TRANSCRIPT

DELL - Ask the Experts Q&A

EVENT DATE/TIME: July 17, 2023 / 1:45PM CT

DELL Technologies

CORPORATE PARTICIPANTS

Jeff Boudreau *Dell Technologies Inc. - President of Infrastructure Solutions Group*

Jeffrey W. Clarke *Dell Technologies Inc. - Co-COO & Vice Chairman*

John Joseph Roese *Dell Technologies Inc. - CTO and SVP*

CONFERENCE CALL PARTICIPANTS

Aaron Christopher Rakers *Wells Fargo Securities, LLC, Research Division - MD of IT Hardware & Networking Equipment and Senior Equity Analyst*

Asiya Merchant *Citigroup Inc., Research Division - Analyst*

David Vogt *UBS Investment Bank, Research Division - Analyst*

Erik William Richard Woodring *Morgan Stanley, Research Division - Research Associate*

Michael Ng *Goldman Sachs Group, Inc., Research Division - Research Analyst*

Samik Chatterjee *JPMorgan Chase & Co, Research Division - Analyst*

Shek Ming Ho *Deutsche Bank AG, Research Division - Director & Senior Analyst*

Simon Matthew Leopold *Raymond James & Associates, Inc., Research Division - Research Analyst*

Steven Bryant Fox *Fox Advisors LLC - Founder & CEO*

Wamsi Mohan *BofA Securities, Research Division - MD in Americas Equity Research*

PRESENTATION

Operator

Good day, and welcome to the Dell Technologies "Ask the Experts Anything" Call. Today's conference is being recorded. At this time, I would like to turn the conference over to Paul Franz. Please go ahead, sir.

Paul Frantz

Hello, everyone, and welcome to today's "Ask the Experts" Technology Q&A. We wanted to spend the next 45 minutes outside our typical financial cadence and focused purely on technology. The major technology trends we're following, our strategy and how we're innovating to support our customers. We plan for this to be the first of several conversations with our Dell leaders to stay tuned for future sessions.

Today with me are Jeff Clarke, our Vice Chairman and Co-Chief Operating Officer; John Roese, our Global Chief Technology Officer; and Jeff Boudreau, President of our Infrastructure Solutions Group. I'll start by sharing our safe harbor statement. Dell Technologies statements related to future results and events are forward-looking statements based on the company's current expectations. Actual results and events could differ materially due to a number of risks and uncertainties, including those disclosed in our SEC filings. Dell Technologies assumes no obligation to update its forward-looking statements. Before we go to Q&A, I'll turn it over to Jeff to share a few thoughts.

Jeffrey W. Clarke - *Dell Technologies Inc. - Co-COO & Vice Chairman*

Thank you, Paul. Happy to be here to talk about technology with you all. We can talk about a number of technology trends reshaping the landscape like data continues to explode and will define our world, multi-cloud by design, taking the right cloud for workloads, cost

optimization, complexity reduction and operational ease. Zero Trust is the needed security shift that puts the good guys back in control. Compute and storage resources are moving closer to where the data is created and modernization, virtualization and containerization of the telecom stack. So I suspect most of our time will be spent talking about the most talked about technology today, generative AI and rightly so.

Gen AI is an inflection. It's disruptive and it's game changing. What an exciting opportunity. Let's widen the aperture for a moment. There are several types of AI, including machine learning, deep learning, computer vision and generative AI, all are growing with new and exciting use cases like computer vision for manufacturing lines, machine learning for supply chain and chatbots that enable greater customer satisfaction. Today's $30 billion AI TAM is primarily driven by the first three. We view Gen AI as a new category of computing in the technology stack has very distinct characteristics. Gen AI doesn't replace anything, and it grows the IT TAM over time. It expands the use of computational machines to open up new ways of automation. The compute will be fed by massive unstructured and object storage infrastructure driven by more and more data. And the models created by Gen AI will need to be protected as some of the most valuable data ever created.

Generative AI powered by LMMs, large language models, has dominated the AI narrative. These are amazing models generalized to answer any question. They are trained over very expansive data sets. The data set can be a one form like text or many forms text, pictures, audio source code, et cetera. The utility of these models with hundreds of billions of parameters is wide in general. They require massive scale, immense computational density to train them.

From our customer discussions over the last 6 months, four generative AI use cases have consistently risen to the top of [CEO's] list. Customer operations, content creation and management, software development and sales. What we are seeing is customers wanting to use their data, processes and business context to train the model. As a result, we see generative AI having two distinct variants. The first being massive, multifunctional generalized models, GPT4 [bar], ChatGPT as examples and domain-specific like [Galaga], PubMed GPT or Dramatron or enterprise-specific models like stable diffusion, Bloomberg GPT and Code Gen.

What's the difference you might ask? First, the core neural network-based language model is essentially the same. The difference lies in the number of parameters and the number of expert systems focused on unique functionality like common sense reasoning, translations, pattern recognition, reading comprehension and code completion. Massive multifunctional generalized models have many more parameters and expert systems. They are trained on a broad and large volumes of data. Domain and enterprise-specific Gen AI use smaller proprietary data sets with fewer parameters and expert systems. These models don't require a massive scale or cost. And in most cases, they can be trained up with a small cluster of servers with GPUs and perform inference on the edge with a single server.

For example, an enterprise-specific model built with open source [Falcon] with 7 billion parameters, on your data can be trained with 4 Dell PowerEdge XC9680s. And once the model is trained, [inference] can be performed on a single Dell PowerEdge 760 XA. To use [Jensen's] characterization, we will have AI factories at the Edge, in the data center and in the cloud. In other words, everywhere, resulting in a range of Gen AI infrastructure solutions. Dell is uniquely positioned with our broad portfolio and services to win in these new categories of computing by helping customer size, characterize and build the Gen AI solution that meets their performance, cost and security requirements. We are very early in the cycle, not dissimilar than 24 years ago with server virtualization or as far back as 40 years ago with the introduction of the PC, disruptive, game-changing, more to learn, more to come. With that, we'll take your questions.

