Semiconductors

Jef U with Transcript: ChatGPT and AI Systems Expert Call

February 13, 2023

Key Takeaway

We hosted a Jef U call with expert, David Driggers, CTO at Cirrascale, which was the cloud partner to OpenAI prior to 2019. Key Takes: 1) Generative AI training systems cost $250m w/ 10K GPUs at the high-end, and $10m w/ 512 GPUs at the low-end; 2) BOM Cost of GPU Server is 70-80% GPUs, less than 10% for CPUs; 3) Generative AI Inferencing is wide-open, likely on unutilized x86 CPUs for ChatGPT; 4) NVDA is solution of choice bc of its HW+SW ecosystem.

NVDA Current Solution of Choice - 80/20 Rule in Generative AI? Driggers views NVDA as the solution of choice for today's generative AI training, and believes that GPUs will be 80% of generative AI training in the future, while other solutions (e.g., Graphcore, SambaNova, Cerebras) will end up being the solution for the 20% at the edges. Driggers believes that AMD has a chance to gain share if it develops its ecosystem.

Interconnect Significant Component of NVDA's Value Proposition... High speed interconnect solutions from NVDA (NVLink, Mellanox Infiniband) enable scaling of nodes/clusters, and are a key part of NVDA's hardware ecosystem. These are the 2nd most expensive component after the GPUs. NVLink switching connects GPUs inside a server box, while Mellanox Infiniband solutions connect GPUs between boxes.

...As is NVDA's Software Ecosystem. Sitting on top of the GPUs is a software ecosystem that includes compilers, libraries, primitives, containers and pre-trained models, that enable high level code to run efficiently on the GPUs. NVDA has taken a differentiated approach by writing its own libraries (e.g., NCCL for communicating between GPUs/ Nodes/Clusters), and optimizing existing HPC libraries (e.g., cuBLAS for linear algebra) for its specific GPUs; thus, NVDA has the most robust software ecosystem in market.

ChatGPT Training Engine Likely $125m-$250m. MSFT recently reported that the supercomputer developed for OpenAI has more than 10,000 GPUs, which would cost $125m-$250m, depending on the generation of GPU. While ChatGPT is a large generative AI engine with over 150bn parameters, smaller, subject specific (e.g., healthcare, legal) generative AI engines are emerging that can be much cheaper, e.g.,: **Generative AI Training System w/ 5-10bn Parameters Est at $50-$100m,** and would consist of 4,096 GPUs at $12.5K-to-$25K per GPU card; **Generative AI Training System w/ 1.5bn Parameters Est at $6m-$13m,** and would consist of a single 64-node cluster, or 512 GPUs. A node is 8 GPU cards + 1 CPU card and is also called a "server" or "box."

For Inferencing, ChatGPT likely uses Cheaply Available x86 CPUs. For text generative platforms, CPUs are good inferencing engines. **Image Generative AI Output Platforms (DALL-E 2) More Likely to Use GPUs... But Inferencing Is Still Wide Open**. Power efficiency considerations become more important as inferencing is done at the edge.

High Power Requirements. The new servers dissipate 10,000 watts. Building a 64 node cluster requires not only the base cost of $16m (64 times $256k per node) plus the power requirements of 64 times 10,000 watts. The power in this example is 2/3 of a megawatt.

EXPERT STUDIES

USA | Semiconductors

GUEST SPEAKER

David Driggers

CTO and Founder, Cirrascale Cloud Services

HOSTED BY

Mark Lipacis

Jefferies Senior Semiconductor Analyst

Link to replay here (until 21-Feb-23)

Links to Related Research:

10-Feb-23: Semis on Fire: ChatGPT, Cloud CapEx, ON and PI Earnings

9-Feb-23: Hyperscale 4Q Commentary

9-Feb-23: NVDA: Raise PT to $275

3-Feb-23: Cloud Instance Dec'22

29-Jan-23: 4Q22 Data Center Leasing

6-Jan-23: Notes from NVDA Q&A

10-Jul-17: 4th Tectonic Shift in Computing

Mark Lipacis *
Equity Analyst
(415) 229-1438
mlipacis@jefferies.com

Brent Thill *
Equity Analyst
(415) 229-1559
bthill@jefferies.com

Janardan Menon †
Equity Analyst
+44 (0) 20 7029 8413
jmenon@jefferies.com

Brian Chen, CFA *
Equity Associate
(415) 229-1478
bchen1@jefferies.com

Jefferies University ‡
Equity Research Team
jefferiesuniversityreport@jefferies.com

Please see analyst certifications, important disclosure information, and information regarding the status of non-US analysts on pages 18 to 24 of this report.

* Jefferies LLC / Jefferies Research Services, LLC † Jefferies International Limited ‡ Jefferies Group LLC

Jefferies University: Insights from Experts

Jefferies University: Insights from Experts

Jefferies University is an executive-level education program that connects independent experts with institutional investors on matters critical to making informed investment choices. Through this global program, Jefferies' clients have the opportunity to meet thought-leaders, innovators and scholars on a wide range of subjects that may be relevant to building investment portfolios. Topics can be tailored to suit your needs and delivered in different formats including private or group conference calls, in-person presentations or via video-conference as well as at our numerous conferences and summits held around the world.

Jefferies University: Exponential Series

Jefferies University courses go beyond investment matters. Our **Exponential Series** also delves into issues that can impact investors on a personal level. Join us for sessions that may unlock meaningful insights relating to your health and well-being, personal development, family dynamics, professional growth and significant cultural issues.

Restricted to Jefferies' clients, the program will continuously adapt to meet your needs and help identify the latest industry trends, investment ideas and macro developments that matter to you most.

Important Disclosure

Any views or opinions expressed by the third-party experts are solely those of the speaker, not Jefferies, and this does not represent an endorsement by Jefferies of the views or opinions expressed.

For more information, suggestions and feedback, please contact JefferiesUniversity@Jefferies.com.

David Driggers

David Driggers is the Chief Technology Officer and Founder at Cirrascale Cloud Services, a premier developer of hardware and cloud-based solutions enabling GPU-driven deep learning infrastructure. Cirrascale architects deploys deep learning infrastructure for customers, as well as manages a GPU as a Service (GPUaaS) offering using hardware and software infrastructure optimized for deep learning algorithms enabling data scientists to focus on application development and optimization. Since Feb 2017, this advisor is Chief Technology Officer at Cirrascale Cloud Services. From Jan 2010 to Feb 2017, this advisor was CEO at Cirrascale. From Jan 2001 to Dec 2009, this advisor was CEO at Verari Systems.

Definitions of Terms Commonly Used in Transcript

- The term node is used interchangeably with box and server.
- One node is a single server, a single unit with 8 GPUs in it. A node is considered the most basic building block that's used.
- A cluster is a collection of nodes that are usually in the same network. They are usually used to split a workload or data among themselves.

Edited Transcript

Introduction and Speaker Background

Event: Jefferies Hosts Conference Call with ChatGPT / GPU-Machine Learning Infrastructure Expert

Date: February 08, 2023

Lipacis: Hi. This is Mark Lipacis, Senior Semiconductor Analyst from Jefferies. Welcome to the call today that we're hosting with on generative AI infrastructure. I'm joined today by my teammate Brian Chen, who co-covers the logic and processor companies with me as well as by our featured guest David Driggers. David is the Chief Technology Officer and Founder of Cirrascale Cloud Services, which is a premier developer of hardware and cloud-based solutions, enabling GPU driven deep learning infrastructure. Cirrascale deploys deep learning infrastructure for customers and manages a GPU-as-a-service offering using hardware and software infrastructure optimized for deep learning algorithms.

The expert, David Driggers, is the CTO and founder of Cirrascale, a cloud service provider focusing on AI and HPC accelerated workloads predominantly using NVDA GPUs

Since February 2017, Dave has been the Chief Technology Officer at Cirrascale Cloud Services, between 2010 and 2017, he was CEO at Cirrascale, and from January 2001 to 2009 he was CEO at Verari Systems. We're going to host a fireside chat with Dave for about 30 to 40 minutes, and then the balance of the time we'll take questions from the listeners. In your web browser interface to this call, there should be a place to enter questions and I will get to as many of those as we can. First I need to read the disclosure and then we will get on with the call. Any views or opinions expressed by the speaker during the conference call are solely those of the speaker, not Jefferies. And this does not represent an endorsement by Jefferies of the views or opinions expressed. Members of the media and the press are not authorized to be on this call. But if you are from the media or the press, please do not register or dial-in.

The content presented on this call is proprietary to and/or subject to copyrights of Jefferies or third parties. As a matter of legal compliance, we remind you that you must not attempt to elicit from any speaker at this event any material non-public information or other confidential information, and accordingly, the speaker may decline to respond to any question in his or her sole discretion. You may not publish or otherwise publicly disclose the name or otherwise identify the speaker's unless Jefferies permitted in writing. By attending this event, you agree to all of these restrictions.

Okay. So with that, first of all, let's get on with the fireside chat. David, welcome and thanks for joining us today.

Driggers: Hi, thank you.

Cirrascale worked with OpenAI up until MSFT invested $1bn in them 3 years ago

Lipacis: David, why don't you just start off and give us a brief overview. What is Cirrascale and what are your roles and responsibilities there?

Driggers: Sure. Cirrascale is a boutique cloud service provider that really focuses on AI and HPC workloads leveraging accelerators. The predominant accelerator using today are GPUs from NVIDIA, and we build clusters or large scale systems that typically are made up of at least 8 GPUs per node, but then we cluster those together in larger amounts, so they can be used by a single customer.

