

AI on the edge

When everything is connected to the cloud

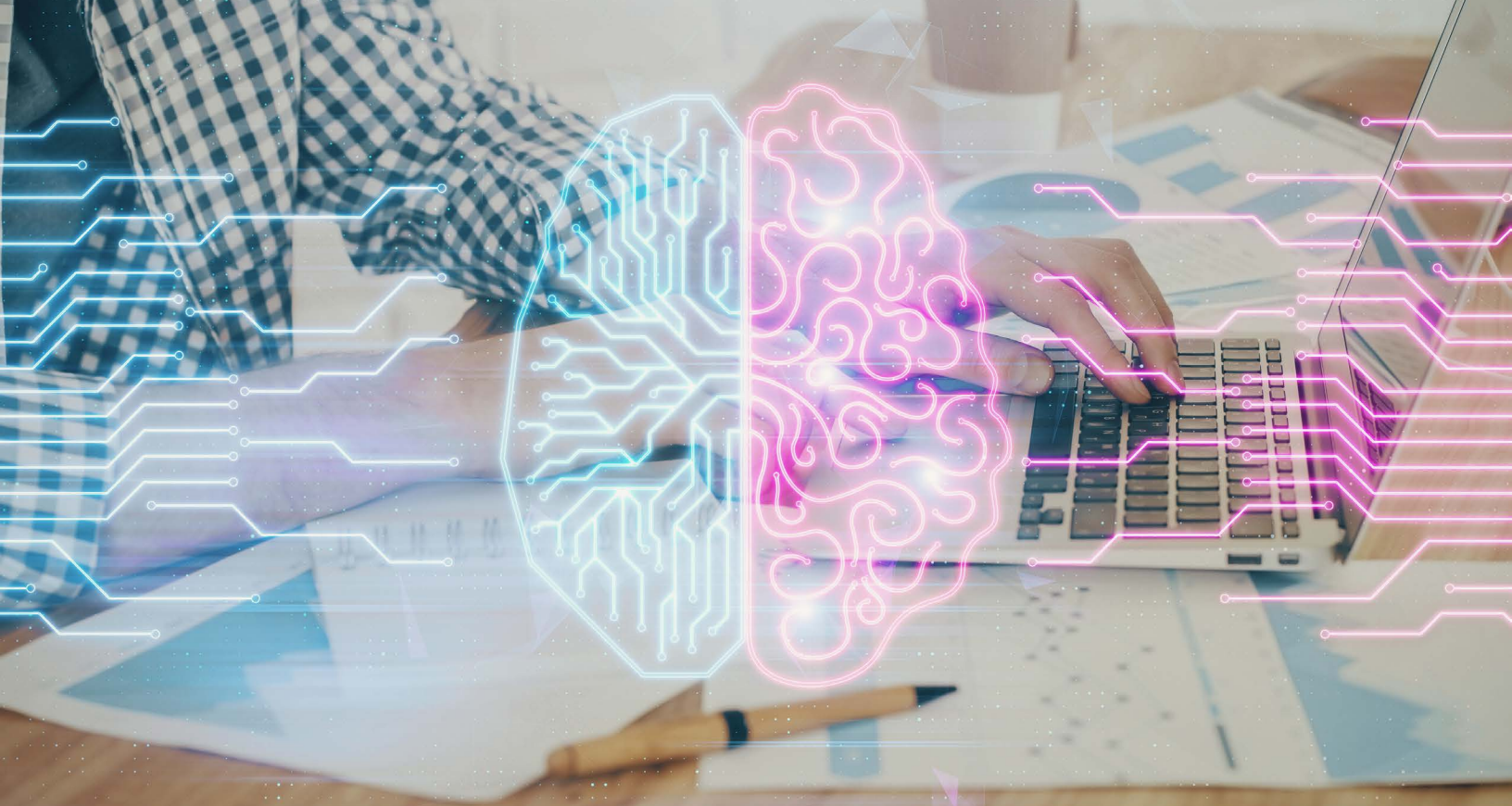
Introduction

While artificial intelligence (AI) has been a formalized research area since the middle of the 20th century, the seemingly rapid recent advancements in generative artificial intelligence (gen AI) have created a new level of interest and investment. While it's hard to say when exactly incorporation of gen AI capabilities inside operator organizations will create a clear, quantifiable return on investment, it is clear that enterprises of all sorts, including operators, are going to explore and spend on gen AI in pursuit of any efficiencies that can be realized.