QUESTIONS AND ANSWERS

Paul Frantz

Thanks, Jeff. We'll ask each participant to ask one question to allow us to get to as many people as possible. Let's go with the first

question.

Operator

(Operator Instructions) And our first question is going to come from Wamsi Mohan, please go ahead.

Wamsi Mohan - *BofA Securities, Research Division - MD in Americas Equity Research*

I guess the question I have is when you think about the technology differences up from Gen AI, it sounded like, Jeff, you're saying that TAM is all incremental. Can you talk about how you would characterize server configuration for Gen AI versus a typical industry standard server that you sell? Both in technology terms and in ASP terms. And I'd also be curious if you think about how the evolution of the interconnect is going to be, for now, seems like [InfiniBand] is fairly dominant. Do you have a view on whether that stays dominant or if it -- Ethernet starts to get incremental traction as well?

Jeffrey W. Clarke - *Dell Technologies Inc. - Co-COO & Vice Chairman*

Thanks, Wamsi. Let me take a run at it, and John and Jeff can certainly fill any holes. I mean, first of all, we believe that TAM is expansive. This is game-changing and disruptive, and we just think it's an opportunity with what this enables. I called it in an inflection point in my talking point, This is really a fundamental change. From how we develop products to how work is done. And with that, we just believe this is incredibly -- an incredible opportunity for companies to reinvent themselves. When I think about specifically your question, what does the average configuration look like? There's actually a correlation of model size to how much memory you need. In your model (inaudible) memory. And we've modeled that characterized to have a sizing capability to help our customers through this. But if you're running a $7 billion parameter model, 20 billion parameter model, you can almost double the amount of do twice as much memory to be able to run the model effectively.

When you look at the ability to run the actual computational side and the configuration today with our 9680 with its H100s are clustered together for groupings of 8, allow you to extend the ASP of our products quite substantially. We think AI drives richer configuration. By the way, all the way down to the PC, doing more through PC drives a richer configuration. So we don't see any scenario that running these algorithms, training these algorithms, using your data to be able to build your models and then ultimately do conference out wherever you deploy it doesn't drive richer configurations and ultimately richer PCs.

In terms of ASPs, they're significantly greater than the average server, just the content of memory and the content of the compute intensity drives that. Jeff, John, anything you would add?

Jeff Boudreau - *Dell Technologies Inc. - President of Infrastructure Solutions Group*

No, Jeff. I think you're characterizing the TAM, it's just broad in general. I mean, I'm going to just add on the second half of that question, if you want, on interconnect. We'd like to talk about compute and we definitely like to talk about storage as components. Obviously, the network broadly in the AI world also gets disrupted to some degree. When you think about a training infrastructure, it's not a singular node. It's a cluster of nodes we're seeing in data centers, both InfiniBand and Ethernet. That's an ongoing -- two technologies competing with each other healthily, and we're seeing them both advance and we play both sides of that equation.

And the reality is we will continue to see demand as computing performance goes up for a node, the I/O path to and from those computing nodes within the cluster will go up. Another dimension of the interconnect though, is the storage interconnect because it's not just the nodes in the neural net talking to each other, it's how fast you can feed the data into those environments. So this is an

area where high-performance storage architectures that we happen to be quite good at are important, having gigabit links from the storage environment into the training cluster is as important as the train cluster itself being able to shuffle data between nodes.

And then kind of a third dimension of interconnect is the IO path into the inferencing infrastructure, which that has profound impact because if we're talking about processing imagery or media, the IO path from the sensor, a camera or otherwise across a 5G network or across an SD-WAN or even within an enterprise environment, also it gets essentially higher utilization. So in general, storage, compute and networking all work in concert to build out these next-gen AI architectures and from our perspective, there isn't necessarily a particular winner around just Ethernet or InfiniBand, there's going to be competition there.

On the storage side, [Fiber Channel] still plays a role along with InfiniBand and Ethernet connectivity. And again, what matters is how performant your storage systems are so they can actually fill those pipes. And then thirdly, broadly in the inferencing, it's going to require relatively high performance from any place where data is created through the place of inference which means that things like our Edge servers, if you didn't notice, are generally coming with a default higher performance [net], higher-performance Ethernet interfaces. And the reason for that is to be able to basically digest the stream of data that has to be inferenced at the Edge. So interconnect is a big deal. It's probably secondary to compute storage, but one of the three pillars.

Operator

And our next question will come from Simon Leopold from Raymond James.

Simon Matthew Leopold - *Raymond James & Associates, Inc., Research Division - Research Analyst*

Great. I guess I want to maybe get your perspective on how you expect the market to evolve and in that I'm imagining that much of the spending is biased towards hyperscalers building infrastructure and training AI clusters today. And that over time, we see essentially the uptake by enterprises, your customers to do more inferencing and to leverage their own data to utilize the systems being built by hyperscalers. And I guess I want to understand whether you think that's a rational way to look at the market or you believe the enterprises will build their own clusters. That's part one.

And part two is really if we're right about enterprise adoption coming later, how do you envision that evolution or timeline of enterprise spending on AI implementation?

Jeffrey W. Clarke - *Dell Technologies Inc. - Co-COO & Vice Chairman*

We'll see if I can unpack that Simon, there's a bit there. I'll take a swing out of here and see if I can answer your question specifically. I mean, for me, the big step back is this notion of inflection that I used earlier. And how game-changing this is and how it's really going to impact every sector and every type of function in every sector. And because of that, we're seeing CEOs of every size company have discussions of how to take advantage of Gen AI. And the reason for that, I think, are reasonably evident growth, it's still on the same page.

The first is it's going to change the way our work is done. Two, it's going to change the way we build products. Three, it's going to change the way we service customers. And probably, four, I think it's obvious, but I think it's worth saying there's a productivity and efficiency element of this. And it's why I believe it's fundamentally additive to our industry. I've been at this -- I think that's (inaudible) a long time, as you know, and how many times has the world dropped a 10%, 15%, 20% productivity improvement upon us. To my knowledge, that hasn't happened in my working lifetime. And that's what's here in front of us. And it's why it's the topic of every CEO and every Board.