Cirrascale builds clusters or large scale systems made up of at least 8 GPUs per node, which is equivalent to a single server

Lipacis: Okay. So we're going to start with a new lexicon for a lot of people, including me on this call. So a node is a cluster of 8 GPUs?

Driggers: So a node is a single server, a single unit. It has 8 GPUs in it. So that makes up a node. So that's the basic building block that's used.

Lipacis: And you had done some work with OpenAI and ChatGPT before. Could you just give us an overview? When did you work with them? What did you do for them? And when did you stop working for them?

Driggers: Yes, we were lucky enough to get to work with OpenAI when they had first started when they had came really as a two man, four man shop out of the brainchild of Elon Musk and a number of others. And we work with them until they were leveraging thousands of GPUs with us. And frankly, up until Microsoft invested $1 billion in them. They were one of our largest customers in our cloud.

Lipacis: Okay. And roughly how long ago was that?

Driggers: Three years ago.

Generative AI and the Training Process

Lipacis: So if we start from the top down, I just want to level set because we have a number of investors, there is wide range of expertise and information, and this is a new topic for a lot of people. What is a generative AI engine?

Generative AI is when AI creates content on its own, either images or texts

Popular Generative AI image engines include DALL-E 2, Stable Diffusion and Midjourney

The most popular language model is from OpenAI called ChatGPT

Driggers: So what makes a generative AI engine different than many of the others is it's a term that's being used to coin when AI makes AI when you actually create content – when an AI creates content on its own versus being directed to. I mean, it is being prompted to create it, but it's actually creating it almost out of the ether, out of a trained model. So, generative AI, the two most popular examples right now for generative AI that I've gotten a lot of attention are image generation which is coming from – OpenAI has a platform called DALL·E 2. There is a platform from Stability called Stable Diffusion that is open source that is allowing a lot of companies to leverage a model to actually create their own content. And then a third company called Midjourney is also right now super popular method of generating these images. And then of course, the one that's been all the rage most recently is ChatGPT, which is the latest language model from OpenAI in Microsoft.

Lipacis: Can you help us understand how do you train these models to do what they do? And if it's different for text versus images, I think that would be helpful for people.

Driggers: Sure. Well in the case of the images, there is actually text associated with it as well. So it's actually more than one model. It's not just a single model doing something. They have a natural language model that actually is used to understand the text that you're trying to generate the image from. So there you have to do the text generation first to understand the text model first to be able to understand the prompt, but then you're inputting billions of images into the training model, so that the AI understands, can associate the text to the image. In a text situation like ChatGPT or GPT-3, you are putting in tremendous amounts of structured text, un-structuring it so that you can understand how words go together, how phrases go together, how sentences go together, ultimately how concepts go together. So massive amounts of different texts, structured texts go into the training of a large language model.

Popular sources for training large language models include Pile, PubMed, Hacker News and the Bible

Lipacis: And maybe just take that one layer or one half of a layer deeper, mechanically, how are you doing the training and what kind of sources would people use to train something like a ChatGPT?

Driggers: Sure. So one of the more popular data sets that's being used right now for training natural language models is called the Pile. And the Pile is an open source collection of a number of different texts from medical texts that are public, PubMed is a large collection of medical texts; people use Hacker News, so it's an open source, emails back and forth, so it's got a different style; the Bible is open source, so often it's used as a reference guide. So, we take lots of different styles and types of texts so that we can understand and then emulate how humans communicate. Is it by verse, is it by small snippets, email, text, things? Do we have older styles of language? Do we have slang that can be incorporated?

So the more diverse the types of text, the more an AI can have a better understanding of how many different ways the text can be structured. So, why it's called the Pile is because it is a pile of different open source texts. If you use something that is copyrighted, you might wind up getting sued by somebody like Getty, for instance.

ChatGPT trained for months on 10s of 1,000s of GPUs

Lipacis: Can you describe the training process, is this something that is continually trained? How long did it take to train? Do you look at a time series or an evolution over time, does the engine adjust to that?

A cluster of 4,000 GPUs can be used to train a 5-to-10 billion parameter model

GPT-3 has 175 billion parameters, and is an evolution of ChatGPT

Driggers: The original training wants to get to a level of accuracy where you have a usable product. So, you will do continuous training until you get to a level of accuracy that is at the desired level. You can then – after a model is fully trained, additional layers or additional information can be added to it through what's routinely call fine tuning. But in general, the first thing you want to do is that you've got a goal of how accurate a model needs to be. And that's that initial training. You've decided, I've got this much data, I've got this many different types of queries, I'm going to try to push against that data. And that can take a long time.

In the case of something like ChatGPT or GPT3, you are talking months of training on thousands of GPUs, if not tens of thousands and many months. I mean, yes, I won't say it's a direct correlation on how long it takes to train based upon how many GPUs there are, because we have a law of diminishing returns on really large models, we have lots of communication between the nodes. So it's, everything doesn't happen in that one node, it's collections of nodes. I mean, we've recently for companies that are looking at doing these types of models we've recently quoted clusters where the base cluster is 4,000 GPUs all tightly connected. And that's to do a five to ten billion parameter. GPT-3 is 175 billion, for ChatGPT– OpenAI has not publicly said how many, how large the actual model is, but it's an evolution of the GPT-3, which was 175 billion. So the numbers of computers necessary and the number of GPUs to train these large models is pretty tremendous.

Lipacis: Okay, let's get to the infrastructure here. So I think what would be interesting for people listening is if you describe the evolution of the buildout of the specific solution that you selected and when we were speaking the other day, I thought the story that you were explaining about how you hit different bottlenecks at different places and then you solved those bottlenecks with the different approaches. Could you take us through that evolution, the story, how did you start, how did you scale it? What were the challenges? What were the solutions to address the challenges of the bottlenecks?

History and Challenges of Scaling and Building out the GPU System for Generative AI

Driggers: Sure. So one of the things that people ask often is where this come from? How come –I had heard of neural networks a long time ago, 20 years, 30 years ago. Why are we able to do this now? And probably the simplest reason why we're able to do it is that very first building block that gives us the right balance of the number of compute resources to a large enough and fast enough chunk of memory is the GPU. GPUs have a lot of cores, the latest GPUs have well over 10,000 cores, that's 10,000 little processing units, and they have significant amounts of memory. So we had a breakthrough about eight years ago as it related to the number of cores and the amount of memory so that we had enough memory and enough cores and everything was fast enough to be able to start building these neural nets which is kind of a model of how our brain works.

Neural Networks have been around for over 30 years, but have only become functional recently, because the GPU finally evolved to a point where it had the right balance of processor and memory resources to train a neural network

We've got lots of different synapses and lots of communication and a decent amount of memory. So GPUs allowed us because it was a significant shift from a CPU, which has very few cores. Even the very latest and greatest CPUs only have 64 cores, whereas GPUs, as they say, have thousands of cores. But GPUs originally had too small of a memory, so it really took GPUs getting to where they had enough memory associated to them, that memory was close enough and fast enough to the cores so that we could start using GPUs for general processing versus graphics processing. And that allowed for a model within the GPU and allowed us to start creating these neural networks. But then ultimately we could only make it as big as that GPU was.

The GPU, with 1,000s of processing cores is better for neural networking as compared to a CPU, which has 10s of processing cores

When GPUs added more memory, they started to get used for general processing and neural networks (NN), but the NN could only be as big as a single GPU

And so what we needed to do was start expanding outside of the GPU so we could break the tasks up and go to multi GPU. And so that was the first, first big step was how do we parallelize and get outside of a single GPUs so that we can do even more tasks. And then ultimately that starts to become a bottleneck. As you scale the number of GPUs in a box, the actual communication back and forth between the GPUs becomes a bottleneck.

When programmers broke tasks up and spread them across multiple GPUs, communications between GPUs became a bottleneck...

And frankly, NVIDIA has done a phenomenal job there by creating the NVLink, which dramatically improves the speed of the communication between the GPUs. And so now the primary building block we're using is eight GPU box that has this tight coupling of the GPUs in the box.

...so NVidia introduced NVLink which dramatically improved communications between GPUs...

...which led to the 8-GPU "node" to become the primary building block of Neural Network training

Over the last number of years, the GPUs have gotten bigger and stronger, and we've gotten more memory on the GPUs. So the individual models that are occurring at that first initial building block have also gotten to be far bigger. And technology in general has been marching at a drumbeat of almost doubling individual pieces every couple years.

The challenge comes when one piece doubles and another piece doesn't double, and then we have a diminishing return of scaling. So we do go in fits and starts. We are at a unique time with the latest introduction of technology in that the hosts are doubling in speed, so the CPUs are faster. We've got a new PCIe bus, which is twice as fast. We've got interconnects, the networking that connects, the GPUs as well as the external interconnects, which connect the servers to each other. Pretty much across the board,

The industry is in a unique time because all the components of a neural network system have been improving performance we're seeing a doubling or tripling of the GPU architecture. And so the technology we're using is incrementally growing quite nicely.

In addition, because of the boom of AI, there's a lot of investment in alternate AI platforms that's happening as well, that are trying to solve these same problems of scaling that we're going to continue running up against as we try to do bigger and bigger and bigger models.

Lipacis: So your solution, you decided to go with NVIDIA, right?

Driggers: Yea, NVIDIA. AI didn't just pop up out of nowhere with using GPUs for AI, it actually came out of HPC. So traditional high performance computing, oil and gas, federal government, media, entertainment, financial services, where they do heavy, heavy computations, lots of math. And those guys have been using GPUs for quite some time. So we already had the building blocks from a system perspective of how to couple these machines and starting to see where the bottlenecks were because of HPC. So we had a ground, we didn't just start AI on GPUs without having already done advanced HPC type applications on GPUs.

