Dell Technologies World 2024

Tuesday, May 21, 2024

Day 2 Keynote Transcript

Jeff Clarke (00:02:06):

Good morning everyone. Well thank you. I need some energy this morning. Welcome to day two of Dell Technologies world, the day where we are going to take from yesterday's announcements and make real. We're excited to geek out with you a little bit today. We're going to talk about some architecture, some forward thoughts. We're going to talk about our products. We're going to talk about our solutions and how this all comes together from yesterday's discussions into what we believe is a defining moment for our company, something very special. We're going to talk about how we blueprint and this blueprint concept that you're going to see throughout the presentations today of what we've learned and how we can help you deploy AI and take advantage of this tremendous technology and opportunity that exists today. And since we were last together, it's been one hell of a year.

(00:03:01):

There's been no slowing down generative AI. In fact, I would argue it's only accelerated since we were last together. And while we're still in the early stages and still the very, very early stages of this, it is clear that this technology is very different than anything that I've experienced in my nearly four decades at this. Now it's happening faster, it's more disruptive. It's game changing. It's changing the basis of competition with unmanaged, unmanaged productivity and impact. The last time the world saw something like this that was so profound that had such a meaningful and lasting impact was nearly 300 years ago.

(00:03:49):

It sounded better. During rehearsal, the industrial revolution, the technology advances were iron. They brought new materials, iron and steel, we had new sources of energy, steam. All of that was combined with new innovative machines that powered a fundamental change in course in history. And over the course of about a hundred years, we transformed from a growing society and a handcrafted economy into one where people migrated to the cities dominated by factories and machines that improved the quality of goods and services lowered the cost significantly. It resulted in significant, I mean signifi, profound productivity improvements, had wide reaching social, economic and cultural changes across the globe. And today we're at the very beginning of a similar change, something that I believe is more profound and will have a long, longer and lasting impact, what we call the AI revolution. And it is happening orders of magnitude faster.

(00:04:59):

Think about it, A year ago we'd been talking about five days to 1 million chat GPT users. That's amazing and it's only accelerated since. And these new raw materials, we talked a little bit about it yesterday. The new raw materials are data and information. The new machines in this era are GPUs capable of massive parallel processing, performing trillions of floating point operations per second. That when coupled with high speed AI fabrics, high speed AI storage, the right models and the right data tools transforms that information and data into knowledge. Knowledge that we've never had before. And everything changes as a result.

(00:05:52):

Bringing us new assistants, agents, if you will, a coworker to everything that we do in the future, redefining customer engagement with new capabilities like autonomous quoting or autonomous sales, intelligent pricing, instant quoting. It's going to redefine how we build supply chains, creating more efficient supply chains with predictive inventory and replenishment, intelligent planning and forecasting, more robotics, more robotic automation, increasing the velocity through factories. And it's going to power all of us as customers with improved services like predictive maintenance, auto healing intelligence, self-serve, and various support agents that will have in all of the interactions across the products and goods that we will buy. And probably most profound for all of us as IT professionals, these generative AI workloads require a vastly different computing architecture. One optimized for vector math with high levels of parallelism, and you're going to hear this consistently throughout the presentations today with high-speed storage over a high throughput AI fabric to feed this generative AI workload with data.

(00:07:05):

Michael said it yesterday, feed the beast. Generative AI, these GPUs, devour data and we need to feed it at an incredible rate to get the outcomes that we're all looking for. And if you think about architecturally what's happening, Jensen touched upon this briefly yesterday. We're moving in this point in time from what has been historically a computational output told more of a cognitive response with context and reasoning. Think about that. We're going from programming and thinking instruction driven, computation driven into context and reasoning, and it's changing the way workflows will be done. We're moving from an instruction workflow into an intention driven workflow. Very profound. The capabilities that we're going to unleash over the remaining part of the decade. Our strategy in this regard is really, really simple, is to accelerate the adoption of AI. For all of you, it's built on five core beliefs.

(00:08:06):

One, data is the differentiator. That shouldn't surprise you. 83% of all data is on-prem. 50% of that data is generated at the edge. Number two, this results in AI moving to the data because it's more efficient, effective, and secure. Three, there is no one size fits all approach here. AI will be implemented in a wide range of ways from locally on devices and the edge all the way to massive hyperscale data centers. Four, And this one's I think very key with the rate of innovation that is happening, you're going to need an open modular architecture to support the rapid innovation. And then lastly, AI requires a broad and open ecosystem to take advantage of all of this rapid innovation and technical advancements that we've witnessed in just the last year. And that's going to continue. Most enterprises will not train their own large language models, but they will use open source models like Llama three, Mistral, and others to utilize GenAI in their businesses. And over time, we believe enterprises will do the five following things. They will use smaller open source models to better obtain performance and efficiency for building their expert systems. They will build these expert systems using a combination of open source software models, your data, and either fine tune or use techniques like retrieval, augmented generation to get the results that you're looking for. A fun one as an engineer is you're going to actually take these expert systems and you're going to optimize 'em over time with techniques of quantization to gain better performance and member utilization of your models and inference. And you're going to build these smart systems, these expert systems, and keep it fresh with data. We're going to build intelligent data pipelines to ensure that the AI is always prime with the correct and freshest data, the most comprehensive data set available when you're doing this work.

(00:10:23):

And then lastly, inference is going to be performed in many, many places. On notebooks, on servers, in the data center, out on the edge, essentially anywhere you're trying to get an AI guided outcome. I hope it's obvious by now with yesterday's announcements are discussions throughout the day kicking off this morning. The traditional data center architecture is ill-equipped for these generative AI workloads. A whole new computing architecture has emerged. We're going to deliver that computing, that new computing architecture through AI factories that you heard about yesterday. Our delivery of this new computing architecture to advance artificial intelligence throughout our customer base is delivered through an AI factory. What's an AI factory? It's not dissimilar than any other factory. It takes raw materials, as I mentioned earlier, your data and it turns it into something useful. Insights, knowledge, tokens, as we heard yesterday. Michael hit this point yesterday, which I think is absolutely essential to what we're seeing.

(00:11:40):

AI factories are fueled by your data and it will be assisted by professional services to help prepare that data to be synthesized efficiently by an open ecosystem running on an open modular infrastructure that is built for a variety of use cases that produces, and I think this is really profound here, produces institutional insights that you've never had before. You're going to gain a wisdom, an understanding about your business that you've never had in your hands. That is the capability and potential that we see with AI as it gets deployed broadly across small businesses, medium businesses, to the largest multinational corporations around the globe. Our Dell AI factory comes in all shapes and sizes. It's not all one size fits all model you can expect us to scale. For example, it can be mobile, it could be a precision workstation running a train model, collecting data to produce insights or that same platform running a Llama three 8 billion parameter model with a multi quarry attention for scalability.

