# NVIDIA Financial Analyst Q&A 2020

## Company Participants

- Colette Kress, Executive Vice President and Chief Financial Officer
- Jensen Huang, Founder, President and Chief Executive Officer
- Justin Boitano, Vice President and General Manager, Enterprise & Edge Computing
- Kimberly Powell, Vice President of Healthcare
- Manuvir Das, Head of Enterprise Computing
- Paresh Kharya, Senior Director of Product Management and Marketing
- Simona Jankowski, Vice President, Investor Relations

## Other Participants

- Aaron Rakers, Analyst, Wells Fargo
- Ambrish Srivastava, Analyst, BMO
- CJ Muse, Analyst, Evercore
- Harlan Sur, Analyst, JPMorgan
- John Pitzer, Analyst, Credit Suisse
- Mark Lipacis, Analyst, Jefferies
- Matt Ramsay, Analyst, Cowen and Company
- Raji Gill, Analyst, Needham & Company
- Stacy Rasgon, Analyst, Bernstein Research
- Timothy Arcuri, Analyst, UBS
- Vivek Arya, Analyst, Bank of America Merrill Lynch

## Presentation

### Operator

Good morning. My name is Michelle, and I will be your conference operator today. At this time, I would like to welcome everyone to NVIDIA's GTC Financial Analysis Event. All lines have been placed on mute to prevent any background noise. After the speakers' remarks, there will be a question-and-answer session. (Operator Instructions)

Simona Jankowski, you may begin your conference.

### Simona Jankowski {BIO 7131672 <GO>}

Thank you. Hi, everyone, and thank you for joining us for our GTC Financial Analyst Event. We hope you're all able to view Jensen's GTC keynote address this morning. We're excited for this opportunity to spend some more time with the investment community unpacking all of our announcements.

Before I go over introductions in today's agenda, let me remind you that during this presentation we may make forward-looking statements based on current expectations. These are subject to a number of significant risks and uncertainties and our actual results may differ materially. For a discussion of factors that could affect our future financial results and business, please refer to our most recent Forms, 10-K and 10-Q, and the reports that we may file on Form 8-K with the Securities and Exchange Commission. All our statements are made as of today, October 5, 2020, based on information currently available to us. Except as required by law, we assume no obligation to update any such statements.

With that, let's -- we are ready to get started. We have six speakers for you today, who will cover the highlights from this morning's announcement and what they mean for our business: NVIDIA Founder and CEO, Jensen Huang, will kick us off with an overview of the NVIDIA computing platform and strategy. Next we'll have four speakers, who will cover important aspects of our data center market platform, and we'll wrap up with our CFO, Colette Kress, before opening up to Q&A. We expect the entire program to wrap up in about two hours.

And, with that, let me turn it over to Jensen.

## **Jensen Huang** {BIO 1782546 <GO>}

Hi, everyone. This is a very special GTC, because many new platforms and products we're working on came together. The strategic themes NVIDIA is driving will reverberate throughout GTC. AI is the most powerful technology force of our time, software writing software, the age of AI, automation of automation, the age of AI will often open large untapped markets.

Accelerated computing is a full stack challenge. It starts with a great chip, but the stack is a lot more complicated than that. Accelerated computing platforms takes on more than a chip just as cloud computing platforms take more than a server. The new unit of computing is the data center, whether because cloud native applications run across an entire data center or because edge computing will have a whole data center on the chip someday.

The iPhone moment of industrial companies is here. AI services will spend cloud to edge. NVIDIA AI is probably somewhere, say, actually 6 today. We also want to bring NVIDIA AI and accelerated computing to Arm, the world's most popular CPU, and offer our architecture to Arm's vast ecosystem. Simply, we are building the computing company for the age of AI.

We're focused on five major domains. The applications in each domain share similarities now within use cases, system platforms, ecosystems or end market.

Despite domains, our NVIDIA RTX, NVIDIA HPC and NVIDIA AI and NVIDIA Enterprise AI and NVIDIA Edge AI. Let me say a few words about each one of them.

NVIDIA RTX is inventing the future of graphics and digital world. RTX is a massive endeavor, reinventing real-time graphics, requiring us to fundamentally change every layer of the stack from top to bottom. RTX was high risk, but it is now clearly a home run. Ampere is our second-generation RTX we expected and prepared for great demand in the 30s series ramp is our fastest ever. Still, the demand is exceeding our expectations.

We also announced (technical difficulty), our physically based simulation and collaboration platform, a platform for creating digital world is an open beta.

NVIDIA AI is in full throttle. Ampere's with the latest MLPerf 20 benchmark, demonstrating our growing lean. Ampere is our fastest ever data center ramp. This is the first generation that required no explaining whatsoever to any of the customers, OEMs or cloud data centers of the need of NVIDIA CPU's and data centers. They only ask when they can take Ampere to market.

Today, we announced the new NVIDIA RTX A6000 AND NVIDIA A40 enterprise and data center Ampere-based GPUs. These are PCI Express base and complement the already in full production NVIDIA A100. We also announced NVIDIA Jarvis, Conversational AI, SDKs and open data. NVIDIA Merlin recommender SDK is an open beta. These are two of the most important AI models in the world today.

And we just also announced the NVIDIA Maxine SDK for cloud AI video process. Maxine will help video conference services take advantage of NVIDIA's GPUs and NVIDIA AI in the cloud. Live video is one of the most active, most busy traffic in the Internet today.

Finally, we announced that our stacks are so popular that we're working to bring all of them inclusively in NGC, our cloud registry, into each cloud marketplace, essentially like a store within the store.

The next wave of AI is the Enterprise. Enterprises will use AI to automate their company and use AI to bring automation to the products and services their companies build. The latter is also called industrial IoT or edge AI, it comes in several names. Imagine a John Deere autonomous tractor connected to a John Deere cloud service or an autonomous Mercedes-Benz connected to their cloud service, where connected air conditioner, street sweeper, or an entire building connected to AI services. This is the iPhone moment for the world's industries.

AI will automate the world's largest industries. Breakthroughs in AI have made the automation software possible. NVIDIA enterprise AI is about helping companies build modern, secure NVIDIA accelerated data centers and offering the software platforms to help each major industry apply AI. Manuvir will talk about NVIDIA EGX

enterprise, the new NVIDIA BlueField data center infrastructure on a chip platform and our vision for enterprise AI.

NVIDIA edge AI is about helping companies build modern, secure NVIDIA accelerated edge data centers in a box with software stacks that let customers operate their network like a fleet and software platforms to help major industries create and operate AI services.

Today, we announced Fleet Command, a software-as-a-service offering to help operate these suites. Another recent example, AI powered connected product and services is on Mercedes-Benz partnership. Our development with Mercedes-Benz is in full throttle. Mercedes will be using our entire stack from infrastructure, AV computer in the car to the driving application. Justin will talk about NVIDIA EGX Edge and the wave of partners joining us.