The HPC community had been using GPUs for years as a computing solution for Oil & Gas, Government, Media, and Financial Services markets

Lipacis: When you talk about a node being eight GPUs in a box, I guess the box is the server, is that that right expression for it?

Driggers: Yeah, the box would be the server and then like you say, once your problem that you're trying to solve gets too big to fit in that one box, you then start trying to parallelize and connect boxes to each other.

A "Node" is eight GPUs in a box or "server"

Larger computation problems require multiple nodes that connect to one another

Lipacis: And then I think a lot of people think about going into a data center and then they see this rack, and then you'll have a place to put a bunch of blades here, and that would be one box, and then there's like another box on top of that. And another box is, is that how the architecture works?

Driggers: Yeah, you typically want them as close as, once you get to the level that we're at, you want them as close as you can get them because the inner communication starts becoming the bottleneck. And therefore you want to have a lot of inner communication. And as the systems get bigger, that communication becomes a much bigger problem to solve. NVIDIA did something really bright a number of years ago in buying Mellanox because that's the primary way that the systems are connected today for larger clusters is using InfiniBand as Mellanox is the only manufacturer of InfiniBand.

NVidia made a smart move when it acquired Mellanox...

...because the primary way GPU nodes and clusters are connected together is through InfiniBand, and Mellanox is the only maker of InfiniBand

Lipacis: Staying on this architecture segment. You talked about memory, the GPUs getting more of memory. Are you talking about memory on the chip or memory on the card or the blade that you would plug in?

Driggers: Memory actually on the GPU.

Driggers: What's really happened in our industry, especially from an AI perspective is the host on a server, which would've either been an AMD or an Intel processor in general, has become just that a host. It's only a method of connecting the GPUs and then how to enable connecting those GPUs to other GPUs in another server. The host has become a lot less important than it was in the past.

Server Bill of Materials (BOM): Percentage of Cost to GPU vs CPU

Lipacis: And sorry for interrupting, I just want to make sure everybody's on the same page. So when you say the host, do you mean the CPU?

The CPU or "Host" has become much less important than the GPU in neural networking

Driggers: CPU, the actual server itself, the CPUs and the memory that are attached to CPU is what we would call the host for the GPUs.

Lipacis: Okay. And so just to visualize this, so you have box, you have eight cards with GPUs on it with memory and then you have a card that's a CPU that you need to have with the eight GPUs. Is that one or two CPUs on that card?

The Bill-of-Materials (BOM) of a server used to be 70% CPUs, but now it is 70%-80% GPUs

Driggers: Yeah, one or two CPUs on the motherboard connect to the eight GPUs. I mean, in the past, to give you an idea what we call BOM, bill of materials, servers in the past were predominantly CPU driven. CPUs would dominate the cost of the BOM. 60%, 70% of the servers cost was related to the CPU.

In the GPU world and AI world, the BOM is dominated by the GPUs. The latest and greatest servers the H100s from NVIDIA H100s 70%, 80% of the BOM is the GPU. The CPUs make up less than 10%. So CPUs have become very insignificant as a part of the dollar perspective of the equation.

Price per GPU card and per H100/A100 box

Lipacis: And when we talk about the bill of materials, I think a lot of people think about the GPU cards being like $10,000 or $20,000 each. Is that a fair ballpark to think about the cost here?

NVIDIA's H100 costs $25,000 per card, more than 2x the previous generation, A100, which cost $12,000 per card

Driggers: They're actually continuing to go up. NVIDIA has a very strong position within HPC and within AI. And so there's been a significant increase in the cost. To give you an idea the A100, the predecessor was about $12,000 a card. And the H100 is more than $25,000 a card.

A brand new, 8-GPU server using H100 cards from NVidia costs $250,000

So it's more than doubled the cost of the individual GPU from previous gen to the current gen, which is unusual. Usually, we see an incremental cost, we see a significant gain in performance, and then a 20% from one generation to another. We typically see about a 20% increase if there's significant performance gains. And that's historical for CPUs, GPUs, all different types of technology. So this is a pretty big jump that we're seeing that, that significant jump in the GPU price.

Lipacis: Right. So a brand spanking new box or node with eight GPUs and a card with CPU on it would be 200,000…

Driggers: With the H100s, we're looking at $250,000 a server.

Lipacis: Okay.

Driggers: Pretty expensive building block.

Value of NVDA's Interconnect Solutions (NVLink and Mellanox Infiniband) - "The Hardware Glue"

Lipacis: But it seems like this is the solution of choice and there must be a reason that people are buying this solution of choice. Can you talk about those reasons and maybe you mentioned two and then the other day you had mentioned some other on the software side, but could you just explain, you got NVLink and then you mentioned Mellanox and InfiniBand specifically what is that bringing to solution, and I imagine these are the elements that motivate people to pay $25,000 a card.

NVLink and Infiniband provide the glue that allow eight GPUs to connect to one another in a fast manner

Driggers: Or motivate them to pay $250,000 for the server, because those things provide the glue, the hardware glue, let's not say software in this case because software's a whole another reason why people choose NVIDIA. The NVLink and the InfiniBand provide the glue that allows you to connect the eight GPUs to each other really in a very fast manner because we don't even use the host processors to do that.

So the CPUs aren't responsible for connecting those GPUs to each other. Its NVLink switching that NVIDIA has developed that allows for a very high speed and very low latent communication between the GPUs and that's inside the box. And then to connect outside the box is the Mellanox, InfiniBand and it's another key feature. Again, we're spending more money on the interconnect than we're spending on the CPU as well. So the next biggest piece of the BOM after the GPUs and NVLink is the Mellanox to connect the servers. So there's an awful lot of NVIDIA content and NVIDIA enablement to allow us to scale these machines, both the single machines and then into very large clusters.

NVidia's NVLink switching allows for high speed and low latency communications between GPUs inside the server or node

The biggest piece of the BOM after GPUs and NVLink is Mellanox

Lipacis: And then how does NVLink manifest? Is that a chip on the board that does the switching in between the other cards?

GPU cards plug into NVLink switch chips inside each server

Driggers: It's chips within the server, within the actual box itself that helps the GPUs to connect. So it is separate chips, NVLink switches, which the GPU plugs into, plugs into a switch. That switch enables the connections to the other GPUs. Within that one box, so all that NVLink does is connect those GPUs to each other, nothing else.

Lipacis: Right. And then the Mellanox, InfiniBand solution that's connecting the boxes between one another?

Mellanox Infiniband switches connect GPU nodes together

Driggers: One box to another box in a very high speed, high throughput manner. So that's what enables the scaling of the cluster from GPU to GPU and then from GPUs to GPUs outside of that box.

Memory Considerations

Lipacis: Okay. And while we're talking about this on the hardware side before we go to the software side, what kind of memory configuration do you have on each one of the cards? What kind of memory do you use?

Driggers: So NVIDIA for the top end platform, think about a high performance race car. Every piece of that car matters. The tires, the transmission, the engine, the fuel, everything matters, because if one thing is too small or one thing is not strong enough, it breaks. And in the case of GPUs and HPC and AI computing, the same thing exists.

As the actual GPU chip gets stronger, we need faster and more memory. In the case of the NVIDIA cards, they're using the fastest memory that's available, that's outside of the chip. Other technologies have started to say, well, we need more balance. So we'll put fast memory inside the chip, really close to the cores.

NVidia uses the fastest memory available for its chips, called HBM3 (high bandwidth memory version 3)

In the case of a GPU, we've got memory that sits right next to the chip. There's more than one type of memory. But what NVIDIA's using in their top end is what's called HBM, which is high bandwidth memory. It's the most expensive separate memory that you can have on a system. And that's what they use on the higher end cards. And they use standard GDDR, which is designed more for traditional graphics on the mid-level and lower level GPU cards. So they've actually selected specific memory for the highest level of compute that they have on a card.

Each NVidia H100 and A100 uses 80GB of HBM3, which consists of 16 memory chips, stacked in 6 separate columns on the GPU, using 2.5D and 3D advanced packaging techniques for low latency connection to the GPU processing cores

Lipacis: And HBM is like a DRAM is that kind of like a DRAM or it's not that right way to think about it.

Driggers: Yeah. Unofficially, I'd say, yes. It is a separate memory rather than it being inside the chip, which is typically called SRAM. This is sitting outside and it is direct attached. It is attached to the GPU. It sits next to the actual GPU itself and feeds the GPU.

Lipacis: Right, okay. And roughly how many gigabytes of this are there?

Driggers: On the latest and greatest cards, and this didn't change from A100 to H100, much to the chagrin of the programmers, because they always like more, we're at 80 gigabytes of memory from A100 or the H100.

Interconnectedness of Entire Cluster; Cost per GPU Allocated to Communication Outside of Box

Lipacis: And then you described the Mellanox InfiniBand connection between the boxes and a rack. How about between the racks? A lot of people have this idea of a top of rack switch and…

There is a 400Gbps Infiniband connection dedicated to each of the eight GPUs per server (node), which costs $2.5K-$3.0K per link

Each GPU connects to each of the other seven GPUs inside the server via NVLink...

Driggers: So it's all to be InfiniBand and so each one of the highest end servers are actually using eight connections. So there's just a discrete connection allocated to every GPU. So we have as much bandwidth as we can from GPU to GPU as well as from that GPU leaving, we have got a very high end – the latest will be 400 gig link per GPU coming out of the box. So we've allocated $2,500 to $3,000 per GPU just for communication out.