(00:12:45):

That's a tongue tie. Or a rag solution on a single PowerEdge R760xa server crunching quality control data and monitoring supply chain visibility at a remote warehouse. Or some of my favorite, the Dell AI factory could be and will be in the data center from an inference and cloud on a single PowerEdge xe9680 to a rack of xe9680s with 72 GPUs. That would be a rack of xe9680 with 72 Blackwell GPUs as we learned yesterday, or a data center with hundreds if not thousands of racks of 9680s acting as one single cognitive computer. And the best way to deploy this new architecture is to separate it from the traditional systems. You have to think of this very differently and optimize traditional data centers the way you've always had and think about building these in a very different way and optimizing for their specific workloads.

(00:13:53):

I think it's been clear by now, but AI workloads start with number one, need unbelievable compute intensity. So accelerated computing is the table stakes. Small accelerated compute to the largest systems that we can build. High speed IO optimized storage for file and object data types, high throughput, low latency networking, fabric four, which doesn't get a lot of discussion, but I think it's important we're creating the most valuable information in the world. The most valuable data is going to be trained and tuned models in the rag. You have to protect that. The data systems need to be integrated into a common data pipeline. And then ultimately we believe you need a broad answer that you have to have the PC to extend AI all the way out to the edge, out to a factory, a hospital, a smart city, what might be all powered by the set of software tools for training, tuning and inferencing.

(00:14:56):

And mind you, these are very, very complicated, highly technical systems. They require high levels of solution engineering and integration to increase the speed of deployment and time to first token. That's why we believe an AI factory and from Dell with the broadest portfolio in the marketplace today is the answer. We can move from the edge into the department, into the data center, into data centers, and span the engineering and solution engineering to deploy at scale and to really get the benefit of the dollars you're spending to maximize performance, so to speak, the return, a deep system level understanding is required to make this stuff work. It's not a opportunity to take apart from here and apart from there and apart from there and hook it in and hope it works. This has detailed what we call systems engineering, the ability to look at the CPU, the GPU, the MPU, the networking topology, the storage subsystems, the underlying drivers and low level software. Tune all of that factory to get the outcome that you're looking for. We believe we're uniquely positioned to do that. And if I think about all of what I just said, it even gets more exciting. As I think about what the world looks like going forward for the remainder part of the decade. My advice is you got to start getting ready now. Fasten the seatbelt strap on the helmet, whatever analogy you like because the right has just begun because here's what's coming.

(00:16:32):

The compute requirement for gen AI is expected to be 27 quetta FLOPS. I know, WTF is a quetta FLOPS? That's 30 zeros, that's a 27 with 30 zeros. That's up from 0.3 today. Two full orders of magnitude increased over the remaining part of the decade of what is needed to compute the given AI workloads we anticipate going forward. Generative AI data center demand will surpass traditional data center demand by 2026 and be 75% of all data center demand by the end of the decade. There will be a rapid shift from training to inferencing by 2030. Only 10% of the compute demand will be for training the rest for inferencing or as I like to call it, AI in production in use. Number four, which I think is insightful in the world that we live in today that we have to overcome. This is 390GW for AI data centers by 2030 plus the 130GW for traditional applications.

(00:17:40):

That's 520GW of power needed by the end of the decade. That's eight times greater than it's deployed today. And then lastly, one of my favorites is by 2030, the entire PC install base will have been refreshed with 2 billion AI PCs making the world's most useful productivity device even more powerful and more capable as we exit the decade. That's what's coming. This is the world we live in. We're excited about it. This is the brave new world as we say. I hope you can tell we're incredibly excited about it. We're ready to help you. And I think it's time that I take the next step and take that architectural fabric that I just talked about and turn it into something real. So I'm pleased to invite one of my favorite fellow engineers, one of my favorite people and my friend to talk more about high speed networking fabric. Charlie Kawwas, President of Broadcom.

Charlie Kawwas (00:18:51):

Thank you very much.

Jeff Clarke (00:18:52):

Welcome to Dell Technologies World.

Charlie Kawwas (00:18:54):

That's phenomenal. That's awesome. Thank you for having me here

Jeff Clarke (00:18:57):

Did you ever think going to engineering school, we'd be in front of talking to this many people?

Charlie Kawwas (00:19:01):

Oh no. This is incredible.

Jeff Clarke (00:19:03):

Yeah, it's something like that. It's not the word I would use, but something like that. I probably mentioned high speed networking, high throughput, low latency networking half a dozen times in my opening remarks. When you think about AI and networking and the fact that we have to synchronize all of these GPUs, what is the role that networking plays today and going forward?

Charlie Kawwas (00:19:27):

By the way, great opening remarks. One fundamental thing that I actually want to make sure we all level set on is networks in today's data centers are the bedrock of the data centers that we've built over the last three decades. And I think this is going to change fundamentally as we move into AI deployments, as you were just showing the AI factories, all of the insights that would come out of this in a couple of ways. One, we have the processors or the compute, which still needs to talk to the outside world to access data, provide insights, all of that is called front end networks and that's the bedrock of the data centers. But these GPUs, these TPUs, these XPUs, they need to talk to each other.

Jeff Clarke (00:20:20):

We have to synchronize them

Charlie Kawwas (00:20:21):

And we need to synchronize them, whether it's training or inference. And that is called the backend networks. And I'm going to say a little bit more things about this as we chat, but if you look at it for the AI workloads, the AI workloads are massive. They call 'em elephant workloads. And the reason they do this is because they actually need many GPUs to run on. And it's not eight, it's not 72, it's hundreds, thousands, hundreds of thousands. AI singularity, as you said, is actually viewed to be a million GPU, which is obviously a huge challenge for the industry to get to. But in order to deliver this, we need to figure out how these backend networks have to deliver on these requirements. And as we do this and as we chat about this, this is going to be the heart of the AI cluster that we both build because ultimately this is becoming a distributed computing architecture.

(00:21:21):

All these GPUs that need to talk have, which are the brains have to go through the heart of the backend. And that piece is super exciting for both of us for our partnership. One, you talked about an open ecosystem, I believe to scale up to that level, to have the brain and the heart, we got to have an open ecosystem. Two, it's got to scale. How do we go from thousands to tens of thousands to hundreds of thousands and three, good luck putting a hundred or 200 or 300,000 of these GPUs in a single data center. We need it to be power efficient. So I'm excited about the partnership we have.

Jeff Clarke (00:21:57):

We're too, if I take that analogy that the GPUs, the brain networking is the heart, I'd argue storage is the lungs pumping the data.

Charlie Kawwas (00:22:05):

Hey, I love that. I love that.