NVIDIA HPC consist of supercomputing centers in industrial HPC. Industrial HPC demand is accelerating a large number of domain-specific applications. For example, healthcare domain representing over a decade of work we created a state-of-the-art suite of accelerating tools to help medical researchers discover lifesaving drug. We announced NVIDIA Clara discovery, Kimberly will talk about the great work we're doing into our discovery and our partnerships there.

This is our biggest GTC ever, over 1,000 sessions, a record number of sponsors, a record number of startups participate. We announced AVSDK. These SDKs are the critical difference between a GPU chip and the accel -- and the NVIDIA accelerated computing platform on GPU. These SDKs run on NVIDIA's 1 billion CUDA GPUs, one architecture, gigantic installed base for developers. These SDKs covered a range of NVIDIA's full stack computing platforms, chips and architectures like CUDA GPU and DOCA for DPU, systems or systems components RTX, DGX, HDX for hyperscalers, EGX for enterprise and edge, AGX for autonomous machines, system software APIs for Windows, Android, QNX, Linux, Kubernetes, VMware for PCs cloud, enterprise (technical difficulty).

Acceleration libraries and engine like the CUDA-X libraries, Magnum IO, cuDNN, TensorRT, Triton Inference Server, the RTX stack, our physics engine, and of course applications and application frameworks, Omniverse, NVIDIA DRIVE, Jarvis, Merlin, Isaac, our robotic stack, or Clara, our computational healthcare and life sciences deck.

All of these decks are optimizing containerized on NGC. The NVIDIA SDKs are created in service of our nearly 2.5 million developers, researchers, software companies who accelerated the 1,800 applications for the billions of computer users, the global computer makers, cloud service providers, solution partners, the 6,500 startups in the inception. We built GTC for them. We built -- our SDKs for them and serviced them. They are what GTC is all about.

To tell you more about our announcements on GTC, we've got some very special speakers lined up for you guys. So let me first introduce Paresh Kharya to tell you about NVIDIA AI. Paresh?

## Paresh Kharya {BIO 21780127 <GO>}

Thank you, Jensen. Jensen talked about how AI is software writing software, and AI is already achieving results that no human written software can. The interesting thing is, this approach is extremely scalable, larger, more complex models, create more capable AI, AI's that's more accurate and applicable for many different types of tasks.

So the chart on your left basically shows the number of days it takes to train a model on a one petaflop supercomputer or a computer and it continues to grow exponentially. It's now doubling every couple of months. Think of it another way. If Moore's Law were still true, it could have delivered every 10 years what AI needs every 10 months. So larger models need larger and larger supercomputers to train on expanding the opportunity for NVIDIA, and the capabilities of these advanced AI models and their potential to transform the industries is immense.

On the right, you can see a few recent examples of it. NVIDIA's natural language understanding AI took the RACE Reading Comprehension Challenge that consists of middle school and high school level test. The average human score on it was 73%. NVIDIA's Megatron-BERT has scored 91%. In addition to reading comprehension, language models are also been used to predict the 3D structure of protein, just by reading the amino acid sequences. This will have a dramatic impact in discovering drugs. Kimberly will talk a bit more about it later.

Facebook AI research developed a very large AI model-based chatbot that exhibits knowledge, personality and empathy, they call it BlenderBot. Half of the users tested preferred chatting with BlenderBot over human. In another work, researchers at Caltech developed reinforcement learning this drone flight control systems. They can smoothly land drones with the payload on different surfaces, a critical problem to solve for making safe drone deliveries. Larger and advanced models bring transformative capabilities and when deployed in applications, they also require GPUs for inference for the economics to work for companies.

Take for instance, a BERT-based model used by Microsoft to improve their search engine Bing. The accuracy of the model resulted in the largest improvement in Microsoft search engine. However, it was impossible to run it on CPUs. With our GPUs, Microsoft hot 800 times higher throughput and they could run these models in real time. This led them to switch to thousands of NVIDIA GPUs running on Azure to power their search.

AI needs GPUs both for training and inference. Our Ampere architecture delivered for the first time a unified platform for AI training and AI inference. It provided 20 times higher performance, a unified computing architecture for data analysis, preventing [ph] an inference and with multi-instance GPU the ability to scale down

50 times in a single server. This is the reason why it was picked up so readily by the industry.

A100, as Jensen talked about, has had the fastest ramp of any data center GPU in our history. Shortly after Jensen announced it in May, over 50 different server models were announced by the world's leading server makers based on A100. All leading hyperscalers announced plans to deploy A100 to their cloud. It's already available on Google Cloud, Microsoft Azure and Oracle Cloud.

Of course, having a mighty GPU architecture is at the foundation of NVIDIA AI platform. However, AI workloads push the limits of every aspect of the data center, computing, storage, networking, and software to run all of this. We are addressing this challenge and democratizing AI with three pillars of NVIDIA AI platform: data processing and training, inference, and AI application frameworks.

First, for data processing featuring engineering and training, our platform can run at any scale from one GPU to multiple GPUs to multiple nodes across the data center running any framework, screening any type of models or being available on any cloud, while some companies are just starting to do deep learning, nearly every company is doing data processing with exponentially growing scale. And our work with RAPIDS is now transforming the data analytics landscape. It now accelerate the Spark, the world's leading data analytics platform used by 16,000 enterprises and 0.5 million data scientists. And earlier today, Jensen announced that Cloudera is now accelerated on NVIDIA AI. Cloudera has 2,000 plus customers and runs on 400,000 data center servers, Manuvir will touch a bit more on this in his presentation.

Second, inference is a great computing challenge. It requires a lot of software to make it work. And we break it down into four steps, each very complex. And starting with pre-dreamed state-of-the-art models for key use cases that are available from NGC, our cloud registry hub for our GPU accelerated software. They provide an easy ramp to enterprise customers to infuse AI in applications. They can then use our transfer learning tool kit to refine these models with their own data sets to optimize for their domains.

NVIDIA TensorRT, our optimizing compiler, then helps optimize these models to run for inference on our GPUs. Finally, Triton Inference Server actually helps in run these models, so applications can just send query and the constraints like the response time they need or the throughput they need to scale to thousands of millions of users and Triton takes care of the plumbing to run these models.

Our third pillar is AI application frameworks. The package of our stack and provide end-to-end workflows for incorporating AI into specific application domains for different industries and use cases. This helps democratize the complex AI development pipeline and help the enterprises jump-start the adoption of AI for their use cases. And our application frameworks target some of the most challenging AI applications in large markets, like self-driving cars, robotics, drug discovery, conversational AI, and recommendation systems. The scale of influence on the cloud

is huge, and we are just at the starting point of AI infused services. So opportunities to service inference is massive. Running AI inference on NVIDIA is the most performant and cost effective and so we continue to see a rapid shift to influence running on NVIDIA GPUs.

