needs of an ever-growing edge.

Wi-Fi 6, the latest generation, includes numerous improvements that make it better suited to large edge deployments than earlier generations of Wi-Fi. These deployments serve conventional computers and much larger numbers of devices like cameras, sensors, robots, and smart buildings. Some of these devices have high bandwidth requirements. Others use little bandwidth but require low latency, meaning they need priority access to the network.

Perhaps the most significant improvement in Wi-Fi 6 is orthogonal frequency-division multiple access (OFDMA), a technology borrowed from cellular networks. Earlier versions of Wi-Fi allow only one device to transmit at a time, but OFDMA allows bandwidth to be divided into multiple, simultaneous transmissions, making networks more responsive, especially with large numbers of users connected.

There are many other improvements in Wi-Fi 6 that improve performance on busy edge networks—for example, by making it possible to use frequencies in use by other devices without interfering.

That 5G is faster than 4G is easy to deduce from the name and is drilled into all of us by telecom advertising. In fact, 5G is capable of speeds many times those of 4G and supports low-latency connections in ways 4G can't.

But 5G wasn't designed primarily for cell phones. The real value of 5G is in Internet of Things (IoT) applications, in which connected devices with limited smarts perform tasks that require communication back to the network—such as on large farms, somewhere on a highway, or spread out through a factory or city. These are applications where Wi-Fi may not make sense, but there are also network applications for which 5G is better suited than Wi-Fi, even if both are available.

In WPA3, even without a password, network connections are strongly encrypted.

The secure edge

Because the edge is where attackers loom, Wi-Fi 6 has been architected to improve the security of Wi-Fi by requiring support for Wi-Fi Protected Access 3 (WPA3). The most important advances in WPA3 are for consumer installations, for which earlier versions of Wi-Fi security were completely broken. In WPA3, even without a password, network connections are strongly encrypted.

Enterprises can use the new WPA3-CNSA (Commercial National Security Algorithm) mode to take advantage of the strongest cryptographic protocols and be recognized by the U.S. Department of Defense and organizations that interact with it as supporting the proven Suite B TLS cipher suites profile for Transport Layer Security, developed by the National Security Agency.

Your own private cellular network

The advances in performance and flexibility in 5G will allow enterprises to build their own 5G networks to maximize efficiency in manufacturing and other complex sectors.

These networks are especially appealing to large manufacturing facilities, which have become extremely high tech and automated. For manufacturers to implement so-called Industry 4.0 technology such as robots, self-driving forklifts, AR/VR glasses, and drones, they will need the capabilities of 5G.

5G networks have both edge and core components. Both are designed to support cloud-native architectures, and both are designed as open standards, much more so than 4G. As a result, many companies have entered the market for 5G core and edge components, where in 4G, there were few vendors.

Network slicing is one feature that provides flexibility unavailable to Wi-Fi networks. Administrators (or software) can divide the physical network into multiple logical networks with assigned portions of bandwidth and specific policies and quality parameters that will be asserted.

In a factory, one slice might serve AR/VR applications that demand higher speeds, while another would serve robots, for which low latency is critical. In addition, network slices can be

allocated to alleviate bottlenecks and improve throughput as workload demands dictate.

New wide-open Wi-Fi spaces

The edge got a lot bigger in 2020, at least in the U.S. and a few other countries, when the FCC issued an order to allow a large range of spectrum, about 1,200 MHz in the 5.925 GHz to 7.125 GHz (6 GHz) band, for unlicensed use.

A large array of tech giants, including Facebook, Apple, Broadcom, Intel, and Hewlett Packard Enterprise, pushed for the new spectrum designation and provided research to back up its viability. The Wi-Fi Alliance has announced a new certification for the interoperability of 6 GHz-capable devices, called Wi-Fi 6E, an extension of Wi-Fi 6.

The 6 GHz band is enormous, low latency, and high speed, including seven 160 MHz channels, capable of high-definition video. Unlike 2.4 GHz and 5 GHz, the 6 GHz band has few existing applications to compete for bandwidth.

Wi-Fi and 5G can work together

In fact, Wi-Fi 6 and 5G have been designed to work together in many ways to optimize performance and user experience at the edge.

Coordination between Wi-Fi and cellular has been happening in a big way for years through Wi-Fi calling, which allows users to make and receive calls through their Wi-Fi connection, secured with IPsec tunnels, rather than through cellular radio. This arrangement often makes calls clearer and more reliable, and offloads traffic from the cellular network, but it's not without its problems.

For example, when a user walks out of a building talking on a Wi-Fi call established in the building and reaches the point where the Wi-Fi drops, a handoff of the call to cellular may fail and the call will drop.

One way 5G addresses this problem is by opening up the standards for the 5G radio access network (RAN), which is the part that the device communicates with directly. Typically, the RAN is a cell tower with antennas and electronics on and nearby it, but a 5G RAN can, in fact, be any local access network, including a Wi-Fi 6 network. It could also use one of a few Industrial IoT network protocols like LoRa, LPWAN, or Sigfox.

This approach is not commercially available yet, but it promises a compelling improvement in user experience and even IoT device performance, by allowing users on Wi-Fi to take advantage of the cellular skill at handing connections off from RAN to RAN.

The edge is delivering

The many applications running at the edge depend on reliable and high-performing connectivity at the edge. Organizations have many options to achieve this level of connectivity at the edge, and increased competition promises continuing improvements.

Wireless network engineering is a complex field. It takes expertise and modern hardware and software to get the best connections at the edge, but it's worth the investment. Modern wireless networks deliver the goods that make new and compelling edge applications possible.

Many team members at Hewlett Packard Enterprise and Aruba, an HPE company, contributed to this article: Jeff Edlund, chief technology officer; Dan Harkins, distinguished technologist; Chuck Lukaszewski, vice president and wireless chief technologist; Eldad Perahia, distinguished technologist; and Onno Harms, product management director. ■

Meeting new security demands

Many organizations are using SASE and SD-WAN to turn their hybrid cloud networks inside out. BY TIM FERRELL

IN THE NAME OF SECURITY, enterprises sometimes engage in a lot of inefficient and convoluted network design. Enterprise resources inside the data center, protected from attack by a strong perimeter at the edge of the network, has been the conventional design for some time. The rise of cloud services and hybrid cloud design, along with the shift to work from home, has made this view obsolete. Enterprises must now manage security out at the edge itself.

Hybrid architecture means that users and software are frequently connecting in and out of the data center. Sometimes they connect from one cloud service to another without needing to connect to the data center. One approach to this problem is a complete architecting of enterprise security.

Modern architectural approaches to edge security propose a pair of strategic shifts: to a new security model called secure access service edge (SASE) and a network management tactic called software-defined wide-area networking (SD-WAN). SASE and SD-WAN provide a higher degree of clarity and control around the management of network traffic and content. But they require significant commitments, including a well-architected WAN, that organizations need to evaluate based on their needs and their readiness.

SASE, a term coined by Gartner, redefines the structure of an organization's IT universe. Instead of the data center being at the center of that universe, fanning out connections to the cloud and business partners, SASE proposes putting a secure network services provider (such as Zscaler, Cloudflare, or Netskope) at the center of the IT network infrastructure. This essentially turns the network inside out, making the data center just another endpoint in the network cloud core.

Users and services making connections through the network must use either IPsec tunnels through the secure network provider or agent software from the provider to authenticate and perform other security functions. In this new model, the interconnect edge network becomes the new cloud. Gartner calls this vision of network-based applications, protected by brokers sensitive to identity and context, zero trust network access (ZTNA).

SD-WAN decouples routing decisions from networking hardware and the physical network. This lets customers set up secure regional zones and direct that traffic where it needs to go based on automated policies for security, cost, performance, and other network conditions in real time.

Edge security challenges

In a modern, hybrid cloud, edge access can come from any type of device on the network, anytime, anywhere by anyone. That means IoT sensors, phones, laptops, digital signage, security cameras, smart watches—you name it. Analyzing all of the characteristics from each connection to determine the risk and then what actions or restrictions to take requires a significant level of intelligence.

The profiling of access risks isn't performed just once at the initial point of access—it needs to be checked continually throughout the life of the connection. The level of risk can change, depending on the user's behavior and any other characteristics that shift during the course of a connection. The initial profiling and the ongoing profiling provide full line-of-sight visibility into what the user is doing and allows for adjustments as the user works.

By moving security evaluations and decisions out into the cloud, the decisions can be made before anything is accessed, and all of the decisions between any endpoints can be considered in the same process.

This changes the way companies perform networking and security. The SASE model puts critical access decisions at the endpoint device and the point-of-presence touchpoint on the network cloud. Once the endpoint is allowed entry to the interconnected mesh, other decisions can be made about routing within the cloud and other particulars of the connection.

As traffic moves through the mesh, an organization can route pieces of it based on the content's priority, source, destination, and other attributes. SASE, combined with SD-WAN, offers a high level of control and availability. Just as important, it gives IT visibility into the traffic that's being moved. From a security point of view, IT can apply sophisticated rules and automation that authorizes and blocks certain pathways. For example, traffic with a credit card number in it can be allowed to follow certain paths in the mesh but can exit only in certain preapproved places. IT can exercise this level of control from Layer 3 in the network protocol stack all the way up to the application level.

SASE also eliminates needless connections through the data center between entities outside of it. The point of these connections is to perform security checks, but SASE makes the process simpler and more efficient by performing the checks outside.

Shifting to an SD-WAN/SASE model

How cumbersome will it be to shift to an SD-WAN/SASE/ZTNA model for edge security? That depends on many factors, including the maturity of a company's WAN infrastructure and the organization's ability to shape its IT culture to a new way of doing networking and security.

If the company has a well-designed WAN, it will take some minor adjustments—perhaps increasing bandwidth on certain links or eliminating links that are no longer needed. It will be a much shorter journey to edge security.

But if a WAN is immature, the journey will be longer. An extreme example of an immature network is one that is big and flat with a single IP address space and all traffic can travel anywhere. If there's weak segmentation and insufficient traffic inspection, it will require a full reengineering of the network architecture to provide solid protection for an ever-increasing number of nodes connecting to outside devices.

The cultural aspect can be even more challenging. Crisscrossing networking and security functions inside an organization forces teams to work together more closely, determine roles and responsibilities, and form a series of best practices going forward. This can be difficult for organizations with entrenched and siloed cultures.

As companies build out their edge development strategies, they need to bring security into the mix.

