Edge AI Computing in Automotive: Scaling for Success

June 21, 2021

By Christophe Couvreur, VP of Product

I recently had the pleasure of joining Dean Harris, Automotive Business Development, NVIDIA; Alexandra Baleta, Manufacturing and Automotive Industry Director, VMWare; and Sunil Samel, VP, Products, Akridata for an exciting conversation about edge computing for AI and machine learning workloads in automotive as part of NVIDIA GTC 21. The discussion, moderated by NIVIDIA's Manish Harsh, focused on the benefits of edge and near-edge compute infrastructure for latency-sensitive applications.

Here are just a few of the highlights from our conversation:

Manish: What are the biggest infrastructure challenges facing the automotive industry?

Alexandra: When it comes to AI, you want to deploy efficiently in a timely manner. You want to have developers quickly deploy resources to run your trials and tests; you don't want to wait for a server to be ready. You also want performance to be really high. That's not only in a pilot phase, but at scale, and this is where we tend to see performance degrading because we don't have the infrastructure underneath. That's why we see up to 53% of projects in the AI space not going anywhere between pilot and production; they are just being stalled because the infrastructure underneath can't cope. That complexity is usually underestimated and a real challenge.

Christophe: What we have seen in the last few years is deep learning technology that has revolutionized the way we approach AI. Deep learning is in one way simple, but also quite challenging to deploy because it requires massive amounts of data and massive amounts of compute power both to train the AI but also to deploy the different AI engines once they have been trained. How do you get that huge network to scale massively to get into every car? That's where we see quite a few challenges that need to be tackled.

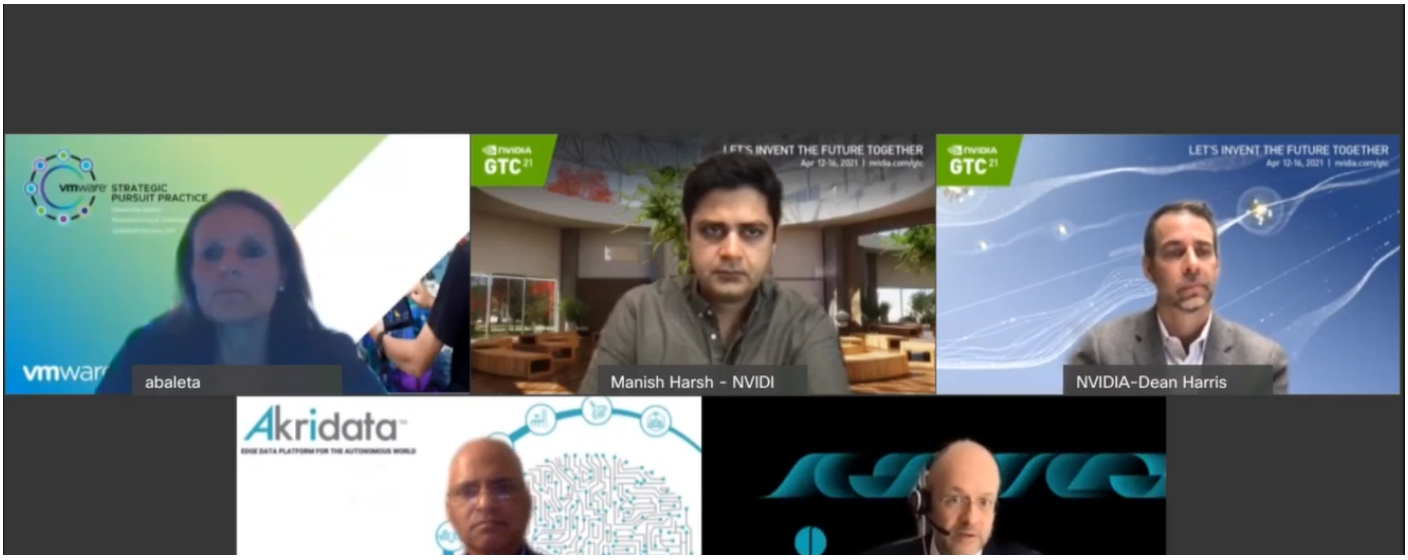
Manish: What is the importance of low-latency capabilities for OEMs?

Christophe: Latency has a huge impact on the driver experience. People don't realize it, but we are very sensitive to timing. If you click on a button, and it takes 50 milliseconds for the button to respond to you, you feel it is slow already. It is not about cost savings; it is about improving user experience. OEMs today are providing an experience for the driver – it is no longer just speed or low miles per gallon. Cars are defined by the experience they provide for the driver, so latency is a key factor. Improving the latency makes the user experience better – making the system fast and able to immediately respond.

Dean: To us, the same challenges with enterprise IT can be seen with edge applications – presence of legacy disparate systems, the need for unified platforms across organizations, need for latency sensitive applications, security – they all play into the challenges of developing applications at the edge.

Manish: What are the key elements of near edge and cloud that bring support to latency in applications especially for AI and connected vehicles?

Christophe: It's important that systems move to edge and cloud seamlessly, using the edge for what can be done there and resorting to the cloud when it is needed. For example, having an in-car virtual assistant respond to inquiries about turning the temperature up does not require the cloud, but for a request like, "what is the Yelp rating for the restaurant to the left?" you would need to go to the cloud.



Manish: How do you see the decision making of hybrid architecture and how do you see that option happening?

Sunil: From our perspective, hybrid architecture is a given. It's just about using the best infrastructure economically in terms of capabilities, and the challenge is in really managing these. In edge compute scenarios, there's always a back end. Part of the back end is there are companies that have an on-prem infrastructure, and they're still in the process of migrating to the cloud. We see that as a given but then how do you manage it? How can you deliver a harmonious experience for the user to be able to get both of these aspects going easily, rather than figuring out how to work with the cloud, work with your data center, work with your edge locations? From our perspectives, the data infrastructure should virtualize these to make it easy for a user to access data regardless of where it is.

Christophe: Cost is only one dimension of the equation; the other one is user experience, so we have to design the architecture and use the near edge, on prem or cloud to deliver the best user experience. This means that we are left trying to go where the data is. Very often what we will find when we design the architecture is that the best place to put the processing is where the data is – having the processing next to the data will often give you the best user experience.

Manish: What's your recommended takeaway for the GPU-tech community focused on building applications for connected everything with the scale in mind?

Alexandra: Consider the infrastructure implications – make sure it's built on a scalable and performing platform.

Christophe: Decide what your architecture is going to be and where you will be deploying the solution upfront and design your full solution from the ground up so that it will be able to scale between the edge and the cloud.

Sunil: Keep the scale in mind.

Dean: Look at the platform as a whole from the beginning and lay out what the scale, security and other needs are.

For more from our conversation, check out the panel discussion here.