I talked about TensorRT a bit earlier, our optimizing compiler for influence. The latest version now has over 2,000 optimizations. TensorRT's has been downloaded 1.3 million times and is used by 16,000 enterprises. In terms of the GPU Compute for inference, we've gone from -- practically negligible four years ago to shipping over 166 extra ops in the last 12 months. This is more than 6 times what we said last year. Since AWS launched their first GPU accelerated cloud instance 10 years ago, every cloud now offers NVIDIA GPU, and the aggregate throughput has been increasing 10 times every two years.

The chart on the right shows the growth of the aggregate NVIDIA GPU inference compute in the cloud. We estimate that the aggregate GPU inference compute now exceeds that of all cloud CPUs. With this trend, in two to three years, NVIDIA GPU will represent 90% of the total cloud inference compute. Any AI application and service can now rely on NVIDIA inference, and we have passed that tipping point. NVIDIA inference is enabling transformative capability for our customers. Microsoft -- earlier today Jensen announced, is adopting NVIDIA AI on Azure to power smart experiences in the Microsoft Office, the world's most popular productivity application. The first features include a smart grammar correction, question and answers, and text predictions. With NVIDIA GPUs, Microsoft was able to cut down the responsiveness to less than one-fifth of a second that enables real-time grammar corrections. GPUs also provide high throughput, so they can efficiently scale to service half a trillion queries they expect to service in a year.

Second, cyber crime cost global economy nearly $1 trillion, about 1% of the global economy. American Express alone does 8 billion transactions in a year, totaling about $1 trillion for their 115 million credit card holders.Using NVIDIA AI, American Express uses advanced AI on tens of millions of transactions everyday and it take just two milliseconds to detect fraud instantly.

AI fraud detection is going to help safe financial industry hundreds of billions of dollars each year and NVIDIA AI makes this possible.

Finally, Tencent platform and content group has numerous recommendation systems to support their applications; video, news, music applications et cetera. They have thousands of models that handles hundreds of billions of queries per day. NVIDIA GPU Inference enables the use of more and more advanced models in production for Tencent.

While I talked about just a handful of customers of NVIDIA AI Inference, today NVIDIA AI Inference is operating services for companies in a broad range of industries from automotive to consumer Internet to cloud-based companies to robotics, medical, retail and financial services, industrial customers and so on. In

many cases, only NVIDIA AI Inference makes it possible to deploy advanced AI for production use cases and in all cases it saves money for customers. NVIDIA accelerated AI Inference adoption has reached the tipping point.

With that, I would like to hand over to our next speaker, Manuvir Das.

## Manuvir Das

Thank you, Paresh. Good morning and good afternoon, everyone. As Jensen mentioned at the beginning of this call, a data center is the new unit of computing. This is because the amount of data that is available to and processed by every application running in the data center is growing dramatically. This means that applications can no longer fit within an individual server and must be spread across the data center.

Paresh has talked about the work that NVIDIA has done over many years now to accelerate a particular classes of applications that are running in the data center. He talked about the three pillars with AI training and Inference, as well as the different frameworks. In this action, I want to talk to you about the next opportunity for NVIDIA, the next phase of our work in accelerating the data center, which applies not just to particular classes of applications, but to every application workload running in an enterprise data center. In particular, we are talking about technology that will be introduced into every one of the tens of millions of servers deployed in enterprise data centers.

If you consider what has been happening in enterprise data centers over the last decade, the infrastructure has moved towards a software defined model as guided by the advancements that have happened in the public cloud. On the right hand side, you can see a variety of data center infrastructure function, which has traditionally been performed by hardware. For example, a firewall located at the edge of the data center or complicated networking infrastructure, as well as operations that have been performed manually by armies of human to manage their infrastructure.

Over the last few years, all of these functions have been moved into software which is now running on every application circle. On the left hand side, I have a picture showing you the stack that is running within every application server. The box on the top represent the actual application workloads running in virtual machines or containers or in a bare metal environment. The box below that represents the infrastructure functions that are now deployed in software that is running on every application server; software defined networking, software defined storage, software defined security, infrastructure management, these are typically deployed in a layer like a hypervisor. A great example of this is VMware, which is used by almost every enterprise customer for their virtualized environment. Every one of these servers is connected to the data center network, communicating and moving data between one server to another. This is accomplished with a network interface cards or NIC.

As you know, NVIDIA is now working with Mellanox as well as part of our family. Mellanox has worked on the ConnectX family of mix for many years now, which are state-of-the-art in terms of mix that transfer data across the data center and they are equipped with powerful acceleration engines to make the networking and IO operations proceed fast.

What is the challenge with this model? Here you will see a representation of these different infrastructure functions that are running in this layer of infrastructure. The challenge we have is that as the amount of data grows, the East-West traffic between servers in the data center is growing dramatically. And so if you look at where we are at today, we already see that more than 30% of the resources on the CPU in every server is being occupied with these infrastructure function, leaving less resources available to the actual applications themselves and this problem is only going to grow, because the amount of data is growing exponentially and therefore the amount of resources required by these different function is also growing over time and there will be less and less room available on the CPU of the server itself to run the application. So a new solution is needed, a new piece of hardware needs to sit alongside the CPU on every server in order to take on this increasing load.

We are therefore introducing the next concept in computing that we refer to as the data processing unit, or the DPU. The role of the DPU is to take the software function, offload them from the CPU and put them into a new kind of chip that we call the DPU. It is an extension of the chip that we already have in the NIC. In particular, along with the acceleration engines that we have on this chip, we have now introduced CPU in the form of powerful arm cores that can host these infrastructure functions. So you can see now from the picture on the right-hand side by moving these functions over to this new chip, the DPU, all of the resources of the host CPU are now available to run the applications, whether they are running on virtual machines or containers of payments [ph].

It's important to stress here that it is not just about moving the functionality to the DPU, because the same situation of the burden growing as the amount of data growing would apply on the picture on the right hand side as well. However, the DPU is sitting within the NIC, it is already processing every packet of data that is flowing through the network and therefore these infrastructure functions that operate on the data can now be much more efficient they can be dramatically accelerated when they perform on the NIC itself rather than on the CPU, leading to the same effect that we obtained on applications when we move them from the CPU to the GPU. It is not just about offloading, it is about a dramatic acceleration which leads to massive reduction in the amount of infrastructure needed to run the same amount of application workload leading to TCO benefits for the customer.

As Jensen mentioned in his keynote this morning, we have introduced the BlueField too as our flagship DPU for this purpose. It provides a variety of acceleration engines for different functionality to do with IO storage, security. It's got 7 billion transistors on it. To give you one example of the power of this technology, if you consider the different activities performed by the DPU and you were to perform them on the host CPU, you would require upwards of 125 x86 cores to perform the same functionality,

which is not really practical and here again is the reason why we believe that as computing moves forward every server will be equipped with one of these GPUs to make all of this processing feasible. Now as we move forward with our journey on GPUs at NVIDIA, we introduced the CUDA platform, which was the software abstraction that allowed our ecosystem of 2.3 million developers to proceed forward with a single API even as we advance the technology inside our GPUs. We are doing exactly the same thing with the GPU.

