NVIDIA Corporation | Event Detail

NASDAQGS:NVDA (MI KEY: 4094286; SPCIQ KEY: 32307)

Detail

NVIDIA Corporation Presents at Piper Sandler Webinar: Networks for AI, Jun-28-2023 09:00 AM

Event Details

Announced Date	27/06/2023 08:25:00
Period Ended	NA
Company Name	NVIDIA Corp. NASDAQGS:NVDA
Source	Company Website
Event	Company Conference Presentation
Advisors	NA
Situation	NVIDIA Corporation Presents at Piper Sandler Webinar: Networks for AI, Jun-28-2023 09:00 AM.

Call Details

Live Phone Number	NA
Live Passcode	NA
Live Other Phone Number	NA
Live Other Passcode	NA
Live Audio Details & Webcast URL	https://pipersandler.zoom.us/webinar/register/WN_E4NFKrynTT2Z-vtJcKw26g#/registration
Replay Phone Number	NA
Replay Passcode	NA
Replay Begins	28/06/2023
Replay Ends	NA
Replay Audio Details & Webcast	https://video.ibm.com/recorded/132877209
Call Description	NA
Host 1	NA
Host 2	NA
Host 3	NA

Company Details

No company details exists for the transcript.

Transcript

NVIDIA Corporation Presents at Piper Sandler Webinar: Networks for AI, Jun-28-2023 09:00 AM

NVIDIA Corporation (NasdaqGS:NVDA)
Wednesday, June 28, 2023 5:00 PM

Executives
Colette M. Kress - Executive VP & CFO
Gilad Shainer -

Analysts
Harsh V. Kumar - MD & Senior Research Analyst

Presentation

Harsh V. Kumar
Okay. Good morning and good afternoon, everyone. Thank you, a huge thank you to everybody that signed on for this webinar. We understand and realize that your time is extremely important. And we appreciate you spending some of that valuable time with us today. Also a special warm welcome to the NVIDIA team, Colette, CFO, who a lot of you know already; and Gilad, SVP of Networking. And also a special things to Simona and Stewart, who made this all possible.

Question and Answer

Harsh V. Kumar
So before we kind of get into the networking piece of it, I did have 1 or 2 questions for Colette based on some sort of speculation that's percolating in the media about U.S. and China. So Colette, maybe in light of last night's press articles regarding potential new export control on AI chip shipments to China, what can you tell us about the potential impact to your business and NVIDIA?

Colette M. Kress
Thank you so much. Thanks for the question. Let me see if I can provide a little bit of understanding. We are aware of reports that the U.S. Department of Commerce is considering further controls that may restrict exports of our A800 and our H800 products to China. However, given the strength of our demand for our products worldwide, we do not anticipate that such additional restrictions, if adopted, would have an immediate material impact on our financial results. We do not anticipate any immediate material impact on our financial results.

Over the long term, restrictions prohibiting the sale of our data center GPUs to China, if implemented, will result in a permanent loss of opportunities for the U.S. industry to compete and lead in one of the world's largest markets and the impact on our future business and financial results is there.

Harsh V. Kumar
Well, that's about as clear as it can be, Colette. And then could you maybe help us -- one of the questions we're getting a lot this morning is some context around the percentage of your data center revenue that is driven by sales to China?

Colette M. Kress
Yes. So historically, it has a little bit of a range in terms of what we've seen historically. We believe the contribution of sales to China has been in the range of approximately 20% to 25% of our data center revenue. Keep in mind, this includes all of our compute products and systems and also our networking.

Harsh V. Kumar
Okay. Great. And then I did have one last one that we've been getting a lot. The switch to -- you guys were able to pivot very quickly to the A800 in a matter of a few weeks. People almost believe that, that was a software change. I just wanted to clarify, when you guys made that switch to A800, was that a software change or was that a hardware change?

Colette M. Kress
It was not a software change or movement to A800, that it was absolutely a hardware change that we made to create the A800 as well as what we did to create the H800.