Jeff Clarke (00:22:09):

But we like that analogy to build around it. And if I think about the announcements from yesterday that Michael made, we had two noticeable announcement, our new power switch product that has Tomahawk in it. That's right. New Tomahawk five where we talk about 51 terabits of throughput. And then we talk about the new THOR 2 NIC, the 400 gig NIC that we can now continue to really drive optimization of the fabric. Tell us more about that and are you excited about those?

Charlie Kawwas (00:22:36):

Two words, great partnership first of all, thank you.

Jeff Clarke (00:22:39):

We appreciate it.

Charlie Kawwas (00:22:40):

We appreciate it as well. And to show everybody what we announced, I brought a few toys if you don't mind.

Jeff Clarke (00:22:46):

I noticed.

Charlie Kawwas (00:22:47):

And I actually would like everyone to know what these are and I don't think you've seen many chips today or maybe this week. So this is a 51 terabyte per second single dye. It's called monolithic dye that actually started production last year, late last year until today is the only chip in the world that actually can deliver 51 terabyte per second in a monolithic dye. This is really, really important. I'm going to tell you why, but first let me pass it to my friend Jeff. This is a five nanometer chip, which is one of the latest process nodes in semiconductors. As I said, it's a single dye and it goes into the power switch 9864, Dell's product that was announced yesterday. This technology actually delivers the lowest power because it's in five nanometer and it also delivers the lowest latency because it's a single chip versus other 51 terabytes per second use multi chipps. This is unique. We have a great product today with Dell that we announced and he's taken that that chip is thousands of dollars by the way. But I tell you what's unique about this chip is also the fact that it works hand in hand with the second announcement, which is THOR 2.

Jeff Clarke (00:24:08):

Did you bring one of those Charlie?

Charlie Kawwas (00:24:09):

I brought this of course. So this is our latest NIC card and it's actually the only NIC card in the industry today that delivers open 400 gigabits per second PCIE Gen five at five nanometers, the same thing as that other chip. And it goes into the PowerEdge Switch, xe9680, which is a great platform. And by the way, the only announcement in the world of this came yesterday from Dell, from Jeff yesterday. So thank you Jeff for the partnership on this.

Jeff Clarke (00:24:45):

We appreciate it. What's interesting, maybe connecting the dots from yesterday and then something I mentioned just moments ago, you talked about open and the value of open networking and building open. I talked about that. And then yesterday Michael announced the 9680 L with 12 PCIE Gen 5 slots, eight of 'em go to synchronize the GPUs, the other four to build incredible bandwidth for us to really bring data into the GPU, correct?

Charlie Kawwas (00:25:15):

Correct. And the cool thing about these two parts, if you don't mind, I'm going to ask you to carry both. So I want you to take it back out of your pocket, that tomahawk switch. I knew he was going to take it. So the tomahawk switch goes on that power switch rack that goes on top of that rack and it's used for the front end networks, but it's also used that same single platform we're building. It's also used for the backend networks that I talked about. And what happens, these NIC cards, 400 gig NIC cards actually connect directly to that switch. So they go hand in hand. And that today is only available actually together between Dell and Broadcom. The cool thing about it is not only the combination of this gives you the lowest power as you scale the number of GPUs that's important.

(00:26:04):

Not only gives you the lowest latency, but one of the coolest things about this is cabling and it is open in the sense that you now can work with Dell and Dell can source any cable that actually fits your AI factory. And the cool thing that we can do here is obviously it supports optical just like other solutions exist today, but that's expensive. This solution is the only solution that actually can support direct attach copper cable, which is the lowest cost. And because of the technologies amongst both of them, it goes as far as five meters, which means you can inside the whole rack, you don't need any digital signal processors, you don't need the timers, you don't need optical transceivers. Ultimately it gives you the lowest cost at the lowest latency in an open platform.

Jeff Clarke (00:26:55):

No, that is cool. Hey, one last question. Another announcement we made yesterday, and I think it kind of paints the picture of where things are going. We announced Enterprise SONiC, our smart fabric manager. I know you're a big proponent of open source. Where does networking go when we think about open source, open systems, open ecosystems, modular design, give the audience your thoughts on that.

Charlie Kawwas (00:27:19):

Actually very, very exciting and I have to tell you, we all of us are very lucky to be living in this inflection point. He's taken both of them now, this is the beginning of a whole set of new innovation, not just in semiconductors for us, but at the system level, at the software level and the AI factory. So one of the things that I can guarantee you over the next three years we're going to have bigger and meaner GPUs, more power, more performance. We're going to have a lot more number of GPUs as we build out these AI factories. And these are going to require innovation. And this is part of what we've done for decades between both companies we've innovated together in the server we've innovated together in the storage, and now we're innovating together in the networking side. So a couple of things I'd like to share, if that's okay.

(00:28:11):

One on the front end networking. It's going to stay ethernet. Ethernet is the defacto standard. Ain't going to change. If anything, we'll be able to run it even faster. We know how to do this. We have 400 gig today, we're going to 800 gig, 1.6 T, three point T, two T, easy. Now let's talk about the backend. Remember the backend is the heart, so it has two components. One is inside an AI node and that is called scale up. How do you scale up and connect 72 today or 64 GPUs going towards 256, 500 and plus? Well, there's actually two standard based technologies to do that. One is ethernet, which we know how to do, and I just showed you one of the products that can do that. Two, actually, we can actually do that using PCIE and today in the powered server that ships today we're shipping PCIE Gen 5.

(00:29:09):

By the end of this year we're going to actually start sampling gen 6. Next year we'll have gen 7 and the year after we'll have gen 8. So we're going to accelerate the rate of innovation in a standard platform with the ecosystem. This is not going to be done by Broadcom alone, it's not going to be done by Dell alone. It's going to be done together with other ecosystem members. Now let's talk about scaling up the cluster. We talked about scaling up the node. To scale up the cluster, we have to go to what we call scale out. And I am pleased to tell you for scale out the defacto standard, the winner is ethernet and that Tomahawk 5 chip is the chip that's used in hyperscalers today to scale up and scale out all of these platforms. And this is where I think we can drive more innovation.

(00:30:01):

As a matter of fact, I have the third toy if that's okay. So inside this toy, it's actually pretty cool stuff, is the Tomahawk 5 chip that Jeff and I showed you before, but you see eight tiles around it, 1, 2, 3, 4, 5, 6, 7, 8. Each of these tiles is a silicon photonics tile that is co packaged with that single dye that I showed you before. So that's unique. That's only available today actually from Broadcom and it's being used with two very large hyperscalers as a proof of concept this year going into production next year. But the cool thing about this is these eight tiles replace a hundred and twenty eight, four hundred gig optical transceivers. Let me just repeat that. 400 gig transceivers times 128 is replaced by these four that I can hold basically in one hand what this means, this means that the power of that PowerSwitch when we go to this technology will drop by 70%, seven zero.