And so, Jensen today announced DOCA, data center infrastructure in a chip architecture, which is our abstraction and interface for developers to access the capabilities of our GPU. It's based on open API, P fall for packet processing, DPDK for the network, SPDK for the storage. These are open APIs on which the developer ecosystem can build a variety of software defined infrastructure that can now run on the DPU as opposed to running on the host CPU. This is just the beginning of our journey.

As you know very well, NVIDIA believes very strongly in the power of AI. The power of AI can be brought to bear on data center infrastructure as well. And so, we are embarking on the journey as Jensen announced to take the silicon of the DPU and add to it tensor cores from the silicon of GPUs as well, so that we can infuse AI into data center infrastructure. And this leads to a complete road map now for NVIDIA from which the BlueField-2 is only the starting point.

Here the chart that shows our road map. At the bottom left, you will see the BlueField-2 which has certain capabilities in terms of network traffic, as well as it's computational power. All these GPUs are based on one architecture which is DOCA, you can see the timeline for advancement. And on the Y-axis we referred to the computational capability. A year from after BlueField-2, we will have the BlueField-3, and then followed by the BlueField-4, two years after the Bluefield 2.

As shown on the Y-axis here, the BlueField-4, which is a combination of the silicon of the DPU, as well as our state-of-the-art GPUs will have 600 times the computational capability of the BlueField-2. It is a significant advancement. But we are not waiting two years to bring this capability. Before we get to the point where we integrate the silicon, we are producing new form factors where we combine the DPU and the GPU on to a single form factor or to a single card. This is referred to as the BlueField-2X, which you can see vertically above the BlueField-2, it will follow the BlueField-2 by only few months, but it has 85 times the computational power of the BlueField-2. It is a significant advancement and it will put AI infused data center infrastructure on a chip in the hands of the customer base much before BlueField-4 is available.

I mentioned earlier that these infrastructure functions that are performed in software today typically live within a layer like a hypervisor. And as NVIDIA and VMware have announced over the last few days, we have a major new partnership with VMware to bring this to the enterprise customer base. As you know, the majority of enterprise customers today use the VMware platform for their virtualized data centers. What we are announcing and the work we are doing together is to take that VMware platform

and to move its functionality onto this NVIDIA Bluefield DPU so that the host CPU can be freed up to perform and run more of the applications.

At the same time, we have also announced a partnership with VMware to bring accelerated AI applications to the VMware platform so that we can truly democratize AI and make AI available in a seamless manner to every enterprise customer. This is a picture of our full stack of collaboration. As you can see on the top here, we are talking about traditional enterprise applications that are already running on the VMware platform, as well as the workload that Paresh talked about that are accelerated by NVIDIA GPU. All of these workloads can now be done on one infrastructure using the VMware platform, which then in turn runs on GPUs, as well as DPUs present on every server to provide now both of these forms of acceleration, both the application acceleration for the domains of workloads accelerated by GPUs, as well as the DPU-based acceleration that applies to every server and every workload.

At the same time, Paresh talked about data analytics as the next frontier of workloads that can be accelerated where we have also announced a partnership now with Cloudera. Cloudera is the predominant platform used by all enterprise customers for their Spark deployments. Cloudera platform will now be powered by NVIDIA, and also accelerated by GPUs, as well as in this model by DPUs.

So with that, I'll hand it over to Justin to talk about the Edge.

## Justin Boitano

Great. Thank you, Manuvir, and good morning to everybody. While the big bang of AI originally happened in the cloud, AI is really about to transform every industry with a wave of new AI infrastructure announced today using the NVIDIA EGX Edge AI platform. And so, I want to talk to you about some of those announcements and put this in context.

Today, the Internet connects billions of people to giant cloud data centers, but in the future there is going to be trillions of devices connected to millions of Edge data centers, and this is going to create an Internet of Things, it's a 1,000 times bigger than today's Internet of people, from smart retail to manufacturing in service robots, self-driving cars, smart streets in cities, computing is going to extend from the cloud data centers to the -- to every corner of the world.

AI will sense, infer, and act accordingly at the Edge. The amount of data generated by high resolution sensors is just simply too much to have the data moved back to the cloud. Some things just can't be done from the cloud as actions need to be immediate. The software powering this new Internet will not be written by humans, but by computers learning from the data. This new way of computing is called AI. And the edge of AI is about driving tremendous acceleration -- is driving tremendous acceleration in demand for computing, precisely the time that Moore's Law has slowed down, and this requires a new approach in computing as legacy

architectures just can't keep up. NVIDIA's accelerated computing platform is really the platform that will power the future of every industry.

Today we made several new announcements with the NVIDIA EGX Edge AI platform. The EGX -- NVIDIA EGX is designed to make it easy for the world's enterprises to quickly stand up, state-of-the-art Edge AI servers. NVIDIA EGX can control factories of robots, perform automatic checkout at retail, orchestrate a fleet of inventory moving robots, and help nurses monitor patients. EGX is a full stack solution consisting of the AI computer, system software, AI frameworks in fleet orchestration and management software.

Security is really a top feature of EGX. Every aspect from security measure of the operating system, protecting data in motion and in rest, securing the applications in AI models with signing an encryption to tamper proof the infrastructure that will automate the future of industries. Today we announced early access of NVIDIA Fleet Command. It's software-as-a-service for deploying and managing AI services at the Edge. Fleet Command simplified setup and management by ensuring a simple one-touch authentication to connect a new node. So you don't need Linux admins floating around in Edge locations. This allows store associates or warehouse managers without IT experience to quickly set up new systems.

What are the most important features of EGX though is the rich ecosystem of partners bringing AI to every industry. EGX servers are certified by all leading OEMs and now include NVIDIA BlueField-2 and Ampere GPUs. This ensures that customers can quickly find and easily buy hardware that's optimized for AI performance and can be securely managed an updated through leading enterprise platforms.

To fully accelerate the run time of EGX service, we're working with leading enterprise platform companies including VMware, Red Hat, Canonical, SUSE and others to test and validate that the performance of CUDA-X can be delivered for AI training and inference through all of their platforms. We're building industry focused AI optimized frameworks to make it easier for developers to bring new innovation to every industry. Every industry can benefit from EGX. With our network of partners, EGX can help companies in manufacturing, in health care, retail, logistics and transportation. OEM partners, software partners, industry focus solution makers can all participate as it's an open platform. This is truly the iPhone moment for the world's industries. NVIDIA EGX will make it easy to create and deploy new edge AI services.

Now, let me share with you the type of customers we've been working with to build and refine the EGX edge AI platform. Previously we spoke about how we're working with leading companies like Wal-Mart, Procter & Gamble, BMW and Siemens to bring a range of new -- in the -- to bring AI to a range of industries, to drive higher business efficiency and -- but let me highlight a few more that we've announced today.

