Transcription:

Steve McDowell: The reason we push AI to the edge is because that's where the data is, you know, we want to do the processing close to where the data is so that we don't have latency. And in a lot of environments, if we're ever disconnected, it's going to shut down my business.

Jason Lopez: The question is, how do you deploy edge resources in real time? In this Tech Barometer podcast, Steve McDowell, Chief analyst at NAND research talks about his paper "Taming the AI-Enabled Edge with HCI-Based Cloud Architectures." I'm Jason Lopez. Our aim in the next several minutes is to discuss how AI impacts edge computing.

Steve McDowell: We've always defined edge as any resources that live outside the confines of your data center. And there's some definitions that say the extension of data center resources to a location where there are no data center personnel. It's remote.

Jason Lopez: But AI, of course, adds complexity. One example McDowell cites is automated train car switching. The sides of train cars have bar codes which are scanned, and a local stack of servers processes where the cars are and where they need to be.

Steve McDowell: I can do this in real time. I can partition my workloads so that, you know, computationally expensive stuff or maybe batch stuff can still live in the core. And I don't have to do that at the edge all the time. So I can really fine tune what it is I'm deploying and managing.

[Related: Slew of Changes Drive VMware Customers to Consider Alternatives]

Jason Lopez: This is important when you consider that AI at the edge differs from traditional edge deployments primarily due to its need for greater computational power.

Steve McDowell: Once we start putting AI in, then suddenly we have to have the ability to process that AI, which often means the use of GPUs or other kinds of AI accelerators. Ten years ago, if we talked about edge, we're talking largely about embedded systems or compute systems that we treat as embedded. Embedded is a special word in IT. It means it's fairly locked down. It doesn't get updated very often. When we look at things like AI, on the other hand, that's a very living workflow. If I'm doing image processing for manufacturing, for example, for quality assurance, I want to update those models continuously to make sure I've got the latest and the greatest.

Jason Lopez: And along with managing fleets of hardware and software in AI deployments at the edge, there's also the issue of security.

Steve McDowell: By treating edge systems as connected and part of my infrastructure, and not as we historically have treating them as kind of embedded systems, if you will, it also

allows me to, in real time, manage patches, look at vulnerabilities, surface alerts back up to my security operations center, my SOC. It makes the edge look like it's part of my data center.

Jason Lopez: Tools like Nutanix allow for this approach, applying a consistent management practice across both core and edge environments. This involves deciding what tasks to perform at the edge versus the core due to constraints like cost, security, and physical space.

Steve McDowell: A key part of the conversation becomes what lives where? And that's not a tool problem, right? That's kind of a system architecture problem. But once you start partitioning your workloads and say, this certain kind of AI really needs to be done in the core, Nutanix gives me that ability and cloud native technologies give me that ability to say, well, I'll just put this kind of inference in the cloud and I'll keep this part local.

[Related: Pivot Past the Enterprise AI and Cloud Native Hype]

Jason Lopez: McDowell's thinking springs from the flexibility afforded by hyper-converged infrastructure. The idea of AI at the edge is part of the whole architecture of storage, network and compute.

Steve McDowell: That can be as disaggregated as it needs to be. So if I need a whole lot of compute in the cloud, I can do that and then put the little bit at the edge and I can manage all of that through that single pane of glass, very, very powerful.

Jason Lopez: Treating edge computing as a part of the data center becomes so interesting because of how the data center itself is being transformed by AI and machine learning.

Steve McDowell: Once we abstract the workload away from the hardware, I've broken a key dependency. I don't have to physically touch a machine to manage it, to update it, to do whatever.

Jason Lopez: The point McDowell makes is how management, not just of the configuration of a node, but across a fleet, is simplified. It enhances efficiency and scalability.

Steve McDowell: We're taking technology that evolved to solve problems in cloud, but they apply equally to the edge, I think. It turns out, it's a fantastic way to manage edge.

[Related: More Reasons for HCI at the Edge]

Jason Lopez: AI at the edge is increasingly adopting cloud-native technologies like virtualization and containers. The shift is to container-based deployments for AI models, sharing GPUs and managing them remotely.

Steve McDowell: If you look at how, you know, NVIDIA, for example, suggests pushing out models and managing workloads on GPUs, it's very container-driven.

Jason Lopez: And McDowell explains why this simplifies edge management.

Steve McDowell: A GPU in a training environment is a very expensive piece of hardware. And giving users bare metal access to that, you know, requires managing that as a separate box. Using Cloud-native technologies, I can now share that GPU among multiple users, very, very simply. That same flexibility now allows me to manage GPUs at the edge with the level of abstraction that works. So I can sit in my data center, push a button and manage that box without actually worrying about what that box looks like necessarily. So I don't need that expertise kind of onsite, right? Which is a key enabler for edge. If you have to have trained IT specialists wherever you're deploying, that doesn't scale. And edge is all about scalability.

[Related: The Future of AI Computing Resides at the Edge]

Jason Lopez: GPUs are typically what power AI, but are are not commonly found at the edge. But inference is a facet of AI that many technologists see value in at the edge. GPUs would be the right fit if at the edge, generative AI is needed. But what's needed now are inference engines, especially around vision and natural language processing.

Steve McDowell: Take, for example, a retail environment where they have intelligent cameras that are positioned all up and down the aisles of the grocery store. And the only job that these cameras have is to monitor the inventory on the shelf across from the camera. And when they've sold out of Chex mix and there's a gap there, it sends an alert, come restock. I mean, it's very kind of data intensive and you don't want to send that to the cloud necessarily.

Jason Lopez: Technology is moving toward managing infrastructure environments seamlessly, such as edge, data centers, and cloud, without changing tools or management models.

Steve McDowell: Nutanix has capabilities for managing AI in your workflow, kind of period, full stop. A good example of this is GPT in a box. Where it's a technology stack and I plug a GPU in and I can do natural language processing. If I want to push that out to the edge. I don't have to change my tools. I mean, the beautiful thing, and the reason that we use tools like Nutanix is that it gives me kind of a consistent control plane across my infrastructure. Now, infrastructure used to mean data center, and then it meant data center and cloud. And now with edge, it means data center and cloud and edge. The power of Nutanix though, is it allows me to extend outside of my traditional kind of infrastructure into the edge without changing my management models. So, as AI goes to the edge, I think the things that already make Nutanix great for AI in the data center are equally applicable at the edge.

Jason Lopez: Steve McDowell is founder and chief analyst at NAND research. This is the Tech Barometer podcast, I'm Jason Lopez. Tech Barometer is a production of The Forecast, where you can find more articles, podcasts and video on tech and the people behind the innovations. It's at theforecastbynutanix dot com.