And because of that, companies are trying to understand how to deploy the technology, how to better understand it, how to use it on their business context, how to apply it to their business models? I don't think that can be done in the broad general foundational models. Is there a lot of great work to be done, is there a lot of applicability without question, we'll continue to see that. But when we want to start talking about your product development community, how you address your specific customers, how to use your marketing data to serve your customers better, how to use your telemetry data to serve and they build a better service model, we believe that type of information is proprietary, unique to the business. And as a result, we see enterprises building out that capability.

Today, we have lots of pilots underway with medium-sized companies, large companies and enterprise scale companies. We see that continuing. What they're asking from us is help them size it, characterize it, get their data prepared and help them operationalize and build these AI factories that we refer to, that's the opportunity. How it's staging today, look, we're seeing a lot of work in the hyperscalers, the massive scale of the models that I described require that sort of scale. But that's not where we see the vast majority of the models being built and whether they're being deployed. John, Jeff, anything you would add to that?

Jeff Boudreau - *Dell Technologies Inc. - President of Infrastructure Solutions Group*

So for me, I guess, also other things that we've talked about in the past [where] our customers are telling us, I would say there's concerns around data privacy and data security in regards to some barriers of everything being in a hyperscaler cloud over time and why they need to bring [on-prem] or in their environment. So Jeff talked about the different use cases and the attributes and the resources that you need to serve those things from LLMs, yes, along those at the hyperscaling [now] because they need massive scale. As you kind of lean into some of these domain-specific areas of the [inferencing], you need a lawless resources to go do that.

And customers are looking at it because of some of the data privacy, the security. Jeff and I talked to you before, I did a lot of (inaudible) around physics and latency and bandwidth matter and the need for real time and [near time] speed is all critical, which actually means into having infrastructure on-prem for the enterprise use cases. That's a big thing that we hear from our customers as well as we go forward.

Jeffrey W. Clarke - *Dell Technologies Inc. - Co-COO & Vice Chairman*

Well, I think the other thing, John, you're a resident expert in this and Simon, maybe this helps build the case, we see the algorithms evolving with smaller data sets that are really honed for specific enterprises and specific, what we call domain and process-specific knowledge. And the fact that you're seeing decentralized AI occurring, which is there's a place where the model is trained, there's a place where the model is (inaudible) and there's where inference is done. Back in the day, that was all the same system, but we're seeing decentralized architectures accelerated. John, if I misspoke, you should correct me.

John Joseph Roese - *Dell Technologies Inc. - CTO and SVP*

No, no. I think it's important, we're getting in this moment in time where we see Gen AI as if it is the first time we've played with AI. And just to remind everybody, we were about to enter what many of us called wave 3 AI algorithms. The first-generation were basic machine learning algorithms, then we moved into the neural networks and deep neural network environments. And these -- what happened between Wave 1 and Wave 2 is things got bigger, meaning we saw this trend where you need a lot of compute capacity to build out a DNN and that was nice, but then the industry started to shift. We realized, hey, to make these commercially viable, we created technologies like transfer learning, which allowed you to take an existing model and only retrain one or two layers of it in the neural (inaudible).

We did things like Federated Learning and that allowed us to distribute the learning process across all the way out to Edge nodes.

Anyway, all that -- that was the narrative by the way, the day before the Gen AI world took off. It was basically a shift to rules engines, transfer learning, all kinds of things that would make the system more efficient and easier to deploy over a larger set of topologies, meaning small entities, et cetera, even out to a PC.

Gen AI popped up and interesting enough, based on the nature of the first generation of it, things like GPT3, GPT4, ChatGPT Bar, et cetera, the problems that we're going after were generalized applications. They are very large scale. And quite frankly, everybody kind of assumed that AI for all eternity was only large things, even though the day before, it was trending towards distribution and smaller -- so our prediction is that now that we have a baseline of the large-scale Gen AI models, which are very good for public generalized services, now we want to take them to the enterprise, more specialized narroscope distributed environments. We'll just begin that journey again. We'll start to figure out ways to make them more efficient, which is code for reduce the burden on infrastructure, reduce the power consumption, reduce the amount of data necessary, but still make them valuable.

And so we go through this periodically in the AI journey of jumping to a new order of magnitude and then figuring out how to optimize it so that we can run that thing not only in the privileged space of a couple of infrastructures, but anywhere the customer wants it. So that is likely going to be one of the big trends that we see as the market evolves, not just who adopts it but the composition of these systems inevitably will start to become more efficient, and that's code for being able to run it in more places and more diversity, which basically will catalyze the industry in our view.

Jeffrey W. Clarke - *Dell Technologies Inc. - Co-COO & Vice Chairman*

I hope that helps, Simon.

Operator

And our next question is going to come from Erik Woodring from Morgan Stanley.

Erik William Richard Woodring - *Morgan Stanley, Research Division - Research Associate*

Jeff, I just maybe want to take a comment you made earlier and maybe expand on it. And that was you obviously -- a lot of people have been spending time on the ISG side about this Gen AI opportunity. I'd love if you could maybe tease out some of your comments on the PC side and the opportunity for CSG to benefit. Will AI at the Edge kind of catalyze PC refreshes? How do we think about the timing of that? What's the incremental kind of componentry you might need to add to your products to handle these AI workloads? Just maybe a couple of incremental details that would help us understand the opportunity for Gen AI specifically in PCs.

Jeffrey W. Clarke - *Dell Technologies Inc. - Co-COO & Vice Chairman*

Sure. I'd be happy to, Erik. Look, any time that we can get a new technology that drives productivity into the best general purpose productivity device on the planet, it's -- we're better off. And when we look at what Microsoft's plans are with AI and future versions of Windows, it's doing just that. It's going to make the workforce more productive. And any time we've seen that with previous versions of Windows, it's driven a substantial refresh cycle. We think that's the opportunity here.

And what's equally important about this refresh cycle, you're going to ask your PC to do more with some form of assistant, some form of, I'd say, do some language modeling or language processing, it's going to do some machine learning with the capabilities that we'll put on our PCs in the future. And that's going to drive a higher ASP. You're going to need a more capable PC to ask it to do more. Then that's good for business. We've seen ASPs increase over the past 3 years. The likelihood that continues, asking more

of it is highly likely.