Jeff Clarke (00:31:16):

Awesome,

Charlie Kawwas (00:31:18):

Thank you.

(00:31:22):

The latency because of the elimination of all these components and pieces from the server all the way to the switch will have the latency, which means we can run better training as well as inference models. And then lastly, we all care about cost. Cost will drop by at least 40% because of the elimination of all these transceivers. So with all of this coming together, Jeff is right, we need a network operating system to glue all of this together. And this is why we've talked about SONiC yesterday. We've been working on it. Actually I'm proud to say that the 10 data centers of Broadcom, our IT data centers run on Dell switches with SONiC network operating systems. So thank you actually for using our technology to help us build our data centers and our look to do the same with our future AI factories with them.

Jeff Clarke (00:32:19):

Awesome. Thanks Charlie. Thank you everybody. Charlie Kawwas, Broadcom. This one doesn't fit. So we're going to move on and we're going to talk about ISG. So we're moving from fabric now. We're going to get into the computational part. We're going to talk about models. We're going

(00:32:38):

To out this AI Factory in depth over the next several presentations. So with that, please welcome to his first ever Dell technology world main stage. Arthur Lewis, the president of our ISG organization.

Arthur Lewis (00:33:03):

Good morning everybody. How are you doing? It's great to be here with you in our AI factory. This morning I'm going to walk you through various components of our factory and then we have some really real cool demos to show you. It's exactly how we're making AI real for our customers. There's no question that we are living in one of the most interesting and exciting times in human history. For years, customers have been on a digital transformation journey, the underpinning of which has always been the data. It's all about the data. The rapid advancements that we see in artificial intelligence will not only accelerate the value that the world's data brings to organizations of all sizes, it will forever change the architecture of data centers and data flows. Silos of the past will be dismantled and everything will be connected. And with our AI factory, we sit at the very center of the AI revolution offering customers the broadest, the deepest, most optimized for AI portfolio in the industry, spanning the desktop to the data center, to the cloud, plus a growing and great ecosystem of partners, plus a full suite of professional and consulting services.

(00:34:24):

Our innovation engine, as you've seen over the last couple of days, is firing on all cylinders. We are running the engine very, very hot. The high level and fast pace of our innovation is being recognized. In March, Dell Technologies had the privilege of being the only OEM named as a leader in Forrester's AI Infrastructure Solutions Wave. Now I want to walk you through our framework and talk about some of the exciting new additions we have to our winning portfolio. And let's start with accelerated compute. We offer customers choice and flexibility ranging from the PowerEdge xe9680 to the four-way, PowerEdge xe9640 and the R750XA/R760XA. platforms. The 9680 is built to offer silicon diversity. It supports GPUs from Nvidia, from AMD and from Intel while also providing a consistent rack level design for operational simplicity.

(00:35:36):

It was built with networking in mind offering 10 PCIe lots to provide customers choice of ethernet or InfiniBand NIC and DPUs for maximum industry throughput. And we are not standing still. Accelerators are becoming more powerful. You heard this multiple times yesterday and customers need to be able to take advantage, full advantage of these new accelerators and it is in that vein that I'm incredibly excited to announce the newest edition to our family, the PowerEdge XE9680L specifically designed for liquid cooling. The first in instantiation of which will be based on NVIDIA's Blackwell 200 GPU.

(00:36:26):

Thank you. Several highlights to cover here. Number one, the density has been improved by 33%, offering eight GPUs in a 4U form factor. Second, leveraging our decades long leadership in liquid cooled technologies. 3rd, and perhaps most importantly, we have significantly increased the network capabilities of the 9680, offering 12 PCIe slots supporting full 400G Ethernet and InfiniBand NICs and DPUs. This will provide the highest level of throughput in the industry. But we didn't stop there to help customers with very large deployments, we are also excited to announce a set of rack scale solutions. These solutions will support air and liquid cooling. In addition, these solutions will be data center cooling neutral, reducing the vast amount of chillers that are needed, ready to deploy and factory integrated.

(00:37:48):

And these solutions will come in three variants. Number one, a 70KW air cooled design to support 64 GPUs including Nvidia H100, H200 and B100 AMDs MI300x and Intel Gaudi 3. Second and maybe one of the highlights of yesterday's keynote, A 100KW liquid cool design built to support 72 Blackwell 200 GPUs. And third, a 130KW design based on the next generation ORV3 21 inch architecture built specifically to support Nvidia Grace Blackwell 200 super chip, but also including x86 variants with Intel and AMD. And the fun and the innovation doesn't stop here. Let's switch over to networking. We've talked a lot about the importance of networking as GPUs become more powerful. The amount of data that is flowing from GPU to GPU and from server to storage is simply massive.

(00:39:06):

And to put some context around simply massive, AI workloads drive 300 times the amount of data throughput that we see in traditional compute 300 times the amount of data throughput. And as you heard just now, we have partnered with Broadcom to provide enhanced fabric capabilities and a full suite of switches and NICs including Tomahawk 5, which supports both 400 and 800 gig switching. THOR 2, a 400 gig NIC that is directly integrated into the 9680 as well as enhanced fabric capabilities with our SONiC operating system and the capabilities that we've built there over the last several years. We have also partnered with NVIDIA to build out a full sweep of switches, NICs and DPUs including Spectrum 4 for Ethernet and Quantum X800 for InfiniBand. Both of these supporting full 400 and 800 gig switching. CX7 to a 400 gig NIC also directly integrated into the 9680 as well as Spectrum X for enhanced fabric capabilities and to help our customers navigate all of these solutions, we have a full suite of professional and consulting services to help customers fine tune and optimize the network for AI.

(00:40:33):

And the fun and innovation doesn't there, it continues. Let's talk about storage. Our software-defined PowerScale was absolutely built with AI in mind, driving maximum speed from the GPU to our AI data platform, offering incredible performance with our GPU direct technology and the right mix of flexibility, scalability, and security. In addition, PowerScale is the very first ethernet storage to be certified on Nvidia super pod and we are making PowerScale even better with the introduction of the F910, a very dense high performant file solution for unstructured data. We have added significant hardware upgrades including DDR5 PCI Gen5, as well as 24 NVMe SSDs all in a 2U rack platform to provide density of up to 1.47 petabytes. With these hardware upgrades and significant software modifications, and with the ubiquity of AI, the demands on storage are only going to grow. GPUs are becoming much more powerful, which will require more data and more throughput, and it is in that vein that we're incredibly excited to announce Project Lightning.

