Piper Sandler Webinar: Networks for AI

Company Participants

- Colette Kress, Chief Financial Officer, Executive Vice President
- Gilad Shainer, Senior Vice President of Networking

Other Participants

- Harsh Kumar, Analyst, Piper Sandler

Presentation

Harsh Kumar {BIO 3235392 <GO>}

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, whom a lot of you know already, and Gilad, SVP of Networking, and also a special thanks to Simona and Stewart, who made this all possible. So before we kind of get into the networking piece of it, I did have one or two 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 Kress {BIO 18297352 <GO>}

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 [ph] 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, we 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 there.

Harsh Kumar {BIO 3235392 <GO>}

Well, that's about as clear as it can be, Colette. Thank you for that. 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 Kress {BIO 18297352 <GO>}

Yeah, 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 Kumar {BIO 3235392 <GO>}

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 believed 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 Kress {BIO 18297352 <GO>}

It was not a software change. Our movement to A800 that it was absolutely a hardware change that we made to create the A800 as well as what we do to create H800.

Harsh Kumar {BIO 3235392 <GO>}

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

Gilad Shainer {BIO 17999373 <GO>}

Yeah, 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 NVIDIA -- Mellanox and now I think three years at NVIDIA more or less. And I started as a network designer. So some of the early InfiniBand devices, I was part of -- I did part of the design of them. And then actually start looking at the entire platform and 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 these statements, our presentation may contain forward-looking statements and please do refer to our SEC filings for risks and uncertainties facing our business. So with that, we can move to the next slide. So we are here to talk about networks for AI, obviously, but when we talking about AI, it's not just a bit 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 a NIC. It doesn't start at a switch. It starts at GPU levels. Memory IO on the GPUs, all the way through the network, and then we essentially started application framework, the application level.

It looks on everything there and that's the great advantage of NVIDIA. If you look on NVIDIA as a networking bar, 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 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 that compute the servers, compute the GPUs, the switches, the NICs, and so forth. Be able to build or essentially the ability to build the entire system, given the opportunity -- unique opportunity to place the rights data algorithms in the right place. There are date algorithms that you don't want to run on the GPU. We actually want to run them on the network and that's what we call the network computing.

There are elements that traditionally you may do it on the network, but if you can, you probably want to do it on the GPU, because it's going to be much more effective. So we are able to move on with this. We're able to build a very effective system that delivers the highest levels of performance at very, very large scales. So with that, understanding that you don't just build the network, actually built the entire data center, let's talk to you about the data centers and then dive into the network. So we all know that the data center is the computer today and in the past CPU was the computer and then the server will become the computer, now the datacenters 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 collection of GPUs, and actually, the way that you connect those GPUs, define the data center. The way that 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 it's open [ph] Cloud, Traditional Cloud. Traditional Cloud, it's a data center that is built to support many, many users and to support variety of workloads and most of them are very small. Even a single node workload, 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 rational. Now, we're actually facing the creation of new kinds of clouds. New kinds of clouds -- 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. AI workloads are not just running on a single node, AI workloads needs to run on multiple GPUs and each one 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 in the clouds actually requires elements to support distributed computing. Now we started talking about latency needs and day latencies and effective bandwidth, those are completely different kind of requirements. So Traditional Internet is still fine for the North-South traffic of Gen AI clouds. We need to use their access to services, control of the cloud, but you need a new class of ethernet to support the new class of foreclosing that new class of cloud, and that's exactly what Spectrum-X is. It's the first ethernet ground-up design for AI. We started interface so people can enjoy or utilize the ethernet ecosystem and it's actually combining both.

Now, if our data center mission is 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 variety of workloads, now we're talking about much less number of users and workloads is going to consume the entire GPUs in that system on a single application. So it's not a matter of how many GPUs you can connect, it's a matter how many GPUs a workflow can consume with that network, it's completely different thing. So now if you want to support workload, it'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, it's the combination of NVLink and InfiniBand. NVLink and InfiniBand are not the same as you think as internet, it's a completely different kind of architecture, designed -- specifically designed for distributed computing, being optimized over the years and that's 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 Internet [ph] and InfiniBand coexist and they will continue to coexist, because in every system that you built, you do need a network to do the user access, you deal with the network to run control and North-South traffic always and Ethernet is great for that, but for AI infrastructure you need a network for East-West. You need a network for the compute -- for distributed computing, this is where NVLink and InfiniBand actually go stronger. 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 understand those terms, so it's going to make life easier, and in particular, NCCL and Sharp. So NCCL -- what's NCCL, so NCCL is the short of NVIDIA Collective Communication Library. It's a software SDK. It's the software SDK for AI communications -- for AI communications, for multiple GPUs. Essentially, this software framework supports mainly two multi-GPU arrhythmics or 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 two operations, reduction operations and all-to-all operations. Now, NCCL, if we want to measure AI networking performance -- AI networking performance or networking performance for AI, NCCL is great --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 all-to-all 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 measures the network with. Sharp, it's a technology