The pop in interest around gen AI comes at a time when many operators are executing on cost-cutting strategies given that 5G has yet to deliver scaled new service revenues beyond what are effectively price increases on consumer subscriptions. The next phase of 5G is focused on the move to Standalone 5G marked by a cloud-native core, and the distribution of cloud computing to the network edge. Taken together, this will hopefully allow operators to deliver more meaningful services to new enterprise accounts in a less manually-intensive manner.

Looking at the intersection of 5G, AI, edge computing and the internet of things, a picture emerges of automated, real-time decision making for a growing number of use cases across vertical industries. Grand View Research and Allied Market Research project a more than 20% CAGR for the edge AI market by 2030 with both putting the future market size at around \$60 billion. This report examines why and how operators are putting gen AI to work, and includes expert commentary on the major benefits and risks of distributing gen AI workloads to the edge of the network—even onto the device—in pursuit of delivering new, high-value customer services while capturing operational efficiencies.





PUTTING GEN AI TO WORK MEANS MAKING A BUSINESS-SPECIFIC SOLUTION

Speaking in April at Dell Technologies World in Las Vegas, company Founder and CEO Michael Dell called out the need for new network architectures and more processing power for AI to deliver on its potential. He also discussed how business-specific, proprietary large language models will become increasingly important as enterprises contextualize their AI investments in terms of security and specialized use cases, e.g. a bank using AI to calculate credit risk or fraud detection rather than a bank, for some reason, using AI to generate poems or images.

"There will be significant proliferation of open- and closed-source large language models for general and specialized uses...The real opportunity is to re-imagine your organization and what you can become given the superpowers AI is unleashing," Dell said. Describing himself as an optimist in terms of the potential negative impacts of broad AI adoption, he said, "Throughout human history we've successfully managed the risk of potentially existential technologies...Our measures of progress are bound to our innovations."

Speaking following the keynote in a group Q&A with media and analysts, Dell called AI "a massive unlock of the power of data...I think this is as big or bigger than the internet, PCs, the smartphone... For us, it's a very comprehensive kind of reimaging of our business."

IBM Consulting SVP John Granger had a similar message at the JP Morgan Global Technology, Media and Communications Conference where he detailed how IBM is "placing its emphasis" on smaller models that "are much more curated."



The Data Platform that Helps Capitalize on Real-time Data at Scale, Without Sacrificing Performance.



EDGE AI PROCESSING



STREAM INTEGRATION



REAL-TIME AGGREGATION



CLOUD NATIVE

Get the STL Partners Report: Revenue opportunities at the intersection of Edge and IoT



VOLTACTIVEDATA.COM

The Volt Active Data Platform enables companies to unlock the full value of their data and applications by making it possible to have scale without compromising on speed, accuracy, or consistency. Based on a simplified stack and an ingest-to-action layer that can perform sub 10-millisecond decisioning, Volt's unique, no-compromises foundation gives enterprises the ability to maximize the ROI of their 5G, IoT, AI/ML, and other investments, ensure "five 9's" uptime, prevent fraud and intrusion, deliver hyper-personalized customer engagement, and save on operational costs.

He continued: "I think what's critical here for businesses is confidence around the quality of the data, the amount of the bias, and so on...That is the fundamental issue I think. We in IBM have always been super focused on enterprise, so what I would say is that at the moment...clients are experimenting with both types, the broader models and the more narrow. When they are starting to think about how they might go into production and scale, they're shifting much more to the narrower foundation models."

NVIDIA, which is an inarguable leader in the AI market, has articulated a vision that speaks to how operators can leverage AI for their own, and their customers purposes, while also using distributed computing infrastructure to open up even more new revenue by offering AI-compute overhead up as a service.