And then you have the subcategory in PCs, one of my personal favorites workstations, engineers, developers, creators, designers, data scientists, working at the Edge using those high-performance PCs with GPUs in them to do more complex AI tests. You'll see next-generation PCs with NPUs in them, neural processors. And those are going to allow us to do some of the basic neural processing that John referenced earlier, that will be in every PC going forward.

It's quite efficient way to do it, optimized for cost and power. We're pretty excited about the new PCs that we will be building on top of the embedded AI services and capabilities that we put into our service stack and our software stack already in our PCs today. Today, we do a lot of work around how to help customers optimize their performance by workloads, the telemetry systems that our service organization has tied into our PCs, we'll be able to extend that customer experience more broadly. So I know it was a long-winded answer. I hope that gave you some context of what we're thinking about, what we think the opportunity is and how we extend more AI to the Edge.

John, anything I missed from a core technology point of view at PC?

John Joseph Roese - *Dell Technologies Inc. - CTO and SVP*

The only thing I'd add is one of the challenges -- first of all, there's three different reasons why you need AI in a PC and you're probably -- you're familiar with them. I mean, one, obviously, is or do you need acceleration in a PC for AI. If you're developing models having a good high-performance workstation with GPU is pretty helpful. We know that with our precision [line], we're doing that for a long time. The second, though, is the proliferation of co-pilots, individually, each co-pilot you run probably doesn't need a massive accelerator. But if you think about the not-too-distant future, you're not running one co-pilot. You'll have a copilot doing transcription, a co-piloting translation, a copilot creating automated imagery, a copilot filling in the gaps and what you're talking about with contextual information.

And as you add more and more "AI workloads" to the system, the idea of having a portion of your semiconductor allocated to do that in an optimal fashion makes a lot of sense. And that's why we can see this inevitable trend towards within the CPU or in other types of accelerators, more and more of that functionality, which kind of brings us to the third area, which is really the user experience, in general, it gets transformed as we do this. We expect just more immersive user experiences on the PC because there's more parties involved in it. These co-pilots actually manifest. They show up in better imagery.

And so it's a pretty profound impact on the PC over the long term because it's the interface the customer has. And it's also the place where you localize many of these experiences around co-pilots. And the combination tells us we're going to need richer user experiences, spatial representation, just greater depth of field to be able to present the kind of data that we're at, and we're going to need both direct processing and co-pilot processing, if you will, as the number of kind of AI is working on your behalf around your increases.

And we don't know exactly what they're all going to be, but we know there's going to be many of them, and we know that the platform they're likely going to run on will be a distributed architecture, which the PC is the kind of personal representation for the user. So it's a journey, but we don't see any -- we don't see any other path other than more and more processing on the PC, more of it dedicated to AI-type tasks and a richer user experience, which you can imagine all of those things are pretty good for us.

Jeffrey W. Clarke - *Dell Technologies Inc. - Co-COO & Vice Chairman*

Thanks for the question, Erik. I hope that helps.

Operator

And our next question comes from (inaudible) from Citi.

Asiya Merchant *Citigroup Inc., Research Division - Analyst*

One of the questions that we get from clients is just as the spending is kind of being allocated towards AI, do you see that as temporarily perhaps taking funding away from a refresh, whether that's on the PCs or further dampening enterprise spending on servers and storage after a very strong calendar '22, if you may? Of course, we have the macro pressures, et cetera. So people are just trying to gauge if this temporary -- the spending on AI is sort of taking it away from some of the more spending that could have happened at least during the second half of calendar '23 on mainstream servers, storage and definitely as we look at the PC ahead of a Windows refresh.

Jeffrey W. Clarke - *Dell Technologies Inc. - Co-COO & Vice Chairman*

I'd probably look at it through that lens, it's a little longer term. I don't think this is at the expense of one or the other. This is additive. Again, if I think about what we're hearing in our discussions with customers before use cases that I described and how CEOs are viewing this in the boardroom, this is really a discussion where generative AI can change the basis of competition. It changes the way we're going to develop products and serve customers that drives a significant productivity increase that I mentioned earlier, but to restate it because I think it's worth restating, how many times have we in this working generation had a 15%, 20% productivity improvement? I can't think of one. So this is additive.

CEOs and boards are looking at this opportunity is not at the expense as that's the real productivity improvement that is out there, how can I not do this? And if I don't do this, I could be left behind, and if I'm the left behind, I may not catch up. This truly changes I think the basis of competition for many companies. It's going to disrupt cost structures. It's going to disrupt again how you serve your customer in a more intimate way if you can figure out how to get ahead of your competitors in any given sector, it's a huge advantage. So our experience with our customers and talking to CEOs and the market research we've done, suggests they're not thinking about, oh, I'm going to add AI and to not do this AI project. I'm going to actually extend PC lives by 6 months (inaudible) I have to invest.

This is that game changing. John uses the word inside our company as an inflection. This really changes the way to how technology is going to be used. And I think he's right. It certainly -- it's certainly in the discussions we're having really [our] leaders and I think how our company as a leader in our company, we're thinking about how do we take advantage of some of the coding assistance or just organizations looking at 20%, 30% productivity improvements based on the complexity of code. We think about how much work is actually language-based depending on whose research you look at anywhere from roughly 60% of all work is language-based. 60%-ish of that could be addressed with generative AI technologies. So when you think of it in that perspective, it's game changing.

And as a result, I know leaders like myself are going, we need to invest. We have to stay competitive. This changes the way we're going to deliver and build products for our customers and serve them. And I know that wasn't quite a direct answer, but it's not about the TAM of this quarter and what's going to happen. It's that much of game changer. And we think of this in terms of this is an industrial revolution. This is the steam engine. This is the assembly line. This is the Internet. This is what the PC was 40 years ago and what it did to productivity. And it's all happening and happening in a much, much faster rate.

Operator

And our next question is going to come from David Vogt from UBS.