Harsh V. Kumar
Thank you, Colette. This was supposed to largely be a networking session. So with that, I think, Simona, if you want to turn over or Aaron, if you want to turn over the presentation to Gilad. We can go ahead and get into the networking piece of things. Thank you, Colette.

Presentation

Gilad Shainer
Yes. Thank you very much. So nice to be here. I'm Gilad, SVP of Networking, came to NVIDIA through the acquisition of Mellanox, and been before at Mellanox almost from the beginning, 20 years plus at Mellanox and now, I think, 3 years at NVIDIA more or less and I started as a network designer.

So some of the early InfiniBand devices, I was part -- I did part of the design of them and then actually started looking on the entire platform, the software capabilities and so forth. So excited of being here, and I think we can move to the next slide.

So I think this is going to be the slide that has most number of words on it and make sure that we have much less words on coming slides. Our presentation -- I need to say this statement, our presentation may contain forward-looking statements. And please do refer to our SEC filing for risks and uncertainties facing our business.

So with that, we can move to the next slide. So we're here to talk about networks for AI, obviously. But when we talk about AI, it's not just a matter of a network, right? It's not just design components. You need to look on the entire system. What are we designing the network for.

And essentially, we want to build or we need to build or to design a full accelerated compute network balance system. It's not started at the niche. It hasn't started the switch. It starts GPU level. Memory IO on the GPUs all the way to the network. And we need to successfully started the application framework, application level and looks on everything there. And that's the great advantage of NVIDIA.

If you look on NVIDIA as a networking part, it's the only networking entity that actually build and design full AI platforms and use those AI platforms. So looking from a software perspective, the platforms, the SDKs, the libraries. There is a huge -- there is a ton amount of software, which is part of the system. There is the case that connects between the network and the GPUs, for example, and then the full hardware capabilities, the [indiscernible], computer DPUs, the switches, the NICs and so forth.

To be able to build or essentially the ability to build the entire system, give you the opportunity -- unique opportunity to place the right data algorithms in the right place. There are data algorithms that you don't want to run on a GPU. We actually want to run them on the network. And that's what we call in-network computing.

There is elements that traditionally, you may do it on the network, but if you can, you probably want to build it on a GPU because it's going to be much more effective. So we are able to move algorithms, we're able to build a very effective system that delivers the highest levels of performance at very, very large costs.

So with that, understanding that you don't just build the network, actually build the entire data center, let's talk about the data centers and then dive into the network. So we all know that the data center is the computer today. In the past, CPU was the computer, and then the server become the computer, now the data center is the computer. We're just putting down data centers to run workflows, to run the applications.

Now if you look at the data center, there is a big collections of GPUs. And actually, the way that you connect those GPUs define the data center. The way did you connect the GPUs define what you can do with those GPUs and define what kind of workloads you can run on those GPUs, essentially define a data center.

And we can look on different examples. The first example, let's look on cloud, traditional cloud. Traditional cloud, it's a data center

that is built to support many, many, many users and to support variety of workloads, and most of them are very small.

You've got a single node workloads, lot of single node workloads. Such a traditional cloud is being connected today with traditional ethernet network. Because traditional ethernet network is good enough for those kinds of platforms. Lots of users, lots of applications, almost all of them are very, very [indiscernible].

Now we're actually facing the creation of new kinds of clouds. New kinds of cloud, new class of clouds, clouds to support Gen AI workloads, cloud to support AI workloads. And AI workloads are different than the traditional workloads that runs on traditional clouds. The AI workloads are not just running on a single node, AI workloads need to run on multiple GPUs. They need to run across nodes.

And even more important, when we talk about AI and AI workloads, we actually start talking about distributed computing. And distributed computing is completely different than disaggregated computing, it's completely different than hyperscale. It's something new. And that's something new to cloud actually requires elements to support distributed computing. Now you start talking about latency needs and day tenancies and effective bandwidth, those are completely different kind of requirements.