Speaking at Computex in May, NVIDIA CEO Jensen Huang said, "Accelerated computing and AI mark a reinvention of computing...We're now at the tipping point of a new computing era with accelerated computing and AI that's been embraced by almost every computing and cloud company in the world." Specific to the telecoms set, Japanese operator Softbank is working with NVIDIA to boost its RAN performance through the use of AI, while also building out excess AI capacity that can be sold.

NVIDIA's SVP of Telecom Ronnie Vasishtha highlighted the flexibility of a multi-use architecture as compared to "telecommunications networks [that] are built for a single purpose...[Operators] have built...for peak demand. So you're over-provisioning the 5G network for peak demand. As new AI applications come in, that peak demand is going to grow, the power required to power this

network is going to grow, the compute requirements for that network are going to grow," he said.

To address this, NVIDIA's concept is to take into account the delta between peak and average utilization, and put the difference to work by standing up "AI factories. In fact," Vasishtha continued, "5G becomes a software-defined overlay within that datacenter and can be provisioned to the use requirements of the 5G network in an automated way. That means that even if you're running RAN at 25% of what you would've done in a proprietary network, you can run it at that rate and the rest of the datacenter is being used for AI...That's very easily said but it's difficult today," he acknowledged. But the vision remains: "5G now runs as a software-defined workload in an AI factory."



HOW IS GENERATIVE AI RELEVANT TO OPERATORS?

Discussion around gen AI has quickly become a mainstay in conference presentations and earnings calls. The predominant thinking is that if you're an enterprise, including a mobile network operator or other communications service provider, you've got to have a plan to put generative AI to work. So, to start, are gen AI tools like ChatGPT relevant to the telecoms set?

VIAVI Solutions Regional CTO for EMEA Chris Murphy explained that solutions like ChatGPT, Midjourney and the like have garnered so much attention because their outputs are "readily tangible and visualizable...but this is just a start...In specialized areas such as telecoms, generative models are starting to have quite an impact and that impact is only going to grow as time goes on." Murphy was also quick to point out that these types of gen AI tools are "part of a whole ecosystem of the technology of machine learning and artificial intelligence."

Before digging deeper, let's take a moment to recap types of classic machine learning algorithms that

have reached a comparatively high degree of maturity, new innovations in machine learning, and new innovations effectively upgrading tech that's been around for decades. Mature machine learning algorithms include things like regression, classification, clustering and dimensionality reduction. Newer machine learning innovations--which are also giving a material boost to classic approaches--include things like transfer learning, reinforcement learning, automated machine learning and, of course, content generation.

As Murphy put it, "There are whole new disruptions in the technology from transfer learning allowing us to train models in a general sense and then apply them to a particular focus area. Reinforcement learning where we can learn to interact with the system in an optimal way, and auto ML where we can automate, to some degree, the generation of the machine learning models and the way we train it to take away some of the expertise that's needed to generate some of these very interesting applications."

This idea of taking a large, general model then honing the focus for a particular type of business or type of business process aligns with how enterprises are actually using gen AI; enterprises are using models trained on domain-specific, proprietary data rather than the collective knowledge of humanity as it exists online.

Back to the relevance of ChatGPT and similar tools to the telecom industry: "We're starting to see operators implementing large language models for helping their to do their jobs in better, more creative ways."



Chris Murphy

Monetizing the future

The big idea behind gen AI is predicting what's next. You give ChatGPT a text-based prompt and it tells you which words come next, whether that's to write a poem, draft an academic essay, supplement web searches or assist in developing code or HR documentation. "When we can predict the future," Murphy said, "we can come prepared for it ahead of time in spinning up the resources required. If we anticipate an increase in demand, we can run what-if scenarios, for example, to measure our resilience to certain scenarios."