David Vogt - *UBS Investment Bank, Research Division - Analyst*

Great. This is helpful. I want to pull back and maybe spend some time on storage. I know you talked about richer configurations on compute. And ultimately, richer configurations at the edge of the network, particularly at the CSG side. But what -- when I think about generative AI and the other flavors of AI that you touched on earlier, what does this mean for storage demand, particularly are we going to see new sort of demand for whether it's more basic levels of storage, that software defined, that's a little bit cheaper to deploy? Are we going to see more all-flash systems deployed? How do we think about the data that's going to be developed and used for inference as whether it's hot data, cold data? Just trying to think about all the different permutations of how this could play out as the enterprise starts to spend more aggressively on storage. And then ultimately, can this, from your perspective, be delivered as a service from the storage side? And is that part of the thinking going forward from Dell's perspective?

Jeffrey W. Clarke - *Dell Technologies Inc. - Co-COO & Vice Chairman*

Excellent question, David. We'll try to unpack it a little bit, I'm going to ask my resident storage expert to come in and help me in a minute. But if you step back and we look at what's happening, clearly, much of this data is unstructured. I mean it's going to come in, in a massive scale. So if we go back to one of the fundamental premises that's been driving the industry, the rate of data creation is not slowing, it's accelerating. It's absolutely the case. It's really accelerating outside of the walls of the data center, as we've talked about many times out on the Edge of the network and the form that it's taking is unstructured. And it's coming at massive scale.

And the types of systems that we're thinking about have to be able to scale that way. They have to be able to respond. And to your point, while we use that [position] inside our company, the hot data at the Edge being handled in real time. Doesn't lend itself to take a long trip up to some cloud and back down. It moves itself being treated right there. The triage has to happen with the data is created. So the algorithm is going to be run. That's where you're going to see some of the micro tuning being done, where you'll see drift detection being done and where we'll be modifying the inference as a result.

I mean, those things all play well to, I think, certainly our strategy, what Jeff and the team have been building in ISG and you think about old flash or we've done with (inaudible). You think about the scaled-out architecture that we've built with our unstructured products, we have a very broad and deep portfolio to meet the needs to where the data is going and the type of data that's being created. Jeff, I know you can make that sound a whole lot better than I did.

Jeff Boudreau - *Dell Technologies Inc. - President of Infrastructure Solutions Group*

Probably not, Jeff. But I'll chime in here a bit. So I think about data growth, I think the data gravity. And I think about kind of where we were and kind of where things are going with Gen AI and all things AI, which is all about data, right, ingesting data and then -- and making sense to that data. So -- and I think about infrastructure, probably going back to the last question, and it's the foundation for me. It's the foundation for all things AI. And it's really important to understand that while compute is at the center of most gen AI infrastructure, that compute will be fed by massive data sets and storage infrastructure. And I think that's why your question is so important and near to me.

In the models that are going to be created by Gen AI, it need to be stored and they're going to need to be protected because this is actually going to become some of the most important data the world's ever seen as we go forward. So it's much broader than compute kind of where you were going. AI will definitely drive demand across, I believe, all parts of infrastructure. It's going to be

compute, it's going to be storage. It's going to be data protection and it's going to be Edge where Jeff (inaudible) a minute ago. Sort of the networking where John was a few minutes before that. And it's even going to expand that client experience that we were talking on with the PCs as well.

I just think there's so much opportunity. Now specifically with storage, right now, the opportunity is both structured and unstructured and full transparency. I know people want to lean to the unstructured. Jeff is completely right. Unstructured is where the growth is coming from. I think of a parallel file system, think about object scale type performance. In the need where Jeff was before is latencies going to matter. (inaudible) box. And so making sure that we have either real-time or near real-time insight is going to be critical. So leveraging things like flash is going to be important as we go further and further into the future.

And by the way, not just flash, other media -- media and network and protocol opportunities as well. I would say also software defined, we nailed it, I think it's going to be in the future, right? A lot of times today, we have purpose-built systems targeted as -- targeted at opportunities, but I think software [defined]. If you think about the massive scale where things are going, especially in the unstructured space, I think software defined is really -- it's going to be at a point where we go from purpose-built into the software-defined world more and more every day. And so customers can scale with their data sets as they scale as well.

And lastly, yes, I do believe it can be as a service as well. So I think data-as-a-service is definitely a right opportunity for us as we go forward.

Operator

Our next question is going to come from Steven Fox from Fox Advisors.

Steven Bryant Fox - *Fox Advisors LLC - Founder & CEO*

Jeff, the company has obviously had a road map to deploy technology on the Edge for a while. You mentioned how AI at the Edge is going to become critically important. And you also touched on manufacturing where obviously, (inaudible) knows a lot. I was wondering if you could sort of pull those intersections together and talk about how AI and your products are going to play in manufacturing and how you envision manufacturing changing with the deployment of AI.

Jeffrey W. Clarke - *Dell Technologies Inc. - Co-COO & Vice Chairman*

Sure. I mean a lot of our work since a year ago with our streaming data platform, some of the early partnerships that we've built in Jeff's Edge organization have all been manufacturing based around how to build modern manufacturing facilities, how to do visioning, how to do higher yields and things of that nature. I think it's absolutely essential to what we're doing out on the Edge and the data is key.

What kind of use models can you imagine happening? We can do preventative maintenance schedules. We can help with production planning. We can think about how to do forecasting, visual inspection, quality management, how to increase labor productivity, obviously, health and safety inside facilities, I think are all aided by AI use cases and that's on the factory floor and with our Edge platform and what we've been doing there. The fact that most of that data tends to be unstructured, that makes the point that we talked about a little bit earlier. I think that kind of gets around your question, Steven. If not, please ask again or maybe more clarification.

Steven Bryant Fox - *Fox Advisors LLC - Founder & CEO*

That was helpful. I was just trying to bring some of the points together since you touched on a bunch of them. But just the one thing you left out is sort of how you envision this playing out over months and years in the future. Like how close are we to seeing some of the more advanced uses for, say, forecasting and preventative maintenance and things like that?

