Evercore ISI Virtual New Mobility & AI Forum

Company Participants

- Paresh Kharya, Sr. Director of Product Management, Data Center Business

Other Participants

- C.J. Muse, Analyst, Evercore ISI

Presentation

C.J. Muse

Well, good morning. Good afternoon everyone, thank you for joining us. My name is C.J. Muse with Evercore ISI, and it is my distinct honor to host the NVIDIA team here today. We have Paresh Kharya, Senior Director of Product Management in the Data Center, and also in the line is Simona Jankowski, VP, Head of Investor Relations and Strategic Finance. So, Paresh, thank you for joining us.

Paresh Kharya {BIO 21780127 <GO>}

Thank you for having me CJ, my pleasure.

C.J. Muse

Excellent. For the audience, there is a chat box where you can ask questions and I'll do my best to integrate those questions into our fireside chat. But, Paresh, you've been busy, new product cycle, so I figured we would just start there, with your new A100 offering. Ampere appears to be a transformative architecture and technology. Can you walk us through what this means for customers in terms of their ability to harness the power of AI?

Paresh Kharya {BIO 21780127 <GO>}

Yeah. So CJ, if you look at, there are two mega trends in AI. The first trend is the exponential growing complexity of AI models. At the launch of A100, when we spoke, we talked about how since the launch of our Volta, the model complexity has grown 3,000 times to train the largest models. In fact, that trend continues. Just a month after we announced the Ampere, open AI introduced a model called GPT3. It's a 175 billion parameter model, and that's 10 times more compute intensity to train these models.

So, that continues, the trajectory continues, 30 times more compute intensity to train the largest model, since we launched Volta. And this requires large scale computers, because otherwise it could just take years to train these massive models. So that's the first trend. The second trend is, because these AI models are now so accurate, there is pervasive AI-powered applications everywhere, from conversational AI to recommender systems and so on, and they service millions of users. And as they service millions of users, each interaction with a user requires a tiny amount of acceleration, so it requires many small-scale accelerated computers. Now, these two are seemingly divergent needs, but A100 for the first time, provided a single solution for both of these divergent needs at the same time. First, it provides a massive performance boost, 20 times higher performance compared to Volta. Second, it provides 50 times higher scalability. So, a single server can be used as one big GPU, working on large scale training problems or can be interconnected over the network to do even larger scale training problems. And secondly, with the capability called multi-instance GPU, a server can be partitioned into 50 different smaller scale GPUs, each running a different inference application.

So, for the first time, A100 unifies data center acceleration for both training and inference and that's why our customers are just loving the A100 architecture and we've really seen a rapid uptake.

C.J. Muse

Okay. It's brings me to the next question, which was, can you share with us the reception so far and I guess I'd love to hear whether it's scale-up, scale-out or other drivers as you think about the adoption for hyperscale versus enterprise versus super-compute, what excites each of those segments perhaps more than -- than something else?

Paresh Kharya {BIO 21780127 <GO>}

Sure. Yeah. So there has been a tremendous uptake. A100 came to the cloud faster than any NVIDIA GPU in the history. It's already available on Google Cloud, on Microsoft Azure, and it's coming soon on every other hyperscale and other cloud providers. There has been over 50 different A100 several models, that were announced by the leading server manufacturers within a few weeks of our launch. And each of these will help the vertical industries ramp-up over the next coming quarters, and the value proposition is really strong. So, for hyperscale with multi-instance GPU, A100 enable's great economics, as we discussed, 50 times higher scalability, as well as a unified training and inference platform. So they can provision instances of different capability from the same infrastructure, and that's why cloud providers and hyperscalers are so excited to bring A100 to their clouds.

For enterprises, this enables unification of acceleration in their Data Center, so they can optimize with a single architecture, while optimizing the utility and utilization, because the same investment to accelerate the Data Center, can accelerate a full range of applications, whether it's data analytics, artificial intelligence, AI influencing, virtual graphics, but all of that can be accelerated by the same infrastructure and the same investment.

Finally, for supercomputing, you mentioned we introduced next-generation of NVLink, and combined with our Mellanox InfiniBand technology, along with our software stack, that we call Magnum IO, we can scale up applications massively, the something that supercomputing really enjoys, multi-GPU, multi-node really large scale applications, all backed up by the same single architecture that they've come to rely upon from NVIDIA, and we've had several design wins in the supercomputing because of this reason. In US, for example, with a supercomputer called Perlmutter from NERSC or in Europe with Juelich supercomputing, with Max Planck Institute and so on. We had several design wins with A100.