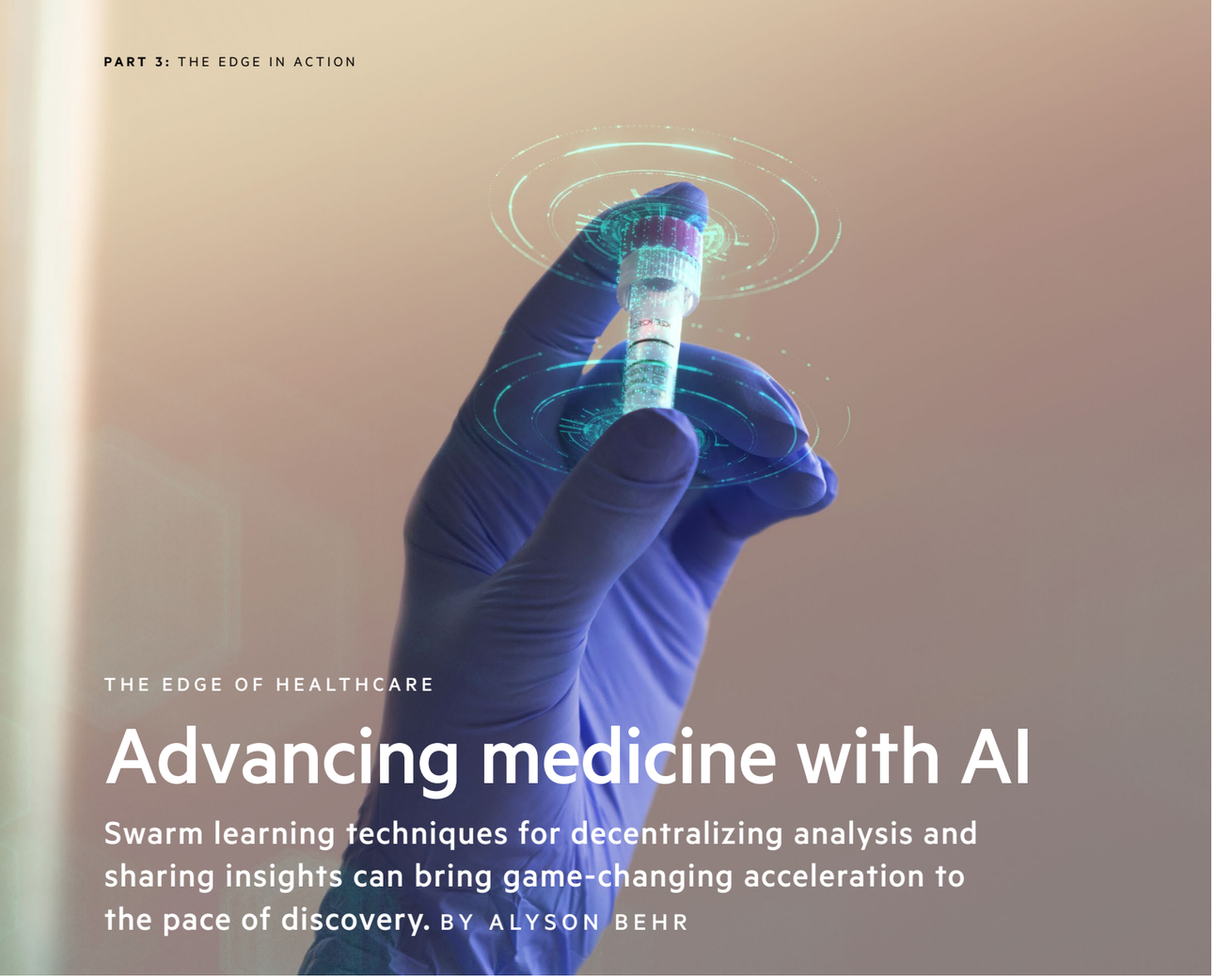
As companies build out their edge development strategies, they need to bring security into the mix. A planning process should determine how much of a lift it would be to shift to a new networking/security model, how the organization can integrate new security tools, and how all appropriate stakeholders can focus on the common goal of edge security.

Starting the journey

It can be a confusing process, but there are steps organizations can take. As the edge increasingly becomes a critical part of an organization's network, it will become inherently more complex and less secure unless changes are made. Organizations need to get ready for a new world of connectivity, information, and scalability. Now is the time to start the journey. ■

PART 3

The edge in action



THE EDGE OF HEALTHCARE

Advancing medicine with AI

Swarm learning techniques for decentralizing analysis and sharing insights can bring game-changing acceleration to the pace of discovery. BY ALYSON BEHR

THE DRUG RESEARCHERS WHO DEVELOPED COVID-19 vaccines in record time were given a crucial advantage: Governments, desperate to quell the pandemic, dropped or modified their normal regulatory and data protection requirements.

If only such speed were the norm.

Concern over data sharing has long slowed drug developments, with regulators understandably trying to protect patient privacy and ensure data security. The challenge is how to develop medicines and treatments at speeds seen during the pandemic without jeopardizing privacy and security concerns.

One potential savior: swarm learning. Swarm learning is a technique for using AI at the edge—say, hospitals or even a doctor's office—to decentralize the analysis of data from multiple locations and then share insights through a learning model that satisfies and bypasses regulatory and privacy concerns. Swarm learning, which is inspired by the swarming phenomena seen in nature, stands to greatly improve how data is collected and analyzed, evolving from a federated, centralized learning model to a decentralized swarm model using trained artificial intelligence.

The current process needs help. As it stands, individual healthcare institutions collect data from research in isolation. Every project is its own discrete source of data, especially when multiple organizations are running the research. With little sharing or coordinating between organizations, the studies and datasets tend to be small and the research often duplicative.

There's a corporate parallel: the so-called silo effect, where teams or even business units fail to benefit from another team's knowledge because it's kept in a silo. Similarly, "my data" ownership attitudes often emerge and result in tedious handling and duplication of large datasets. Moreover, the datasets are rarely standardized because the data is structured for the specific research or research entity.

The downside of the status quo is twofold: It creates regulatory and privacy headaches when different entities want to share their data, and the insights gained are less accurate and less comprehensive because they are not derived from the largest possible base of data.

The key here is that doctors can't share data because of privacy and regulatory concerns. However, they do want to be able to share the insights derived from that data, quickly and effectively.

The beauty of a swarm

Swarm learning makes sharing "insights only" possible, offering agility and elegant simplicity to data mining while respecting medical requirements such as patient confidentiality as well as regulatory and legal privacy concerns. The result is more accurate diagnoses and better treatment decisions, based on larger datasets of healthcare insights, without compromising patient privacy. This is possible because in a decentralized learning swarm, the original organization—be it a research team, pharmaceutical company, or hospital—retains control of the data; it does not move, nor is it duplicated. This ensures that privacy and sovereignty are maintained.

Technologies at play

Patients and providers are driving the digitization of healthcare, and people are choosing to digitize themselves through a variety of sources. Those sources include smart watches and exercise monitors, bedside medical devices, or any telemetry system that collects information from patients and inserts it into their medical records. All these devices are considered edge devices.

"When we say edge to cloud, we're looking at the full spectrum. The edge is on the patient's wrist; the edge is the technology applied in the hospital. Telehealth care really brings those two worlds together," says Rich Bird, a healthcare and life sciences expert at Hewlett Packard Enterprise. "The improvements we've seen over the last 12 months in telehealth care—not only in people being paid to do it, but also the technology available to provide it—is something that's unparalleled."

> "I like to think about swarm learning within the context that if I go to the doctor, she has downloaded all of the lessons that the other doctors in the world learned yesterday."
>
> **RICH BIRD**
> Healthcare & Life Sciences Expert, HPE

The pandemic will leave an important legacy when organizations start to think about inpatient care, clinical settings, and research settings. Documentation management and workflow management tools have become the major applications healthcare providers use to record and track information. Healthcare professionals also add their interpretation, expertise, knowledge, and empathy into the documents, making it possible for insights and lessons learned to be included in the shared data.

"I like to think about swarm learning within the context that if I go to the doctor, she has downloaded all of the lessons that the other doctors in the world learned yesterday," says Bird. "So I'm not just talking to the one doctor's experience that I have in front of me. I'm talking to the collective experience that's been generated, collected, and shared via the swarm."

AI is the perfect technology for swarm learning. Properly trained AI and machine learning can find insights on larger datasets that might escape traditional data modeling techniques with only limited access to training data. It works well as the basis for precision medicine and is in line with medical requirements such as protecting personally identifiable information and other sensitive data.

"The beauty of AI is that instead of shifting the data around, you shift the algorithm to where the data resides," says Hartmut Schultze, a business architect in the data lifecycle practice at HPE.

That's how it makes machine learning "democratic." It's one of the fundamental swarm principles: It searches out equal and like-minded partners on the network, protecting the data and providing security at the source where the data resides on the edge. Machine learning accesses the larger dataset, is less susceptible to bias, and provides higher resiliency.

Becoming an early adopter

Beginning the process of moving to swarm learning involves an organization-wide commitment. The heavy lifting doesn't rest just on the IT department's shoulders. Since the first steps can be done in the cloud, Schultze recommends beginning with a centralized model and then decentralizing it.

"Like any organization, a hospital or research institute is a central organization at its core," he says. "Follow cloud-native structures. Containerization is the key to having the right abstraction that everybody agrees on. For example, every orchestration is standardized on Kubernetes. It doesn't matter if it runs on an edge device or on an IoT device, on a cloud or in a hospital, on a large machine or on a scaled-out machine."

Be prepared to experience many challenges along the way, such as interoperability issues between the major global applications, like Epic, Cerner, and Meditech. The problem is each

has data stored in proprietary systems. Says Bird, "When we're designing systems like this for use in swarm learning, we also need to design a set of agreed-upon principles on the formats in which the data is shared."

The next challenge is getting good curation that renders good data. The question is how much preprocessing must be done? And can it be done locally, producing only quality data for analysis, and then be shared to a swarm?

And as soon as the information is prepared for decentralization, there is now a harmonized and democratized way of sharing insights from that data. The benefits of the data sharing start to become obvious. That, in turn, "motivates the data owner or the data possessor where the data is generated, while incentivizing that one who's creating the model for the algorithm and for the pattern recognition as well," says Schultze. "So everybody contributing is benefiting."

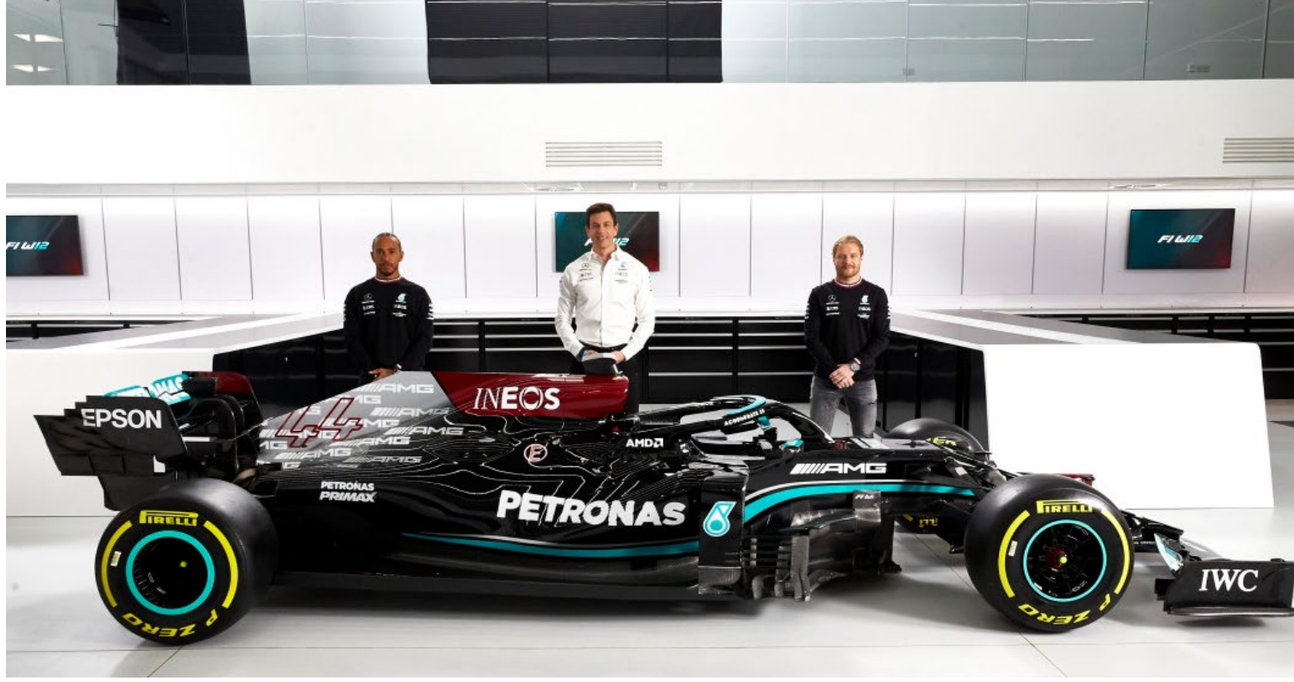
Medical and research professionals agree that people want to share if it's possible to do so using anonymized data. But for this to happen, standardization is critical. Once in place, this will lower hurdles, making it easier for everybody to participate.