Lipacis: Per GPU?

Driggers: Per GPU.

Lipacis: Okay.

Driggers: Because when we start building these large networks, one GPU is talking to another GPU and that's either inside the box or outside of the box. It's not eight talking to eight, it's one talking to another one, talking to another one, talking to another one. There each one talks to one other GPU.

...and each GPU has a 400Gbps Mellanox InfiniBand connection to enable it to connect to any GPU in the cluster outside of the server

Lipacis: When you see the pictures of data centers, like in a traditional one would be top of rack switch, end of row switch. So is that a way to conceptualize this? Or is this something different?

In a cluster of nodes, each node connects to a Mellanox InfiniBand switch, and those switches connect to every other switch...

...in a large cluster, every GPU will connect to every other GPU through these switches with either one, two or three hops

Mellanox InfiniBand connections are needed to eliminate bottlenecks as your clusters grow to 100s or 1,000s of nodes

Driggers: Yeah, I mean it's similar to that, but they're all interconnected. Every switch talks to every other switch as well. So rather than one switch connecting to one switch, connecting to one switch, in the types of the networks that are built for the physical networks that built to do these clusters, the nodes connect to a switch, those switches connect to every other switch. So everything connects to everything else with as few hops as possible. We start with one, and then we go to two. And the entire cluster is built the same way. So every GPU will connect to every other GPU with either one connection or two hops or three hops. We wouldn't have a situation where one has three hops and one has one hop. As soon as you leave the box, they all have to connect to each other in the same fashion, the same number of hops to keep the whole cluster because it'll be as fast as the slowest link. So we want all the links to be the same.

Lipacis: And so are these – so are these all Mellanox InfiniBand connections?

Driggers: For the larger scale clusters, that's really what's needed in order to eliminate the bottlenecks as you grow to clusters of hundreds or thousands of servers The Mellanox InfiniBand is definitely the preferred method to link them together.

Number and Price of GPUs and Hardware Required to Train Generative AI Platforms like ChatGPT

Lipacis: Okay. From a hardware side, before we go to the software side and the ecosystem, what did you end up with? At the end of the day how big is this? Sounds like a monster. How big did this monster get?

Driggers: Ultimately, you build it as big such that you don't yet again create another bottleneck. And where we're at today is about 64 servers, 64 servers makes a really nice building block and that with the latest technology that's coming, that's also a lot of wattage. The new servers are 10,000 watts a piece. So to build a 64 node cluster, which 64 nodes doesn't sound like a lot, but ultimately you're talking about very expensive because it's 64 times $250,000 a piece, plus 64 times 10,000 watts. So you're at two-thirds of a megawatt, which is a tremendous amount of power. So that's really the nicest building block that's being used and being consistently quoted right now.

A "cluster" is a standard building block of 64 servers, or 512 GPUs...

...but each server costs $250K and dissipates 10,000 Watts, which sums to 640 kilowatts per cluster

Lipacis: So that's roughly 512 GPUs per – is that called a cluster or…

Driggers: Yeah, that's called a cluster.

Lipacis: Okay. And then how many of those did you have at the end of the day when you were at kind of at the peak maximum capacity?

A cluster used to be 16 servers, until InfiniBand switching speeds and capacity expanded to make a 64-server clusters the standard

Driggers: Well, so in the past, we weren't able to build them that big and have them flat. We had to build smaller because the interconnects themselves weren't as large. The previous sizes, you were really looking more at 16 nodes to have them flat. A 16 node cluster was more typical. We've started moving to the bigger clusters as the InfiniBand has increased. But people take these and even put these together into bigger clusters. Well, that's a building block. That's not where people stop. That's the idea of building block today. Because then you can certainly go far larger than that for bigger, more complex clusters.

The next step to scale neural networks is to connect eight 64-node clusters together, which would be 4,096 GPUs - all connecting together with a maximum of two hops

Lipacis: Right. So if you have, let's just say, 64 servers in a cluster, is that where you reach a law diminishing returns? Or do you scale it up for ChatGPT did you scale it up from there? It's like how many GPUs did you have?

Driggers: So we did not do ChatGPT, but yes, you would then scale it up. Typically you'd go from 64, you'd do eight of those and then connect those on that next layer up. So you'd add another layer of switching and then you'd have blocks of 64, eight blocks of 64 into the next size up cluster. So you'd have a 512 node cluster, which would've 4,096 GPU's, all talking to each other within two hops.

Lipacis: Okay, so 4,096 GPUs in this cluster – for these kinds of generative AI engines do you think that's where the industry is? Do we just say, hey, we're going to do this engine, we're going to get funding, we need 4,000 GPUs and that's going to cost us whatever, $100 million, $200 million? Do you think that's where the average generative AI engine is?

Stability trained its "Stable Diffusion" text-to-image neural network on 256 NVidia A100 GPUs and has 890 million parameters

Driggers: Well, we've certainly moved. Stability has proven that you don't have to go all the way to a GPT-3 have a viable product. And we're starting to hear this from more and more companies that they're starting to identify a sweet spot that they've been able to prove with something like GPT-3 from OpenAI that you can use a natural language model to do really productive work. But we don't need something that big. And in general, we may need something smaller that's more specific to our use case.

And we're hearing a lot of companies that are believing that models that are in the 1.5 billion to 5 billion parameter range, so much, much smaller than GPT-3, much smaller than ChatGPT, are viable for doing real world tasks. The generative AI engines that, that Stability made available to the public wasn't trained on a gigantic cluster. It was trained on a significant cluster and took months to train. But it wasn't anything like GPT-3.

ChatGPT has 20 billion parameters,

and GPT-3 has 175 billion

Many companies believe that models with 1.5-to-5 billion parameters are viable for doing real-world tasks

So, we are quoting to these well-funded startups, clusters that are quite a bit smaller, because they're not trying to do 170 billion, they're trying to do 5 billion, and with 5 billion, if they've been very selective about their data sets, or they're selective about what they want their generative AI to do can produce tremendous results. I mean, Stability if you've messed with Stable Diffusion or if you've been on discord with Midjourney**,** the outputs are pretty spectacular. And those are not big models. Those are, billion, billion to 2 billion parameter models.

Lipacis: So can you drive the numbers that you're tossing around for parameters into the investment in the hardware architecture? So if ChatGPT-3 is 178 billion, what does that tell you in terms of the number of GPUs you need to try and versus 1.5 billion for Stability or something like that? Yeah, if you could give with that rules of thumb, that would be great.

A 64-node cluster (512 GPUs) is a viable platform for a 1-to-5 billion parameter model...

...which might be appropriate for a language model that is focused on a specific vertical market, like medical, which would be trained on medical specific text

Driggers: Yes. And that's what I'm saying in this 1 billion to 5 billion a 64-node cluster is a very viable platform. It can train the models and it can do it in a quick enough manner that can do multiple trainings. I mean, you can have these gigantic models that try to do everything within one model, or you can have smaller models and then really focus on fine tuning that model to the data set you're trying to solve for. For instance, if you're doing natural language, a natural language model, and it's focused on medical information, you don't have to train a model on the Bible, you're not going to find a whole lot of data in the Bible. That's going to be very specific to medical data.

So you can make a smaller model with more specific information, often it's hard to get that information. And that's one of the biggest challenges to building these bigger models, is getting the data and getting what is called labeled data. So the data itself has tags that lets us know what it is. Medical is really hard because of personal, because of privacy laws. So, we don't get nearly as much medical data as we'd like to train on because of the privacy.

Getting "labeled data" to train neural networks is often a problem...

...medical data is challenging because of privacy laws

Lipacis: And then 170 billion parameter model like, what do you think the consensus is the technical community about how big the current ChatGPT training engine is right now in terms of nodes, yeah.

A 170 billion parameter model, like GPT-3 would require many thousands of nodes to train

Driggers: Big. I mean, I would say it's thousands, many, many thousands of nodes doing the training on hundreds of billions of parameters.

Lipacis: So thousands of nodes, each node being eight GPUs and the associated Mellanox infrastructure and like how many generative AI engines are out there do you think are using that kind of an infrastructure or are on the comp so to speak?

Hyperscale companies are building systems to train 170 billion parameter models...

...but companies are popping up with specific use cases that require much much fewer parameters

Driggers: There is quite a bit of work in at that size, but they're being done by the really big, really big entities like a Google or somebody who already has massive compute capability. They're definitely building these big, big models, but they have general, a more general need for any of these natural language models, they have to deal with multiple languages, they have to deal with technical. So anybody who's dealing with all types of data is going to have to have a really big general model that's huge in order for it to be effective across all of their use cases. But what we're seeing from the marketplace is companies popping up that have more specific use cases. So they don't need 175 billion, they need more specific models.

The other problem, when a model's that big, the inference engine to run it has to be big. So by making the model smaller, the inference engine that needs to run it, if it's a more specific model, what comes out of that model is smaller as well. You don't have to have as much memory to hold all the parameters necessary to leverage that model for inferencing. Because the other task is inferencing, training is just creating the AI getting everything to a tighter, smaller engine that can be used for inferencing. But a big model has a big inference requirement.

The larger the parameter set, the larger the inferencing requirement

Value of NVDA's Software Ecosystem

Lipacis: I have another three hours of questions, and we only have 15 minutes left on the call. So, I want to jump over to the software ecosystem. I think a lot of people are going to be interested to understand, given this huge investment - you must be making it for a good reason. So you talked about the hardware architecture that NVIDIA has delivered that is solving solutions for this. Can you talk about the software ecosystem? What was valuable? What has NVIDIA done that other people have not done on the software side?