John Joseph Roese - *Dell Technologies Inc. - CTO and SVP*

Yes. Maybe I can jump in a little bit. There's a really important intersection. Jeff just listed a whole host of AI use cases in a factory. Now the question is, do you want each of those to run on its own discrete infrastructure? Or do you need to build an Edge platform so that in the multi-cloud world, those are all just software assets. And so the reason, if you look at our Edge strategy, I mean, it's got 3 layers, the foundational layer is the hardware. We launched all kinds of new edge servers and you'll notice something about them, that most of them have more accelerators than CPUs. And there's a reason for that because we expect them to be the landing point for lots of AI processing.

The second layer, though, is the native Edge announcement that we made, which basically says we really need to separate the logical and physical edges. We need to have a physical Edge platform, which is the capacity pool where things like Zero Trust and Zero Touch and the base level of capacity lives but we need to treat the Edge workloads, an image recognition package that's going to monitor an assembly line for quality assurance issues, or a quality assurance mechanism that's going to look at sensor feedback on the voltage level of the production systems themselves as just software packages. They're containerized code running on a platform.

And the trick is with native Edge, which is the uniqueness about it versus other offerings out there is we've turned in horizontal. We can orchestrate and deliver code from whichever clouds and upstream services you want as containers on that common platform. And in manufacturing, that's really one of the first places where that materialized because the diversity of digitization that's going on in the factories is just spectacular. Everything from HVAC monitoring to power conditioning to visual inspection of the production systems, these are all what we'll call, apps that live out in that environment.

Some of them are connected to public cloud tool chains, many of them are delivered by industrial companies. We work with all of them. And most of those companies are working with us and others to refactor their code, to run this containerized code. And we see this convergence that the digital factory of the future, yes, is heavily AI-powered, but more importantly, it has the constraints of being in the real world. So it needs to be on a highly efficient platform, which is not just the hardware, but the ability for you to kind of do whatever you want in the AI world as just a software function which really lends itself nicely to the native Edge story.

So we see this intersection between the digitization of the factory, the need for an Edge platform architecture as opposed to a whole bunch of mono edges and a new cast of hardware platforms that are actually optimized to run AI workloads as their default behavior. And those are kind of the tick marks of what we did with our ecosystem, what we did with native Edge, what we did with the earlier announcements where we launched a whole bunch of new Edge platforms that were optimized for this. So it's a very important space, and it is a leading indicator. From an AI consumption perspective, manufacturing is one of the first markets to move. They were doing it before Gen AI, and they're going to do it even faster after Gen AI, and I think we're pretty well positioned for that.

Jeffrey W. Clarke - *Dell Technologies Inc. - Co-COO & Vice Chairman*

Maybe a little self-serving into that, Steven, is the Dell supply chain has been digitizing for years now. We are using AI to do our production planning, we're using AI to do scenario planning, which is how we made it to the COVID and got faster decisions made up in realtime. This is what we're doing in our logistics network to improve our delivery accuracy to the hour of the day, which is going over well with our customers.

Operator

And our next caller is Samik Chatterjee.

Samik Chatterjee - *JPMorgan Chase & Co, Research Division - Analyst*

I guess on the ISG side, I had a couple of questions and firstly, I mean, you referred to this as well in your prepared remarks, power consumption of AI data centers is a major concern right now towards a large scale deployment. Just in terms of how you're thinking about Dell participating in that solution, how you're working with your suppliers and addressing that. Secondly, investors do want to see association with NVIDIA for most of the companies in the ecosystem. But outside of that, what are you seeing in terms of interest from enterprise customers and having a wider portfolio when it comes to like AMD or Intel?

What are you seeing in terms of engagement or customer willingness to sort of look at those, evaluate those solutions? And what's the current sort of breakdown of your portfolio on that front between NVIDIA-based servers versus AMD or other diversified suppliers?

Jeffrey W. Clarke - *Dell Technologies Inc. - Co-COO & Vice Chairman*

Sure. Maybe a couple of thoughts on power consumption, cooling and then the ecosystem (inaudible) clean up for me. It's why we believe designing these things from the ground up are important. We worked on the 9680 for years with NVIDIA. We were able to build a system that took advantage of our own iDRAC capabilities so we could drive power efficiency across all of the PCIe components. It allowed us to do -- put some pretty robust controls and that we can actually work through the Temp Transient during various AI workloads. We [can tune it] in other words. We've customized the air flow. We knew this was going to be a challenge at 700 a watts of throw times 8, there's a lot of energy being dissipated. We've designed systems that cannot be more predictive or we can, if you will, we can manage the acoustics as just getting a bunch of clients fans blowing across these things deafening people. So we've actually looked at the acoustic design along with the thermal.

We continue to look at new technology and Jeff's organization around liquid cooling and how we use cold plates on GPUs and CPUs and other PCIe components. It's how do we build different standardization around the cooling interconnect. So the systems are efficient in their heat transfer. We're looking at new technologies and their cooling -- memory cooling, some pretty advanced engineering that goes along in Jeff and John's organizations that allow us to really think how we provide a system that could be cooled on the score that it goes into the data center.

We know these are computation intense systems, and we've designed them accordingly with a lot of forethought. In terms of our customers asking for alternatives, we'd love to see a rich and vibrant ecosystem of AI accelerators and NPUs. We will, they're under development. Clearly, NVIDIA has a lead. It's a wonderfully capable performing product and certainly has an investor's attention. And it has the industry's attention, which means people are trying to develop alternatives. John's team is engaged with -- well, correct me because I'm sure I'll misremember the number, over 50 different silicon companies developing purpose-built accelerators and we see the trend going from general purpose accelerators to purpose-built accelerators.

We see some of the folks trying to use int4 and int8 sort of algorithms to actually simplify the calculations doing faster without sacrificing accuracy. So there is a ton of fascinating architectural work on the table and the broad supply base even as much as taking some of these simple algorithms and printing them on silicon, if you will, making them incredibly efficient from a power point of view. Or mapping this, John and the team continue to work through this. Jeff's team, all of the engineering accolades that I talked about are what we're

building these optimized systems. It's why we don't think everybody can build these. It's why we're selective with our choices. And it's why when we obviously put our Dell brand on it, we believe at scale and why we built the services around this, so we can deploy multi-clustered nodes to help enterprises deploy AI workloads. And they will be reliable and they will work. Jeff, I gave in (inaudible) so if I missed something, fill-in, please.