Fast Health Interoperability Resources is a standards body trying to assist in standardization via its Health Level 7 standards accreditation for healthcare interoperability. One major benefit of global standardization is that it can ensure privacy for companies by shielding their source code from competitors.

Starting from scratch

How much of this capability is currently in place with today's technologies? "Not much," according to Schultze. "Medical organizations and governments discuss their data-sharing agreement on a case-by-case basis and if and how it's done, then the paperwork gets drawn up. That takes about three months, and then they exchange data peer to peer. Then they have another consortium and do it all over again."

Not surprisingly, then, there's no clear roadmap. However, almost all of the technology necessary for the back end of a complete swarm learning solution is already available to the potential customer. The technology side of the industry is ready for regulatory bodies and healthcare organizations to make this move. And this is the move that will make research at the speed at which COVID-19 was addressed the norm, rather than the exception. ■



THE EDGE OF AUTOMOTIVE

Mercedes-AMG Petronas Formula One Team proves the value of the edge

Succeeding in the high-tech world of F1 means making use of millions of data points. BY DAVID RAND

LEWIS HAMILTON WASN'T EXPECTED TO WIN when he stepped into a Mercedes-AMG Petronas Formula One Team car for the 2021 season opener of the F1 racing season in Bahrain in March.

By his own admission, the champion driver did not have the fastest vehicle. The Red Bull Racing Honda cars piloted by rivals Max Verstappen and Sergio Perez had logged better times in preseason testing and were considered early favorites. But Formula One teams like Mercedes F1 and Red Bull have an edge—literally.

For example, in the days and minutes leading up to the event, Mercedes F1 employed a mix of high-performance edge computing, analytics, and sensor technologies to crunch millions of data points about every car and driver in the competition. It examined thousands of race scenarios, looking at everything from ideal pitstop times to tire selection to arrive at the optimal playbook for its two cars and drivers in the race. And as the contest got underway, both remote and on-site staffers continued collecting, examining, and adapting to the data to improve performance.

In the end, all of the edge-driven preparation paid off with a masterful victory at the checkered flag for Hamilton and the Mercedes F1 team, which defeated the closest competition, Verstappen and the Red Bull team, by a mere fraction of a second.

"It was one of the most thrilling, nail-biting, chest-bursting experiences that I've ever had at a racetrack," says James Allison, technical director for the Mercedes-AMG Petronas

Formula One Team, in a video recap. "When he finally crossed the line, there was an outbreak of mass hysteria in our garage with the utter, utter delight of it."

Beyond the racetrack

Formula One racing has long been one of the most technologically advanced sports on the planet. But with the recent rise of edge computing empowering more local data processing, analysis, and actionable intelligence, just about any motor vehicle could soon benefit from the possibilities it brings.

Indeed, edge computing in the automotive market is poised to grow 27.5 percent each year through 2026, according to Mordor Intelligence.[1] That momentum is expected to be largely fueled by the emergence of autonomous vehicles and connected cars, which will need to rapidly process huge amounts of data to be safe and reliable while providing a wealth of information and entertainment to motorists.

"The value of placing computing at the edge is really due to the way it reduces latency as the data makes its way from the vehicles and back again," says David Witkowski, a senior member of IEEE and CEO of Oku Solutions, a wireless professional services firm. "That turnaround time has to be very low for autonomous or connected cars to work, which is why you're seeing increased automaker interest."

Witkowski says the industry is reaching a sweet spot where a marriage of edge computing and emerging 5G wireless capabilities could radically transform motor vehicles as we know them. From in-cabin augmented reality safety alerts and directions tied to nearby sensors to Lidar systems in self-driving cars that detect oncoming objects (or people), he sees this technology making unimaginable differences.

But it's not just the impact of edge technologies on the end product. Improvements in edge devices, especially in areas of the Industrial IoT, mean that manufacturing vehicles, already a highly automated process, will gain greater intelligence and flexibility as AI and autonomous edge technologies become a more integral part of the manufacturing process.

Industry making a pitstop?

But Witkowski says such dreams could be a ways off because, despite investments from automakers that are racing cars, the infrastructure doesn't yet exist to support it. And until that happens, "automakers won't have much of an incentive to do anything" more than they are today.

Matt Arcaro, an IoT and automotive analyst with IDC, agrees.

"Edge computing is both an interesting and challenging proposition for the automotive industry," he says. "There is an incredible amount of data coming from an increasing number of sensors and connected systems. But vehicles can travel anywhere, including into areas with limited, slow, or unavailable network coverage. As a result, automotive manufacturers and suppliers will need to work with technology companies and ecosystem vendors to ensure that their solutions support the rigorous requirements for automotive."

While that will happen, Arcaro says, automakers still face a range of ongoing and increasing challenges. For example, he notes that the current semiconductor shortage, which has hurt the industry, has spurred automakers to consider more software-defined architectures to drive compute and storage consolidation. That, he says, "will not only reduce the number of distinct processing nodes in the vehicle but will align with the adage of vehicles being 'data centers on wheels.'"

> "There is an incredible amount of data coming from an increasing number of sensors and connected systems."
>
> **MATT ARCARO**
> Automotive & IT analyst, IDC

The digital dance with humans

Dr. Eng Lim Goh, CTO for high-performance computing and artificial intelligence at Hewlett Packard Enterprise, says automakers could and should watch what's already happening on racetracks for an example of that. Formula One cars, in particular, could be considered the ultimate edge devices on wheels. They are chock-full of sensors and probes that interact with high-tech edge computers, software, and storage right in the midst of all the action.

But he notes there's also a digital dance happening between all this technology and human beings. This includes the driver, whose athleticism, experience, skills, intuition, and rapid decision-making ability all play significant roles in wins or losses. Split-second decisions by Hamilton and the team in holding off a final challenge from Verstappen, for example, were widely hailed as key to reaching the winner's podium in Bahrain. The team also includes IT staffers who are constantly evaluating digital data and making real-time adjustments in the race, like pitting a car early to achieve a performance advantage from fresh tires that they hope will put them ahead when a rival driver goes into the pit (a strategy called undercutting).

Michael Taylor, IT director for the Mercedes-AMG Petronas Formula One Team, says that's exactly what happened in Bahrain.

"As Lewis says, we didn't have the fastest car on the day, yet we made use of all the capabilities available to us to maximize our performance," he says. "This enabled us to take the flag when, on another day, we probably wouldn't have been able to do that, to be honest. There was a human element involved, but it was that combination with edge technology that enabled us to win that race."

Adjusting to regulations

Edge technology has become pervasive in Formula One and racing at large. "It's been a bit of an arms race over the last decade," Taylor says. "With the major advancements in processing capability, chips, and memory, all teams have made significant investments in technology to improve their chances on the track.

"But it's also influencing day-to-day operations in factories making these cars, through material science [design] advancements and production," he adds. "If you look at the shapes and surfaces of Formula One cars going back 15 years, you see they're vastly different. Digital technology has underpinned all of that, and I would expect that to accelerate with edge computing."

Taylor notes that, because of the opportunity for one team to outspend another on such technology, the governing body for motorsports, the Fédération Internationale de l'Automobile, implemented financial regulations to level the playing field. As such, the Mercedes F1 team began adopting what it saw as the most advanced technology, within allowed limitations. And it chose partners and consultants that "understand the intricacies of our business" to help optimize the team's use of those systems.

Going down this path provided yet another edge over rivals for the Mercedes-AMG Petronas Formula One Team.

"Although all of our competitors buy the same technology, having access to experts to help us with it brought out the best in our team," Taylor says. "Time is the biggest challenge in Formula One. The more of it you give to your experts, engineers, and technicians, the better the decisions they'll be able to make." ■

HPE is a sponsor of the Mercedes-AMG Petronas Formula One Team.

1. "Edge Computing in Automotive Market: Growth, Trends, COVID-19 Impact, and Forecasts (2021-2026)," Mordor Intelligence, April 15, 2021

THE EDGE OF RETAIL

Lessons for retailers: Finding opportunities in challenging times

The pandemic forced retailers of all sizes to rethink their businesses. One solution: the 'hyper-aware' store.

BY DAVID RAND

AS RETAILERS CONTEMPLATE what business will look like as they fully reopen their stores, it's a different perspective for retailers that have seen the changes forced on their operations as a chance to transform or adapt their business workflows.

While some retailers hunkered down and limited spending, others decided to go in the opposite direction. Many companies knew their customers depend on their products to make a living and took the opportunity to expand their services, while more general retail providers looked to keep up their revenue streams. Delivering a broader range of edge-focused services, from implementing faster shipping and delivery to offering new ways of contactless product pickup, the industry has seen an increased focus on the edge of the retail experience.

Selling on the edge

All these digital efforts have helped improve customer experience and shorten the time it takes people to order online for delivery or to grab their items from a local store. Many retailers are using edge computing technology and IoT devices to become even more relevant today.

Deloitte says digital acceleration is a top priority for 88 percent of retailers surveyed in 2021.[1] In-store tech funding already reached $2.2 billion in the first quarter of 2021—four times what it was at the same time last year, a global CB Insights report found.[2] Funding for e-commerce technology, in particular, jumped 73 percent from the last quarter of 2020 to $11.7 billion, the research firm says.

> "Meeting the needs of today's shoppers can be challenging. They have higher expectations than ever before for touch-free, interactive, and personalized services."
>
> **GERRI HINKEL**
> Director of Solutions & Vertical Markets, Aruba

According to Gerri Hinkel, director of solutions and vertical markets at Aruba, a Hewlett Packard Enterprise company, much of this investment has been driven by pandemic necessity. But she notes retailers are also taking the long view and deploying edge-driven technologies they believe will propel them through whatever things look like in post-pandemic years.

"Meeting the needs of today's shoppers can be challenging," she says. "They have higher expectations than ever before for touch-free, interactive, and personalized services. I think you'll see a continuing retail appetite, therefore, for smart technologies that facilitate shopping online or in physical stores, making them as smooth, efficient, and enjoyable as possible—for shoppers and workers alike."

Retail becomes 'hyper-aware'

Hinkel predicts that, before long, all retail channels will become "hyper-aware." By that she is referring to a model where applications are cognizant of the contextual status of an environment, occupants, energy consumption, service needs, security, and safety, with IoT becoming the eyes and ears of retail organizations. It can generate logical representations of a wide variety of physical data, such as:

- Which products are on store shelves, and are they in the right place?
- How many people are in the store, and why?
- What are the environmental settings, and are they comfortable for the vast majority?

For example, imagine a loyalty customer, a woman, sitting at home searching a retailer's website for the perfect electric toothbrush. Rather than buying the item online, she decides to go into a local outlet to see and compare it with similar offerings. As she walks up to the front door, a camera tied to location-based services and facial recognition technology notes her arrival.

The store's database system is immediately alerted and consulted. Artificial intelligence technology reviews her recent browsing history and determines she is likely there to shop for electric toothbrushes. The AI automatically checks to see if the item of interest is in stock and, after discovering it is not, assembles a list of other electric toothbrushes. An associate carrying a wireless device is then alerted to greet the woman near the entrance, confirm what she's looking for, take her to the shelves with the items, and offer any one of them at a discount given the inconvenience. If the woman wants the item, the associate could then check her out on the spot from a portable point-of-sale (POS) device.