part of E-network computing, NVIDIA E-network computing. It's a technology that implemented in the InfiniBand switch ASICs. It's not something that runs on a CPU, it's embedded within the switch ASIC. That enable the switch network to perform data reduction operations on the data as the data has been transferred within a data center.

Now, previously those data reduction operations which is part of NCCL were done on the host, and running them on the host has a big toll on the host, and this is going to be part of NVIDIA advantage, the ability to move algorithm from one side to another side and to run them in the right place. Moving to 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 is better than an 800 gigabit per second end-to-end network without Sharp. It's amazing capability of InfiniBand and that's one of the elements that enables InfiniBand or making InfiniBand the go standard for AI factories. So if you look on what is 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 a switch network compares to the best other network you can actually beat. So if would -- if you compare to the best theoretical performance on 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 for AI factories.

So now we can go back and talk about the different networks and we will start In the Cloud and then we go to Spectrum-X. So In the Cloud itself -- In the Cloud, there are two kinds of Ethernet networks, essentially two walls of Ethernet. There is a network that is doing the North-South connectivity. It is the controlled access and it's the user access. Those are the cloud services. Cloud services or user access, those are Loosely Coupled Applications. So you typically use TCP for the traffic, jitter is fine. In the 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 heterogeneous traffic. If 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 clouds, there is no much East-West traffic because most of the workloads, most of the user running on a single node, and therefore in a 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 do clouds for generative AI applications, now East-West network needs to be completely something else, because now East-West networks need to deal with disaggregated computing. And distributed computing is very sensitive to latency, but even more 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 500 GPUs, if one GPU communication is going to be late out of the 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 bandwidth is important and you want to provide constant performance, you cannot have changes in performance levels and you need to deal with burstiness [ph], so the requirements for distributed computing are completely different, I would say the opposite of what you need in 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 did Spectrum-X. That's the reason we designed Spectrum-X because we needed a new class of ethernet for this kind of infrastructure. Next slide, please. So now let's -- let's look on Spectrum-X. So Spectrum-X, on the left side you see all starts, 51.2 tera, the number of ports essentially and so forth. And on the right side, we can see a snapshot of the software that it's being developed for Spectrum-X. There is tons of software and SDKs. There is -- DOCA is the case that runs on the GPU and BlueField to provide the network utilization [ph], the isolation between the application infrastructure -- the applications in the infrastructure.

There are -- the Spectrum is the key for the switching. Magnum IO is the SDK that includes the NCCL framework that I mentioned before. Then you have the operating system that runs on Spectrum for switches, which are SONIC and Cumulus and other aspect like that. So there is tons of software with it. Now, Spectrum-X would essentially design ground-up for AI and we build 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 to drop packets. Dropping packets mean you're creating jitter and creating jitter and now you're reducing AI performance, so you don't want to 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 flow late adaptive routing in ethernet switches in traditional ethernet, which is -- flow by flow means that you need to -- you run the stream of data and you don't change the path of this stream of data before the stream end, that's not good for AI.

In AI, you want to do fine grain adaptive routing. You want to do packet-by-packet adaptive routing. So there'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 the buffers -- not the buffers. Their Ethernet options are, for example, they're sometimes referred to as fabric, not ethernet because sometimes they don't run actually ethernet, and those depend on the buffers. Big buffers in the switches to be shock observers. So, if there is congestion that can kind of hold data and stuff like that, the buffers mean long-day latency. Long-day latency is not something that is very nice for AI workloads. You don't want the buffers. So, now the idea here is combining lossless ethernet, find good [ph] adaptive routing and shallow buffers, that's the combination. This 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 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 we can react to hotspots before they can impact performance of applications. And this is important because this is the key to provide traffic resolution. 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 are running on the same network, but you want to make sure that you isolate the noise from the small workloads and they will not impact the AI workloads and that's exactly what we're doing with congestion control, an inventory-based congestion control and the capabilities on the different latency changes and identify hotspots before they actually can do a negative impact. What does it give us? It give us 1.6X higher AI Fabric Performance over traditional Internet.