KION Group is the largest logistics and automation supply chain solutions provider globally and operates over 6,000 automated warehouses worldwide. With the Matic [ph], they're developing smart cabinets and adaptive speed conveyors to increase distributions in our efficiency and throughput. We feel they are developing automated forklifts. KION is looking to simplify the management and deployment of AI applications to fleets of GPU accelerated systems used for optimizing warehouse efficiency using NVIDIA Fleet Command.

If we look at the retail industry, they lose 1.5% of sales or over $60 billion per year to shrinkage. AI can instantly detect and scan the checkout. Kroger is one of the largest superchain markets -- or super chain -- or supermarket chains in the US and can operate -- and operates close to 2,800 stores. With ever seen a vision AI application, they are reducing customer errors, providing faster customer checkout to improve the shopping experience. And every healthcare provider in the world wants to reduce operating costs, while improving patient care. At Northwestern Medicine, Whiteboard Coordinator is being used to operate a network of thousands of sensors, cameras and microphones with perception in conversational AI to help nurses monitor patients, reducing the load on nurses while improving care.

So that's our EGX platform. It's a full stack open platform, state-of-the-art computing, designed for security from the ground up. We have a broad ecosystem of partners to help enterprises around the world to create AI services. This ranges from software vendors like VMware, Red Hat, Canonical or SUSE to industry focused assays, including IBM and Accenture, as well as every major OEM and ODM, such as Dell, HP, Cisco, Lenovo, Fujitsu and many more.

We're working with hundreds of ISVs who are leveraging the power of our AI frameworks to build industry-focused AI for every industry. Every enterprise company understands the power of AI. They no longer need to be convinced why they did, but they need a broad partner ecosystem to show them how to become an AI-driven enterprise, powered by the world's most widely adopted accelerators and a broad range of ecosystem partners, now we're working to bring AI to every industry with NVIDIA EGX edge AI platform.

Now, let me hand it over to Kimberly Powell, who runs our Healthcare business.

## **Kimberly Powell** {BIO 22145194 <GO>}

Great. Thank you, Justin. And I hope everyone as well this morning. The healthcare industry is at an extraordinary moment in time. The global pandemic has created the biggest threat to humanity in our lifetime. The race to discover new therapies has never been more critical and today the healthcare industry is producing more biomedical data in a couple of months than the last several hundred years. This is a perfect storm to catalyze AI in drug discovery. This is why we're so excited to announce NVIDIA Clara Discovery, a suite of state-of-the-art tool to tackle the most pressing and future challenges in drug discovery.

The end-to-end drug discovery process is incredibly complex starting with biology to understand human disease and why we get sick in the first place, then to chemistry to combine molecule that can inhibit or enhance biological behavior next to patients and uncovering biomarkers that are the medical science associated with our disease or our response. The pharma industry is huge at $1.5 trillion large, it is still very much an unsolved problem taking well over 10 years, costing $2 billion and still has such a high failure rate of 90%.

Drug discovery draws on every computer science domain from accelerated computing, data science and machine learning, deep learning and natural language processing, so Clara Discovery brings together more than a decade worth of work and working with the industry's most popular GPU accelerated tools like Schrodinger for computational chemistry and were there weren't any tool we built our own with NVIDIA Clara Parabricks for genomics, Clara Imaging for pathology and radiology, BioMegatron and BioBERT for natural language processing and NVIDIA RAPIDS for GPU accelerated machine learning.

This suite of tools is powering the next generation of computational drug discovery to accelerate discovery from months to minutes and using AI in this neutral of data to improve success rate of discovering new lifesaving drugs. The UK is an epicenter for healthcare research. Cambridge is home where -- to Francis Crick and James Watson discovered the structure of DNA -- in its home to the world leaders in the pharmaceutical industry with the rich university and startup ecosystem focused on healthcare. Researchers and scientists in the UK need to say they are computing infrastructure, because there is no more important time NVIDIA is building the UK's fastest supercomputer we're calling Cambridge-1.

It's a 400 petaflops AI performance supercomputer based of NVIDIA's DGX SuperPOD. It will become the fastest supercomputer in the UK and it will be in the top 30 on the top 500 and also the top 3 in the Green500. Cambridge-1 will host our collaborations already underway with the UK AI and healthcare researchers in academia, industry and startup. Our first partners are GSK, AstraZeneca, King's College London, Guy's and St Thomas' NHS Foundation Trust, and Oxford Nanopore Technologies, all which are already using NVIDIA GPU computing.

Cambridge-1 let them to do experiments too large for their infrastructure or resource while they're building up their own. Cambridge-1 will accelerate the use of AI in the vast and wide healthcare ecosystem.

Also to exciting to announce today is GlaxoSmithKline is leading the way in the pharmaceutical industry in adopting artificial intelligence in data driven drug discovery. We're partnering with GSK and one of the world's first AI Drug Discovery Labs. GSK has been pushing the frontiers in drug discovery and data driven drug discovery for years using genomics to improve target selection early in the drug discovery process and a recently established GSK's London based AI hub. GSK and NVIDIA together will expand the use of biomedical data in the field of digital pathology, radiology, genomics, natural language processing using Clara Discovery to optimize computational discovery applications. In addition to GSK's investment in

DGX A100 system, GSK will also be able to access NVIDIA's new Cambridge-1 supercomputer.

And just to conclude, we are in a perfect storm for AI in healthcare with a race against time in the global pandemic, the explosion of biomedical data and the utility of AI, we can accelerate healthcare research and discovery from months to minutes, 16x across many domains used in drug discovery and harness the biggest AI breakthroughs in natural language processing to tap into invaluable biomedical literature and clinical data, healthcare vocabulary is domain specific, complex diseases, proteins and drug names the case in point as in COVID-19. So this is a tipping point for AI and healthcare and we're delighted to be building the industry's computational platform and partnering with the world leaders in healthcare. Thank you.

## Colette Kress {BIO 18297352 <GO>}

Okay. This is Colette Kress and I'm just going to do a quick overview for you about what GTC Fall 2020 is all about. First, you've seen us talk so much about NVIDIA's AI and NVIDIA's overall momentum in terms of gaining across so many different pieces. One thing that bounds us altogether is really about CUDA and the overall development platform. We have more than 20 million CUDA downloads a year, more than 6 million in those last year.

What we are seeing also is an expansion of NVIDIA and being able to power so much of the overall enterprise. We have seen managerial discuss in terms of what we're seeing with our collaboration with VMware and Cloudera as well as the introduction of the NVIDIA DPU data center on a chip architecture software as well. You also got to hear about NVIDIA's edge AI for the Internet and then the trillion things out there on the edge and our fleet command reoccurring revenue service that we will be adding later.

And then here with Kimberly, we've heard about our focus in terms of NVIDIA healthcare as a very key part in both drug discovery and key partnerships with important areas in the UK, such as GSK. This is allowing us to broaden our overall ecosystem or customer adoption every server, every storage OEM, hundreds of ISPs, thousands of enterprise, and just keep in mind here at GTC, we are also exposed to more than 2 million -- 2.3 million developers focused on our overall computing platform.