Such hyper-aware scenarios might sound relatively easy to pull off, but they have proved elusive until now because there are so many data-driven technologies involved and those technologies do not talk to one another very well. What's more, trying to move so much IoT data from core networks out to remote locations and back again—without much latency or lag time—is a huge challenge.

"Ensuring a consistent shopping journey across different interfaces—such as online, in store, and so on—is a key challenge for the application of digital technology in stores," says Filippo Battaini, a retail industry analyst at IDC in Europe. "Retailers need to effectively harness customer data gathered from different sources by breaking data silos and employing the right analytics tools to achieve a single view of the shopper."

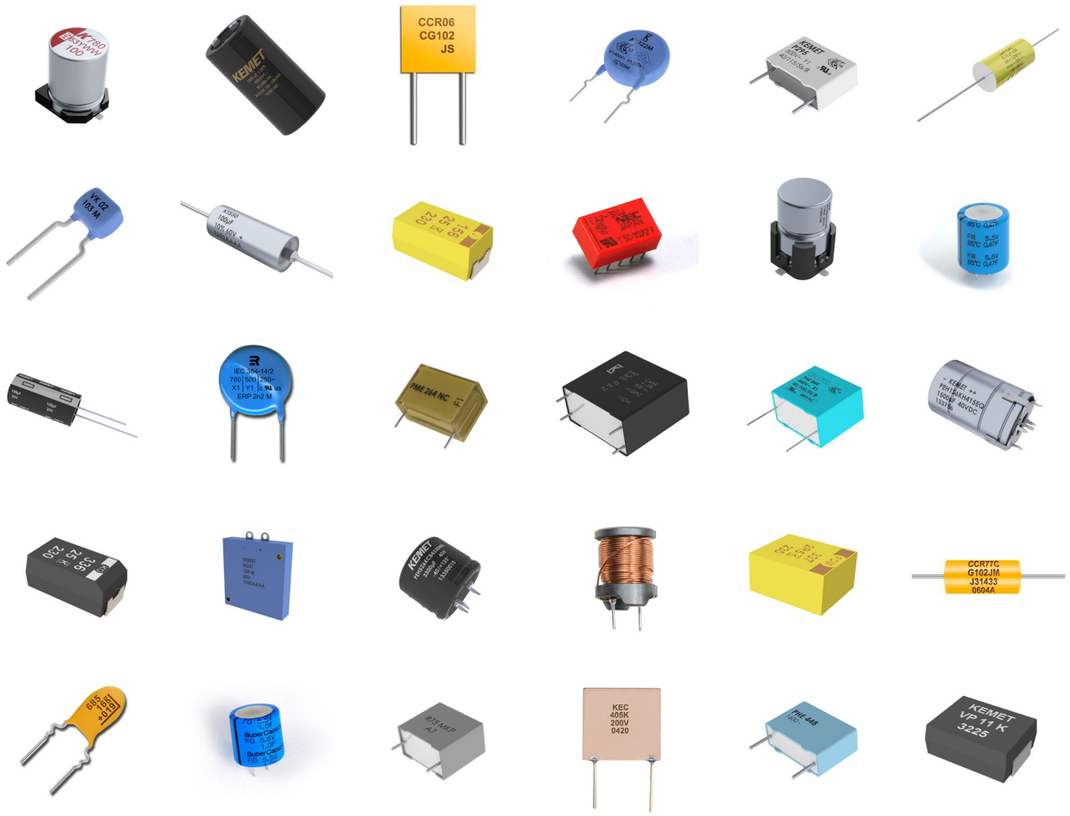
Edge overcomes IoT challenges

This is part of the reason retailers are increasingly considering underlying edge computing management and security capabilities. "Edge computing provides the unified infrastructure retailers need to connect POS, scanners, kiosks, and shopper devices together while bringing management, security, and troubleshooting onto a single platform," says Hinkel. "It enables retailers to implement zero trust scenarios to ensure people accessing networks are who they claim to be. And it integrates AI operations to automatically optimize network performance and reveal issues before they can impact customer satisfaction or store revenues."

Many retailers have had those types of results from their pandemic experiences using edge computing, and from the public's reaction, such new or modified retail experiences have been largely seen as positive changes to the customer experience. ■

1. "2021 Retail Industry Outlook: The New Rules of Retail," Deloitte, 2021

2. "State of Retail Tech Q1 '21 Report: Investment & Sector Trends to Watch," CB Insights, May 6, 2021



THE EDGE OF MANUFACTURING

KEMET Electronics confronts aging infrastructure by embracing digital transformation

Transformation demands a modern infrastructure and the latest technology. That was an issue for this electronic components supplier. BY JAMES P. WOODS JR.

WHEN KEMET ELECTRONICS, a 16,000-employee global manufacturer of high-end electronic components, sought to modernize its operations, it faced a common challenge: Even as an engineering-driven company with an excellent grasp of technology, its infrastructure was decades old. To move forward, that had to change, and the company's IT team knew it couldn't do it alone.

"We were relying on 10- to 20-year-old infrastructure in many of our plants and facilities," says Chris Hall, senior vice president of IT and CIO at KEMET, which has 20-plus manufacturing facilities across the Americas, Asia, and Europe. "It was just not possible to make the digital transformation changes we needed using legacy technology. Modern cybersecurity, IoT, AI/ML, and modern business intelligence platforms rely on lower latency and higher bandwidth than traditional systems like ERP. Attempting our transformation without improving our base infrastructure would be like renovating your house without fixing the rusted-out pipes."

To meet the demands of digital transformation, KEMET realized that automation was also key within its environment. "You can have the best application platform in the world, but if you can't keep up with the business from a change perspective on the infrastructure and software side, does it really matter?" Hall says.

Everyone has a role to play

"It's a balancing act," he says. "You need to ensure you have a plan for upgrading the digital operations and the network infrastructure and then understand your technical challenges as you deploy automation. You strike a balance when you also get support from the business and your IT team."

KEMET's infrastructure solution included deploying Aruba indoor access points, switches, cloud-based network management tools, and network access control to bring the company's networks into the 21st century.

It's critical to show fearful workers that automation will not take away their jobs but instead create efficiencies that can improve their work, Hall says. "You need to communicate the benefits of the new infrastructure and automation to get the buy-in from the top down and the bottom up."

Hall promotes measuring performance gains along the way. His team has shown where processes have improved with the new infrastructure and automation versus the old environment and way of doing things. At KEMET, the comparisons have allowed everyone from the C-suite to the sales reps to the line workers on the factory floor to see the results and realize the benefits. "Even our CEO and COO...can't imagine going back to the way it was," Hall says.

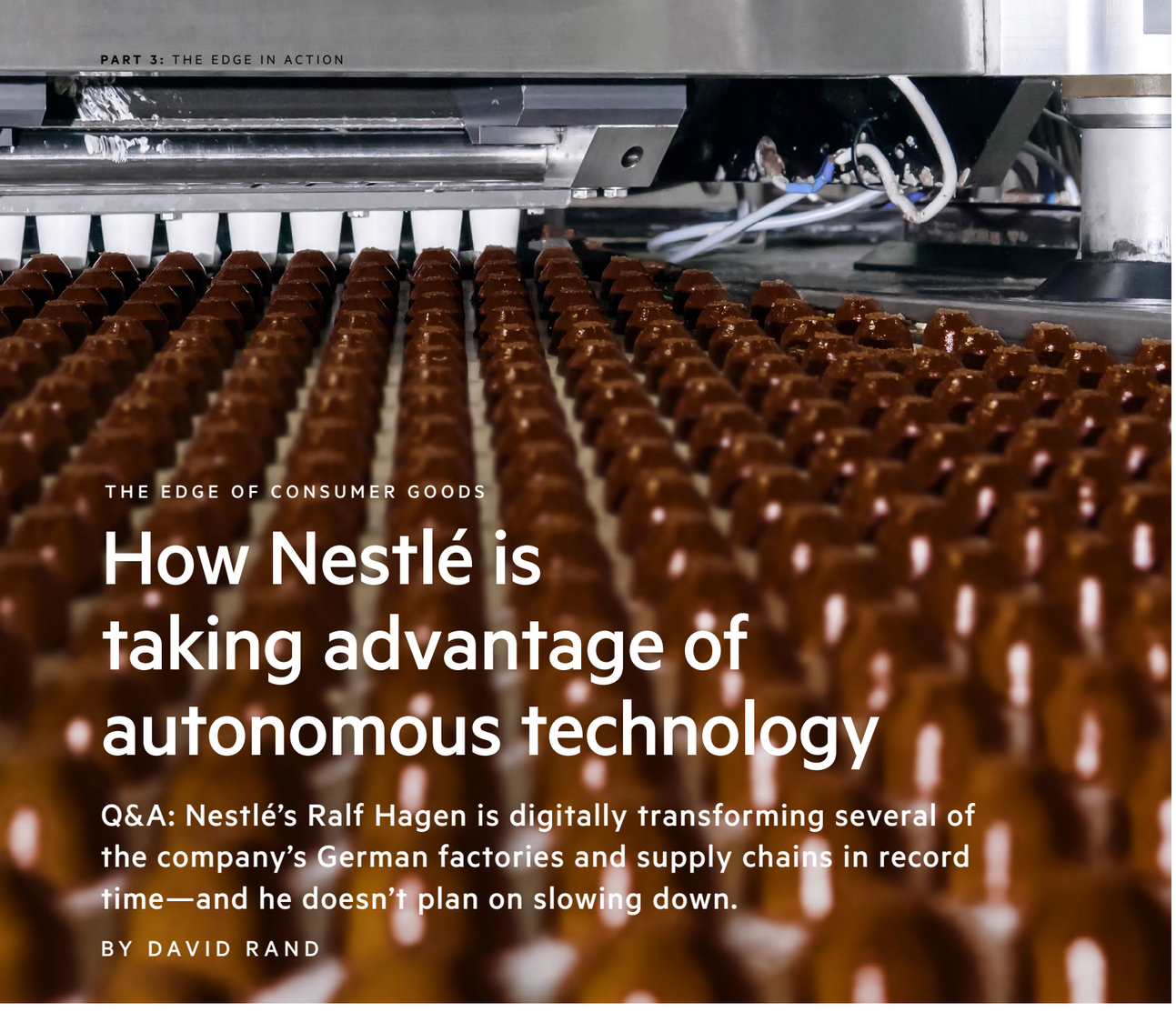
> "You can have the best application platform in the world, but if you can't keep up with the business from a change perspective on the infrastructure and software side, does it really matter?"
>
> **CHRIS HALL**
> Senior Vice President of IT & CIO, KEMET Electronics

Collaboration and cost savings

"A lot of the things that we have enabled are due to cloud-based computing and a network infrastructure that supports an edge-to-cloud environment like ours," Hall says. "The biggest impact is how we operate as a global company."

Significant benefits include "the collaboration that we have begun to enjoy during the pandemic and our ability to quickly and securely deploy new IoT solutions," he adds. "One of the biggest items to impact our business to date is what we're doing with IoT. It is going to save us eight figures annually in the coming years. The projects we've implemented have already paid for the underlying technology in the first year of existence, and the rest is just savings potential for the foreseeable future."

Based on the success of these ongoing automation efforts, KEMET's parent company, Taiwan-based Yageo, has tasked Hall's global IT team with leading the digital transformation for other parts of the organization. Working in partnership with Aruba, a Hewlett Packard Enterprise company, Hall is confident that the same benefits that have taken KEMET's networking platform to the next level will soon be shared by more of the larger Yageo Group. ■



THE EDGE OF CONSUMER GOODS

How Nestlé is taking advantage of autonomous technology

Q&A: Nestlé's Ralf Hagen is digitally transforming several of the company's German factories and supply chains in record time—and he doesn't plan on slowing down.