(00:42:13):

I was waiting for that. Coming next year. Project Lightning is a game changing parallel file system built for unstructured storage specifically for AI. What do I mean by game changing? When we look at it versus our nearest flash only scale out competitors, we will see performance increases of 20x and throughput increases of 18.5x truly game changing performance and the innovation doesn't stop there, it continues. The rate and pace of innovation in AI is extremely challenging and customers are looking for simple proven solutions. We offer to help customers accelerate the adoption of AI. We offer more than 40 turnkey solutions tested and optimized against a myriad of AI use cases. Let me give you a couple of examples. As you heard from Michael and Jensen yesterday, it was only a few weeks ago at GTC that we announced in collaboration with Nvidia, the Dell AI Factory.

(00:43:21):

And in addition to collaborating with Nvidia on all things infrastructure, we are also working together with Nvidia to provide full stack deployment automation to hasten the setup of AI environments. This full stack deployment automation will deliver up to an 86% reduction in time to value. And when coupled with Nvidia inferencing microservices, we will see the time from delivery to actually running inferencing jobs reduced even further. We are also adding AMD MI300X single and multi-node validated designs specifically for inferencing. We are also working with Intel on collaborating on their Intel developer cloud to offer Gaudi 3 for flexible testing and reserved instances. We are working with Red Hat on their enterprise Linux for AI solution for hybrid cloud deployments, leveraging our APEX cloud platform. This deployment will allow customers to deploy open source models using bootable containers, which will greatly streamline the development process for the development community.

(00:44:32):

And we are constantly building out the great ecosystem of partners with solutions from Palo Alto Networks, Run AI, and Lamini coming soon and many, many more in the hopper. I think we can all agree that the pace and rate of innovation in this space is simply astounding. Customers are looking to leverage AI to their advantage and to do it the right way, and that means that they want to bring AI to the data and they want to do it on infrastructure that's open and flexible because they want the flexibility on how, when and where to run the models. And to help us show exactly how we're making this real and some of the fabulous innovation that our partners are driving. I'd like to invite Ihab Tarazi, our Senior Vice President and CTO of AI Compute and Networking to the stage. Thank you.

Ihab Tarazi (00:45:36):

Thank you. Hi everybody. Excited to be here. AI innovation is accelerating at a very fast pace. And open source models are driving the most impactful innovations. I'm excited about our deep collaboration with Meta from the early stages of the Llama 2 release, and I would like to invite Sy Choudhury, Director of AI Partnerships at Meta to the stage. Thank you for joining us Sy, and congratulations on the release of Llama 3. Can you tell us a little bit about how Meta continues to invest into Llama?

Sy Choudhury (00:46:22):

Yeah, thank you for having me here. A year ago we open sourced Llama 2, and more recently Llama 3, and it's really because we want to democratize the use of AI. If you think about it, very, very few companies out there actually has the talent to design these models, all the GPUs to train these models. And a very important last part, the fine tuning and reinforcement learning to align the models as it's called, and all this work should not be within just a few companies. And so we took that bold step to open it up to the wide community and open sourcing Llama 2 and then more recently, Llama 3. By the way, this actually helps all of us, including us, in the community. By running the models, companies and researchers alike find errors and hallucinations which are fed back into our GitHub repo so we can fix them in the next version of the model.

(00:47:16):

In addition to that, you have other things like safety benchmarks that we're working with the community to improve and catch different ways that the model can be tricked and to improve upon that. And finally, we've working with the Silicon community, a lot of the partners you talked about here who are optimizing their software stacks to make sure that the Llama 3 family of models run in the most efficient way. But frankly, that's not enough and that's one of the reasons why we're working with industry leaders like Dell to make sure the optimization happens also at the system level on your PowerEdge systems and to be able to do so in a very easy one click manner. The last thing I'd like to just point out is Llama 3 has a very permissive license, so this allows many of you in industry to be able to use it just as-is or to be able to fine tune it on your data, keeping the model to yourself.

Ihab Tarazi (00:48:14):

Yeah, thank you, Sy. We've looked at Llama 3 and we really like the features that come out of it. First of all, it's trained on a much more expanded set of tokens, 15 trillion tokens. It has a new tokenizer, an improved tokenizer with more vocabulary and it has a bigger context input of 8,000 tokens. What do these things mean to our customers?

Sy Choudhury (00:48:44):

Yeah, so I think everyone can realize that the revolution in large and small language models is happening at an extremely fast pace and with Llama 3, even compared to the models from just a year ago, you're seeing better reasoning, more accurate and sentient type of responses, better multi-turn conversation, and frankly even better coding abilities - all of these things in just one set of models that can be utilized for a wide variety of use cases. Many of you out there are in a product or business type of role. And so what Llama 3 will allow you to do is deploy this in a wide range of use cases including, for example, an AI agent or chatbot, for example, customer support. You can build some of the most innovative search and analytics tools, for example, to go through your sales data by using this technology. You can build some really interesting coding assistants to help even your internal developer productivity.

(00:49:49):

All of these kind of use cases are unlocked by models like Llama 3. Now if you're on the engineering side, your product leads might be asking you go ahead and get me one of those Llamas. You don't have to go to Peru to get that. You can download the model from our website, but more importantly, we've built a wide ecosystem of tools from things that can help you in prompt engineering to get real-time data sources all the way through to fine tuning the models very easily. And the real amazing part of this is you can utilize this in a wide variety of use cases. You can actually use the models as-is just with some really intelligent prompt engineering. You can actually go the next step where you vectorize your data, where you use something called retrieval augmented generation. I think everyone's heard this term RAG and you actually vectorize your data and you actually make the model consume much more intelligently to be able to produce that answer. Or go the final bit where you actually take and fine tune the model based on your data sets. So you get kind of the lingua franca version of Llama 3 that's tuned to your enterprise. And so the most important thing I would want to leave everybody with is as you're building your AI Factory along with Dell and other ecosystem partners, as you're doing that, keep in mind that using technologies and open models like Llama 3, you're able to actually have your data and your model in your AI Factory.

Ihab Tarazi (00:51:24):

Yeah, that's a great story and we really enjoy the innovation with you. Let's talk about the concept of using AI agents to stream complex business processes like product development. Using AI agents is gaining steam and it's only possible because of the capabilities of Llama and the commercially available open source model. In this example, if we can look at that example of highly tuned Llama 3 AI agents working together to assist every step of the product development process. This process does include market research, includes product design, code development and validation. All of that can happen with these agents working together in this complex environment, resulting in a whole new product. This whole new approach enables customers to leverage vast amounts of data and be able to make important decisions, but more importantly, they can have their highly skilled resources achieve results much faster. These capabilities of Llama 3, optimized on the Dell AI factory is really accelerating business outcomes in a whole new way. It's really exciting and this is what competitive advantage looks like in the AI era. Thank you for joining us, Sy.

(00:52:55):

So as we've seen with Llama 3, which is really exciting, let's talk about how we make it easier for the developers to do these exciting innovations. So I'd like to invite Jeff Boudier, Head of Product at Hugging Face to join us.