C.J. Muse

Congratulations on that. I was hoping to pivot to your system level approach, which I think is something that is perhaps still under-appreciated for NVIDIA, I'm thinking, less so as every day passes. So obviously acceleration of computing starts with the GPU, but it continues to system design, system software algorithms and optimized applications. Can you speak to the importance of this platform approach for NVIDIA?

Paresh Kharya {BIO 21780127 <GO>}

Yeah, the platform approach is critical. We made great GPUs, Tensor cores now in its third generation is unrivaled in terms of performance. Our NVLink provides really massive speed-up in a multi-GPU environment and so on, but it's the 2 million NVIDIA developers that are rapidly growing. That's a testament to our platform approach. So, we provide our fully end-to-end platform to our developers, our CUDA, which is a single programing model that sits on top of our GPUs and CUDA-X, which is our layer of domain-specific libraries that provide the map that's needed for the different domains of applications, whether it's AI training, inferencing, running larger scale jobs and so on, it's that layer of CUDA-X, which is millions of lines of code and it has evolved for over 15 years now, that's really important and AI frameworks they sit on top of it, supreme works like TensorFlow and PyTorch. The reason why they're able to perform so well on our GPUs is because of the layer CUDA-X, which contains all the acceleration libraries that make these frameworks work well. But, we are not stopping there. We also have now created application frameworks that are really important for enterprises to deploy and create applications, whether it's conversational AI applications with our framework called Jarvis or recommendation applications with a framework called Merlin and so on. And all of the software stack that we have is delivered through NGC, which is our hub for GPU-optimized software.

So the net effect is, with the full platform approach, we are able to sustain a virtuous cycle of adoption, because the more developers, because they find it easy to develop on our platform, it implies creating more applications and more deployed endpoints, because there are more deployed endpoints, it attracts even more developers on our platform. So because of our full platform approach, we are able to sustain this virtuous cycle of adoption that's so vital in this industry.

C.J. Muse

I guess playing Devil's Advocate here, talking to some folks in the industry that clearly, I think this is more one to two years ago, there is a notion that, if you can just support one or two AI frameworks, that's enough and that CUDA-X isn't necessary. So, can you speak to, why you disagree with that notion?

Paresh Kharya {BIO 21780127 <GO>}

Yeah. So, the reason why frameworks like TensorFlow and PyTorch work so well on our platform, is because of CUDA and CUDA-X. These frameworks sit on top of our CUDA and CUDA-X, and they automatically benefit from all the capabilities that we built in our hardware. For example, we launched our third-generation of Tensor core GPUs with Ampere architecture, and we introduced the new precision called TF32. There is zero code change that is required from the frameworks, because these frameworks use our libraries like cuDNN and so on, and automatically the benefit of TF32, the new precision that we introduced comes to developers. So the CUDA and CUDA-X are really important in accelerating all the frameworks.

This is the reason why, frankly, our customers confidently invest in our platform, because they've seen our commitment with our libraries and our single CUDA programing stack, they know that with every generation of our GPU, they'll automatically benefit from the architecture, sometimes without changing a single line of code, and that's really the benefit that you get, when you have platform in layers with the industry framework sitting on top of things like CUDA and CUDA-X.

C.J. Muse

Great and helpful explanation. I'm just curious as you think about again platform approach, are you seeing any differences in terms of what is required by hyperscalers versus your enterprise customers, and I know that we're early in the enterprise ramp for the A100, but would be curious if you're seeing any sort of differences there?

Paresh Kharya {BIO 21780127 <GO>}

Yeah. So with our platform approach, the great advantage we have is, we are able to meet where our customers are. Meaning, on one hand we create our end-to-end solutions like DGX for enterprises. But at the same time, we open up our platform with AGX which is the guts behind the DGX that hyperscalers deploy as they create their infrastructures. As well as our OEM server make our partners work with us and that layer. So, that's on the hardware front. And if you look at the software stack, with the hyperscalers, our approach is to deeply integrate into their software stacks, into the frameworks that they are promoting.

So for example, let's take TensorRT, it's our library for optimizing trained models for inference, it's deeply integrated with TensorFlow. Similarly, it's deeply integrated with Microsoft ONNX Runtime. So, with the hyperscalers, our approach is to deeply integrate our software stack. On the other side, hyperscalers also contribute to our software stack. Just last week for example, Microsoft announced DeepSpeed. It's used for extreme scale training of models on thousands of GPUs, for models that reached trillions of parameters, really important contribution by Microsoft, and it benefits, because it's an open source project, it benefits every NVIDIA customer. So, that's our strategy with hyperscalers.