So we're talking about not just 95% effective bandwidth at skill and under loads, but keeping that performance constant, predictive performance, keeping that performance constant even that you have a lot of other workloads running in the same environment because we did it with cloud. I think the security, the virtualized network, everything is part of that. So now Spectrum-X actually bring 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 through to support larger-scale of AI workloads, this is where we go to InfiniBand. InfiniBand, starting to 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. Ethernet was built for wide-area networks and over time within a data center more and more algorithms were designed for Ethernet -- more and more algorithms were designed for Ethernet.

PFCs for example and BGP is more moderate and that were designed for Ethernet. So Ethernet is a very complicated protocol. It's a very complicated protocol. And when you build an Ethernet network, you need to actually 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 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 ports, but no much of good performance for distributed computing and just supporting kind of cloud interfaces. There are switches with buffers in order to support sometimes DOCA and service applications that come in issues of day latencies and reduced number of ports and so forth, so you need to choose between features, performance, and other stuff. The Spectrum-X -- Spectrum-X is actually designed to have the right elements (inaudible) 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're not using the same architecture.

InfiniBand was designed from the beginning to support distributed computing. And for 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, kind of freaking terms of Ethernet, there is leaf and spine, and in Ethernet you try to build two-level of networks, two-level of switches, and don't go beyond that two-level switches and stuff like that. It does not exist in InfiniBand. There is no such thing in InfiniBand. InfiniBand, you can use as many switches if you want. Even more -- most of the large-scale systems out there are using three levels of switches in InfiniBand. Some systems even use four. If you want to use five, use five. There is no performance penalty there. There is no issues around that, you can build any size of system that you want. It's like, you're designing here a Formula race car. So if you are designing a Formula race car, how many seats are going to put in this car, who cares. So it's a different kind of design.

And if you look at three-level of switches with InfiniBand, which is what most systems use today, that can go all the way to 65,000 GPUs. And if you go to four levels, we have several four levels and multiple four levels already out there, you can go to two million GPUs in InfiniBand network. You want to go five, go five. 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 point. So there is no limit there and that's why InfiniBand is gold standard for large-scale AI. Now, InfiniBand pioneered RDMA obviously, so there's lot of elements in RDMA but InfiniBand pioneered RDMA within [ph] full computing. And we saw the impact of Sharp. Sharp gives you 1.7X on NCCL, compares to the best Ethernet network you can build. And it's a Pure software-defined -- Pure software-defined network. It was designed in SDN before people knew what SDN means or what SDN is, which means 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.

Reconfigure the network of course [ph] is down, ports are up, you can configure it very, very quickly, there's a huge amount of benefits in Pure software-defined network. So what that gives us? If you look on the total performance, small in 2X, being gracious here, is many GPUs as you want, and building a network that have the loss latency, again under -- in large scale and under loads. Very short latency, extremely short latency. We know the impact of Sharp on NCCL operations, nearly 100% effective bandwidth at scale. It's an amazing network. It's really amazing network, and it's being developed over -- more than 20 years, right? It's every generation bring new capabilities. The upcoming Quantum-3, the things that we were 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 tonnes of software, right? We have the SDK elements there. Magnum IO and NCCL are two, obviously, management of the network, able to simulate everything. There's 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. 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 highest scalability out there. Again, you can build any size of system that you want with it, 3

levels, 4 levels, 5 levels, there is unlimited number of GPUs you can connect together. And then if you'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 its enabled the Ethernet ecosystem, right, so if you want the joint ecosystem and you need performance for AI it's Spectrum-X. And then if you want to build the system that's going to go to scale, if you want to get the highest level of performance, you can also bring InfiniBand to the 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, even if someone is going to offer me traditional Ethernet for free, completely free, it's not going to be good enough, right, or they're actually paying good InfiniBand because I'm going to get much better -- much better from that, right? So, essentially building an AI infrastructure network that is essentially free. So with that next slide, I think this is the last one, yeah. So if I look -- looking on networking revenues -- NVIDIA networking revenues, 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 tripled, so it's growing, growing very fast, continue to grow. We will continue to grow. Then Spectrum-X, Spectrum-X is new class of Ethernet -- new class of Ethernet that is needed for new class of clouds. And therefore, Spectrum- X will boost the cloud AI network market and will increase the market and will increase the Ethernet revenues as we're moving forward.

Overall, we believe -- 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 if they are not accelerated. We used to be in a situation that we got 2x performance every two years, just doing nothing. That doesn't work anymore, that doesn't work anymore. So you know if you want to be able to increase capabilities, it's accelerated computing and therefore, every data center will become an accelerated datacenter. Every server will have a GPU processing unit, every datacenter will have an element there.