Ihab Tarazi (00:53:21):

Thank you for joining us, Jeff. Hugging Face has had tremendous progress over the last few years. Can you tell everybody about your journey and focus?

Jeff Boudier (00:53:31):

Of course. Thank you, Ihab. It's great to be here today. Hugging Face is the leading open platform for AI builders. Our mission is to democratize good machine learning. We make it easy for companies to build their own AI with open models and open source. Today the Hugging Face Hub hosts over 1 million models, data sets, AI applications to process and generates text, images, videos and more. And when it comes to our customers, open source AI is the only way that they can build models that they truly own, that they can host themselves in their own secure environment without compromising customer data.

Ihab Tarazi (00:54:21):

Yeah, we've enjoyed the collaboration over the last six months. Can you talk about what we're solving for our customers?

Jeff Boudier (00:54:29):

Of course. We are hearing two things loud and clear. First, they need more security, privacy and compliance when working with open models. Providing free access to a million models in your environment is just out of the question. And the second thing is that for developers going from a model repository to an on-premises, production deployment is still very hard. You have to deal with containers, you have to deal with quantization with out-of-memory errors, which means weeks of trial and error. So today I'm super excited to share the results of our collaboration, the Dell Enterprise Hub. With the Dell Enterprise Hub, it's easy to build, train, and deploy GenAI applications on-premises, on Dell platforms using the latest open models like Llama 3 from Meta.

Ihab Tarazi (00:55:31):

Thank you Jeff. This is a major step forward. The ability to use Hugging Face models quickly, securely in an automated fashion on-premise. This is a big step. It'll bring significant benefits to Dell customers. We have created ready-to-deploy, curated containers for the most popular models - you don't have to choose - and we also optimized those containers specifically for each Dell platform. Then we enhanced these containers so you can start training, inferencing and fine tuning with no code changes. So the Dell Enterprise Hub provides secure access to models from trusted sources and this is for both hosted data sets and models. So why don't we show them what we're talking about here?

Jeff Boudier (00:56:32):

I would love to. This is the Dell Enterprise Hub. It looks and feels like the Hugging Face Hub, but it's really designed from the ground up for enterprise customers. You're going to sign in with your Hugging Face account so your entitlements will carry over for access to Meta models, Mistral, Google, and more. We have a variety of outbound models to choose from. You can filter them on size, you can filter them on their license type. And behind each of those model cards is a container that is configured to run optimized on different Dell PowerEdge platforms. For instance, we're going to pick here the XE9680 for this model and our first job is going to be to fine tune this model with our data. Alright, so now we're looking at the model. All we have to do is set the path for the training data sets. We copy the code snippets into a terminal in the Dell server, and everything happens out of the box. And next we're going to deploy that model, same process. We grab the code snippets, run it within the Dell environment and voila, we have a model deployed as an API endpoint we can start hitting with our applications. Speaking of which, why don't we just chat with the model?

Ihab Tarazi (00:57:55):

This is very cool. So this usually takes weeks, but in just few clicks, you’re up and running quickly on Llama 3 models on Dell infrastructure – simple, and more importantly, secure. The best part of it? Dell Enterprise Hub is available now.

Ihab Tarazi (00:58:28):

As you can see, our innovation engine is humming. We're making it to real for AI customers. Our end-to-end portfolio supports customers from data centers all the way to the PC. And speaking of the PC, to show you how AI is augmenting the most productive device in the modern workforce, I'd like to invite to the Dell AI Factory Sam Burd, the President of our Client Solutions Group.

Sam Burd (00:59:13):

You know Ihab is a pretty smart math guy and he is absolutely right. The PC is the world's preeminent productivity tool and we've been building AI into our most powerful devices for years. For example, our workstations, which Jeff talked about earlier, have shipped with tensor cores since 2017 and they're helping power users do incredibly complex work. Now imagine being able to bring an LLM or small language model directly to mainstream PCs and running inferencing straight on the device. That's a big deal. It puts an extraordinary amount of power at the fingertips of your employees. And on a PC, you'd be doing that cost-effectively at low latency while keeping your proprietary data secure. When I think about it, that's exactly the power we're unlocking with AI PCs. In the future, they're going to become your true digital partner, enabling software developers, content creators, sales makers, knowledge workers, and everyone in between to solve problems faster and focus on the most meaningful strategic work.

(01:00:51):

That's exactly why we've gone all in on AI PCs. We're driving NPUs, GPUs and powerful CPUs deep into our portfolio. From gaming and premium consumer PCs to mainstream business laptops and powerful workstations so everyone everywhere can unlock their potential today. Companies ask me “what should I do?” We say the opportunity cost of waiting is just too high. Companies that win invest in technology and AI PCs will be central to that. They are the ultimate edge in the Dell AI factory. One company that that is Deloitte, and I'd like you to hear directly from them. So please join me in welcoming Dounia Senawi, Chief Commercial Officer of Deloitte, to the stage. Hey Dounia, thrilled to have you today. We've been talking with a few of our friends here about the power and benefits AI PCs can deliver today. I know that's something Deloitte's experiencing firsthand and I would love to hear about it.

Dounia Senawi (01:02:25):

Yeah, so first we had an idea…what if we could empower our developers with a local application to make their day-to-day jobs even better from testing to code completion and create better outcomes for our clients? We saw that AI was following the data to the device and we really wanted to make sure we were consistent with that flow.

Sam Burd (01:02:48):

We agree, and that's really important.

Dounia Senawi (01:02:50):

There is so much promise power even without cloud or internet connectivity, protected proprietary data and the portability of a smaller language model on the device, plus a better experience for our developers doing the work and the clients that we serve. We see this as the future, and we really want to lead now. So Sam, we said to your team, let's test this hypothesis together and we built an out-of-the box AI application runs code Llama 7B on the device and we're excited about 7C next. We gave our developers across offices and verticals Dell Latitude AI PCs to use right in the flow of their day-to-day work, and we're now testing both the experience and the impact.

Sam Burd (01:03:40):

That's really exciting and I know as we've been talking about this, Dounia, you're seeing some really meaningful results. Some that you expected and some that ended up an added bonus.

Dounia Senawi (01:03:50):

Yeah, so great initial results in both process and quality, speed, productivity, reduced errors and improved privacy and security. And as an added bonus, we expect that we'll see sustainability and energy savings as we look to right-size workloads where it makes sense. We really see many use cases and meaningful applications of a AI PCs to make a huge difference in mainstream corporate computing, both for us and our clients in industries like healthcare, the public sector, in-field services, and as we talk about this potential, I think what's most important is that we keep the human experience at the center. How do our clients, our developers, our customers, actually feel and experience this?