So traditional ethernet is still fine for the north-south traffic of Gen AI clouds. We need the user access, the services, the control of that cloud, but you need a new class of ethernet to support the new class of workloads in that new class of cloud, and that's exactly what Spectrum-X is. It's the first ethernet ground-up design -- ground-up design for AI. We start interface, so people can enjoy the -- or utilize ethernet ecosystem and it's actually combining both.

Now if our data center mission, it's actually to run massive large-scale workflows, massive large-scale applications, large language -- large LLMs, large language models and complex training -- to deal with complex training. This is a new kind of data. This is a different kind of data center, right? Now we're not talking about many, many, many users and a variety of workloads. Now we're talking about much less number of users and workloads that are going to consume the entire GPUs in that system on a single application.

So it's not a matter of how many GPUs it can connect. It's a matter of how many GPUs or workflow can consume with that network. It's completely a different thing. So now if you want to support workflow, that's going to consume tens of thousands of GPUs or hundreds of thousands of GPUs. The only option -- the only option, and that's why it's become gold standard is the combination of NVLink and InfiniBand.

NVLink and InfiniBand are not the same architecture. It's a completely different kind of architecture designed -- specifically designed for distributed computing being optimized over the years and has the network that can support running workflows on a large scale of GPUs.

Okay? So this is where we have different kind of networks. It's not one network fits all. And this is, by the way, shows you that both ethernet and InfiniBand coexist, and they will continue to coexist because in every system that you build, you do need a network to do the user access. You do need an effort to run control in north-south traffic always and ethernet is great for that. But for AI infrastructure, you need a network for East-West. You need the network for the distributed computing. This is where NVLink and InfiniBand are actually gold standard.

So if we go to the next one, in the coming slides, I'm going to refer to a couple of terms. We want to make sure that everyone understands those terms. So it's going to make life easier. And in particular, NCCL and SHARP.

So NCCL, what's NCCL? So NCCL is a short of NVIDIA Collective Communication Library. It's a software SDK. It's a software SDK for AI communications for multiple GPUs. Essentially, this software framework supports mainly 2 multi-GPU [indiscernible] communications. One of them is reduction or reduced operation and the other one is all-to-all communications.

So NCCL essentially enables the connection between the GPU side and the network side to support those 2 operations, reduction operations and auto operations.

Now NCCL, if we want to measure AI performance -- AI networking performance or networking performance for AI, NCCL is a great option to test the performance with. So you can look on what's my performance for NCCL reduction operations, what's my performance for NCCL [indiscernible] operations, for example, and that will demonstrate the impact of the network. So it's a good

way actually to test the network or to measure the network with.

SHARP, it's a technology part of in-network computing, NVIDIA in-network computing. It's a technology that's implemented in the InfiniBand switch ASICs. It's not something that runs on a CPU or some other, it's embedded within the switch ASICs that enable the switch network to perform data reduction operations on the data as the data is being transferred within the data center.

Now previously, those data reduction operations, which kind of part of NCCL were done on their host, and running them on their host takes a big toll on their host. And this is kind of part of NVIDIA advantage, the ability to move algorithm from one side to another set and to run them in the right pace.

Moving the data reduction operation to be run on the switch network reduces the amount of data that you need to send on the network by half. It's a huge impact. It means that a 400 gig end-to-end InfiniBand network with SHARP, it's better than an 800 gigabit per second end-to-end network without SHARP. That's amazing capability of InfiniBand. And that's one of the elements that enables InfiniBand or make InfiniBand the gold standard for AI factories.

So if you look on what's this NCCL, the impact of NCCL with SHARP, you can see that on the right. We're gaining 1.7x higher performance because of SHARP, because of running NCCL, because of running reductions on the switch network compares to the best other network you can actually build. So if you would compare it to the best theoretical, the performance on the ethernet, it's 1.7x. So this is one of the key things that actually make InfiniBand, again, the gold standard for large-scale AI and for AI factories.

So now we can go back and talk about the different networks, and we'll start in the cloud and then go to Spectrum-X. So in the cloud itself, in the cloud, there are 2 kinds of ethernet networks, essentially, 2 worlds of ethernet. There is a network that is doing the North-South connectivity, gets the controlled access, gets the user access. Those are the cloud services.