I'm going to take this next opportunity to discuss an important piece of the high level view that seizes up our market opportunities for NVIDIA Concur [ph]. First, NVIDIA RTX targets the large and growing markets of gaming and professional visualization. Our trailing 12-month revenue in these two markets is close to $8 billion, and representing an 18% CAGR over the last five years. Computer graphics is the first and holistically the largest application of NVIDIA's GPUs. Our graphics growth going forward will be fueled by the expanding universe of gamers, traders and professionals, which already have a number over $1 billion around the world.

One day we expect every human will be a gamer or connected to others in virtual worlds.

Our RTX platform and the Ampere architecture launched this year was a giant step to making that future a reality and a foundation for our growth in graphics over the next decade. While gaming was historically our largest revenue driver, last quarter for the first time, it was eclipsed by our data center platform, driven by AI. Last quarter was also the first to include our Mellanox acquisition. So let me take this moment to update you on our total addressable market opportunity for data center, which is significantly expanded with the inclusion of Mellanox.

We see a $100 billion TAM by 2024 across the four main markets within data center, including high performance computing, hyperscale and cloud, enterprise and edge. As you may recall, last year we sized our data center TAM at $50 billion by 2023. So let me help you understand the drivers for this expansion.

First with Mellanox, we are now addressing the large data center networking market with a particular focus on high performance hyperscale and software-defined environments. This adds over $20 billion to our TAM. Second, as you heard from Manuvir's presentation, we are introducing a new class of processor in the data center called data processing units or DPU. The DPU offloads a substantial amount of the processing currently done by the CPUs, as well as processing done by other data center infrastructure today, in other words, data center (inaudible). This new process outs more than 10 billion to our estimated TAM in this period.

And third, you heard from Justin's presentation, we are enabling emerging edge AI market with our EGX computing platform. This adds approximately 10 billion to our TAM. Each of these opportunities is uniquely enabled by the combination of NVIDIA Compute and Mellanox networking, and we are delighted to have Mellanox team on board.

So that concludes our prepared remarks for today's presentations. We will now turn it back over to the operator, and we will open up the line for Q&A.

# Questions And Answers

## Operator

(Operator Instructions) Your first question today comes from Aaron Rakers from Wells Fargo. Your line is open.

## Q - Aaron Rakers {BIO 6649630 <GO>}

Yeah. Thank you for taking the questions and doing the presentation. I just want to, Colette, if I can just touch on briefly the TAM assumptions that you're making. I guess, first of all, just kind of the general kind of GPU TAM, can you give us any kind of framework of how you're thinking about attach rate on GPUs for the data center?

And kind of a similar question as you think about the time horizon, the 10 billion opportunity on data processing unit, what's the underlying assumptions of kind of the industries move, I mean every server incorporating some form of a DPU in them? Thank you.

## A - Colette Kress {BIO 18297352 <GO>}

Yeah. Thanks, Aaron, for the question. First, let me start off with, it's a great place to look out in terms of the overall server environment out there, and the use of overall GPUs and acceleration, as well as AI in many of those servers. Nothing has changed in terms of our view over the future. Very similar to, everybody would be a gamer, we do expect everything in the overall data center environment to be accelerated and the movement of AI is getting us there. So it is early in these days to talk about that attachment. We have discussed already our continued growth in that overall attachment, but coming after a very, very, very small base of folks that have been focused on that AI.

Our expansion that you've seen us do here in today's presentation in terms of GTC as overall is really expanding to all of the different types of workloads that are out there, as well as the different customers, whether they be enterprise, whether they be focused on the edge or in the core of quality workloads within that data center to really have a stronger attachment as we go forward. So in summary, nothing has changed. Our goal in terms of getting all of the servers to be accelerated with the use of AI is still there.

## Q - Aaron Rakers {BIO 6649630 <GO>}

Thank you.

## Operator

And your next question will come from CJ Muse from Evercore. Your line is open.

## Q - CJ Muse

Yeah, good afternoon. Good morning. Thank you for the presentation today and thank you for taking the question. I guess, if I could ask two. First, for Jensen, you discussed the number of AI platforms in your keynote this morning. And just to help us prioritize where we should be focused, where do you think the biggest revenue opportunity is looking out over the next one to two years?

And then perhaps a question for Kimberly. You announced a partnership with GSK, working with AstraZeneca, you got Cambridge-1. Really curious how you think about your go-to-market strategy and the revenue model for your medical business? Thank you.

## A - Jensen Huang {BIO 1782546 <GO>}

Yeah, CJ. AI started in research, and the first five years of the work that we did and most of our conversations that we had was related to the groundbreaking work that

was being done, superhuman image recognition, superhuman speech recognition, and now near human natural speech synthesis. The ability to process data at a scale no humans can so that it could predict recommendations on conversational AI, because of the recent breakthroughs and natural language understanding. The first five years was really focused on groundbreaking work and the early works of self-driving cars, robotics, et cetera.

The first wave of economic growth of AI, the economic impact of AI was in the cloud. And I would expect the next couple of years is still -- unquestionably still be in the cloud. The vast majority, not -- I would expect that the next couple of years, not only will the cloud grow in very significant percentages, but often quite a large base now. And it's a multi-billion dollar business. And as we said, it has now passed the tipping point where any service, any application can take advantage of NVIDIA GPU inference, because it's in every single cloud and it's in such large abundance.

The amount of computation -- aggregated computation of NVIDIA GPU for inference now exceeds and cross -- surpassed CPU. And if it's growing at a factor of 10 per couple of years, in a couple of years 90% of the world's total inference compute capacity would be GPU accelerated -- NVIDIA GPU accelerated. And so, I think that the -- as you've seen in other type of platforms, once it reaches the tipping point, the acceleration of adoption actually goes up and -- for obvious reasons, because people feel really, really safe now that they could take advantage of it, because they can always count on the NVIDIA architecture in every cloud, in abundance of it in every single cloud. And so, I think the next couple of years you will see NVIDIA AI growing in the cloud and service providers into large numbers and ever larger numbers.

The next wave is enterprise. And enterprise, we've described in two ways. Enterprise is helping to automate company which is a lot of the work that Manuvir was talking about. It requires us to re-architect the data center, the system software has to be re-architected once, and the reason for that is because enterprise software and the enterprise data center infrastructure is very different than that of our clouds. And so we have to work with partners, particularly VMware to do a lot of computer science around the current stacks. And then of course the data analytics applications, we're going to grow in enterprise even before we finish with VMware, because many of the early adopters are perfectly comfortable building multiple infrastructures, but that's going to get turbocharged incredibly when the work that we do with VMware over the next several quarters come to market. Meanwhile, we're preparing the ecosystem as we speak.