Driggers: So NVIDIA took a really heavy handed approach to deliver the software plumbing underneath that allows the data to move from memory place to another memory place to get to the processors, to get to another processor and really designed the software glue for the programmers. So the programmers didn't have to figure out how to actually get underneath the hood. They don't need to know how the server's built. They don't need to know how the server's connected to another server. They program at a much higher level.

NVidia itself developed the "software plumbing underneath" that allows data to move between memory and GPUs automatically...

And this is really important frankly, really important for AI versus HPC. In traditional HPC, often the computer programmer really understood what was happening inside the box. In the case with AI, a lot of these guys are data scientists. They're not traditional computer programmers. They could be analytics people, they can be mathematicians. They can be statisticians. So the data scientists are just that they're data scientists. They're not first and foremost computer programmers. So making that easier and for them allowed for a much bigger ecosystem of people to use the hardware.

...which is more important for AI vs HPC, because in HPC the users were "programmers first" and understood the server architecture...

...but in AI, many of the programmers are data scientists, who do not have an in-depth understanding of the hardware, and can only write software at a high level

And NVIDIA took a very heavy hand on this, rather than rely upon the open source community to actually build this, which was more traditional within HPC. They actually took the open source platforms and enabled GPUs to be optimized for them. So the whole CUDA ecosystem was built to take a heavy handed approach [by NVIDIA itself] to building, leveraging open source software, but making it far more optimized for GPUs. And then they went deeper than that and they started grabbing the key libraries that existed.

Where the HPC industry relied on the open source community to build its software ecosystem...

...NVidia took the opposite approach, and developed the software ecosystem itself...

Again, these are little computer routines that the open source community had written to do routine tasks like linear algebra. There are a set of linear algebra equations that have been built within the open community. And NVIDIA went and "CUDAfied" them. GPU optimized them, and it's called cuBLAS. So the BLAS library exists for CPUs and general processing. But they actually created an enhanced version of that specific for GPUs called cuBLAS.

...starting with rewriting the well-used and understood libraries and tools that were developed by the HPC community, and optimizing them to run on NVDA GPUs

And they went down the line picking these different open source libraries and made them optimized for GPUs and then made them available to anybody who's using NVIDIA GPUs. I mean, the libraries themselves are optimized for CUDA, which is optimized for NVIDIA GPUs, not just, not AMD GPUs or Intel GPUs. These have been optimized specifically for NVIDIA. But they did that versus relying upon the open source community to make those links and make those connections.

One library NVidia optimized for its GPUs was BLAS, a linear algebra library, NVidia's version is cuBLAS

In addition, they designed the communication libraries that enabled the programmer to not worry about where the GPU is or how it's connected to the next GPU. They have a library, the most important library to make it easy for us to build large scale clusters is called NCCL. And that's NVIDIA's communication library to enable the programmer to not have to know where the GPU is. The library enables the communication from GPU to GPU and finds the best path forward.

NVidia also created new software libraries to make its GPUs easy to use...one key library is NCCL, a communication library, which enables GPUs to communicate with one another and is critical for building large scale clusters

Lipacis: So I'm going to play back what you said, and I'm hoping that you would correct me if my understanding is incorrect. A lot of people have a picture in their minds of a stack. So you might have the GPU at the bottom, you might have the high level programming language, the deep learning framework at the top, Keras, Tensorflow, PyTorch or whatever.

And then in between, what NVIDIA has done is optimized what the programming community has used in the past. Libraries such as BLAS and then even added something more, which is NCCL, which is a communications libraries that enable these clusters to scale. And so NVIDIA has invested a lot in those software layers, primitives, libraries or so on. And that's a big part of the value that you're effectively paying for. Is that fair?

NVidia's "full stack" solution has its GPUs at the bottom, third-party high-level programming "frameworks" on the top, and a lot of NVidia-customized libraries inbetween that make the high-level sofware run efficiently, and enable these clusters to scale

Driggers: Absolutely. And they've kept it open source so other third parties can actually write to it as well. So that was one of the other really nice things that NVIDIA did, is they built it on open source and they made it open source. So third parties can actually build libraries that can also be used on GPUs and optimized for GPUs. So they've made this. CUDA is open source and they've made it so that the ecosystem can still go on its own and build things on its own. So you do have third-party libraries that focus on specific workload enhancements for their Fortran libraries that have been written specifically for NVIDIA, it makes it easier. So if somebody is needing to use Fortran, they don't have to go back and relearn how to program Fortran to a GPU, it's sitting there waiting for them those connections.

While NVidia's software ecosystem is proprietary, and it did a lot of heavy lifting to develop it, another smart thing NVidia did was to make it open-source, so that it harnesses the horsepower of a broader 3rd-party ecosystem

I can't tell you how important it is, it's one of the moats that makes it harder on new companies coming along. Now, the beauty is with AI, the new technologies really have to focus on that top level, the compatibility to PyTorch and the libraries associated to PyTorch. The TensorFlow, right now we've really got two primary methods that AI's programming and its PyTorch and TensorFlow. Those frameworks have really dominated and won the framework battle. Google is starting another one. Jax, it's getting some attention but the new guys can really focus on enabling PyTorch and TensorFlow versus having to get deeper into C or deeper into Fortran or deeper into the lower level languages because most of these guys that's they're programming with, they're programming very high level.

NVidia's software ecosystem is a moat that makes it difficult for new companies to break into the market

Lipacis: Got it. And is anybody close in this middle layer, the libraries, is anybody else doing that work as comprehensively as NVIDIA? Maybe one way to ask question is when you think about the value of the architected solutions that you have, to what extent is it the hardware versus to what extent the ecosystem has been developed?

No one is close to NVidia's software ecosystem...

Driggers: Nobody is close right now. Yeah. Nobody is close to NVIDIA from end-to-end solution, hardware, software, interconnect. There are novel technologies that are coming out that are targeting workloads that are difficult for NVIDIA to do. Yeah, we do support quite a few of those other technologies. We've got Graphcore, SambaNova and Cerebras. They very much understand they have to have the software to enable the hardware and make it easy on the programmer.

...but some novel technologies are coming out that are targeting workloads that are difficult for NVidia to do, like Graphcore, SambaNova and Cerebras

Lipacis: Okay. I want to shift over. We got eight minutes left. Can you talk what does inferencing infrastructure look like once it has been trained?

Hardware Considerations for Inferencing

Driggers: So inferencing is highly independent. Training hardware looks pretty much the same, it's all based upon how big the model is. Inferencing on the other hand is totally different. Inferencing is really reliant upon what the tasks needs to be. Inferencing can be little tiny ARM chips or even in your phone. Your phone does inferencing all day long. Voice recognition is inferencing. It can also be something quite big. If the content especially on

While training infrastructure is largely similar, inferencing infrastructure is highly dependent on the tasks and number of parameters

Generative AI, the content that's being created is rich, you're going to need a big engine. In the case of like the images that are being created by Stability or DALL-E 2 or Midjourney, we're often using a GPU to actually do the inferencing like T4… that's an NVIDIA GPU inference card that has 16 gigabytes of memory. It looks, smells, acts just like a normal GPU that would be for gaming. People are inferencing for stability. You inference on your laptop or in your desktop and you use a GPU to do that. We can't use something small because the number of parameters to deliver that image are too large to fit into something small. Whereas text is smaller. So inferencing on text tends to happen on much, much smaller inference engines.

Voice recognition inferencing can be done on your phone...

...but generative AI inferencing needs a big inferencing engine...

...and images created by Stability or DALL-E2 will need an inferencing GPU like NVidia's T4

Lipacis: What would you guess, given that the investment Microsoft has made into ChatGPT? What would you guess that the inferencing architectures that they are using? Do you think they're just using open capacity X86 servers that just happened to be in Azure network?

Under Microsoft, OpenAI's generative AI engines like ChatGPT and GPT-3 are likely using unutilized x86 CPU cycles to do inferencing

Driggers: That's most likely what's being used. So, CPUs are very good and very general purpose from an inferencing engine because what makes CPUs nice is they have enough cores. You're not typically so core constrained, you're typically more memory constrained. And CPUs can have a lot of memory, a lot more than a GPU. And there are tons, if you're a cloud provider like an AWS or Azure or Google you always have millions of CPU resources that are unused, that are waiting to be used.

And since inferencing is so unstructured versus training, which needs one big cluster to train one thing all at a time. Inferencing is lots and lots of tiny little tasks that are just happening over and over and over that take could be seconds, could be microseconds, so very easy to distribute those workloads. So inferencing in general is a more traditional hyperscale cloud type of an operation because it's not training where you get on, you stay on, you run for hours, days, months continuous communication.

Inferencing is a more traditional hyperscale cloud operation

Lipacis: So let me just bounce this back to you. Is it fair to say that when people are thinking about Generative AI engines that the architecture that you built, which was focused on the NVIDIA solution and the whole ecosystem, do you think that is the solution of choice for training Generative AI engines? Number one, Is that fair to say?

NVidia is the solution of choice today for generative AI training engines

Driggers: It is today, I think there's a pretty big move to have pre-trained models available or, and GPT-3 is a pre-trained model that companies can go to and put their data in and fine tune. We're going to see more of this from even the hardware vendors. You've got companies like Hugging Face, that are service providers that are putting out pre-trained models. You've got the hardware vendors themselves. SambaNova and Cerebras are talking about enabling models ahead of time.