Cloud services or user access, those are loosely coupled applications. So you typically use TCP for that traffic. Jitter is fine because there is user access, there is jitter and jitter is okay. Latency is actually not critical. Predictive and constant performance of bandwidth, it's not important as well. What's important for you is to deal with heterogenous traffic. You need to deal with multiple loosely coupled processes and process them and enable them to run on the network.

And this is where traditional ethernet is being used, right? This is where ethernet was designed. This is kind of the cloud network that we all know about. Now in the cloud, there is a second network, which is the compute network what we call East-West. In traditional cloud, there is not much East-West traffic because most of the workloads, most of the users are running on a single node.

And before, in the traditional cloud, you can take the same North-South network and use it for East-West network and that's fine, that's okay, that works. Now if you want to host AI workloads, if you want to build clouds for generative applications, now East-West network needs to be completely something else because now East-West networks needs to deal with distributed computing.

And this distributed computing is very sensitive to latency, but even more, it's sensitive to tail latency. In distributed computing, you run application across multiple GPUs, many, many GPUs in a sense. If one GPU communication is going to be late, only one, let's say that I'm running on a 500 GPUs. If one GPU communication is going to be late out of that 400, just one, the entire workflows will be delayed, the entire workflow will be delayed. So tail latency is a critical element for AI performance. It's completely not relevant for North-South traffic, but it's critical for East-West.

Effective balance is important and you want to provide constant performance. You cannot have changes in performance levels and need to deal with burstiness. So the requirement for distributed computing are completely different, I would say, the opposite of what you need in the North-South. So now you cannot use a traditional ethernet for East-West, you need to do something else. You need to have a different class of network to support the new needs of AI applications in the cloud. That's the reason we need Spectrum-X. That's the reason we design Spectrum-X because we needed a new class of ethernet for this kind of infrastructures.

Next slide, please. So now let's look on Spectrum-X. So Spectrum-X on the left side, you see all the starts, 51.2%, the number of ports essentially and so forth. And on the right side, by the way, you can see a snapshot of the software that is being developed

for Spectrum-X. There's tons of software. The SDKs, there is [indiscernible] SDKs that transform the DPU in BlueField to provide the network virtualization, the isolation between the application infrastructure -- and the applications and the infrastructure.

There are -- the Spectrum is the key for the switches. [indiscernible] is the SDL that includes the NCCL framework that I mentioned before, the newer of the operating system that runs on the Spectrum-4 switches, which are SONIC and Cumulus and other aspects like that. So there's tons of software with it.

Now Spectrum-X, what essentially designed ground-up for AI and we built new capabilities, actually designed new capabilities for ethernet. And some of those capabilities are including first lossless ethernet. Now what's interesting here is essentially the combination of those elements. So I'm going to go through them. First, lossless ethernet. You don't want drop packets. Dropping packets mean you're creating jitter, you're creating jitter and now you're reducing AI performance. So you don't want the drop packets.

On top of those -- of lossless ethernet, you want to support adaptive routing. Now not a flow-by-flow adaptive routing. We do see flowlet adaptive routing in ethernet switches -- in traditional ethernet switches. Flow-by-flow means that you need to -- you run a stream of data, and you don't change the path of that stream of data before that streaming ends. That's not good for AI. In AI, you want to define grain-adaptive router. You want to do packet-by-packet adaptive routing.

So that's an element that it's enabled by actually doing lossless, but even more, even more. You want to do the packet-by-packet adaptive routing on lossless network with shallow buffers, not deep buffers, not deep buffers. There are ethernet options out there, for example, that sometimes referred to as fabric, not ethernet because sometimes they don't run to actually ethernet. And those depends on big buffers, big buffers in switches to big shock observers. So if there is congestions that can kind of hold data and stuff like that. Deep buffers mean long-tail latency. Long tail latency is not something that is really nice for AI workloads. You don't want deep buffers.