The next wave is edge and autonomous. And so, the -- now that the enterprises and the companies are comfortable and have mastered processing large amounts of data, they are also collecting a giant amount of data and that data is connected to sensors or webservices or applications or products and before you know these things that are out all over the world today, it could be anything, a lawn mower, a refrigerator, an elevator, air conditioner, you name it. They're all going to be connected to either 5G or WiFi and that allows them to collect more data, in turn all of these products into essentially smartphone connected smart device. And so the

iPhone moment is coming. That's the wave after that. And so we're preparing for each one of these waves and we hope that they happen just bam, bam, bam, bam, and it just keep on happening over the course of the next five years, surely. But we're -- you're looking at one of the largest computing opportunities ever.

Kimberly has got the next question. Go ahead, Kimberly.

## A - Kimberly Powell {BIO 22145194 <GO>}

Yeah. Thank you. Thanks, Jensen. CJ, thanks for your question. So a bit about healthcare's go-to-market strategy, it's much like all of NVIDIA's enterprise go-to-market strategy and industries. If you think about what we're building with Clara, we are building a domain specific computing platform for healthcare, for computational healthcare and we deliver what we call a full stack silicon systems and software and that software, if you think about, what I described in Clara Discovery, there are aspects of drug discovery where there is still just an insatiable demand for compute. Doing that -- looking at exabytes of genomic data in the near future and trying to use artificial intelligence to extract associations out of these very large populations of genomic data or looking at the vast chemistry space of 10 to the 60 potential combinations and using a combination of search to docking to simulation, all done in silico today, we still can totally consume the world's fastest supercomputers in doing that work and we're doing that with COVID today. The systems of COVID are more complex than we've ever simulated. There is still an insatiable demand there.

And then as we move into these new budding areas of natural language processing, as Paresh talk to you about, the fact that they are growing in complexity, in size to train these models every 10 months. That's no different. I mean these models biomedical specific natural language models take tens of thousands of GPU hours to train. So we have just an incredible amount of opportunity ahead of us and that's what we're really inspired to do with the Clara platform. We hope to catalyze and quickly disseminate the capabilities through building out of Cambridge-1, enabling the industry leading researchers with it and then longer term, over time, we have plenty of opportunities to monetize across all three; systems, silicon and software.

## Q - CJ Muse

Thank you, both.

## Operator

And your next question will come from Vivek Arya from Bank of America Securities. Your line is open.

## Q - Vivek Arya {BIO 6781604 <GO>}

Thanks for the presentations, and thanks for the question. I actually had two as well; one for Colette first, and then one for Jensen. Colette, I was hoping if you could give us a sense of the supply situation for Ampere and -- on both the data center and the gaming side? Good to have a lots of demand, but just how is the supply situation working out? And there, if you could talk specifically to that gaming side as well?

And then Jensen, my question is, where does Arm fit into your data center vision? Because from what we heard today if more of the workload and value are going to the pre-processing or the DPU or the smart NIC, which I realize contains some embedded CPU, but more of the value is going to that DPU and various kind of accelerators and GPU, does it matter to you one way or another, whether it is Arm or x86, that's a CPU architecture of choice. And, in other words, is it really critical to owning Arm or do you think you can achieve similar levels of success by just optimizing DPU and accelerators, because that's where most of the value in the data center is shifting anywhere?

## A - Colette Kress {BIO 18297352 <GO>}

Vivek, so thanks for the first question to discuss in terms of supply. We're very comfortable with the supply and where we are in terms of that supply for our outlook that we have provided. When we turn to our overall data center and we look at the overall Ampere architecture, keep in mind the A100 and that going forward is a very complex product. I mean, it will probably take multiple quarters to really work through all the supply needs and get that to market in its full capacity.

You also focused in terms of Ampere architecture for overall -- for gaming, we're in the initial stages of ramping that. It will take some months for us to fully supply to the channel, but we are right on track in terms of providing that. Sure, we'd love to have more supply sooner when we're ramping, but we're also executing to our plan, so we feel very comfortable with the supply and what that means to our outlook.

## A - Jensen Huang {BIO 1782546 <GO>}

We can achieve extraordinary success and all of the success we've talked to you guys about without Arm. However, with Arm there are some really exciting things we can do, let me highlight two of them. The first one is extending NVIDIA's architecture accelerated computing to the Arm ecosystem. You might notice that we've -- the accelerated computing is surely here, it has passed the tipping point and everybody acknowledge that this is the way to go forward, that Moore's Law has ended, it's ended last week, it didn't revive this week. And in order to extend computing further, we have to bring accelerated computing to all device all computing devices, including Arm.

The benefit of owning Arm is that we could also offer NVIDIA accelerated computing in soft IP form, not just hardened IP form but soft IP form, NVIDIA is an IP company, we're not a chip company, NVIDIA is an IP company, because I'm pretty sure that TSMC makes the chips and I'm pretty sure that when we deliver to them, that was effectively an email as their completion of a multi-billion dollar project and so NVIDIA is a soft IP company, we're a IP company and the benefit of having Arm is to have this team that has a vast network to their ecosystem of Arm and all the devices that they are in, the billions and billions of units that are sold every year and we can extend Arm with accelerated computing, AI computing that they are is -- renowned for.

Second, we are going to bring a lot of platform technology to ARM in a whole bunch of new data center environments. It could be high performance computing, cloud

data centers, enterprise data centers as we were talking about earlier with GPUs, edge data centers, a whole bunch of new technologies that we're rolling out. As we turn the CPU core of ARM, which is world class, I mean, this is absolutely the most energy efficiency CPU core in the world. As we turn to CPU core into computing platforms, we're going to bring up -- we're going to deliver a lot of value to Arm, we're going to create a lot of value in Arm beyond the mobile device. These are all markets that I'm talking about that are really nascent and as we create almost had value around the Arm platform that would be great to own it first. And so I think we have two enormous opportunities to extend NVIDIA's accelerated computing, to a large ecosystem around the world. And, secondarily, we're going to create a lot of value around data centers and servers, computing platforms that are very nascent to Arm and we would love to own it as we create the value around it.

## Q - Vivek Arya {BIO 6781604 <GO>}

Thank you.

## Operator

And your next question comes from Timothy Arcuri from UBS. Your line is open.

## Q - Timothy Arcuri {BIO 3824613 <GO>}

Thanks a lot. I had two as well. I guess the first question is for Colette. How much of the $100 billion in TAM is China? That will be my first question.

And then the second question for Jensen, and really it's expectations around your share of that $100 billion TAM. Is there a way to sort of handicap what would be a reasonable share assumption within that non-gaming TAM? I guess, I asked because if I sort of take what your non-gaming and your non-Mellanox revenue as this year relative to the TAM that you set forth a few years ago, it seems like you're maybe mid-teens to like 20% of that TAM. So I guess the question is, would you be disappointed with say 20% share of that TAM by 2024? And I guess, ask a different way, is it really that you expect to gain share within that rising TAM where the story really is riding the growth in the TAM? Thanks.