Sam Burd (01:04:41):

I think that's really insightful. Like you, and I know you're passionate about this, we believe AI has great potential to enhance user experience and ultimately productivity.

Dounia Senawi (01:04:54):

Yeah. I'm going to give you one recent example. Our developers were excited to get their laptops. We all were. One very optimistic after initial testing, she's now using the AI plugin and a process that would take about an hour for her to get up to speed on an existing project, build on the code, and then test it is now taking a half an hour. So improved speed and productivity for sure, but the excitement, the enthusiasm, the energy from our developers, that's what makes all the difference.

Sam Burd (01:05:26):

That's a really amazing story. You think about it for that team here, taking a 7 billion parameter AI model, running it in your workflow on our current a AI PCs to empower your workforce today. Then on top of that, thinking about the opportunity to take that to your clients across industries and transform how work gets done. It's really incredible.

Dounia Senawi (01:05:51):

Yeah, we are loving it. Thank you so much, Sam, and we're super excited about what's on the horizon for us together.

Sam Burd (01:05:58):

Hey, thanks for being here.

Sam Burd (01:06:06):

Deloitte is an excellent example of how companies can lean in to AI PCs today using Intel Core Ultra processors to drive transformation. And I'll tell you, AI is just one example of where we're innovating. We're investing in intentional innovation that delights the end user, simplifies IT, and advances sustainability. Many of you are our customers and you've seen this. We've been leading the industry in displays for a decade. We pioneered infinity edge screens with near zero bezels. We have the world's first five-star rated eye comfort monitors that minimize harmful blue light. We lead the industry in high resolution displays with amazing products like our first-to-market, and my favorite, 40-inch curved, 5K monitor. As many of you can testify, once you have a Dell Display, it is hard to look at any other screen.

(01:07:15):

For IT, we make security and manageability easier than anyone else in the industry. Hopefully the group here really likes that. We're creating self-healing PCs that leverage telemetry and AI to fix issues without human intervention or disruption. We're the first to offer fleet-wide BIOS configuration natively in a cloud-based management console. And we're doing all this sustainably as the first in our industry to use bioplastics and reclaim carbon fiber in our products. First to have recycled cobalt batteries in our laptops and first to have 50% recycled steel in our monitor and desktop chassis. That is an impressive list, all requiring a big shout out to our engineering teams who drive that every single day. It takes advanced engineering and it takes a mindset in our company to do things differently and do things better. To add to all that, we have never seen the ecosystem so excited and buzzing with innovation. Just yesterday we announced five new AI PCs powered by Snapdragon X processors.

(01:08:54):

These devices have four times more powerful NPUs, a level of performance, which you will soon see across our portfolio with our ecosystem of silicon partners. The exciting thing for me with this new level of performance, our software ecosystem partners are driving innovation like never before. So, let's get hands-on to see what these new PCs can do. To help us do that, I'd like to invite Microsoft's Corporate Vice President of Windows, Matt Barlow, to the stage.

Sam Burd (01:09:37):

Hey, so a busy couple days and a big announcement yesterday around Copilot Plus PCs. I've personally been using one of these for a couple of weeks and I'll tell you, Matt, I am blown away by the contextual understanding that it has. It helps me get ready for meetings, find content instantly, but don't tell my boss how much more productive I am. It's really amazing.

Matt Barlow (01:10:00):

I can only imagine Sam Burd being even more productive, man. I think these Dell Copilot Plus PCs are just incredible. I mean, you see them here on screen. They are an entirely new class of Windows PC. They're beautiful and they're powerful and they're designed to really extend our vision for AI at work, which I think is amazing. And even more powerful, I think they're going to really unleash this hybrid AI era that we're heading to. Taking that compute all the way from the cloud, now down to the Windows edge with Dell. I think it's unbelievable. You've highlighted a pain point that we're really all going through right now and, I don't know all of you, but I'm sure you're going through the same thing - finding things is brutally difficult. You think we're intelligent and we know where it is - we don't. And I think that with Recall, the feature we talked about yesterday only on Copilot Plus PCs, anyone can quickly access anything from their past activities on their PCs. I mean, it's awesome.

Sam Burd (01:10:57):

Yeah, it sounds great to me. Let's take a look with some of our friends.

Matt Barlow (01:11:01):

So yesterday I was working on Outlook and really responding to emails and I got an IM asking for my flight confirmation number. And the difficulty here is I couldn't remember when it was sent or the way it was sent to me. So instead of searching through a bunch of applications on Windows, I just click the Recall icon on my task bar and then the Copilot Plus PC remembers where it was, recalling it just like that. So with Recall, I can easily review my recent history using search or just typing in my flight information like you see there. And there it is. All the details I need in text and visual formats. It's awesome. And I actually forgot that these details were sent on my personal email, so I used Screen Ray and I was able to copy all that information without having to look at Outlook or search for that exact email that you see there as well. It found what I needed, virtually instantly, and it was so much faster finding these things versus looking through a bunch of accounts, opening tabs or browsing history.

Sam Burd (01:11:55):

That looks like a huge time saver, Matt. I know before this I often felt like I was spending a day a week searching for things and getting ready for meetings. So really impressive. Now as you were going through that, I noticed you got interrupted with a ping on Teams… can Recall help get you back in your flow.

Matt Barlow (01:12:14):

We all get interrupted like that for sure. Recall certainly can do just that. The email that I had open really requested some updates, but it didn't include a link. So now I don't even need to remember that file or even where I saved it, Recall just makes remembering easy. It does it for me. Just type in a keyword like future initiatives that you're going to see, and just by typing in that – “future initiatives,” I can find the right slide, not the file, but the right slide itself. You see here, right here, it's coming up - future initiatives. I don't have to look through folders, I don't have to look through files. It's just quick and direct. It pulls up that slide I'm looking for and gets me on track faster than ever before.

Sam Burd (01:12:51):

Nice. Very good.

Matt Barlow (01:12:53):

And then when I get back on track with this PowerPoint here, it actually helps me get into my workflow even faster. I can summarize the deck, I can turn it into a Word doc. It saves us all time.

Sam Burd (01:13:04):

Another question for you Matt. What about security and privacy? What information gets captured? How's that work?

Matt Barlow (01:13:11):

You're going to see it here too because we built Recall with data protection built right in. We've got enhanced data controls that really put you in charge. You're going to see them here. You can simply turn off Recall any time. You see the slider? You can choose the applications and the web content that you want to ignore or block out. It's easy to do. And frankly, we're going to make their control available to IT as well so they can manage these Copilot Plus PCs the way that they want.

Sam Burd (01:13:35):

Yeah, I would say Matt, that level of control for IT is a must, a great feature. How about what happens when an end user gets offline?