So now the IP here is combining lossless ethernet, fine-grained adaptive routing and shallow buffers. That's the combination. That combination does not exist in traditional ethernet, completely does not exist. This is one part of Spectrum-X advantage.

Second part is doing congestion control. You need to eliminate hotspots. And we designed in Spectrum-X congestion control, which is based on first telemetry information, but also have unique capabilities in the network in order to identify latency changes so you can react to hotspots before they can impact performance of applications.

And this is important because this is the key to provide traffic isolation. This is the key to eliminate noise to make sure that noise cannot impact AI performance. In the cloud, you run many, many workloads. You want to make sure that those workloads, especially the small scale workloads will not impact the large scale workloads.

They're running on the same network, but you want to make sure that you isolate the noise from -- on the small workloads that they will not impact the AI workloads, and that's exactly what we're doing with congestion control, [indiscernible] congestion control and the capabilities to identify latency changes and identify hotspots before they actually can do a negative impact.

What it gave us, that gave us 1.6x higher AI fabric performance over traditional ethernet. So we're talking about not just 95% effective bandwidth at scale and under load, but keeping that performance constant, predictive performance, keeping that performance constant, even that you have a lot of other workloads running in a semi environment because we did it with cloud. Having the security, the virtualized network, everything is part of that.

So now Spectrum-X actually brings the speeds and feeds that you need for AI, but it does it with an ethernet interface. So people can leverage the ethernet ecosystem for services that were built for ethernet, for cloud services and things of that sort. But now they actually have an ethernet that was designed for AI.

Next slide, please. Now as we look in to support larger scale of AI workloads, this is where we go to InfiniBand. InfiniBand that you see on the left side, on the latest generation there. But one thing that you need to understand, InfiniBand is designed based on a different kind of architecture versus ethernet.

Now ethernet was built for wide area networks. And over time, within the data center, more and more algorithms were designed for ethernet -- more and more operatives were designed for ethernet. PFCs, for example, [indiscernible] that we're designed for

ethernet. So ethernet, it's a very complicated protocol. It's a very complicated protocol.

And when you build an ethernet network, you need actually to choose between features and performance. You need to choose between features and performance. That's why in ethernet, there is no one switch fits all. You see a variety of switches coming from different kind of entities. And the reason is that no one switch fits all.

There are switches with shallow buffers and more pores, but not much of a good performance for distributed computing and just supporting kind of cloud interfaces. There are switches with deep buffers to support sometimes [indiscernible] service applications, but that comes in issues of day latencies and reduced number of ports and so forth. So you need to choose, you need to choose between features, performance and other stuff.

In Spectrum-X, we actually designed that to have the right elements that you need for AI. And actually created things that doesn't exist in traditional ethernet. But InfiniBand, when you look at InfiniBand, this is a different kind of architecture. They are not using the same architecture. InfiniBand was designed from the beginning to support distributed computing.

And from that reason, InfiniBand protocol is very simple. It's lightweight. It's very, very simple. And because it's very simple, there is no meaning in InfiniBand for leaf and spine, and in terms of ethernet, there is leaf and spine and in ethernet, you're trying to go through level of network, 2 level of switches and don't go beyond that 2 level of switches and stuff like that. It doesn't exist in InfiniBand. There is no such thing in InfiniBand.

If anything, you can use as many switch layers that you want, even more. Most of the large-scale systems out there using 3 levels of switches in InfiniBand. Some systems even use 4, if you want to use 5, use 5. There is no performance penalty there. There is no issues around that. They can build any size of system that you want. It's like you're designing here a Formula race car. So if you're designing a Formula race car, how many seats I'm going to put in that car, who cares?

So it's a different kind of design. And if you are looking on 3 level of switches, with InfiniBand, which is what most systems use today, that can go all the way to 65,000 GPUs. And if we go to 4 levels, and we have several 4 levels or multiple 4 levels already out there, you can go to 2 million GPUs in InfiniBand network. You want to go 5, go 5. There is no limit of how many GPUs you can connect together, and it's even more. We didn't see a limit of how many GPUs you can use for a single workflow. That's the important part.