He continued: "We can generate synthetic data, which is tremendously valuable, for example, for training our models using realistic data without having to collect that realistic data in large volumes. And this all leads into things like digital twin, which is an enabler for all sorts of training scenarios of optimization to drive our telecommunication networks to the next level...ChatGPT is in the collective consciousness, but I guess what I'm saying is that this is just one small piece of what's going on in the technology, and there's a lot more that will be at least as important as the large language models are for telecommunications."

For more on this idea of creating a digital twin of the network, read "From real to lab to live—continuous testing in the era of AI."



AT&T TAPS GEN AI FOR EMPLOYEE EFFECTIVENESS, EFFICIENCY AND CREATIVITY

Using OpenAI's Chat GPT functionality combined with "special sauce" and Microsoft Azure compute power, AT&T in June announced Ask AT&T to support internal operations. In a corporate blog post, Chief Data Officer Andy Markus wrote: "Out of the box use cases are helping our coders and software developers across the company become more productive. It's also helping to translate customer and employee documentation from English to other languages as well as to simplify that documentation and make it easier to use."

AT&T is also looking at use cases, including network optimization, "upgrading legacy software code and environments," customer care, HR support and automated summarization and action items from meetings.

What AT&T is doing, and how it's doing it, also speaks to the focus on taking open large language models and more precisely tailoring those solutions for a specific company.

As Markus wrote, Ask AT&T "incorporates AT&T knowledge and processes that focus the system and responsibly deliver accurate results." With regard to Microsoft's role, Ask AT&T "runs in an AT&T-dedicated Azure tenant that's been pressure tested for leakage. AT&T employees can bring company data and information into Ask AT&T without worrying about that material leaking into the public domain."



Generative AI at the edge

Big picture, 5G is continuing to evolve and continuing to become more complex. Virtualization and disaggregation are happening in tandem with deployment of network workloads in hybrid cloud environments, including distributed clouds deep in the network at radio sites or even at customer premises. This is prompting organizational overhauls wherein operators are establishing DevOps workflows and setting up CI/CD pipelines to take advantage of the network's flexibility. And all of this is in service of delivering new types of (largely) enterprise-facing services that set the stage for new lines of revenue and more effective network monetization.

Against this dynamic, complex backdrop, data is king. "The data can be aggregated in different ways, everything from subscriber level, the individual user level through to aggregated by network function and network components," Murphy said. "So we've got a lot of data. That's good news because we've got some interesting models, but we have to be careful about how we use the data... If we're making decisions about how to schedule our data or how to adapt our MIMO, these decisions need to be made right at the edge because those are very short-term decisions based on the channel conditions, for example, and we need to be able to react to those changes."

With the goal of leveraging data to predict the future then prepare to react accordingly, the edge is an important focal point in the larger context of 5G monetization. 5G is designed to support low-latency applications over

the air interface, but to do this requires moving compute out of centralized data centers and closer to the people and devices creating the data. "There's an interesting challenge here of how to deal with the data that we have to squeeze the maximum value out of... So the data that we use for training, it may be generated at the edge and be in large volumes. And we might want to be careful about transferring [that data] to the datacenter for training models which then need to be operated at the edge." Hence the importance of AI optimized for edge deployments.

In summary, "We're lucky to have a lot of data that we can draw on to build the best models, train in the right place, in the optimal place, operate in the optimal place to deliver performance improvements and the ability to deliver the interesting and disruptive services," Murphy said.

DIGITAL TWIN SOLUTIONS

Simplifying Testing in an Increasingly Complex Network

The VIAVI Digital Twins combine RAN and Core emulators, assurance solutions, realistic traffic scenarios, and cyber threats to mirror an operator's network in the lab:



Run 'what if' analysis without experimenting with real subscriber data.



Test cyber-attacks and mitigation strategies to see how they impact the network and subscribers.



Test the automation of network operations, while making smart decisions based on business goals to maximize ROI.