Jeff Boudreau - *Dell Technologies Inc. - President of Infrastructure Solutions Group*

Actually, I think you did a good job, but in the spirit of the kind of the short term, you talked about both what we've done around air cooling, but also around direct liquid cooling. But we're also working on a power size for our customers and our partners. So actually, they can go to our customers and actually see what the most efficient, most effective, most sustainable infrastructure can be for what they're trying to deploy. So if it's something large or something much smaller, they can actually lean in to the right architecture and technology to support their needs.

Jeffrey W. Clarke - *Dell Technologies Inc. - Co-COO & Vice Chairman*

I know we're running a little late. That's okay. Paul, sent us a note, we're going to stay a couple extra minutes. So next question, please.

Operator

Our next question is going to come from Sidney Ho, Deutsche Bank.

Sidney Ho - *Deutsche Bank AG, Research Division - Director & Senior Analyst*

Great. Thanks for taking -- doing the call and taking the questions. I think we can all appreciate why customers wanting to deploy AI capabilities on-prem and set up to the public cloud. But maybe you can touch on that a little bit. But for on-prem or maybe Tier 2 cloud, what are some of the reasons your customers buy from Dell instead of directly from the GPU supplier or from some other hardware suppliers? In other words, what does Dell offer differentiate from competitors? And along the same lines, how much customization are your customers asking for? And what are the opportunities to generate ongoing [revenue] from the same customers?

Jeffrey W. Clarke - *Dell Technologies Inc. - Co-COO & Vice Chairman*

That's a lot of questions. Let's work our way through that. I would start with -- the systems we built are built from ground up with our technology partners to be able to be deployed at scale, sort of the ending point of the previous question. So we've designed all sorts of thermal characteristics into these products, how to dissipate the heat that we were just talking about, performance attributes, how these can be managed. We are experts in clustered systems, multi-node systems. That's what we build, have been building for a very long time to be able to drive some of these workloads or the example that I gave in our opening remarks of a cluster of [490s] 680s, which would have 32 GPUs in it we interconnected, to do that is quite complicated. We've engineered that, in that case, working with NVIDIA.

And equally important, what I think is a tremendous amount of value add is on our services layer, where we have professional services, consulting services that allow us to help customers through their challenges today. I think about that as sort of the following way, if you will, how do we help simplify the generative AI design so they can get these things deployed and put in at scale. At the minimum, we have to scale the compute from 1 to 64 nodes, Jeff and his experts, these things scale, we need to be able to scale from terabyte systems to petabyte systems. We need to have the bandwidth to be able to scale across that vast CPU network and the multi-cloud

cluster management that goes along with it.

And then we ultimately are trying to help our customers work through the entire AI life cycle. So how do they do inference training machine learning Ops, how do we help them with the software, the hardware, the support and service that goes along with it, all the way to the point where we actually may provide a managed service to help our customers. Jeff, John and I were actually talking this morning, we were talking about an AI APEX offer, where you can imagine we extend this as a managed service.

So I think that (inaudible) examples gives the breadth of capability our company has from purpose-built portfolio of accelerators to the largest storage portfolio in the marketplace, to the understanding of how to build multi-clustered systems, to the service capabilities, professional service and consulting service that goes along with it and the fact that our AI strategy has been around for a bit of time when we talked about artificial intelligence in our products, on our products, helping our customers deploy them for inside our company, what we're doing in the partner network. And we have the resources behind it and the investment behind it to bring it to fruition.

That's how I'd answer that question, Sidney. I mean, Jeff and John, if I gapped something I know you guys will fill it in.

John Joseph Roese - *Dell Technologies Inc. - CTO and SVP*

Yes. I think one -- easy is important for customers because this is hard. They need -- the amount of -- there are very few customers that have all the capability to do an AI project all by themselves at the component level. That is not even a logical place to start. And so obviously, we know our friends in the hyperscaler world have a focus on one platform to work with, and they get an advantage in some cases, as being easier. When you get to the non-hyperscalers, we're fairly unique in the sense that, relatively speaking, we are as easy to work with in the sense that we can address the entire system. We can actually codevelop with the customer, we can deliver it as a service. And there is no AI project in the world that is based on a single piece of technology. It just doesn't exist.

It's a storage and networking, a compute problem, it's an integration problem. By the way, it's also a security and a trust problem. If you're going to implement AI in critical infrastructure, for instance, and you're going to use it to control the power systems, that control the power grid, guess what, to run it on it has to meet certain security specifications. It has to be able to operate potentially in a Zero Trust environment, which we just launched Project Fort Zero to go address that issue.

And so we find ourselves not only having the breadth of technology that is equivalent to almost anybody else in the industry and definitely bigger than any other non-cloud service, but we also have the ability to deploy that technology in almost any topology the customer wants. Remember, we're not anti-cloud. We worked with the cloud. In fact, Jeff is building a lot of software-defined offerings that sit in the public cloud. If you want to do it, they're great, we can help you do that. If you want to do it on-prem. We definitely can do that. If you want to do it at the Edge, we can do that. And more importantly, if you want to do it in a multi-cloud hybrid system, we're almost unique in being able to do that.

And if you follow the narrative, almost every large-scale AI system in the world is trending towards becoming a distributed architecture that will be hybridized, inferencing at the end, trading in the core, and that puts us in a pretty strong position, which I think customers see that. They don't want to deal with 1,000 companies. They want to deal with an expert that can actually address real world AI. And not only do we have the product (inaudible) and the one (inaudible) to [choke], but we also have this multi-cloud strategy and the ability to be essentially able to exist in whichever topology you want that piece of the AI system to work in.

So I think we have pretty good assets there. And we do struggle to find traditional competitors that can do that. That isn't really something that most of our traditional legacy competitors do. They're most of them are about a single product or a single part of the solution, which I think gives us an advantage over the traditional competitors, and our openness gives us an advantage potentially over even the hyperscalers in terms of how to navigate this.

Paul Frantz

We're going to take 2 more questions here.

Operator

And our next question is going to come from Aaron Rakers from Wells Fargo.