BY DAVID RAND

NESTLÉ SA, the Swiss multinational food and drink giant, may be more than 150 years old, but that doesn't mean the household brand is stuck in the past when it comes to manufacturing processes.

In fact, unlike many major manufacturers still clinging to traditional automation, Nestlé is quietly doubling down on the digital transformation of its manufacturing and supply chains. Its goal is to create competitive gaps through data, artificial intelligence (AI), and predictive analytics. In many factories, the company is beginning to deploy Internet of Things (IoT) connected devices with remote-sensing technologies, along with autonomous vehicles and collaborative robots.

The company believes digitalization will help it generate efficiencies, create agility, and provide new platforms for growth. Achieving all of that, however, will take more than the systems themselves. Having true factories of the future will require underlying intelligence and localized or edge computing so everything can run autonomously as more connected endpoints come online.

We recently met with Ralf Hagen, engineering manager for Nestlé Deutschland, whose team is five years into transitioning several German factories from automated to autonomous systems. Here's what he had to say about how the program is progressing and where he sees it going.

Q: Why is Nestlé on this autonomous technology journey?

HAGEN: Our vision is to use autonomous technology to handle the challenges we are facing in a more localized and reliable way than we have in the past.

Automation has been central to our operations for decades, as it has for many manufacturers. We've had classic client-server structures driving the business. But as digital technology becomes more prevalent, there are more limitations in that approach because of the dependency on server farms located away from factory floors. This central computational approach is no longer benefiting us.

For example, one of our server farms recently experienced a failure of a hard disk array and an un-activated monitoring [system] that led to an avalanche of failures for all mounted devices. We had to rely on our computing and software vendors to help get everything back on track. And we experienced 30 hours of downtime in the process. Fortunately, this happened on a Friday afternoon, and by Sunday, we were back on track. But it was a problem that this occurred at all. Had it hit earlier in the week, it could have taken down an entire factory. This was an awakening for us—recognizing these system dependencies and realizing that autonomous technologies could help tackle the problem.

Q: How does autonomous technology solve these continuity challenges?

It frees us from relying entirely on automation and centralized systems. It helps ensure contextual data about systems and operations is always available, even with the failure of a disk array. It helps spread contextual data around instead of placing it [in a server farm] and then sending it forward or upwards in a linear direction. Autonomous systems give us the opportunity, even in a hacking event, to continue operations. At the same time, we avoid the kind of waste that occurs anytime you need to shut down to correct, say, a PO. By keeping things going, we also avoid the waste and mechanical wear and tear that can occur whenever you hit stop.

To accomplish this level of autonomy, you must move your technology and compute power close to where plant floor operations data is generated. You're still talking about the same power, but it's situated differently. Hardware isn't just centralized in a room somewhere. It's in cabinets closer to the machines. This is really a step forward that we are trying to take. We are putting the brains back down with the lines where they should be.

Q: Have you already shifted computation power from server farms to factory floors?

At the moment, it's only logically there, not physically. But we look forward to the types of advances the independence of autonomous systems will provide.

For example, with OT [operational technology] security, you need a good foundation in IT security policy. The problem is IT security can be performance killing. Performance requires energy. It requires time. And you have to balance that with your need to secure systems. We are considering how autonomous technology might provide that equilibrium in the background without us having to do much along the way.

> "Had [the failure] hit earlier in the week, it could have taken down an entire factory."
>
> **RALF HAGEN**
> Engineering Manager, Nestlé Deutschland

Q: It sounds as if that might be one of the main attractions of autonomous systems for you.

Correct. We are surely not going to have a sufficient number of skilled systems administrators in our company to do this because we are a production company. We are not an IT company. We are not an infrastructure company. We are just a simple producer of food. We want systems to be like the iPhone. Everyone has one in their pocket but doesn't have to think about how it works. It just does. On the industrial side, we can learn from this consumer experience. We want to remove the complexity. It is not beneficial. And this is the kind of autonomous state we would like to achieve.

That's why we are on a journey to restructure [our operations]. We are trying to re-layer our network layers, for example, according to functionalities. This means a lot of data generation and sharing will happen on the factory floor instead of somewhere else.

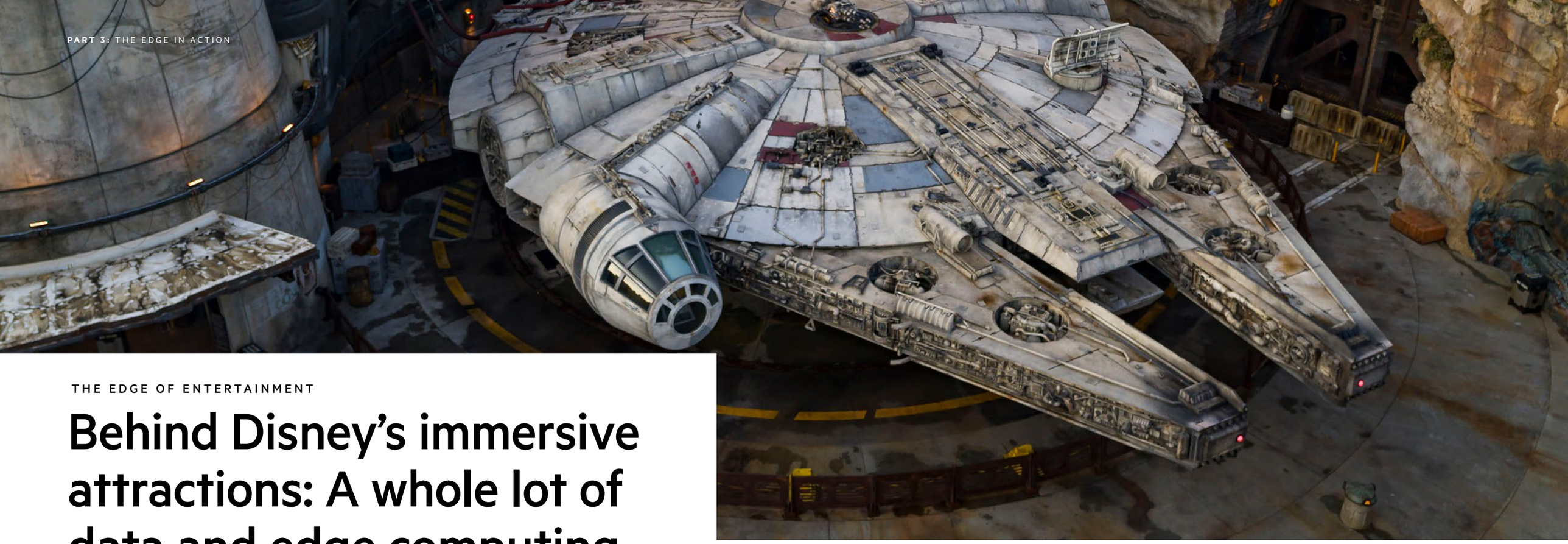
I remember not long ago people were talking about "sensor to cloud." This makes absolutely no sense to me. You want to pull reports directly from sensors in real time without having to send data to the cloud only to retrieve it again five minutes later, when it might not be relevant. With autonomous systems, the sensor [information] stays put, and you only send information to the cloud that really needs to be in the cloud for some reason.

Q: We've heard autonomous systems can help better control production quality to, say, adjust the mix of ingredients in candy bars or the printing on labels. True?

Yes, the idea is that your MES [manufacturing execution systems] for tracking and documenting the transformation of raw ingredients into finished products has access to the information needed to make such improvements. As opposed to basing necessary information in monolithic central servers, you arm your local systems to access all the parameters—locally or from SAP, ERP, and other databases—they need to autonomously complete a PO. This is what we're striving to achieve.

Q: Has COVID-19 affected the drive to implement autonomous systems at Nestlé?

It has played a big role in our drive to digitalize. With the pandemic, we have been told that delivering this [autonomous technology] is not just urgent, it's over-urgent and overdue. A lot of things have been made possible the last six or seven months that would have taken two or more years to accomplish previously. Hopefully, we can maintain that rate of change. ■



THE EDGE OF ENTERTAINMENT

Behind Disney's immersive attractions: A whole lot of data and edge computing

The engineering director who helps make the magic happen shares how Disney uses enterprise technology in its attractions—and how working in aerospace has served him well. BY PAUL SLOAN

WHEN PEOPLE VISIT DISNEY PARKS, one thing they're sure to appreciate but likely know little about is the degree to which enterprise technology plays a role in making the Guest experience so thrilling and immersive. Technologically advanced attractions like *Millennium Falcon*: Smugglers Run and other popular Disney theme park experiences are original designs using edge computing, powerful hardware, advanced robotics and control systems, and much more.

We spoke with Michael Tschanz, director of engineering technology and analysis at Disney Parks, Experiences, and Products, a Hewlett Packard Enterprise customer, to discuss what goes into bringing Disney attractions to life. Tschanz and his team run engineering technology and analysis for all parks around the globe, so if it involves engineering technology, Tschanz likely helps support it.

Q: To start, how have things changed over the past few years for Disney from a technology standpoint?

We combine Disney storytelling with some of the latest and groundbreaking advancements in technology. These have always been, and will continue to be, some of the ways we create the most immersive and magical experiences possible for our Guests. When you think of a place like *Star Wars:* Galaxy's Edge, for example, and everything that goes on there, there's a lot. There's lighting control, lots of different show items, advanced robotics and control systems—and all of these different things have to come together in a way to create this ubiquitous experience in a way that keeps the storyline real. To create these experiences, the technology has had to advance over the years. And it's important that we have servers that can handle processing that volume of data. HPE provides servers that process data at the edge.

Q: Can you tell us about a tool you use in your position to help keep Disney attractions running smoothly?

Our goal at Disney Parks, Experiences, and Products is to provide each Guest with the most immersive experience possible so that our Guests become part of the story. We're always bringing in new methods and new processes to understand exactly what's happening in our attractions. We bring in large volumes of data to inform us of the state of our attractions, but bringing in that data can sometimes be challenging because of how much there is. Edge computing helps us overcome this challenge.

> "[Simulation] is how all the new aerospace, spacecraft, and everything in between is being developed now."
>
> **MICHAEL TSCHANZ**
> Director of Engineering, Disney Parks

Q: How important is edge compute?

Edge compute is one of the main pieces that lets us connect to that data stream I mentioned. When we have everything interconnected, we are able to create these more engaging, immersive experiences. But all these pieces have to work together. That means edge computing is absolutely one of the primary pieces of that pipeline enabling us to manage data.

With edge computing, there's often a discussion about whether you should be doing more in the cloud or doing things more locally. What we've come to understand is it's not just about doing one or the other; it's about a hybrid way of doing things. It's managing which application or solution works best. Does it work best in the cloud? Or does it work best in a local location as far as an edge computing solution? And you need to make that decision based on good, sound business judgment.

Another question is, what's the best way to run that data? Is it better to leverage a whole bunch of cores in a cloud perspective or keep it focused at that edge computing space? Because if you need it, you need it now, and you need it in real time. And you need to think about that. You can't just have one blanket

solution. Working together and keeping those parts together allows everything to work—and edge computing is absolutely a very important part of that overall compute solution.

Q: Can you talk about a technology that you either recently adopted or created that helps your team support Disney attractions?

There's a number of them. Take our simulation engine and our simulation architectures. They are one of the ways that support how we develop long-term growth and how we build things. Plus, they help reduce costs and contribute to a faster time to market for new attractions and experiences.