AWS TALKS GEN AI FOR OPERATORS— A PHASED JOURNEY

Building a foundation, training models, operationaliza- tion

In June, AWS announced a \$100 million investment to develop a center to help enterprises use gen AI, the technology behind OpenAI's Chat GPT chatbot. The AI innovation center will offer generative AI testing and training services and will be staffed by data scientists, engineers and solutions architects. To find out more

about the company's broader approach to generative AI and how telcos might use it, RCR Wireless News spoke with Ishwar Parulkar, the chief technology officer for the telco industry at AWS.

While Parulkar acknowledged that gen AI does indeed represent a "big transformation," he also pointed out that AWS is no stranger to AI and has, in fact, been working in this space for the past 15 to 20 years, both internally in a customer-facing capacity via offerings like the code generator tool Amazon CodeWhisperer and Amazon Bedrock, which makes it easier for developers to build foundational generative AI models.



Looking ahead, though, Parulkar said that gen AI will "touch pretty much every industry" in one way or another and that nearly every customer the company works with is interested in finding ways to put this tool to use. For telcos, specifically, Parulkar sees the value of gen AI unfolding across the following three phases:

The foundational phase

The first phase involves the foundation models and capabilities that exist today, such as text summarization and generation, as well as image generation. “So we’re looking at things like chatbots that can be used for customer care,” he explained further, adding that this of particular importance to telcos because customer churn is a top metrics they use to track success. Therefore, they are looking at using AI to improve the customer experience, as well as business applications like revenue assurance.

Additionally, the existing manuals that field technicians use to install equipment and troubleshoot network equipment can be used to train models that can create a more real-time interactive interface to guide them through the installation and troubleshooting processes.

For the most part, all of these capabilities are available today, according to Parulkar.

The training phase

The second phase, shared Parulkar, is about training foundational models to do “new types of activities.” One of those standout activities for telcos is configuring networks. However, Parulkar said that in order for this to be possible, the existing foundational models must be tuned with new data, and a lot of it, and so this phase will take some time to come to fruition.

The network-based phase

The final phase is “a little way out,” Parulkar confided. These would be foundational models that are “network-based” and “geared for the network.” Such models, he continued, could be used for several network applications “from the designing of [the] network to maintenance of networks to all operational aspects of networks.”

In closing, Parulkar warned that gen AI still faces notable challenges, particularly around security, data privacy and ethics and that it is crucial for telcos – or anyone looking to use this tool, really – to remain vigilant and only operate in environments where the AI can be managed with responsibility. “This will only become more and more important because gen AI is based on large amounts of data from multiple sources,” he added.



THREE BENEFITS OF DEPLOYING AI AT THE EDGE

Running AI workloads at the edge enables better economics, faster decision making and automation

If you look past the hype, look past the technological complexity, look past the protracted proofs of concepts, 5G is all about leveraging a high-bandwidth, low-latency air interface to move data, analyze data and action on data in as close to real time as possible. With that capability in place, enterprises of all stripes could realize every possible operational efficiency, automate things that can be automated, and see the benefits on their balance sheets. To hasten decision making and automate where possible, AI has a clear

role to play. And, given trends in distributed network architectures and distribution of cloud compute/storage infrastructure, AI residing at the edge, where data is created, brings numerous benefits to enterprise users.

Andrew Keene, head of product management at Volt Active Data, made the case for AI at the edge in a recent webinar hosted by RCR Wireless News. He noted that AI at edge is not just advantageous but “sometimes critical for the viability of the use case.” But why?



Andrew Keene

"Some of these use cases generate vast volumes of data, much of which in itself is relatively useless," Keene explained. "But it all must be processed for the use case to work. Backhaul and transmission costs to the cloud can be prohibitive, but if you can process the data at the edge, and only send valuable consolidated data to the cloud for further processing, that kind of solves the problem."

Beyond the economic reality of moving data, timescales are critical—Volt Active Data tends to measure things in single digit milliseconds. "If you could make... decisions much closer, at the edge, to where the events are happening, that again solves the problem of ultra low latency responses." Another implication here is around data sovereignty/security; enterprises dealing with proprietary or regulated data need to hold that data closely. "So a distributed tiered data platform that can run these [machine learning] models at the edge and only send consolidated, invaluable data to the cloud, mitigates many of these issues and actually makes some of the use cases viable that otherwise wouldn't be."