There is a move for companies to make pre-trained neural networks available to customers, and then let their customer "fine-tune" those neural networks

This could evolve to a "Models-as-a-Service" business model for hardware companies

So that somebody can start with, they don't have to do that heavy first lift and they can just do fine tuning on the models. I think we're going to start seeing more models-as-a-service as opposed to just, here's your hardware and you've got to figure out how to train your model yourself. We're going to see more and more models-as-a-service, not unlike how OpenAI does it, but I think we'll see even hardware vendors starting to make models-as-a-service versus platforms-as-a-service.

Lipacis: And then on the inference side, is it fair to say that it's just wide open?

Driggers: Absolute wild, wild west. And a lot of times you'll actually even go back and optimize your training, so you can optimize the inference engine. Because I mean, if you're

Inferencing is not standard, rather workload dependent. For example, inferencing on your phone needs to be power efficient doing something for a mobile phone, that inference engine needs to be power-efficient. How much power it uses is actually pretty important. So first you make it so that it actually works, then you go back and optimize for whatever is the critical function. In the case of inference engine on a phone, it's power.

NVDA as the 80% Market Share Player within an 80/20 Framework

Lipacis: Got it. And then we only have two minutes left. I want you to put your long-term strategist hat on and think about the computing eras of the past, the mainframe era, the minicomputer era, the PC era, the handset era, the server era, classically it's been like one hardware, software chip ecosystem that captures 80% of the value, 80% of the share. There's this 80-20 rule where one ecosystem captures 80% and then 20% is fought over by other players. Do you think that this computing era that we're entering right now call, we call it the Parallel Processing Era, some people call it an AI computing era. Do you think that this one will be similar?

Historically, one integrated HW+SW ecosystem captured 80% of the value of each computing era...

...GPUs dominate 80% of the generative AI engines...

Driggers: It certainly is right now. I mean GPU clusters are certainly dominating that 80%. And it's really the 10% on either end that GPUs don't do well, where the data is really, really sparse. So that's lots of tiny bits of data. So there's too much communication to, there's too much communication required or really rich data where the data itself is so rich that we need a lot more memory to the compute side. So, we're out of balance because of the amount, the building block's not big enough. So building block, not big enough communication, not enough communication. So, but right now, GPUs fit in that sweet spot. The 80% in the middle as what we want to do, pivots to either direction, richer or sparse or data that window could move. But right now we need different hardware for that. The hardware needs to be built.

...and NVidia is dominating GPUs in these applications because their ecosystem is a moat...

Lipacis: And I want to get you to confirm or deny just specifically, you're talking about GPUs. It sounds like you're talking about NVIDIA and their ecosystem. Is that what you're talking about? Or do you think that, it could be any GPU that that hits that 80% at all?

...but these networks are huge and the moat could disappear if price is too high...

Driggers: It could be any GPU. I mean their [NVidia's] ecosystem is a moat, but price, price, price when it comes to, when you're in general purpose, price starts becoming the king ultimately, and clusters got this way, compute always goes this way. Eventually that moat disappears if you start charging too much because the market will fix pricing.

NVidia and AMD are the big GPU players, but Intel's GPUs are coming and Intel has a rich software capability

Lipacis: And there's two GPU vendors for the most part. Is that fair?

Driggers: I mean, there's really three, there's two big boys, which is obviously NVIDIA and AMD. But Intel's coming and Intel has a very rich software capability. So they have the ability to help the market bridge that gap.

Lipacis: It sounds like you're saying today NVIDIA has a good moat. AMD and Intel have ways to go to close that gap. Is that fair to say?

Driggers: Absolutely. Absolutely.

Lipacis: Great. Dave, we ran over. Thank you very much. I think, we're going to have to ask if you could come back and chat again, as it's extremely and informative, insightful. We really appreciate you joining us today. And thanks for everybody on the call for staying on. With that I wish everybody a great day. Bye-Bye.

Please see important disclosure information on pages 18 to 24 of this report.

Company Valuation/Risks

Advanced Micro Devices, Inc.
Our $90 PT assumes P/E of 11x applied to 2025 EPS Power of $10 discounted at 10%. Risks include lower-than-expected synergies and diluted growth from Xilinx and Pensando acquisitions, consumer cyclicality via weaker sales of PCs and consoles, mis-execution of product roadmap and strong competition from Intel and Nvidia.

Microsoft Corporation
Our Price Target of $310 is based on a DCF. Key risks include the PC cycle and risks related to revenues and margins as MSFT transitions to the Cloud.

NVIDIA Corporation
Our $275 PT assumes 25x P/E on our 2026 Non-GAAP EPS power of $15, discounted by 10%. NVDA is on a roadmap to grow its EPS by CAGR of 40%+ over the next five years as it builds its ecosystem starting with chips, switch fabrics, software, and AI computing systems. Downside risks include demand destruction from macro slowdown; slower PC Gaming growth; and competition from INTC, AMD, or new entrants to deep learning market. Upside risks include faster adoption of DL applications in Datacenter and Auto.

Analyst Certification:

I, Mark Lipacis, certify that all of the views expressed in this research report accurately reflect my personal views about the subject security(ies) and subject company(ies). I also certify that no part of my compensation was, is, or will be, directly or indirectly, related to the specific recommendations or views expressed in this research report.

I, Brent Thill, certify that all of the views expressed in this research report accurately reflect my personal views about the subject security(ies) and subject company(ies). I also certify that no part of my compensation was, is, or will be, directly or indirectly, related to the specific recommendations or views expressed in this research report.

I, Janardan Menon, certify that all of the views expressed in this research report accurately reflect my personal views about the subject security(ies) and subject company(ies). I also certify that no part of my compensation was, is, or will be, directly or indirectly, related to the specific recommendations or views expressed in this research report.

I, Brian Chen, CFA, certify that all of the views expressed in this research report accurately reflect my personal views about the subject security(ies) and subject company(ies). I also certify that no part of my compensation was, is, or will be, directly or indirectly, related to the specific recommendations or views expressed in this research report.

I, Jefferies University, certify that all of the views expressed in this research report accurately reflect my personal views about the subject security(ies) and subject company(ies). I also certify that no part of my compensation was, is, or will be, directly or indirectly, related to the specific recommendations or views expressed in this research report.

As is the case with all Jefferies employees, the analyst(s) responsible for the coverage of the financial instruments discussed in this report receives compensation based in part on the overall performance of the firm, including investment banking income. We seek to update our research as appropriate, but various regulations may prevent us from doing so. Aside from certain industry reports published on a periodic basis, the large majority of reports are published at irregular intervals as appropriate in the analyst's judgement.

Investment Recommendation Record

(Article 3(1)e and Article 7 of MAR)

Recommendation Completion	February 13, 2023 , 15:44 ET.
Recommendation Distributed	February 13, 2023 , 15:44 ET.

Company Specific Disclosures

Steven DeSanctis owns shares of Microsoft Inc. common shares.
Jefferies Group LLC makes a market in the securities or ADRs of Microsoft Corporation.
Jefferies Group LLC makes a market in the securities or ADRs of NVIDIA Corporation.

Within the past twelve months, Jefferies LLC and/or its affiliates received compensation for products and services other than investment banking services from non-investment banking, securities related compensation for client services it provided to Advanced Micro Devices, Inc..

Within the past twelve months, Jefferies LLC and/or its affiliates received compensation for products and services other than investment banking services from non-investment banking, securities related compensation for client services it provided to Microsoft Corporation.

Explanation of Jefferies Ratings

Buy - Describes securities that we expect to provide a total return (price appreciation plus yield) of 15% or more within a 12-month period.
Hold - Describes securities that we expect to provide a total return (price appreciation plus yield) of plus 15% or minus 10% within a 12-month period.
Underperform - Describes securities that we expect to provide a total return (price appreciation plus yield) of minus 10% or less within a 12-month period.
The expected total return (price appreciation plus yield) for Buy rated securities with an average security price consistently below $10 is 20% or more within a 12-month period as these companies are typically more volatile than the overall stock market. For Hold rated securities with an average security price consistently below $10, the expected total return (price appreciation plus yield) is plus or minus 20% within a 12-month period. For Underperform rated securities with an average security price consistently below $10, the expected total return (price appreciation plus yield) is minus 20% or less within a 12-month period.
NR - The investment rating and price target have been temporarily suspended. Such suspensions are in compliance with applicable regulations and/or Jefferies policies.
CS - Coverage Suspended. Jefferies has suspended coverage of this company.
NC - Not covered. Jefferies does not cover this company.
Restricted - Describes issuers where, in conjunction with Jefferies engagement in certain transactions, company policy or applicable securities regulations prohibit certain types of communications, including investment recommendations.
Monitor - Describes securities whose company fundamentals and financials are being monitored, and for which no financial projections or opinions on the investment merits of the company are provided.

Valuation Methodology

Jefferies' methodology for assigning ratings may include the following: market capitalization, maturity, growth/value, volatility and expected total return over the next 12 months. The price targets are based on several methodologies, which may include, but are not restricted to, analyses of market risk, growth rate, revenue stream, discounted cash flow (DCF), EBITDA, EPS, cash flow (CF), free cash flow (FCF), EV/EBITDA, P/E, PE/growth, P/CF, P/FCF, premium (discount)/average group EV/EBITDA, premium (discount)/average group P/E, sum of the parts, net asset value, dividend returns, and return on equity (ROE) over the next 12 months.