There's a big push in the world—and not just in what we're doing, but for any company that does things for simulation-based products or model-based designs. This is how all the new aerospace, spacecraft, and everything in between is being developed now. You can't land a booster in the middle of the Atlantic Ocean without doing design and development of this type. I was doing it in aerospace when I was back at Texas Instruments in the early '90s.

You've got to think about how to build something in a space virtually before you actually put it together, because if you don't, then you're going to end up building something that doesn't meet the necessary requirements. Or, in our space for Disney, it may not meet our creative intent. If we're able to build in a virtual space, the digital pipeline shows how it's going to function with really rigorous engineering, physics, and math behind it. Then, when we build it, it's going to look like what we expected from the beginning.

We have been working for some time to build that simulation and model-based design pipeline, and it takes a lot of processing power to do that. So, going back to edge compute, we need a lot of edge processing power. It affects how we think about sharing it with the cloud, too. Sometimes we need those answers right away. Overall, I would say that the simulation engine that we've built to help design our attractions has been one of the most significant leaps forward that our team has done.

Q: We talked about edge technology and other industries. Will edge technology grow in importance for you in the coming years?

Absolutely. And this whole conversation—and thinking about how things either run on the edge or in the cloud—makes me think about discussions that were taking place back in the mid-'90s, like having terminals as opposed to a computer desktop capability or something that ran in a data center. These discussions around technology have been around for decades.

As this played out, what we've learned overall as technologists is that there's never a perfect case. There's always going to be a local requirement to run a very high-compute-intensive application or process other data locally. There's never going to be enough bandwidth, and there's never going to be a use case that's 100 percent cloud-based or 100 percent edge-based.

Again, it comes down to looking at the business decision and what makes the most sense. And I think people are now reconsidering other options as opposed to when there was a prevailing wisdom that we would push everything to the cloud. When we did that, people were looking at things like high-performance image processing, audio processing, or high-data-rate things, and they quickly realized that you couldn't really do those kinds of use cases in the cloud in real time. That's where edge processing almost always comes in, because there will only be so much bandwidth that you can ever work with to be able to manage things outside of the edge. ■

HPE is the official server of Walt Disney World® Resort.

THE EDGE OF BUILDINGS

Hyper-aware smart buildings: The future of automation

Standardized IT protocols and edge integration will drive the evolution of building automation.

BY MICHAEL TENNEFOSS

AN OVERHEAD CRANE MOVES ON A COLLISION COURSE toward an unsuspecting worker. A water pump feeding a building's air conditioning system starts vibrating abnormally. A carbon dioxide leak in a brewery fermentation system spews gas into occupied areas. Each of these scenarios represents a point at which disaster could either strike or be averted. Which way things play out hinges on the level of situational awareness that exists at the edge of each site.

Situational awareness is knowledge of the current environment and projections of future environments across time and space. Successful decision-making relies on situational awareness, and its absence can put life and property at risk. In the three scenarios above, actionable situational awareness requires both direct observation of the physical world (such as CO_2 gas detection by a sensor) and a combination of declared and inferred contextual data from the logical world (such as location, identity, and applications in use). Edge IoT data originates in machines, while contextual data comes from edge data networks that link people and machines.

The conjunction of edge data and context is called hyper-awareness, and its business value stems from connecting,

protecting, and analyzing the interactions of machines and people. Hyper-aware metadata can be shared across business applications, allowing new use cases just by tapping into available data, without replacing any infrastructure.

The building automation and IT industries have devoted huge resources over decades to create connected facilities that touch every machine in the building. Connectivity is essential for sharing data, but is not by itself the endgame. The edge networks that connect us also generate their own contextual information, and if that goes untapped, a connected building can't achieve hyper-awareness or deliver the safety, efficiency, productivity, profitability, and other benefits derived from situational awareness.

Hyper-aware buildings are a fusion of edge IoT systems and network-generated context. IoT devices are the eyes and ears of the facility, and IT systems provide contextual information as well as serve as a backbone for facility-wide communications.

The evolution of automated buildings

Building automation has evolved alongside factory, consumer, and process automation systems. Starting in the early 1980s, the work was focused on better ways to reliably connect machines at a price point that was practical even for low-priced sensors and actuators. At that time, telephones, fire alarms, access control, closed-circuit television, power management, lighting, and comfort systems operated independently from one another over dedicated cabling.

As IP networks emerged, facilities managers saw the value of converging systems into a common backbone and standardizing on ways to share data. Many building automation standards emerged from this process—including BACnet, LonWorks, KNX, DALI, and others—utilizing gateways as on-ramps to the IP backbone. Sensors, actuators, displays, and other building systems were embedded with these technologies, and the promise of truly smart buildings was one step closer.

The problem was these were all competing, noninteroperable standards with different physical layers and protocols. Since they could not interoperate, customers typically selected and stuck with one technology. Intense price competition meant that few vendors made money from selling devices, so instead, they embedded manufacturer-specific features that locked customers into using their brand of products. That included adding manufacturer-specific extensions to open standard protocols like BACnet, so that the full range of features could be accessed using only that vendor's gear.

Securing building automation networks was an afterthought, at best, so systems were often physically and logically isolated to protect them from attack. The combination of proprietary protocols and isolation made the data they carried inaccessible to other systems. The result was islands of isolation.

The rise of hyper-aware buildings

The innovations that enable buildings to become situationally aware didn't arise from smart building control vendors; they came from the IT industry.

Enterprise IT buyers have long mandated edge cybersecurity, open data exchange, application awareness, and specialized location services, which the building automation buyer was, until very recently, unable to obtain. Advanced IT cybersecurity systems must identify every user and device before granting permission to access the edge network. This identity data can be shared with other authorized applications, enabling a rich suite of services based on who—or what—is on the edge network, how network resources are being used, and the real-time security posture of network users.

The innovations that enable buildings to become situationally aware didn't arise from smart building control vendors; they came from the IT industry.

Since IT networks are the backbone of businesses, business applications that optimize inventory and monitor time and motion need accurate edge location data about assets and people. This location data, like the other services described above, is accessible to authorized applications.

While large building automation companies still employ vendor lock-in based on proprietary solutions, newer vendors have bucked the trend by producing open standards-based IoT sensors and actuators. EnOcean, iBeacon, Eddystone, and other protocols are open and accessible, and many manufacturers have latched on to them to create open, interoperable building control devices and systems.

When identity, applications in use, security posture, and location data coming from edge IT networks are mashed up with open IoT sensor and actuator data, the building becomes hyper-aware. The richer the set of available edge data, the more readily new applications can be added by just repurposing existing data flows. The infrastructure essentially becomes a future-proof platform on which new services can be built. The IT technology to accomplish this exists today, and the good news is that even legacy IoT devices can be tapped for data if the payloads can be interpreted.

The most sophisticated IT networks go one step further by interfacing with wireless edge IoT devices directly from radios embedded within or plugged into Wi-Fi access points. That means IoT sensors can be deployed virtually anywhere within a facility on an as-needed, "lick-and-stick" basis using existing IT infrastructure and without installing any new cabling. Instead, the Wi-Fi access points serve as secure edge gateways.

Key benefits of hyper-aware smart buildings include enhancing human productivity, energy reduction, health compliance, and physical safety.

The first step of the journey

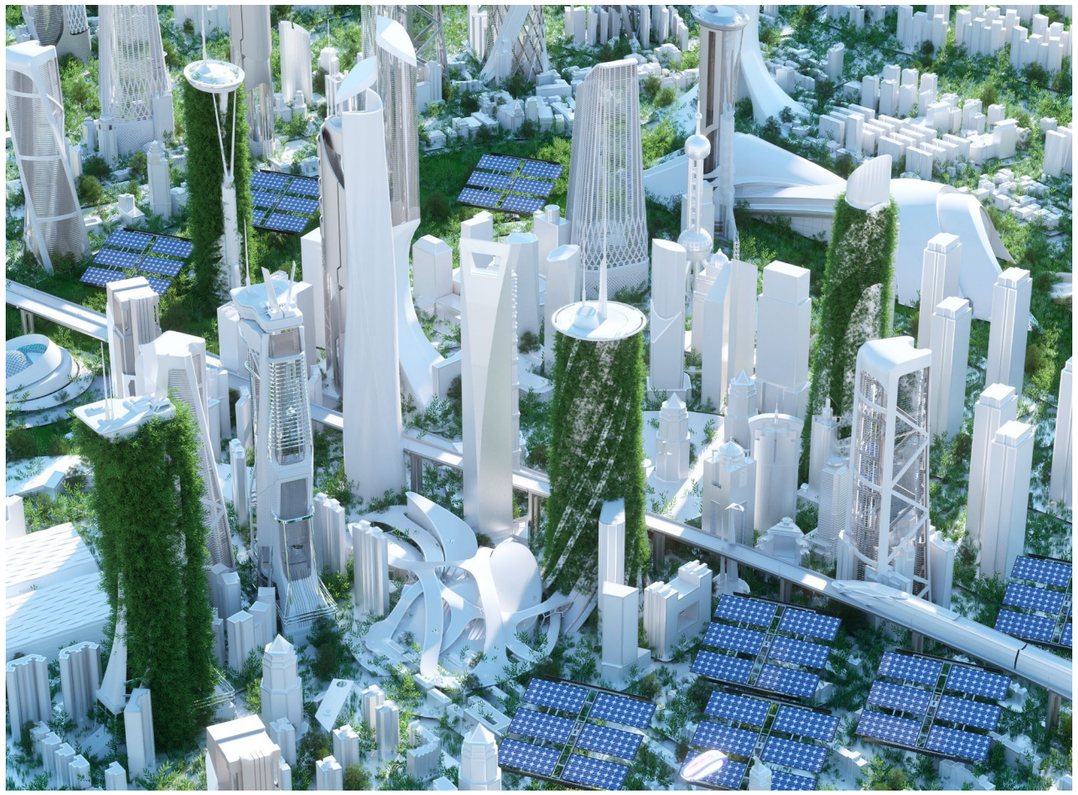
No organization will be able to achieve hyper-awareness for every building project overnight. The journey starts by identifying the strategic goals you are trying to achieve for your organization. Goals will vary by organization and department, so it is important to clarify from the outset the gains expected from a hyper-aware facility. Every facilities planner should focus on this target because the business benefits are just too compelling.

Next, identify what key milestones need to be achieved. One of the reasons many IoT projects fail is because companies become enamored with shiny new technology and lose sight of the business objectives. Do not invest in system and technology until you understand the steps, timing, and stakeholder support needed to drive your project to success.

Be certain not to fall in love with a solution to the exclusion of interoperability and openness. With interoperability comes flexibility in your choice of devices and vendors.

Finally, consider where computing needs to be done and ensure that your solution supports this decision. Are IoT workloads moving to the cloud for processing, or is processing happening at the edge or according to a hybrid model? Locating compute at the edge speeds system response times, whereas cloud services are massively scalable. Think through which architecture works best for you.