Matt Barlow (01:13:42):

Yeah, so that's the neat thing too about how Copilot Plus PCs really work is that you don't need to have a cloud or an internet connection to get a lot of these great features to deliver a benefit. They all run locally with small language models and the NPU. So for example, on my flight here, I caught up at work by looking at a recorded teams meeting that I had missed and the meeting was held in multiple languages. Basically, Copilot Plus PCs use live captions, and that's only available on these new devices that we've got with Dell. And I watched the video itself translated into English subtitles in real time, even when I was in airplane mode not connected to the internet. It was awesome. So check out what live translation can do on Copilot Plus PCs.

Sam Burd (01:14:39):

That is really impressive and I can see how that is going to transform how all of us do work and how things get done. So hey Matt, really appreciate the collaboration work with Microsoft in bringing the Copilot Plus PC to life.

Matt Barlow (01:14:55):

Can't wait. The Dell Copilot Plus PCs are great. Check them out today.

Sam Burd (01:15:00):

Hey, so as you can see, once someone experiences a day with an AI PC, they are not going to want to live without it. Soon everyone is going to expect an AI PC, just as everyone expects an internet browser or wifi on PCs. We know that the highest performing companies are those who invest in technology. So as you think about the future of your business, consider the role PCs play in an end-to-end AI factory. And with that, let's bring Jeff back out here. Jeff, you get your world-class AI Factory back.

Jeff Clarke (01:15:57):

Thanks Arthur, Sam, Ihab and our partners. Thank you for your patience. I know we're running a little long, we're saving the best for last. When we put this all together, it's so fast. Help me welcome Zak Brown, CEO of McLaren Racing.

Just back from Italy, Miami, Italy?

Zak Brown (01:17:09):

Yep. It's been good. We won Miami and we were seven tenths of a second away from winning in Italy.

Jeff Clarke

We're going to talk about tenths of seconds in a few minutes or seconds, but our partnership is long. This is the fourth time over the years you've been with us.

Zak Brown (01:17:14):

Yeah, I love joining you guys.

Jeff Clarke (01:17:15):

We have a lot of fun together.

Zak Brown (01:17:16):

We certainly do. And we go fast together.

Jeff Clarke (01:17:18):

Speaking of fast, let's talk about fast. When we think about AI and all of what we've spent the last hour plus talking about, tenths of seconds, I think Bill McDermott mentioned it yesterday, tenths of seconds are the differences between winning and losing. You mentioned it again. How is AI helping you on the track be competitive?

Zak Brown (01:17:36):

It's critical. I mean, we Formula One teams, we live in kind of tomorrow and we get a tremendous amount of data. We have very much an efficiency game, a performance game, and we need the AI, the AI Factory, our partnership with you to make us go faster. It's working. Look at where we started in our journey, I think together in 2018. Now we're running at the front all the time and it's fantastic. So without AI, without all the technology that we use with you, our partnership goes well beyond just the AI Factory. But we have a factory to put together the race car and then we have the AI Factory to help us put the race car together.

Jeff Clarke (01:18:16):

We were talking this morning about that factory that you use to really take the technology and the insights to compete to the point where you take the technology stack of an AI Factory, our infrastructure server, storage, networking, the software stack, and then ultimately you're running simulations using digital twins to predict performance, to tweak performance over the course of a race. And over the season, I assume.

Zak Brown (01:18:43):

Our industry is covered by 2.6%, the best to the worst. So we're working in the smallest of margins. And so it's critically important. Our world starts in CFD, we kind of live digitally before we actually produce some interesting stats. We pull down about one and a half terabytes of data a weekend, 300 sensors on the car, 50 million simulations. And all that goes into running a race weekend where we have to make split second decisions.

Jeff Clarke (01:19:13):

That's a weekend?

Zak Brown (01:19:15):

That's every weekend, 24 times a year all over the world, about 1,200 people in total. We'll have about a couple hundred people at the racetrack, about 80 back at the factory working real time in how we run these races. And when, I mean we have a split second to make a decision, we will know we need to run a certain tire compound and it'll say take 20 seconds to get the tires ready. We're running these race simulations and we might have the car as two seconds to make a decision on when to come into the pit. So you can imagine on pit wall, we don't have a lot of time to go, well, what do you think? You literally have two seconds. You've seen teams get it wrong before where you come in and tires aren't ready or prepared. It’s because literally you've got two seconds to decide, are we pitting, are we not? Which tire compound do we want to use? And that's where all the data, the AI, all come into play.

Jeff Clarke (01:20:04):

You told me this morning, if you didn't learn through this process, the car that you start the season at the beginning of the year does what over the course of the season without this work?

Zak Brown (01:20:16):

It would be dead last. So you take the car that's on pole, the first race of the year, if it was untouched, it would be last by the end of the year. That's the pace of the development of all of our racing teams. So we've got some pretty awesome competition.

Jeff Clarke (01:20:27):

And if we bridge that from how you run McLaren, the company, there are many uses beyond the racetrack of AI inside your company. What are you exploring and using today?

Zak Brown (01:20:38):

Everything points to performance, but whether that's our marketing, our commercial team, our HR department, and getting the most out of our people. Whether that's our commercial department, whether that's our finance department, because there's now a cost cap in Formula One. We used to be able to spend our way out of problems. Now we can't. We have a maximum amount of money we can spend. So we have to be so unbelievably efficient with every dollar that we spend or pound, because we’re based in England.

Jeff Clarke

Racing companies have a budget?

Zak Brown

We have a budget. We used to have an unlimited budget. So you would just spend, spend, spend. But now you have to make sure that what kind of starts in the digital world works when it goes on the racetrack. Because if it doesn't, you've wasted money and you can't kind of get that spend back. So everything is about performance and we see AI as powering our people, powering our team. It's additive to what we're already doing.

Jeff Clarke (01:21:30):

And the types of jobs they're doing, the value creation they have as a result has gone up tremendously.

Zak Brown (01:21:35):

Our performance and our throughput is so much stronger today than it was yesterday and we're already working on the 2026 car, our 2026 regulations. So I think that will be the most AI-developed race car we've ever seen. So pretty exciting times.

Jeff Clarke (01:21:53):

Awesome. Thank you so much for joining us today. Zak Brown, thank you. All the way from Italy, McClaren Racing.

(01:22:02):

AI Factory in action. I know I'm on borrowed time. I got one slide. Let's get to the one slide. This is what we spent the last hour and 20 minutes talking about. This is the AI blueprint. This is how we're going to help you adopt AI and how we'll be with you every step of the way. So the entire discussion has been blueprinted of how we take AI Factories and the extension of our partnerships to open a modular infrastructure, to build AI systems and solutions for all of you. Thank you for your time today. It's about Dell AI Factories making it easy for you. Enjoy the rest of Dell Technologies World. Thank you again.