So there's no limit there. That's why InfiniBand is called standard for large scale AI. Now InfiniBand pioneered RDMA, obviously, so there is lot of elements in RDMA, but InfiniBand, pioneered and laid with [indiscernible] computing. And we saw the impact of SHARP. SHARP gives you 1.7x on [indiscernible] compared to the best ethernet network you can build and it's a pure software defined, pure software defined network.

It was designed as an SDN before people knew what SDN means or what SDN is, which means is that you can control the entire routing from a single place. You can optimize the routing to the workflows, you can build different kind of network topologies. You can treat with changes in the network quickly. You're reconfiguring the network, ports is down, ports are up, you can verify quickly. There is a huge amount of benefit in a pure software-defined network.

So what that gives us, if you look on the total performance, it's more than 2x, kind of being gracious here, as many GPUs as you want, as many GPUs as you want and building a network that have the lowest latency, again, in large scale and under levels. Very short in latency, extremely short in latency.

We know the impact of SHARP on NCCL operations, nearly 100% effective balance at scale, it's amazing network. It's really amazing network. And it's been developed over more than 20 years, right? It's -- every generation bring new capabilities, the upcoming generation, Quantum 3, the things that we're planning there are amazing, completely amazing. Those will take InfiniBand to a completely next level compared to anything else.

Now on InfiniBand, also there's tons of software, right? We have the SDKs elements there. Magnum IO, NCCL, [indiscernible], obviously, management of the network, be able to simulate everything, there is tons of software as well. That's why it's important to do end-to-end. That's why we're doing end-to-end design. Next slide, please.

So this is where we look on the impact of the network. The network is essentially a small part of the data center. Very small part

of the data center expense. It has a huge impact, a huge impact on AI performance. And essentially, the network pays for itself, the network pays for itself. InfiniBand offers the high scalability out there. Again, you can build any size of systems that you want with it, 3 levels, 4 levers, 5 levels. There is an unlimited number of GPUs that you can connect together.

And then if we're looking on performance, and we took NCCL here again because NCCL is a good indication of the network performance for AI. So first, you see Spectrum-X. It's completely different design for ethernet and it's enabled the ethernet ecosystem, right? So if you want to the ethernet ecosystem and you need performance for AI, it's Spectrum-X.

And then if you want to build a system that's going to go to scale if you want to get the highest level of performance. We can also bring and fit them into cloud. There is no reason why not to. And if you look at InfiniBand, that's kind of amazing on top of that. And if you look on the impact of the total AI performance, the network essentially is free, completely pays for itself.

So if you need -- someone's going to offer me traditional ethernet free, completely free. It's not going to be good enough, right? I heard you actually pay to get InfiniBand because I'm going to get much better, much better from that, right? So essentially, I'm building an AI infrastructure, the network there is essentially free.

So with that to the next slide, I think this is the last one. So if you're looking on networking revenues, NVIDIA networking revenues, so the revenue more than doubled since the Mellanox acquisition. And within that, you can see the breakdown between InfiniBand and ethernet and other.

InfiniBand, more than more than tripled. So it's growing very fast, continue to grow, will continue to grow. But then Spectrum-X, Spectrum-X is new class of ethernet, that is new class of cloud ethernet that is needed for a new class of clouds. And therefore, Spectrum-X will boost the Cloud AI network market and will increase the ethernet revenues as we're moving forward.

Overall, we see that essentially, we believe that every data center will become an accelerated data center in the future. There will not be data centers that are not accelerated, right? We used to be in a situation that we got 2x performance every 2 years just doing nothing. It doesn't work anymore. That doesn't work anymore.

So if you want to be able to increase capabilities, accelerated computing, and therefore, every data center will become an accelerated data center, every server will have a DPU persistent unit, every data center will have an element there. And as such, we're talking about the $60 billion market opportunity for NVIDIA on the networking side. And with that, first, thank you for listening. It took some time. And happy to answer questions.