It's time to reassess how the facilities professional thinks about smart buildings. Edge connectivity is necessary but not sufficient to achieve strategic corporate objectives that depend on situational awareness. Hyper-awareness is the solution of choice for smart buildings because it provides a future-proof foundation on which an almost limitless set of services can be built without ripping and replacing infrastructure. ■



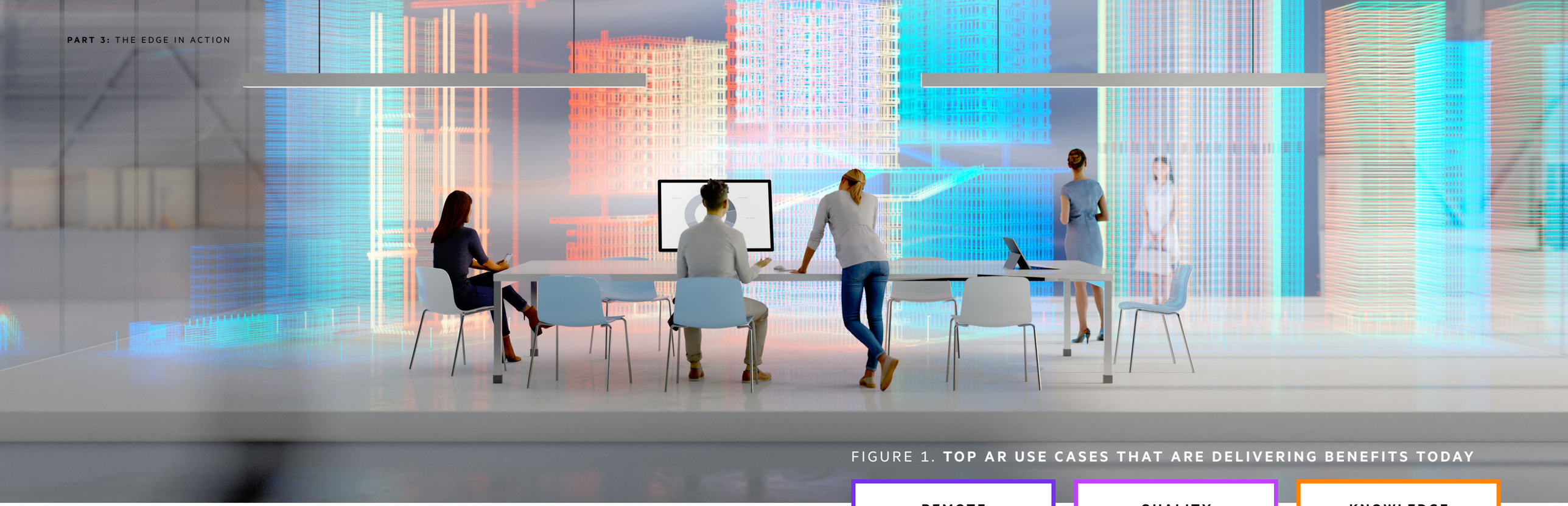


FIGURE 1. **TOP AR USE CASES THAT ARE DELIVERING BENEFITS TODAY**

REMOTE TROUBLESHOOTING

Accelerate time to resolution for technical problems by brokering access to the right expertise, faster

QUALITY ASSURANCE

Ensure work performed in field locations matches specification and quality criteria

KNOWLEDGE TRANSFER

Scale scarce expertise wherever it is needed to accelerate organizational skill development

THE EDGE OF REALITY

A growing augmented reality

Need to quickly tackle a problem or train people on complex systems from afar? Augmented reality can prove invaluable.

BY GARRY ORSOLINI

ENGAGING AUGMENTED REALITY APPLICATIONS illustrate powerful ways of connecting the digital and physical worlds—whether that's placing a virtual sofa in your living room as part of an immersive retail experience or enabling predictive maintenance guided by real-time data and an overlay of step-by-step visual instructions.

Yet, the complexity of AR technology and the limits of some early applications can belie the simplicity of what most businesses want today: a way to quickly get remote colleagues and customers connected to geographically dispersed experts. The goal: to let both parties see what the other sees as part of coordinated and secure workflows. We're talking about pairing smart phones, tablets, and wearables devices with visual collaboration platforms to reinvent training and enable product experts to assist from afar. The aim is to more effectively enable remote inspection, repair, and maintenance of products already deployed.

Think of it as looking over someone's shoulder to collaborate or get training without being together on site. Such a shift has potential for massive savings in travel time and related expenses. It's also a way to minimize downtime and keep business going when travel is simply not possible—a reality the world is now all too familiar with.

Thanks to continuous innovation and more accessible product offerings, industry experts have high hopes for AR, in both the consumer and business space. According to a report by Grand View Research, many industries, including retail, manufacturing, healthcare, and education, are quickly adopting AR technology; the market is projected to reach more than $340 billion by 2028, growing almost 44 percent per year.[1]

Most companies are in the early stages of adopting AR. An IDC Worldwide CIO Agenda Predictions 2021 survey found 45 percent of respondents were in either an early deployment or early pilot stage, while only 18 percent were in the midst of limited or full rollouts.[2] Remote support dominates the AR landscape, with 43 percent of respondents citing such use cases, followed by employee training (39 percent) and knowledge capture (33 percent).

Think big, start small, get ROI

Creating and deploying advanced enterprise AR applications presents challenges. One of the challenges has been to show how hands-free, head-mounted devices like Microsoft HoloLens, Google Glass Enterprise Edition, and RealWear HMT-1 can use AR experiences to solve real business issues. Consumer virtual reality and AR applications have historically set a high bar for immersive user experiences but have been costly to develop. While there are packaged AR platforms and products, as well as the Unity open source gaming engine that many use to build DIY AR applications, creating and updating AI/AR models can be expensive.

1. "Augmented Reality Market Size Worth $340.16 Billion by 2028; CAGR: 43.8%," Grand View Research, Feb. 15, 2021
2. "IDC FutureScape: Worldwide CIO Agenda 2021 Predictions," IDC, October 2020

Many companies in industries such as manufacturing and aerospace are excited by AR's potential yet concerned about serving up intellectual property and other crucial enterprise information to the cloud. This has been a requisite for many AR applications powered by real-time data and video AI models. That is why, for both performance and security reasons, enterprises are frequently turning to edge computing for AR solutions.

As with many technologies in earlier phases of adoption, quick-win projects that demonstrate early ROI are key to larger scale adoption. As shown in Figure 1, AR is providing provable benefits in several areas:

The market is projected to reach more than $340 billion by 2028, growing almost 44 percent per year.

REMOTE TROUBLESHOOTING. The IDC survey found that manufacturers lose more than 14 hours a week due to employees searching for or re-creating information that can't be found. With AR, workers on a plant floor, at a wind farm, or on an oil rig can ask questions of remote experts to complete complicated procedures and help troubleshoot potential problems. Case in point: Using smart glasses and a video collaboration platform, a large energy company paired technical experts with plant floor operators to examine faulty equipment and provide real-time remote assistance and specific feedback. The AR-enabled remote support was crucial to the plant's ability to minimize downtime and keep operations running smoothly.

Other organizations that have put remote support to work have reported significant ROI as well. The ability to see things in tandem can significantly reduce the time it takes to come to a common understanding of an issue.

REMOTE QUALITY ASSURANCE INSPECTION AND GUIDED ASSEMBLY. Equipped with hands-free, heads-up displays or tablets, operators working on an oil rig or another industrial asset can use AR for training directly in the work environment, reducing the need for in-person classroom training. AR capabilities can also help engineers and plant floor operators compare production output to original design specs, verifying that the right parts are being correctly assembled, confirming inspection, and flagging any potential quality issues.

TRAINING AND KNOWLEDGE TRANSFER. Making remote expertise more widely accessible—an area AR facilitates—will remain important as manufacturers regroup for the post-COVID era. Specifically, these pragmatic AR use cases can help businesses maintain continuity of operations and customer support in the face of ongoing remote requirements as the nature of work continues to shift.

AR can also provide a cost-effective way to scale hard-to-find expertise to far-flung employees and transfer valuable institutional and technical knowledge to and from workers in the field and among various product groups. Typically, many handoffs must occur to funnel intelligence about a particular support problem back to the engineering groups—if the insights ever make it back to them at all. With AR, remote engineering teams can experience problems directly and have access to snapshots, recordings, and pictures of work in progress. As such, they can issue systemic fixes more quickly and even make adjustments to subsequent product designs.

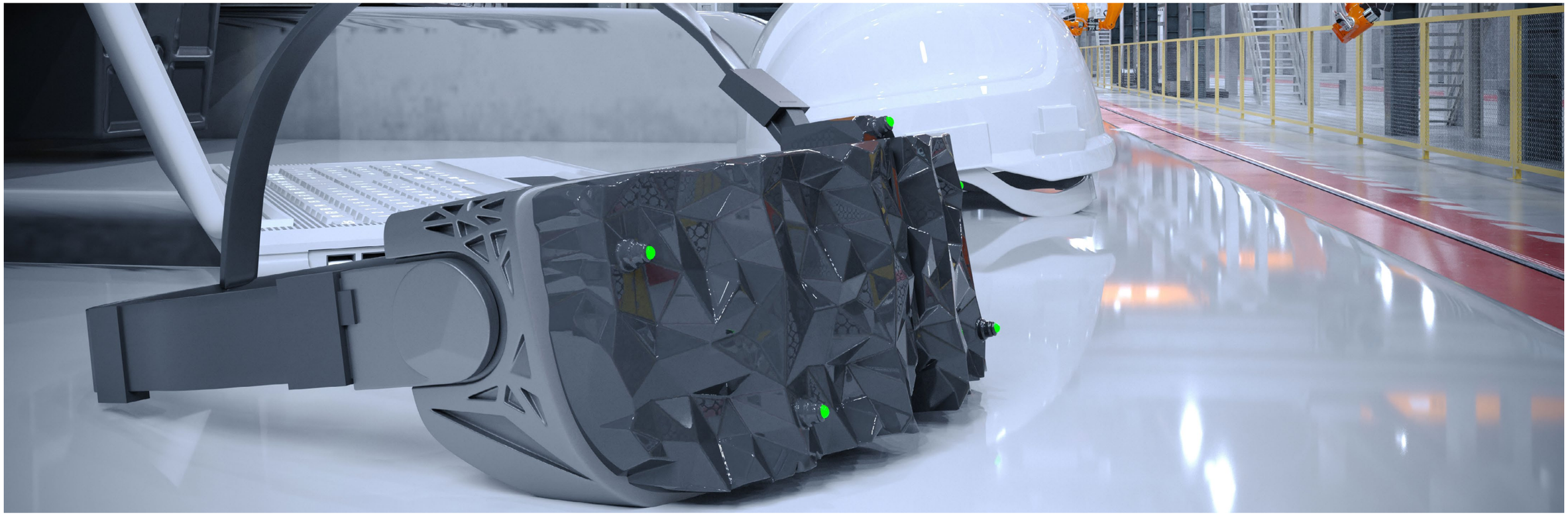
Closer to the edge

Many companies ready to take the leap with AR aren't able to do so because they are restricted in their ability to share data in the cloud. While they can use visual collaboration tools like Zoom or Microsoft Teams for standard communications, they can't leverage the same cloud-based technologies for sensitive enterprise workflows such as training, maintenance, and remote support due to concerns about data privacy and data sovereignty.

Performance is another barrier for AR applications using cloud-based collaboration tools. Advanced applications such as real-time video or image analytics can be stymied by latency issues.