Back to that idea of deploying AI at the edge to speed up decision making—in addition to doing that, AI can also continuously get better as it has access to more and more data. "We are seeing a big and increasing interest in machine

learning models...to enhance many different real-time data processes and use cases across a variety of industries to automate continuous model improvements," Keene said. This lets the enterprise user achieve a better outcome and potentially pass on improved outcomes to their customers.

Keene summarized: "What really makes for a powerful solution is when these machine learning models not only continually are updated to improve their accuracy rating by learning from real outcomes, but when you execute them at the edge on a distributed platform to ensure that optimal response times and to reduce unnecessary backhaul of costs of transmitting potentially prohibitively large quantities of data to some centralized Cloud hosted platform."



INTERNET OF THINGS

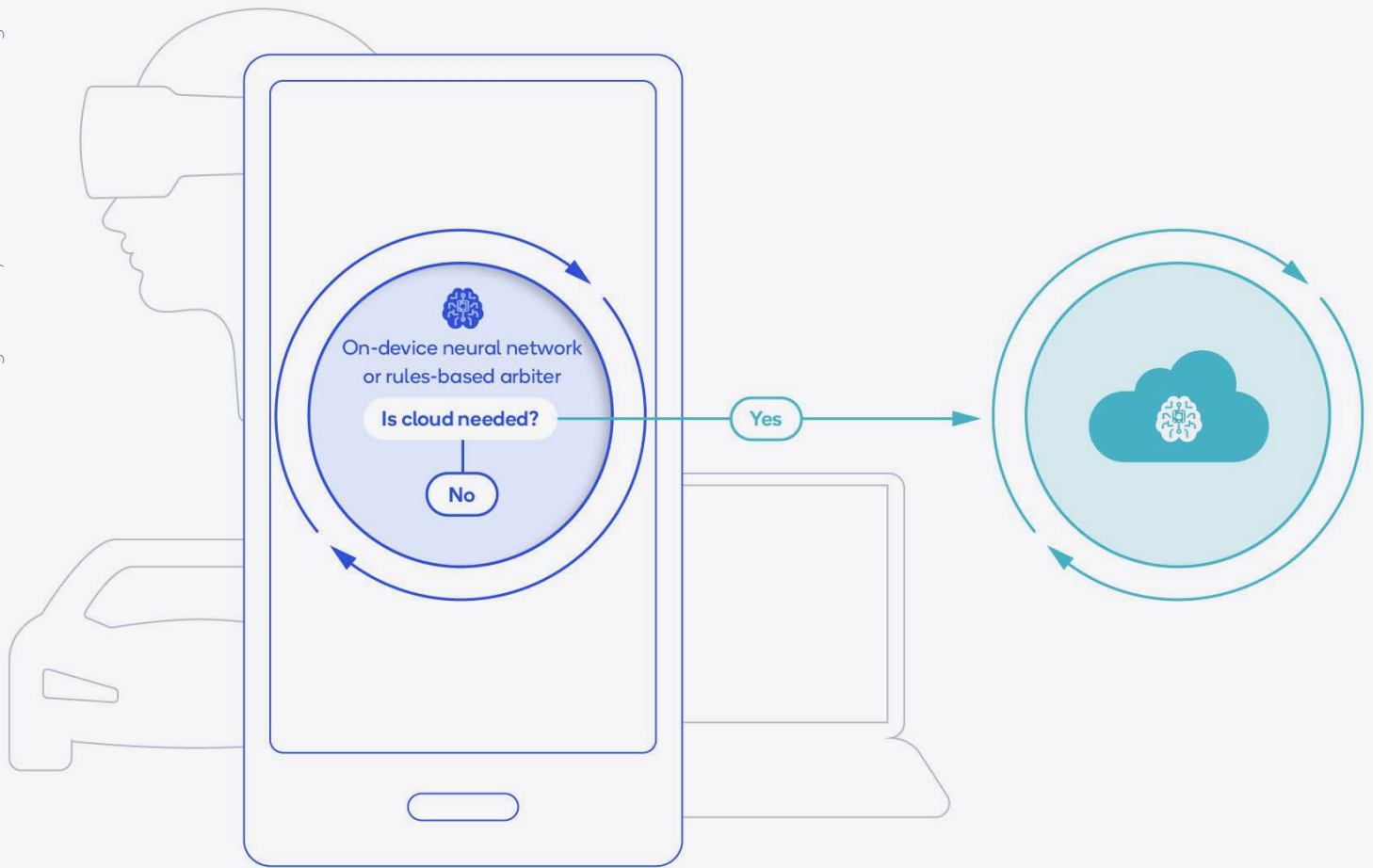
Edge AI to support IoT device management