Question and Answer

Harsh V. Kumar
That was extremely informative. You've actually answered a whole bunch of the questions that I had before. One of the ones that I do get is investor concern around the fact that they already come to NVIDIA for compute. And then they come to NVIDIA for networking now based on the merits of what you, for example, just talked about. So we get a lot of questions. We're basically tied to NVIDIA a lot.

And I think people, as you know, in semiconductor business and IT, they always want options. Could you maybe talk about what's -- is there a workaround to that? Or are you the only ones that makes InfiniBand or is it farmed out to other places that do it for you?

Gilad Shainer
Yes. Well, InfiniBand, it's a standard technology, right? It's not a proprietary technology, standard. It's the same like ethernet. Ethernet is also standard in that sense. So companies can definitely create InfiniBand devices. And actually, there are some companies that build InfiniBand devices for different kind of applications.

There is a company building devices for long haul connectivity. There is InfiniBand elements for FPGA things and so forth. So of course, InfiniBand is open, so everyone can use that. Now there's always a [indiscernible] for networking, right? If you don't want to use InfiniBand, you use ethernet. If you don't want to use ethernet, you can use InfiniBand. You can always choose between them.

But the question is on standard, but essentially, especially when you look on AI, when you look in AI, AI requires a dissenter scale. And if you look on that, then actually, you want to have the right elements inside, right?

I said it before, we used to get [indiscernible] performance every 2 years. Now it's not the case. And the 4, we're going to see more and more specialized technologies and actually more users of accelerated computing and more use of technologies that it can enable you to achieve the goals, achieve your goals. So optimizing AI workload performance cannot be performed by discrete compute or networking device level.

We want to look on a full stack approach. So now -- what's important essentially, I would say that it's time to market, time to solution. Customer considers total cost of ownership, performance, [indiscernible], time to build and deploy the large scale architectures. And this is what NVIDIA delivers. So we build a full platforms. We're doing a huge amount of optimizations and our customers can take it as a whole. But if customers want to take pieces and where they wanted to take business of that and mix and match other things that happen in the market, that exist in amount.

Harsh V. Kumar
Great. And Gilad, one more for you. You guys are sort of the gold standard as a company in the accelerated data centers. As it comes to InfiniBand network adoption, have you noticed the big difference in metrics for training versus inferencing, for example, for InfiniBand networks, either in terms of ports or in terms of any other metrics that you think you can talk about because investors generally feel like inferencing is on the come, and that's going to be a huge opportunity. So I wanted to address that?

Gilad Shainer
No, yes. It's definitely a good question. So training requires very large scale clusters that are tightly coupled and optimized for massive, massive data and compute. Inferencing typically require much smaller scale clusters.

But what's happening now is that generative AI is becoming mainstream. And for the number of the separate jobs running, inferencing will dramatically increase. Therefore, inferencing will require a larger number of accelerated service and the flexibility, essentially to do that.

And we're probably going to see people that are going to deploy a system and they will want to use those systems for both training and inferencing, both training and inferencing. And therefore in such a case, obviously, InfiniBand is a great option for that.

Now if someone is just going to interesting and doesn't need to go to the large ones, of course, they can use Spectrum-X for that. But we're going to see probably a system that I'm going to use for both make sense to build system inferencing or both, InfiniBand is a good option from them.

Harsh V. Kumar
Great. And I have one more for you, Gilad. There is a perception in the investment community even in the people that know Generative AI very well, that InfiniBand only works with NVIDIA's GPUs. Is that accurate? Or listening to you talk, it seems like that's not the case, but I wanted to ask you since you're the expert on the topic.

Gilad Shainer
Yes. No, InfiniBand is open to be used with any other accelerated and non-accelerated compute platforms. At NVIDIA, we do develop the full stack platform. And our customers can choose to take it as a whole if they want to take our design and copy that design is all, but they can actually take pieces of it. They can take our GPUs and use them with other networks. They can take our network and use it with other compute elements. It's free to use in any platform. It's definitely not the time.

Now obviously, end to end, there's a lot of benefits into that, right? We invested a lot of effort, and we're investing that so our customers will have much faster time to compute, much faster time to solution, much faster time to build their system. When you build a supercomputer, AI supercomputer, you don't want to spend 9 months to build it. That's 9 months after the lifetime of very expensive system.