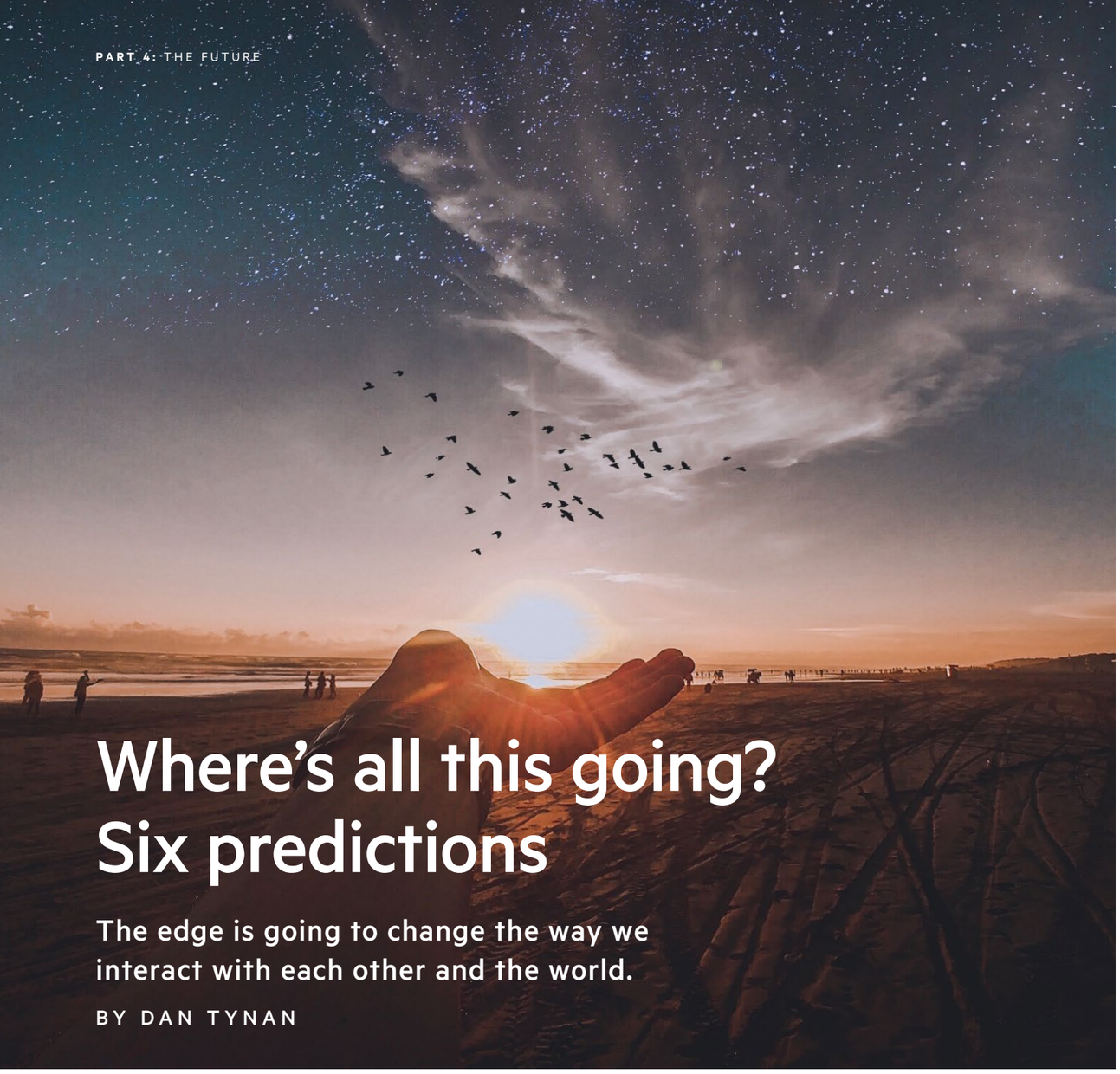
AR systems that can tap into edge computing resources can address issues of data privacy, sovereignty, and performance by taking advantage of workload-optimized compute resources near the point of access and behind firewalls, which offer an extra layer of security along with guardrails for data privacy protections.

Instead of waiting for all the technology pieces to fall into place, businesses would be well served to consider the benefits of AR when building out their three-year plans. Already, companies using AR to solve today's remote collaboration problems are benefiting in ways not possible in a pre-AR world. ■



PART 4

The future



Where's all this going? Six predictions

The edge is going to change the way we interact with each other and the world.

BY DAN TYNAN

THANKS TO EDGE COMPUTING, the world is about to look much different.

Within a decade, the edge will boast more computing power—and produce far more data—than the cloud does today, says Lin Nease, HPE Fellow and chief technologist for IoT.

The edge is where the Internet of Things, artificial intelligence, and ultrafast 5G networks are converging. And that will change our lives dramatically over the next few years.

Here are six major trends we can expect to see over the coming decade.

1. Every large facility will have its own edge data center

By the year 2025, the number of connected devices in the wild is expected to exceed 56 billion, according to IDC.[1] And as IoT sensors proliferate through public and private spaces, the volume of data they produce will grow exponentially.

Organizations will need to process that data locally so they can act on it in real time, Nease says. Instead of shuttling that data to the cloud, they'll be operating their own mini data centers on site. Modular, self-contained units like the HPE Edge Center, which can be placed wherever needed, will become common across a wide range of industries.

"Within three years, all operations facilities, including retail stores, hospitals, warehouses—any place the physical operations of the company occur—will have data room capabilities to ingest data from cameras, microphones, and environmental sensors," says Nease. "These little edge clouds will resemble the private clouds companies already have in their big data centers, but with less than a dozen servers."

And instead of deploying increasingly intelligent end devices, many organizations will opt for inexpensive sensors that simply collect data, which is then processed locally in their small private clouds.

"The low-cost approach is to buy cheap sensors and have software running somewhere in the facility that can do pattern recognition or infrared sensing," Nease says. "If you're a large retailer, for example, you will do this in your store in about three years. You will have to if you want to compete."

> **"Within three years, all operations facilities will have capabilities to ingest data from cameras, microphones, and environmental sensors."**
>
> **LIN NEASE**
> HPE Fellow & Chief Technologist for IoT

2. Smarter spaces will enable frictionless transactions

Data captured by edge devices in public spaces will help provide context around who people are and what they came there to do. That will change both how these spaces are designed and how people interact with them, says Partha Narasimhan, CTO at Aruba, a Hewlett Packard Enterprise company.

"Today, some of the context is split between the digital world and the physical world," he says. "Those two will absolutely come together."

For example, when you walk into your bank, smart cameras can identify you as a regular customer and make an intelligent guess as to why you came in, says Narasimhan. If you were researching interest rates on the bank's website the night before, a loan officer could be expecting you, with your financial records already called up on screen.

Similarly, when you step into your doctor's waiting room, edge systems can automatically check you in, pull up your electronic medical record, take your copayment, and alert the nurse you've arrived.

"As you journey through your course of care, hospitals will know who you are, where you are, and what assets, like wheelchairs or X-ray machines, you are using," notes Christian Renaud, research director for IoT at 451 Research. "They'll have lower operational costs, you'll have a faster diagnostic experience, and clinicians will get to spend more time with patients and less time entering data into electronic health records."

And, as we gradually return to the workplace, edge systems in the building will know that we're there primarily to collaborate with colleagues. We may be assigned a different workspace each day, depending on the people we need to collaborate with that day. Offices will be redesigned with more meeting rooms and fewer cubicles, better teleconferencing gear, smarter whiteboards, and more ways to intelligently capture the content of conversations.

"To some people, it may feel a little creepy that someone already knows who you are and why you're there," says Narasimhan. "But so long as you manage that data in a secure way that respects people's privacy, most of us will overcome that feeling because it's so convenient."

3. The robots will be watching us—and learning

The killer app for AI-powered cameras at the edge won't be security surveillance or autonomous vehicles, Nease says. Instead, they will be helping many of us become better at our jobs.

"People will be analyzing the physical environment of their operations using computer vision," he says. "But instead of employing it as a management tool, they'll be using it to redesign processes to figure out how to make them more efficient."

Companies like Drishti are already doing this on manual assembly lines for companies like Ford and Honeywell, using smart cameras to quantify processes, identify quality control issues, and train employees more effectively.

A major luxury car maker is studying how to use computer vision and robots to mimic the work of its craftspeople, notes Dr. Eng Lim Goh, CTO for high-performance computing and artificial intelligence at HPE. The goal is not to replace human workers but enable the creation of highly customized bespoke vehicles, built to the exact requirements of each customer.

"These robots will not only learn from humans but also from each other," he says. "And if you have factories in different countries, they'll be able to share their learnings, without sharing local data, across borders."

4. Edge devices will begin to learn on their own

Today, edge devices infer decisions based on machine learning models that have been trained in the cloud and then

1. "How You Contribute to Today's Growing DataSphere and Its Enterprise Impact," IDC, Nov. 4, 2019

pushed down to the edge. But as the battery life of IoT devices improves and their computing power increases, these devices will begin to learn on their own, Goh says.

"The beauty of this is that the devices collect the data and then immediately learn from it," he adds. "Imagine the latency reduction in your decision-making."

To avoid bias that can be introduced by relying on limited sets of training data, edge devices will be able to aggregate learnings from multiple sensors, a concept known as swarm learning. Because each edge device shares only the insights gleaned from its own data, the local data itself remains private and secure, Goh adds.

For example, connected X-ray machines at a hospital that treats a lot of patients with tuberculosis can share insights with another location that sees more cases of pneumonia. The ability to analyze lung X-rays at both facilities improves, while no patient data is shared.

"We'll continue to see more robust compute and analytics get further and further out in the network until they approach the point of origination, whether that's a camera, a sensor, a drill press, or an MRI machine," Renaud says.

5. Augmented and virtual reality will become actual realities

Today's AR and VR apps have proved how quickly nascent consumer technologies can deliver value in enterprise settings. As edge compute and connectivity continue to advance, we will see a continued evolution of experiences that are more immersive, with equipment that is less intrusive. That will drive continued applications of AR and VR into use cases that today are considered impractical.

With greater compute and storage capability at the edge, images and video can be cached locally and served up instantly via ultra-low-latency 5G networks.

"The edge will enable new services like AR and VR that require content to be closer to where users are," says Narasimhan. "Virtual experiences will be enhanced to the point where they're easy and intuitive to consume. And as they get more refined, they will make it easier for people to meet virtually." (See "A growing augmented reality," page 58.)

Beyond reducing the need for business travel, AR and VR will be deployed in a wide range of industrial and commercial settings to improve workflows, bridge expertise supply and demand imbalances, and transform customer experiences.

Factory workers can use AR glasses to view 3D schematics as they assemble parts. AR mirrors inside clothing stores will let customers digitally try on different outfits. We will see more widespread use of the technology by doctors to perform remote surgery a thousand miles away from the operating room. Visitors to theme parks will use it to interact with life-size holograms of their favorite characters.

"AR will allow employees to play with different datasets and explore them from different perspectives," notes Ross Rubin, principal analyst at Reticle Research. "It will be useful for any application that frees people from having to look at a screen at a particular time and overlay information in a way that sparks insights that might not otherwise be available in that environment."

Organizations that hope to realize the full benefits of edge computing will need to solve the privacy problem.

6. The big privacy issues will eventually be solved

The ability to deliver digital services to anyone in any location carries with it the ability to track everyone's behavior everywhere. Organizations that hope to realize the full benefits of edge computing will need to solve the privacy problem.

"We've done more than 50 consumer and enterprise surveys over the past five years, and privacy always comes up as one of the top concerns," Renaud says. "Hopefully, enlightened regulation will moderate the ambitious potential of the technology to get to a point where our privacy is safe."

In the European Union, for example, data sovereignty rules limit the physical locations where information can be stored, and some countries require connected cameras to automatically blur faces or license plate numbers.

Technologies like swarm learning, which allows insights to be shared without needing the underlying data, can help alleviate data sovereignty concerns. HPE and others are currently working on new data standards that would allow people greater control over how their data is used, Nease says.

"I think these standards will emerge and people will give permission for their data to be used, just as they do with Google every time they perform a search," he says. "But there is no stopping the privacy problem. It is now upon us." ■

enterprise.nxt

Hewlett Packard
Enterprise

a00115606enw, June 2021