Jefferies Franchise Picks

Jefferies Franchise Picks include stock selections from among the best stock ideas from our equity analysts over a 12 month period. Stock selection is based on fundamental analysis and may take into account other factors such as analyst conviction, differentiated analysis, a favorable risk/reward ratio and investment themes that Jefferies analysts are recommending. Jefferies Franchise Picks will include only Buy rated stocks and the number can vary depending on analyst recommendations for inclusion. Stocks will be added as new opportunities arise and removed when the reason for inclusion changes, the stock has met its desired return, if it is no longer rated Buy and/or if it triggers a stop loss. Stocks having 120 day volatility in the bottom quartile of S&P stocks will continue to have a 15% stop loss, and the remainder will have a 20% stop. Franchise Picks are not intended to represent a recommended portfolio of stocks and is not sector based, but we may note where we believe a Pick falls within an investment style such as growth or value.

Risks which may impede the achievement of our Price Target

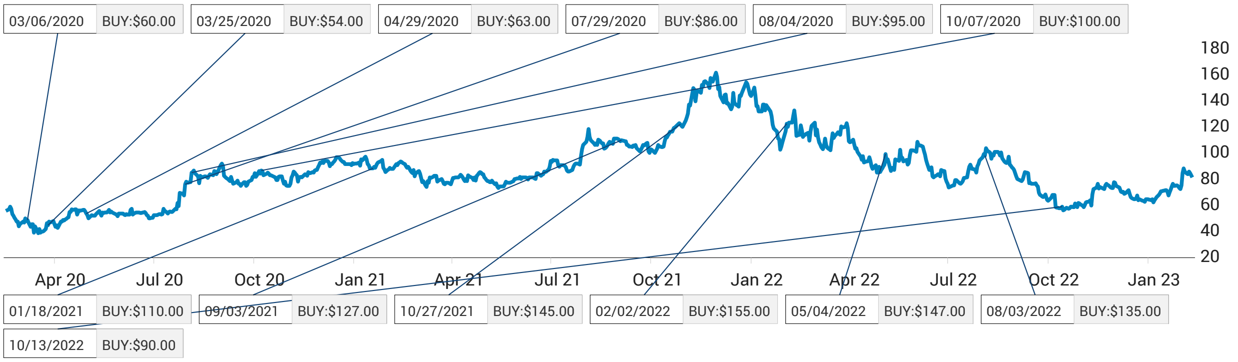
This report was prepared for general circulation and does not provide investment recommendations specific to individual investors. As such, the financial instruments discussed in this report may not be suitable for all investors and investors must make their own investment decisions based upon their specific investment objectives and financial situation utilizing their own financial advisors as they deem necessary. Past performance of the financial instruments recommended in this report should not be taken as an indication or guarantee of future results. The price, value of, and income from, any of the financial instruments mentioned in this report can rise as well as fall and may be affected by changes in economic, financial and political factors. If a financial instrument is denominated in a currency other than the investor's home currency, a change in exchange rates may adversely affect the price of, value of, or income derived from the financial instrument described in this report. In addition, investors in securities such as ADRs, whose values are affected by the currency of the underlying security, effectively assume currency risk.

Other Companies Mentioned in This Report

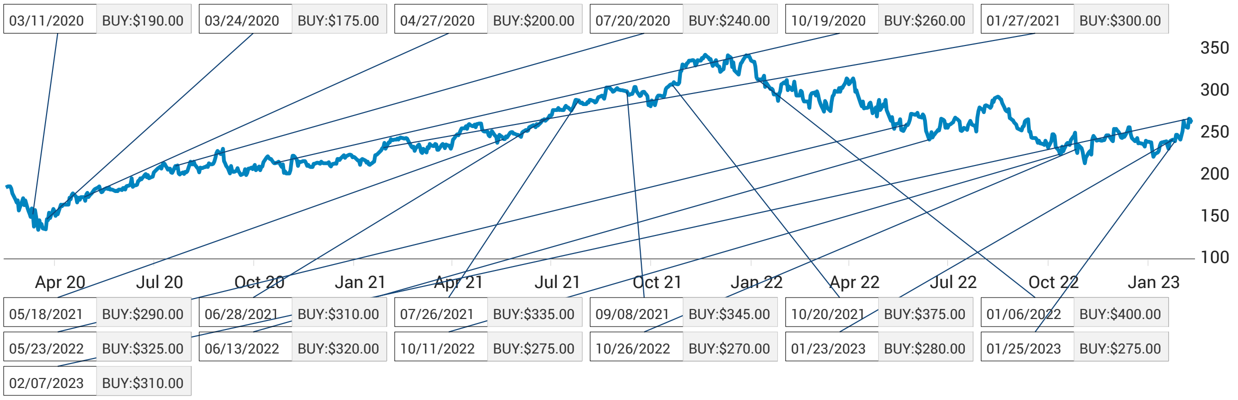
- Advanced Micro Devices, Inc. (AMD: $81.48, BUY)

- Microsoft Corporation (MSFT: $263.10, BUY)
- NVIDIA Corporation (NVDA: $212.65, BUY)

Rating and Price Target History for: Advanced Micro Devices, Inc. (AMD) as of 02-10-2023

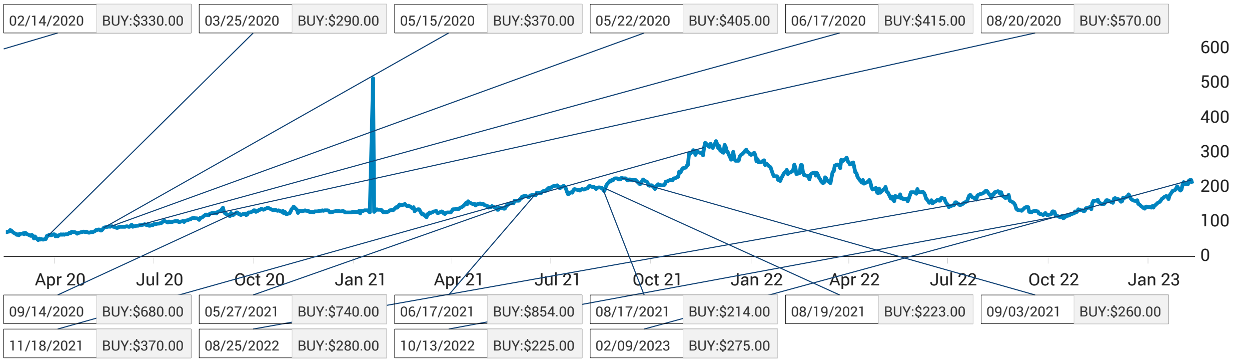


Rating and Price Target History for: Microsoft Corporation (MSFT) as of 02-10-2023



Steven DeSanctis owns shares of Microsoft Inc. common shares.

Rating and Price Target History for: NVIDIA Corporation (NVDA) as of 02-10-2023



Notes: Each box in the Rating and Price Target History chart above represents actions over the past three years in which an analyst initiated on a company, made a change to a rating or price target of a company or discontinued coverage of a company.

Legend:

I: Initiating Coverage

D: Dropped Coverage

B: Buy

H: Hold

UP: Underperform

Distribution of Ratings

Distribution of Ratings						
			IB Serv./Past12 Mos.		JIL Mkt Serv./Past12 Mos.	
	Count	Percent	Count	Percent	Count	Percent
BUY	1916	58.36%	60	3.13%	15	0.78%
HOLD	1171	35.67%	9	0.77%	0	0.00%
UNDERPERFORM	196	5.97%	2	1.02%	1	0.51%

Other Important Disclosures

Jefferies does business and seeks to do business with companies covered in its research reports, and expects to receive or intends to seek compensation for investment banking services among other activities from such companies. As a result, investors should be aware that Jefferies may have a conflict of interest that could affect the objectivity of this report. Investors should consider this report as only a single factor in making their investment decision.

Jefferies Equity Research refers to research reports produced by analysts employed by one of the following Jefferies Group LLC ("Jefferies") group companies:

United States: Jefferies LLC which is an SEC registered broker-dealer and a member of FINRA (and distributed by Jefferies Research Services, LLC, an SEC registered Investment Adviser, to clients paying separately for such research).

United Kingdom: Jefferies International Limited, which is authorized and regulated by the Financial Conduct Authority; registered in England and Wales No. 1978621; registered office: 100 Bishopsgate, London EC2N 4JL; telephone +44 (0)20 7029 8000; facsimile +44 (0)20 7029 8010.

Germany: Jefferies GmbH, which is authorized and regulated by the Bundesanstalt fuer Finanzdienstleistungsaufsicht, BaFin-ID: 10150151; registered office: Bockenheimer Landstr. 24, 60232 Frankfurt a.M., Germany; telephone: +49 (0) 69 719 1870

Hong Kong: Jefferies Hong Kong Limited, which is licensed by the Securities and Futures Commission of Hong Kong with CE number ATS546; located at Level 26, Two International Finance Center, 8 Finance Street, Central, Hong Kong; telephone: +852 3743 8000.

Singapore: Jefferies Singapore Limited, which is licensed by the Monetary Authority of Singapore; located at 80 Raffles Place #15-20, UOB Plaza 2, Singapore 048624, telephone: +65 6551 3950.

Japan: Jefferies (Japan) Limited, Tokyo Branch, which is a securities company registered by the Financial Services Agency of Japan and is a member of the Japan Securities Dealers Association; located at Tokyo Midtown Hibiya 30F Hibiya Mitsui Tower, 1-1-2 Yurakucho, Chiyoda-ku, Tokyo 100-0006; telephone +813 5251 6100; facsimile +813 5251 6101.