Aaron Christopher Rakers - *Wells Fargo Securities, LLC, Research Division - MD of IT Hardware & Networking Equipment and Senior Equity Analyst*

I think this question is going to build off of the prior question a little bit is that the complexity involved and Dell's expertise. One of the areas, I think, a month or so ago, you announced Project Helix with NVIDIA. And part of that stack strategy was leveraging the AI enterprise software suite that NVIDIA offers. And really, I guess my question is, as you're engaging with enterprises, given their own expertise, is a software layer for AI embedded in enterprise infrastructure is a requirement? Are they needing a layer-like enterprise AI software suite from NVIDIA? Or are there alternatives in terms of how they're developing their own internal AI strategies?

John Joseph Roese - *Dell Technologies Inc. - CTO and SVP*

I'll take a stab at that. No. There isn't a universal software layer that's going to materialize. AI is a very diverse area where there are many, many different use cases. Like by building out an AI system to automate imagery for quality insurance in my factory, the tool chain I use is likely the software that I'm going to use to do that even the models I use are very likely going to come from kind of industrial-centric OEMs and partners, people that we work with in that space. If you then pivot over to building a chat bot, which is very popular right now, as you know, with the large language models, you'll go and find the best tool trying to do that.

Now the NVIDIA software is fantastic because what they've done is they've created a collection of base models to address a number of use cases and we think that in terms of speed to execution, the advantage the customer gets by maybe starting their chatbot design with NVIDIA is that they get a ready-made model, they get a turnkey architecture, they get a system that can actually accelerate it. They can do it on their data in their data center, under their control. So they avoid a lot of the regulatory and compliance obstacles. And that's great.

But if you take that exact same architecture and went after some very specialized industry-specific AI model, the NVIDIA software might not be the best choice. Maybe it is, maybe it isn't, but we expect it not to normalize to one kind of monoculture about all AI projects are delivered via the same software framework just like they're not delivered by the same model. In fact, we love the fact that we're seeing incredible expansion of the number of large language models that are available because each of them do things in different ways. Some are more efficient, some are more performance, some are optimized for multimodal versus single modal.

And so I think our view pragmatically is we need to have things to get people started. And absolutely, the NVIDIA software stack does that and it does it for a bunch of very important and useful use cases. But over the long term, that's one of many tools that the customer is going to use to make sure that they can execute their AI projects across their very diverse set of functions in their enterprise. The interesting thing is even though there's that diversity of the software layer, they got to run it on their infrastructure, and it would be -- and it's definitely in their advantage for that infrastructure to be highly reusable, standardized and common, which is really where we play.

So software complexity will probably continue to be high for a reason because of the diverse set of use cases. Hopefully, we can normalize infrastructure complexity and make it that simple, but we don't expect there to be one master software suite for all things AI anytime soon. Even though NVIDIA is a fantastic way to get started, and it will address many of the use cases in a very easy way for customers so they can move fast.

Operator

And we'll take our last question from Mike Ng from Goldman Sachs.

Michael Ng - *Goldman Sachs Group, Inc., Research Division - Research Analyst*

I was just wondering if you could talk a little bit more about Dell's go-to-market for generative AI from a product perspective. it sounds like it's leaning more on probably project Helix today and then over time, it will be Power Edge servers with other types of compute. What are you going to do from a networking perspective? Are there any gaps in the current portfolio that you need to fill with other partnerships or through incremental R&D and proprietary networking?

Jeffrey W. Clarke - *Dell Technologies Inc. - Co-COO & Vice Chairman*

Yes, sure, Mike. I'll start and Jeff can come in at the end here. But look, our AI offer doesn't start with Project Helix. Our AI offer starts with a broad storage portfolio that runs at massive scale, scale-out architecture, particularly structured and unstructured data as we talked about, the data protection portfolio that can protect those valuable models, and then we can scale that out to the edge. And then we purpose built on our last generation of servers, our sixth generation of servers, a purpose-built AI set of servers. We talk a lot about the XE9680, but let's not forget about the XE9640 and the XE8640 and the R760xa, great inference machine. And you'll see us continue to build a broader set of AI offers across Jeff's business. So storage, compute. John talked about partnering in all of the different fabrics that exist. We work with all of the different fabrics. Obviously, we know a little bit about fiber, know a little bit about most of the interconnections that interconnects our storage subsystems.

We'll continue -- we know how to cluster. We've been building multi-cluster designs for a long time. That's part of our offer. I think what's an equally important part of our offer, which we try to hit on is the service capabilities. So our ability to help customers ultimately size characterize what their AI needs are, how to help them with their data and get their data where it needs to be. What workloads can be accelerated and then ultimately delivering the service -- I love John's word earlier, easy.

Project Helix is an example in a particular type that makes it easier. It helps enterprises scale, design, build and deploy AI systems. Combination of our capabilities and NVIDIA's capabilities to be able to use customers' proprietary data to build their models. They look at what's happening in the open source world and how fast this is moving and how to tap those open source model communities with their libraries or data sets, the transformers that exist there that allow you to really take advantage of this capability very quickly. That's what we're trying to help our customers through, is great interest, the 4 use cases that I talked about in the opening comments aligned to our portfolio and helping them go fast and making it easy, I think, is a good way to, at least in my comments then. Jeff, anything you would add to that?

Jeff Boudreau - *Dell Technologies Inc. - President of Infrastructure Solutions Group*

You already -- you hit on our strategy with regards to AI on [4 and with], which I think is critical to everything you just said. If I think about modern AI stack, it really has 3 layers, right, which is an infrastructure layer that's both hardware software and OSs. There's a platform layer where a lot of the tool chains plug into and then there's application layer. I think that creates a tremendous opportunity for us to, I

guess, go to market and win in the hardware layer, the software layer and the services layer, where Jeff was a few moments ago. (inaudible) to add, Jeff.

Paul Frantz

We'd like to, again, thank Jeff and Jeff and John, and I appreciate the question, and that wraps up today's call. We'll see you next time for our Q2 earnings.

Operator

And this concludes today's call. Thank you for your participation. You may now disconnect.