For enterprises, on the other hand, they don't always have in-house AI or software capabilities. So, our approach is to continue to provide higher level stacks and value prop, NGC is a critical part of that. With NGC, we provide pre-trained models, so enterprises don't have to start from scratch, with training models and so on. They can just take pre-trained models that are state-of-the-art, whether it's for conversational AI or recommenders and so on, and they can apply their own proprietary data to customize it to their needs, using what's called transfer learning. We also provide application frameworks which are again targeted for enterprises, which have end-to-end workflows, like how we were describing whether it's Isaac for robotics or Clara for healthcare.

So, while hyperscalers led adoption of AI and accelerated computing, enterprises are now on that same adoption curve, and for many it's an existing imperative, frankly. And we've seen tremendous growth in the vertical industry, our Data Center business derives close to half of our revenues from enterprises.

C.J. Muse

That's helpful. So you've made a splash announcement, I guess a week back, with the potential ARM acquisition, curious as you think about full stack solution, end-to-end, how does that asset fit into your go-to-market strategy and how excited are you to potentially have CPU for the Data Center in your portfolio?

Paresh Kharya {BIO 21780127 <GO>}

Yeah. So as you pointed out, first, it creates a premier computing Company for the age of AI, by combining our AI computing platform, with ARMs vast CPU ecosystem. Secondly, it expands ARMs IP licensing portfolio with NVIDIA technology in several large end-markets like mobile and PC and thirdly, it turbo charges ARMs servers CPU roadmap pace, which can then further accelerate the Data Center and Edge AI and IoT opportunities for us together. And finally, it expands the reach of our computing platform from 2 million developers that we have today to 15 million developers on ARM. And if you look at from an endpoint perspective, we sell about 100 million or so chips per year, ARM architecture based chips sell 22 billion per year. So, ultimately, the benefit is the virtuous cycle, more endpoints and applications will attract more developers, and more developers will result in even more applications and more deployed endpoints and so, that's the exciting possibility here.

C.J. Muse

That's great, I appreciate that. Hoping to pivot now to different business verticals. I think something that's special to NVIDIA is that you're clearly taking a customized approach to each vertical. And I guess, can you speak to which businesses are currently the fastest to adopt accelerated computing for AI and how are they benefiting from your platform?

Paresh Kharya {BIO 21780127 <GO>}

Yeah. So we're seeing adoption across retail, industrial, automotive, healthcare, video analytics, as you were discussing, close to half of our Data Center business is now now driven by vertical industry and the use cases and the value prop has been tremendous. If you look at retail, for example, Walmart deploys NVIDIA AI for optimizing their supply chain. The problem is immense, for customers like Walmart, they have 100,000 different products that are going into thousands of their stores, just in US and if you look at how much they have to stock to minimize the stock-outs and optimize the shelf space utilization and so on. On a weekly basis, they are predicting demand for half a billion item-by-store combinations and each of that combination is impacted by a number of factors. So, they have a massive sort of data analytics and AI training problem, using NVIDIA platform, they're able to increase their data analytics by over 100 times and that results in faster delivery to stores and so on.

Similarly, if you look at on the manufacturing side, we talked about how we are working with BMW, where BMW produces 10,000 cars everyday, and the challenge is, the challenge of customization. They offer 40 different car models, each with 100 different options per car. So, the logistics problem is just humongous. They have 10s of millions of parts that are coming everyday from hundreds of suppliers, and they have to deploy that in their factories to produce 10,000 build-to-order cars everyday. And in order to keep all these production lines humming and operating smoothly, the parts need to arrive just in time, just in sequence, and for that, they selected our Isaac robotics platform to enhance automation in their automotive factories. So, really amazing possibilities with what the enterprises are now able to do and transform their business models, and operations with AI.

C.J. Muse

I guess, sticking with the vertical and really digging a bit deeper into the data analytics and data science side, in the past, you've spoken about coordination of CPU servers, creating a bottleneck with the exponential data growth. Can you walk us through this dynamic and how does supporting Spark 3.0 solve this?

Paresh Kharya {BIO 21780127 <GO>}

Yeah. So, before a model can be trained, the data needs to be prepared for it, and this is a process that's called as ETL in the industry, which stands for Extract, Transform and Load, meaning, data is coming from various sources and you're preparing that data, readying it, so it can be used to train these models. But this is a very computationally intensive process. In fact, 70% to 80% of data scientists' time today is just spent just preparing for data, not even training, just preparing for data. So, data analytics is a big computing challenge and Spark is the most popular data analytics framework for distributed computing, for distributed data analytics. There are 16,000 enterprises and 0.5 million data scientists that use -- that use Spark for data analytics, primarily running on CPU servers.