India: Jefferies India Private Limited (CIN - U74140MH2007PTC200509), licensed by the Securities and Exchange Board of India for: Stock Broker (NSE & BSE) INZ000243033, Research Analyst INH000000701 and Merchant Banker INM000011443, located at 42/43, 2 North Avenue, Maker Maxity, Bandra-Kurla Complex, Bandra (East), Mumbai 400 051, India; Tel +91 22 4356 6000.

Australia: Jefferies (Australia) Pty Limited (ACN 623 059 898), which holds an Australian financial services license (AFSL 504712) and is located at Level 22, 60 Martin Place, Sydney NSW 2000; telephone +61 2 9364 2800.

This report was prepared by personnel who are associated with Jefferies (Jefferies International Limited, Jefferies GmbH, Jefferies Hong Kong Limited, Jefferies Singapore Limited, Jefferies (Japan) Limited, Tokyo Branch, Jefferies India Private Limited), and Jefferies (Australia) Pty Ltd; or by personnel who are associated with both Jefferies LLC and Jefferies Research Services LLC ("JRS"). Jefferies LLC is a US registered broker-dealer and is affiliated with JRS, which is a US registered investment adviser. JRS does not create tailored or personalized research and all research provided by JRS is impersonal. If you are paying separately for this research, it is being provided to you by JRS. Otherwise, it is being provided by Jefferies LLC. Jefferies LLC, JRS, and their affiliates are collectively referred to below as "Jefferies". Jefferies may seek to do business with companies covered in this research report. As a result, investors should be aware that Jefferies may have a conflict of interest that could affect the objectivity of this report. Investors should consider this report as only one of many factors in making their investment decisions. Specific conflict of interest and other disclosures that are required by FINRA and other rules are set forth in this disclosure section.

* * *

If you are receiving this report from a non-US Jefferies entity, please note the following: Unless prohibited by the provisions of Regulation S of the U.S. Securities Act of 1933, as amended, this material is distributed in the United States by Jefferies LLC, which accepts responsibility for its contents in accordance with the provisions of Rule 15a-6 under the US Securities Exchange Act of 1934, as amended. Transactions by or on behalf of any US person may only be effected through Jefferies LLC. In the United Kingdom and European Economic Area this report is issued and/or approved for distribution by Jefferies International Limited ("JIL") and/or Jefferies GmbH and is intended for use only by persons who have, or have been assessed as having, suitable professional experience and expertise, or by persons to whom it can be otherwise lawfully distributed.

JIL and Jefferies GmbH allows its analysts to undertake private consultancy work. JIL and Jefferies GmbH's conflicts management policy sets out the arrangements JIL and Jefferies GmbH employs to manage any potential conflicts of interest that may arise as a result of such consultancy work. Jefferies LLC, JIL, Jefferies GmbH and their affiliates, may make a market or provide liquidity in the financial instruments referred to in this report; and where they do make a market, such activity is disclosed specifically in this report under “company specific disclosures”.

For Canadian investors, this material is intended for use only by professional or institutional investors. None of the investments or investment services mentioned or described herein is available to other persons or to anyone in Canada who is not a "Designated Institution" as defined by the Securities Act (Ontario). In Singapore, Jefferies Singapore Limited (“JSL”) is regulated by the Monetary Authority of Singapore. For investors in the Republic of Singapore, where this material is prepared and issued by a Jefferies affiliate outside of Singapore, it is distributed by JSL pursuant to Regulation 32C of the Financial Advisers Regulations. The material contained in this document is intended solely for accredited, expert or institutional investors, as defined under the Securities and Futures Act 2001 (Singapore). If there are any matters arising from, or in connection with this material, please contact JSL, located at 80 Raffles Place #15-20, UOB Plaza 2, Singapore 048624, telephone: +65 6551 3950. In Japan, this material is issued and distributed by Jefferies (Japan) Limited to institutional investors only. In Hong Kong, this report is issued and approved by Jefferies Hong Kong Limited and is intended for use only by professional investors as defined in the Hong Kong Securities and Futures Ordinance and its subsidiary legislation. In the Republic of China (Taiwan), this report should not be distributed. The research in relation to this report is conducted outside the People’s Republic of China (“PRC”). This report does not constitute an offer to sell or the solicitation of an offer to buy any securities in the PRC. PRC investors shall have the relevant qualifications to invest in such securities and shall be responsible for obtaining all relevant approvals, licenses, verifications and/or registrations from the relevant governmental authorities themselves. In India, this report is made available by Jefferies India Private Limited. In Australia, this report is issued and/or approved for distribution by, or on behalf of, Jefferies (Australia) Securities Pty Ltd (ACN 610 977 074), which holes an Australian financial services license (AFSL 487263). It is directed solely at wholesale clients within the meaning of the Corporations Act 2001 (Cth) of Australia (the “Corporations Act”), in connection with their consideration of any investment or investment service that is the subject of this report. This report may contain general financial product advice. Where this report refers to a particular financial product, you should obtain a copy of the relevant product disclosure statement or offer document before making any decision in relation to the product. Recipients of this document in any other jurisdictions should inform themselves about and observe any applicable legal requirements in relation to the receipt of this document.

This report is not an offer or solicitation of an offer to buy or sell any security or derivative instrument, or to make any investment. Any opinion or estimate constitutes the preparer’s best judgment as of the date of preparation, and is subject to change without notice. Jefferies assumes no obligation to maintain or update this report based on subsequent information and events. Jefferies, and their respective officers, directors, and employees, may have long or short positions in, or may buy or sell any of the securities, derivative instruments or other investments mentioned or described herein, either as agent or as principal for their own account. This material is provided solely for informational purposes and is not tailored to any recipient, and is not based on, and does not take into account, the particular investment objectives, portfolio holdings, strategy, financial situation, or needs of any recipient. As such, any advice or recommendation in this report may not be suitable for a particular recipient. Jefferies assumes recipients of this report are capable of evaluating the information contained herein and of exercising independent judgment. A recipient of this report should not make any investment decision without first considering whether any advice or recommendation in this report is suitable for the recipient based on the recipient’s particular circumstances and, if appropriate or otherwise needed, seeking professional advice, including tax advice. Jefferies does not perform any suitability or other analysis to check whether an investment decision made by the recipient based on this report is consistent with a recipient’s investment objectives, portfolio holdings, strategy, financial situation, or needs.

By providing this report, neither JRS nor any other Jefferies entity accepts any authority, discretion, or control over the management of the recipient’s assets. Any action taken by the recipient of this report, based on the information in the report, is at the recipient’s sole judgment and risk. The recipient must perform his or her own independent review of any prospective investment. If the recipient uses the services of Jefferies LLC (or other affiliated broker-dealers), in connection with a purchase or sale of a security that is a subject of these materials, such broker-dealer may act as principal for its own accounts or as agent for another person. Only JRS is registered with the SEC as an investment adviser; and therefore neither Jefferies LLC nor any other Jefferies affiliate has any fiduciary duty in connection with distribution of these reports.

The price and value of the investments referred to herein and the income from them may fluctuate. Past performance is not a guide to future performance, future returns are not guaranteed, and a loss of original capital may occur. Fluctuations in exchange rates could have adverse effects on the value or price of, or income derived from, certain investments.

This report may contain forward looking statements that may be affected by inaccurate assumptions or by known or unknown risks, uncertainties, and other important factors. As a result, the actual results, events, performance or achievements of the financial product may be materially different from those expressed or implied in such statements.

This report has been prepared independently of any issuer of securities mentioned herein and not as agent of any issuer of securities. No Equity Research personnel have authority whatsoever to make any representations or warranty on behalf of the issuer(s). Any comments or statements made herein are those of the Jefferies entity producing this report and may differ from the views of other Jefferies entities.

This report may contain information obtained from third parties, including ratings from credit ratings agencies such as Standard & Poor's. Reproduction and distribution of third party content in any form is prohibited except with the prior written permission of the related third party. Jefferies does not guarantee the accuracy, completeness, timeliness or availability of any information, including ratings, and is not responsible for any errors or omissions (negligent or otherwise), regardless of the cause, or for the results obtained from the use of such content. Third-party content providers give no express or implied warranties, including, but not limited to, any warranties of merchantability or fitness for a particular purpose or use. Neither Jefferies nor any third-party content provider shall be liable for any direct, indirect, incidental, exemplary, compensatory, punitive, special or consequential damages, costs, expenses, legal fees, or losses (including lost income or profits and opportunity costs) in connection with any use of their content, including ratings. Credit ratings are statements of opinions and are not statements of fact or recommendations to purchase, hold or sell securities. They do not address the suitability of securities or the suitability of securities for investment purposes, and should not be relied on as investment advice.

Jefferies research reports are disseminated and available electronically, and, in some cases, also in printed form. Electronic research is simultaneously made available to all clients. This report or any portion hereof may not be reprinted, sold or redistributed without the written consent of Jefferies. Neither Jefferies nor any of its respective directors, officers or employees, is responsible for guaranteeing the financial success of any investment, or accepts any liability whatsoever for any direct, indirect or consequential damages or losses arising from any use of this report or its contents. Nothing herein shall be construed to waive any liability Jefferies has under applicable U.S. federal or state securities laws.

For Important Disclosure information relating to JRS, please see https://adviserinfo.sec.gov/IAPD/Content/Common/crd_iapd_Brochure.aspx?BRCHR_VRSN_ID=483878 and https://adviserinfo.sec.gov/Firm/292142 or visit our website at https://javatar.bluematrix.com/sellside/Disclosures.action, or www.jefferies.com, or call 1.888.JEFFERIES.