Volt's data management solutions designed to support a range of real-time applications, Keene said, calling out fraud and threat prevention, hyper personalization, real time private network SLAs, traffic management, fleet management, charging and policy, IoT device management, compliance and regulatory reporting, edge optimized federated decisionmaking, and active digital twins. But, he noted, "We provide the enabling technology...not the end application," which is built by their customers and partners. The company's focus is on "applications that require

massive scale, low latency, accuracy and resiliency and are well suited to executing machine learning models at scale."

Keene gave the example of using 5G in dense urban areas to improve traffic management through proactive routing; this is in contrast to legacy traffic management solutions like timed traffic lights set to vary through peak, off peak and other static configurations. By introducing 5G and AI-powered processing at the edge, "Smart cities have the ability to take feeds from hundreds of traffic cameras or other endpoints and use machine learning as a form of AI to predict the behavior, spot anomalies as they happen, and to route traffic away from problem areas in real time."

Another example that would fall under the Industry 4.0 umbrella is IoT device management. Keene laid out a hypothetical operation where 5G SIM cards are used to connect, track and monitor various people and assets. "We've seen experiences of enterprises that have problems with fraud" in the form of people stealing SIMs or assets with SIMs in them. "The solution uses geofencing to bar or alert when a device moves outside its normal operating area and also spots unusual traffic data usage patterns of device changes. The machine learning model learns the regular, predictable usage patterns and operating geographical areas. So if it's normal for a particular device to go offsite every Thursday afternoon to collect something, it won't flag it as fraud, whereas any other device that never moves outside its area gets flagged straight away if it happens to stray."



WHAT IS HYBRID AI?

Benefits of hybrid AI include data security, context and efficiency

With gen AI is poised to permeate virtually all types of consumer and enterprise applications, network architecture becomes increasingly important. If the goal is to use AI to quickly make a decision and effect an action, latency is a major consideration. And with increasingly performant end user devices, hybrid AI architectures that distribute AI workloads between centralized clouds, distributed edge

clouds and the device itself bring a number of benefits.

Tantra Analyst Founder and Principal Prakash Sangam, discussing the subject on a webinar hosted by RCR Wireless News, said hybrid AI supports improved security and privacy, context-aware decision making, scalability, and cost efficiency.

To the security/privacy point, "Every organization...will share so much information with AI systems, and storing all that away at a distant service on which users have no control seems really dangerous and worrisome...If it can be stored on the device or at an edge cloud that you have control over, then that makes it that much safer and gives security."

Sangam also discussed the importance of context in effectively using generative AI which mirrors messaging from major AI companies around the need for enterprises to use their own data to develop their own models for their own specific use cases. While the versions of tools like ChatGPT used by you or me contain enormous amounts of generic data essentially sourced by crawling the internet, a domain-specific tool would provide more immediate business value.



Prakash Sangam

"Domain specificity is important," Sangam said. Open models use "huge amounts of generic data. When you use that data for predicting something very specific, obviously errors are prone to happen. So, because of that and especially for enterprises, if you are trying to use AI for any of your applications, it makes sense that you use the data specific to your domain so that you get very accurate results. It is better to do that on an edge server somewhere or on-device." This also has obvious security/privacy implications for proprietary corporate data.