Now with Spark 3.0, we are bringing GPU acceleration to this vital part of our end-to-end AI development pipeline, and the speedups are tremendous. The Databricks, for example, is offering GPU-accelerated Spark on their platform, Google is offering data on cloud Dataproc, the way this benefits is two-folds. First, this faster performance. So, now you can do all of the data analytics much faster, you're able to, as customers, our customers are able to iterate faster and they are able to train their models faster. We talked about the Walmart example earlier, they are predicting -- let's say they're predicting their supply chain on a weekly basis, but now, if you're able to optimize your models and analyze them faster, what if you can bring down that prediction to a day. Now you can optimize supply chains even further.

Secondly, because now you have the same infrastructure to both, analyze the data, as well as train, you can reduce infrastructure cost. Same infrastructure can do data analysis and training. Adobe for example, talked about, how using Spark 3.0, they were able to achieve seven times higher performance and at the same time 90% savings in the cost. So, Spark 3.0 and GPU acceleration is really transformative. It's a very important part of the overall AI and data analytics workflow that we are now able to accelerate for our customers.

C.J. Muse

And if you think about the economic benefit to NVIDIA, is it more selling GPUs for the e-tail process or is it TAM expansion for AI workloads?

Paresh Kharya {BIO 21780127 <GO>}

That's a really great question. The way to look at this, CJ, is because the data analytics can be done much faster, you can do a lot more iterations. So instead of having some having training jobs run on a weekly basis, for example, you can do it on a daily basis, you can do it in some cases even in real time. So the impact is really expansionary because you are now able to do more of the training and automation, so it expands the overall opportunity, at the same time, because it helps customers optimize their supply chains and optimize their business operations, they are able to invest even more into the infrastructure in order to do the combined data analytics in training.

C.J. Muse

And I guess just a last question on Spark 3.0, how does that fit with NVIDIA's overarching goal in making the Data Center, the computing unit?

Paresh Kharya {BIO 21780127 <GO>}

Yeah. So Data Center is fast becoming the unit of computing and what's meant by that is, developers today are writing applications that are not limited to the confines of a single server. The modern applications that are being deployed today, these are micro services based, and they run everywhere in the Data Center. In fact, in some cases, they are running on the hybrid cloud as well. So the on-premise, the cloud and the edge is no longer siloed. There are hybrid applications that are taking advantage of each as needed. So, Spark in particular, is a great example of distributing data analytics processing and similarly Kubernetes is used to manage all of these clusters, at the Data Center scale.

And so, the Data Center needs to be software defined, because we are now able to accelerate Spark, we are able to accelerate a very important framework that is used for distributed processing in the Data Centers. And so, it accelerates this transition towards the modern Data Centers of the future, that are optimized for the full Data Center scale. We've also talked about how this modern Data Center requires three pillars, the CPU, there is GPU and there is DPU. A CPU for running sequential tasks and hosting the operating system, GPU for accelerating all these modern applications, whether it's data analytics or machine learning or AI and DPU for processing the data in transit and securing all of that communication.

C.J. Muse

Very helpful. On the enterprise side and I guess maybe focusing a bit on the virtualization side, enterprise computer is still a vast majority of the computing market. Can you talk about what you're doing to make enterprise computing look more like how hyperscalers run their cloud?

Paresh Kharya {BIO 21780127 <GO>}

Yeah. So enterprises as you know, enterprises run 75% of their applications on-premise. However, they are being impacted in really profound ways by cloud computing and so they are now finding themselves at an inflection point, and there are three sort of major changes, if you will, that are driving this modern -- enterprise infrastructure modernization. First, AI and data driven applications, enterprises are moving from few experimental AI applications to a future where every enterprise application will be AI-infused and constantly improved with the data driven insights.

Secondly, all the monolithic applications are giving way to micro services based modern applications. And finally, we talked about the notion that on-premise, cloud and edge are no longer siloed. There are hybrid applications that are seeming -- seamlessly taking advantage of each, as needed for the application. So the way we are addressing that is, on the server and the hardware side, we have DGX SuperPOD, which is basically Data Center as a product, so enterprises can stand up, a world-class AI First Data Center in a matter of weeks. Secondly, for mainstream and edge computing, we have EGX, which is helping enterprises do the processing closer to the source of data.