And as such, we are talking about a $60 billion market opportunity for NVIDIA on the networking side. And with that, first thank you for listening. That took some time. And we're happy to answer questions.

Questions And Answers

Q - Harsh Kumar {BIO 3235392 <GO>}

Yeah, hey, Gilad. Thank you so much. That was extremely informative, actually answered a whole bunch of 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 are basically tied to NVIDIA a lot and I think, people as you know in semiconductor

business, in IT, they always want options. Could you maybe talk about what is -- is there a work around to that or are you the only ones that makes InfiniBand, or is it farmed out to other places that do it for you.

A - Gilad Shainer {BIO 17999373 <GO>}

Yeah, well InfiniBand, InfiniBand it's a standard technology. It's not a proprietary technology standard. It's the same like Ethernet. Ethernet 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 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 is always no guarantee 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 -- and the question is and understanding what's in about it. Essentially -- especially when you look on AI, when you look in AI, AI requires datacenter skill. And if you look on that then actually you want to have the right elements inside. I've said it before we used to get 2x performance every two years. Now it's not the case.

And therefore, we're going to see more-and-more specialized technologies and actually more uses 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. You 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, or availability [ph]. Time to build and deploy it in large-scale architectures. And this is what NVIDIA delivers.

So we will deliver full [ph] 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, so whether they wanted to take pieces of that and mix-and-match other things that happen in the market -- that exist in the market.

Q - Harsh Kumar {BIO 3235392 <GO>}

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

A - Gilad Shainer {BIO 17999373 <GO>}

No, yeah it's definitely good question, so training -- training requires very large-scale clusters right that are tightly coupled and optimized for massive, massive data

compute. Inferencing typically required much smaller -- smaller scale clusters. But, what's happening now is that generative AI is becoming mainstream. And therefore, the number of the separate jobs running inferencing will dramatically increase. Therefore, inferencing will require a larger number of accelerated servers and the flexibility essentially to do that. And we probably going to see people that going to deploy system and they will want to use those system 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 just going to inferencing 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 system that are going to use for both, makes sense to build system that are used for both, InfiniBand is a good option for them.

Q - Harsh Kumar {BIO 3235392 <GO>}

Great and I have one more for you, Gilad. There is a perception in the investment community even the people that know generative AI very well that InfiniBand only works with NVIDIA's GPUs, is that accurate or listening to your talk, it seems like that's not the case, but I wanted to ask you, since you are the expert on the topic.

A - Gilad Shainer {BIO 17999373 <GO>}

Yeah, 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 as a whole. 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 is definitely not tied[ph] Now obviously end-to-end is a lot of benefits into that right? We invested lot of effort, a lot of effort and we investing that so our customers will have a much faster time to computing, much faster time to solution, much faster time to build their system.

When you build a supercomputer, an AI supercomputer, you don't want to spend nine months to build it. That's nine months out of the lifetime of very expensive system. So we're doing what we are doing, 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 within the other compute element and so forth.

Q - Harsh Kumar {BIO 3235392 <GO>}

Wonderful, and I know, you're on the road, so I want to be very mindful of your time. We've got two minutes, so I'll just ask one final question. In a typical setup, let's say, as you guys go and deploy accelerated AI datacenter. Do you typically find the entire datacenter 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?

A - Gilad Shainer {BIO 17999373 <GO>}

Yeah. So first obviously there are entire Ethernet systems out there, we all know that. There are system 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 Ethernet 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 Ethernet 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 speeds and each [ph] that 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 took in large-scale then we need systems that are essentially combining both InfiniBand and Ethernet, it's not one versus the other, it's completely -- they are 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 -- for that interface, we have the Ethernet. And then for the compute fabric, once you want to connect large amount of GPUs in thousands to tens of thousands to hundred of thousands of GPU in a single workflow, InfiniBand actually gives you or the combination of NVLink and InfiniBand gives you the connectivity -- the connectivity there. So if you look on systems that we design and system that for example we recommend and if you look what we did and coping that will enable to leverage everything we designed, our system includes both InfiniBand and Ethernet completely coexit. It's not that one replaces the other. I think, you have those networks, they both exist and each one has its own purpose.

Q - Harsh Kumar {BIO 3235392 <GO>}

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 early on. Simona, Stewart, thank you for your help and pulling this together. With that, until next time. Thank you.

A - Gilad Shainer {BIO 17999373 <GO>}

Thank you very much.