Sangam also talked through the business logic around where AI workloads are run, and how that relates to ease of scalability. For AI models, training helps a model learn from data and is followed by inference where the AI makes decisions based on its training. The centralized cloud lends itself to training due to the sheer compute and memory capacity necessary; development of edge computing capacity could enable some training to happen at the edge, but certainly inferencing will happen at the edge—as well as on the device in a hybrid architecture. "And for companies providing gen AI," Sangam said, "it'll be very cost effective" to use a hybrid architecture. "They don't have to invest in all of this infrastructure to do the gen AI."

In terms of the on-device part of a hybrid AI architecture, Qualcomm sees itself in a "unique place," as Qualcomm CFO Akash Palkhiwala described it at the JP Morgan Global Technology, Media and Communications Conference.

"This is a rapidly evolving industry...over the last several months and it's going to continue to be that over the next year or two," Palkhiwala said. "If you think about the hyperscalers, there's a tremendous focus and effort on having large language models running in the cloud, but also reducing the size so you can run it on the device." Given that divergence in the size of LLMs and Qualcomm's proven capabilities around on-device AI, that's important "because you could run the smaller models on the device with very good accuracy and performance which we can take across our ecosystem."

Key point here is that, for Qualcomm, on-device AI maps to the company's ongoing strategy to diversify beyond smartphones. As the company looks to grow its consumer and industrial IoT businesses, as well as its booming automotive business, on-device AI tech developed for handsets can be tweaked and ported across all lines of business.

Palkhiwala explained: "From an AI perspective, our view is as large language models come into play, a lot of the inference is going to happen on the device rather than in the cloud...The cost is definitely way cheaper on the device side," not to mention considerations around data privacy and security, and application-specific latency needs. "If you can run a model, inference-wise, on the device...that's a huge advantage for us...We have the opportunity to expand the capacity of this low-power engine [developed for smartphones] and apply it to large language models...That's what creates an advantage for us going forward...It's something that creates a competitive advantage for us across all edge devices."

THE NEED FOR HARDWARE/ SOFTWARE CO-DEVELOPMENT TO SUPPORT ON-DEVICE AI

A new report from the tinyML Foundation, a nonprofit professional org focused on nurturing ultra-low power machine learning tech and edge intelligence, and tech-focused digital media firm Wevolver, looks at the edge AI benefits around latency, bandwidth and security among others, while also acknowledging challenges that emerge as edge AI scales.

The report authors focus on the device endpoint as a primary “edge,” and dig into the Internet of Things development that would come along with AI embedded at the myriad devices making up this ultimate edge, a key piece of a hybrid AI architecture. “Edge AI essentially functions as an extensive sensory systems, continuously monitoring and interpreting events in the world,” the report authors wrote.

To get to that point, the authors call for “an integrated technology approach” where on-device AI enables use cases that require rapid decision making. “The successful deployment of AI at the edge requires a harmonious blend of hardware and software co-design, informed by a deep understanding of the constraints and potentials of both domains.” That essentially speaks to delivering on-device compute power without sacrificing form factor or power draw. This is playing out in the real world with advancements in semiconductor design/fabrication, and co-development between semi firms and OEM customers.

As for the role of 5G, the report authors called out the system-level ability to support a much higher device density, along with improvements in latency and bandwidth as compared to LTE. “5G networks facilitate edge AI applications to access large amounts of training data and to operate without disruptions in crowded environments and device-saturated contexts. As 5G networks continue to expand, we can expect a proliferation of AI-enabled intelligent edge devices that will perform complex tasks and make autonomous decisions in real time.”

Conclusion

There’s a strong case to be made for operators and enterprises to consider their investments into 5G, and other connectivity mediums, and distributed compute/multi-cloud in the context of optimizing where AI workloads are run. If the goal with AI is to speed up accurate decision making for process optimization, the workloads need to reside as close to the source of data as possible. Similarly, to get the best results, enterprises need to leverage their internal data rather than relying on open models. To end we return to Michael Dell: “If you aren’t already thinking about AI and how it will change every part of your business, then you are already behind.”