And on the software side, we are -- and the platform side, we are working with the leading partners in the industry. So on one hand, we're working with Kubernetes ecosystem, to make GPUs the first class citizen here, working with downstream partners like Red Hat Open shift to make GPU acceleration natively supported. And also with our partners like VMware for mainstream enterprise computing. Last year, we announced virtual compute server, that enables VMware vSphere to be used as the management layer for AI, and data analytics workloads that are GPU accelerated. And finally, we make all of our software stacks available, has containerized stacks for enterprise deployment, to help with this transition to the modern infrastructures and enterprises.

C.J. Muse

So, (inaudible), we have about five minutes left. So, I figured I'd kind of conclude with maybe some larger picture question. So, where do you think we are in the evolution of AI and what are some of the most interesting applications that you have seen over the last three to six months?

Paresh Kharya {BIO 21780127 <GO>}

Yeah. So, AI is still very nascent. And you can see that it's still nascent is, because it continues to advance exponentially, compute complexity is up 30,000 times over just last four years, for example, and it's opening up new markets, as these giant models can now be trained and some wonderful use cases are now possible. From our roadmap, we are driving our roadmap and that opens up the possibilities for developers, in terms of interesting applications, conversational AI for example, it's highly accurate now. It's really changing the way we interact with applications and machines.

One of the companies closer to the investment community for example is Kensho, it's a S&P global company. They have a speech recognition product or ASR product called Scribe, that transcribes earnings calls, but it customizes these models for the financial jargon. So, 10s of thousands of earning calls, management presentations, acquisition calls each year can be transcribed with really high accuracy and it helps improve the coverage of all of these calls.

Similarly, the other interesting use cases are with recommendation systems, for example, Alibaba, the largest e-commerce company, they have massive volumes, just mind-boggling. On their Singles' Day event for example, they did $38 billion in sales. That's two times the Black Friday and Thanksgiving online shopping combined, and because they were able to deploy advanced recommendation systems powered by NVIDIA GPUs, they had six times more complex models that they were able to deploy for recommendations, that improved their click-through rate by 10%. So, phenomenal value.

On the healthcare side, for example, we are looking at cold and flu season is upon us, we are still in the pandemic and so, we work with NIH to create an AI that can take the chest CT scans and help identify or classify the cause of pneumonia. Pneumonia can be caused by bacterial infections, fungal, viral, you want to be able to detect whether a pneumonia is caused by COVID-19 or by other causes, and we working with NIH, we are able to create this model that detects with over 90% accuracy, when -- when pneumonia is due to COVID-19 and we made this model available on NGC for anyone to deploy.

So use cases are immense and we are still in the nascent phases, as we drive our roadmap, it opens up new possibilities for developers and what's possible. So we're very excited about this.

C.J. Muse

So we get about 30, 50 seconds left in our last question for you in where you want to go, but what do you think is under-appreciated about either AI or NVIDIA or both?

Paresh Kharya {BIO 21780127 <GO>}

Sure. I would say there are three things. First, the pace of adoption and the scale of GPU usage in everyday applications, is really under-appreciated. One quick example is, Microsoft recently talked about how BERT, which is just a model that came a year and a half ago, now powers all the search in Bing, their search engine, and it's only made possible by GPUs, thousands of GPUs are being used to serve the search results. So from research to adoption, the pace is just mind-boggling.

Second, AI is not just used in the consumer Internet and hyperscale space, enterprises are adopting AI very rapidly. We talked about several examples. Half of our revenues now come from enterprise segment.

And finally, the importance of full stack is under-appreciated. It starts with a great chip, but it needs a full software platform, to make all of that work and happen and at NVIDIA, we continue to innovate on the architecture, while our software platform enables developers to access all of that hardware capability seamlessly, as we go from one architecture to the other, and our NGC simplifies the development and deployment to all of our customer base.

C.J. Muse

Well, Paresh that was fantastic, unfortunately we've run out of time. But thank you. Really appreciate it and I wish you the best and great health in these still crazy pandemic times.

Paresh Kharya {BIO 21780127 <GO>}

Thank you, CJ.

C.J. Muse

Thank you.

This transcript may not be 100 percent accurate and may contain misspellings and other inaccuracies. This transcript is provided "as is", without express or implied warranties of any kind. Bloomberg retains all rights to this transcript and provides it solely for your personal, non-commercial use. Bloomberg, its suppliers and third-party agents shall have no liability for errors in this transcript or for lost profits, losses, *or direct, indirect, incidental, consequential, special or punitive damages in connection with the furnishing, performance or use of such transcript. Neither the information nor any opinion expressed in this transcript constitutes a solicitation of the purchase or sale of securities or commodities. Any opinion expressed in the transcript does not necessarily reflect the views of Bloomberg LP.*