So we're doing what we do, so they can build the systems in weeks, not in months and they can take the full performance out of it. But again, people can choose our -- take our components, they can use our network with any other compute elements and so forth.

Harsh V. Kumar
Wonderful. And I know we -- I know you're on the road, so I want to be very mindful of your time. We've got 2 minutes, so I'll just ask one final question. In a typical setup, let's say, as you guys go and deploy accelerated AI data center, do you typically find the entire data center to be either or is it all InfiniBand or is it all ethernet? Or is there a possibility to mix and match some of your offerings, depending on what the lines are supposed to do?

Gilad Shainer
So first, obviously, there are entire ethernet systems, are there, right? We all know that. There are systems that are full just ethernet. And in such systems, there are different kind of ethernet. And we created Spectrum-X in order to bring the right class of fitter for the AI compute fabric. So you can definitely have just ethernet systems. If you want to build a Gen AI cloud system, and you want to leverage the current ecosystem for some of the elements.

So you don't need to develop all the software yourself in the cloud. Then Spectrum X is a good answer. It gives the feeds that's needed for AI and give you the ecosystem friendliness of ethernet. So those systems are definitely going to exist.

On the other side, when we talk in large scale, then we have systems that are essentially combining both InfiniBand and ethernet. It's not one versus the other. It's completely -- they're completely going to coexist in a large AI factory, large system that runs, large language models or doing training, you have ethernet for the North-South access.

InfiniBand was not built for user access. That's not it's meant. That's not its purpose. And for that interface, we have ethernet. And then for the compute fabric, once you want to connect a large amount of GPUs, thousands to tens of thousands to hundreds of thousands of GPU in a single workflow, InfiniBand actually gives you the combination of NVLink and InfiniBand, gives you the connectivity there.

So if you look on systems that we design, the system that, for example, we recommend kind of look on what we did and copying that will enable to leverage everything we design. Our system includes both InfiniBand and ethernet, completely coexists. It's not that one replaces the other. I think you have those networks that both exist and each one has its own purpose.

Harsh V. Kumar
With that, we have come to the end of this presentation. Gilad, I cannot thank you enough for your time, particularly, I know you're on the road. Colette, thank you again for your time, and appreciate your comments and thoughts earlier on. Simona and Stewart, thank you for your help in putting this together. With that, until next time. Thank you.

Gilad Shainer
Thank you very much.

such damages. S&P Global Market Intelligence's opinions, quotes and credit-related and other analyses are statements of opinion as of the date they are expressed and not statements of fact or recommendations to purchase, hold, or sell any securities or to make any investment decisions, and do not address the suitability of any security. S&P Global Market Intelligence may provide index data. Direct investment in an index is not possible. Exposure to an asset class represented by an index is available through investable instruments based on that index. S&P Global Market Intelligence assumes no obligation to update the Content following publication in any form or format. The Content should not be relied on and is not a substitute for the skill, judgment and experience of the user, its management, employees, advisors and/or clients when making investment and other business decisions. S&P Global Market Intelligence does not act as a fiduciary or an investment advisor except where registered as such. S&P Global keeps certain activities of its divisions separate from each other in order to preserve the independence and objectivity of their respective activities. As a result, certain divisions of S&P Global may have information that is not available to other S&P Global divisions. S&P Global has established policies and procedures to maintain the confidentiality of certain nonpublic information received in connection with each analytical process.

S&P Global may receive compensation for its ratings and certain analyses, normally from issuers or underwriters of securities or from obligors. S&P Global reserves the right to disseminate its opinions and analyses. S&P Global's public ratings and analyses are made available on its Web sites, www.standardandpoors.com (free of charge), and www.ratingsdirect.com and www.globalcreditportal.com (subscription), and may be distributed through other means, including via S&P Global publications and third-party redistributors. Additional information about our ratings fees is available at www.standardandpoors.com/usratingsfees.