## A - Colette Kress {BIO 18297352 <GO>}

So let me first start with the question regarding our TAM and a regional breakout. We don't have the ability at this time to really look at that opportunity by region, but keep in mind, for each of the areas that we're focused on, we have the opportunity to take that to each and every single region. We can look at it by the additions that we have added with Mellanox, with the GPU, and with the Edge or we can look at it in terms of the types of customer markets that we will also be addressing. And this is an opportunity for us to focus on both our hyperscales and the massive expansion in terms of the cloud. That will be a considerable portion of our overall TAM as a whole.

Additionally, you can think about the overall enterprise opportunity. A lot of growth just announced today in terms of key areas of the enterprise that we can have a focus on. High-performance computing has been a big part of us for more than 10,

12 years. So again, expansion in terms of bringing further acceleration in AI to that is also an area. And now the Edge and you can think about those devices, even in all of the regions has the opportunity to grow. So we've expanded each and every single one of those different areas today in terms of our increase to $100 billion with the Mellanox, the DPU, and the overall Edge. And yes, our region, such as China or the US can definitely benefit from that.

## A - Jensen Huang {BIO 1782546 <GO>}

We address the entire TAM except for x86 CPUs. That's really the simplest way to think about that. And in almost all computing platforms that we serve, accelerated computing is the most important part of it. And going forward, if we believe in the thesis -- if we believe in the thesis, two important thesis. The first one, all applications in the future will be infused by AI. For example, something like Microsoft Office will be infused by AI. Then I would suggest that AI would be in every computer, and I would suggest that every computer will be accelerated. I am certain every computer will be accelerated in fact, just as I was convinced in 30 years ago that every PC will be accelerated with GPUs. And so, I am certain of that. I am certain of the fact that the CPUs alone will not do the job. That is complete certainty.

And so I believe in the thesis of AI, I believe that AI will be in every application, and AI will open new markets and create new applications that weren't writable before that no humans know how to right. And so that I believe in and therefore accelerated computing is going to be everywhere. And in all the platforms that we're in, accelerated computing the GPU and the networking frankly dominates the vast majority electronics inside those computing platforms.

Second, I believe in the world of zero trust. I believe that protecting data centers at the perimeter is historic. It's like building a big wall. It makes no sense. That the future of security is about trust -- is about zero trust and it's about securing every single transaction, every single node, every single application. And therefore, you need to take what it used to be security appliances at the perimeter of the data center and put it into the servers, every single one of them, every single network, that's the reason why the networking chip is going to be the most important security chip in the future, because that's where all the input and output comes from. That's exactly where you want to put it, and that's exactly why the GPU is invented.

Security is going to force every single computer on the planet to have something like a DPU. And so, I believe that every single server node will be accelerated by a DPU. And every single server application will be accelerated by a GPU. And, therefore, the vast majority of the world's TAM minus the x86 CPUs will be our opportunity.

## Q - Timothy Arcuri {BIO 3824613 <GO>}

Thank you, Jensen. Thanks.

## A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Thank you.

## Operator

Your next question comes from Stacy Rasgon from Bernstein Research. Your line is open.

## Q - Stacy Rasgon {BIO 16423886 <GO>}

Hi guys. Thanks for taking my questions. I have two as well. First, to go back to your chart showing GPU influence workloads in the cloud versus exceeding all CPUs. How does, I guess, given that trend, how does your cloud revenue today split between inference and training? And how are those each -- like, what were the trajectories of growth that are relative for those two, relative to each other?

And the second question, can you tell us a little more about how the revenue model for the VMware partnership works. Is it just working to incentivize quarter GPU use in enterprise or is there a more direct revenue share? Something about that. Thank you.

## A - Colette Kress {BIO 18297352 <GO>}

So let me give you --

## A - Jensen Huang {BIO 1782546 <GO>}

(Multiple Speakers) both of them.

## A - Colette Kress {BIO 18297352 <GO>}

Yeah. Why don't you give it a shot?

## A - Jensen Huang {BIO 1782546 <GO>}

Okay. The reason why our inference -- well, first of all, you know that our inference -- hello.

## Q - Stacy Rasgon {BIO 16423886 <GO>}

Hello. Yes.

## A - Jensen Huang {BIO 1782546 <GO>}

Yeah, we've got some reverb. That's okay. The first thing is, of course, Stacy, you know that our inference acceleration business has been growing very, very quickly and -- for all the reasons that we've already talked about. We turbo charged it even more this time with Ampere, because Ampere is our first universal GPU. We used to have essentially three GPUs in data center. And one of them would be very heavy duty training systems to build these large models -- AI models.

The second is the cloud training platform, which is based on PCI Express. Like the new GPU that I just -- that we announced today, the NVIDIA A40 is going to go into all the clouds and easy to deploy. It's easy for them to put into all of their PCI Express

servers, because it's based on PCI Express. The A100 is based on SXM2, a different type of networking and a different type of system architecture, in fact, very different.

And so, the second which is cloud-based -- cloud-based training was our second GPU. And our third was our inference GPU. Well, what we did with Ampere is we combined it into one architecture, and so, you could literally now use the Ampere, the A100 SXM base for both training, as well as cloud training, as well as inference, one single architecture. And then -- and if you like -- if you would like to have PCI Express versions, because your cloud data center is able to facilitate a lot more PCI Express servers, we have the A40 GPU which now allows you -- the A100 and the A40 GPU depending on sizes that allows you to do both training and inference. So we now have an architecture that's universal. It does training, it does influence, it does computer graphics. It does all the things that we -- that you know that we do very well, all into one architecture.

And so, that's the reason why our inference performance is going to not only continue to grow at the historic rate, which was really, really high, both the number of units and the fact that we're increasing throughput by a factor of 20 generationally. We now have the ability to put a lot more unified aggregated GPUs and cloud that are inference based. And so -- so that strategy was a really good strategy.

You know Stacy, what people want is, they want to make sure that if they were to use a cloud evolve software forward a type of technology they develop software for capability, it could be video decoding, it could be whatever it is, x86 or Arm or in our case and there is accelerated AI. Whatever capability they use in a cloud is available in every cloud, so they have the flexibility, which -- that is established. And number two abundance of capacity and that's why the tipping point of our aggregated AI Inference throughput is such a big deal. And I think at this point because of the rate that we're growing, it's a forgone conclusion it's going to be a vast majority of computing cloud.

Colette, what was the second question?

## A - Colette Kress {BIO 18297352 <GO>}

Second question was on VMware (multiple speakers) (technical difficulty).

## A - Jensen Huang {BIO 1782546 <GO>}

Yes. Yeah, VMware, there are three ways that we benefit and then I'll leave the most important one last. The first way we benefit, of course, is VMware running on -- VMware, as you know, is the data center operating system. They represent 70% of the world's data centers. This operating system is really computationally complex these days and the reason for that is because of software defined data centers, the networking stack, the storage stack, the security stack, the virtualization stack, it's all running in VMware. And so the first thing that we're going to is we're going to offload and accelerate and isolate the beta playing from the application play and that all flowed alone creates really fantastic opportunities for our DPU. It is cheaper, it is purely faster and unquestionably more secure to have VMware running on x86

plus of DPU instead of x86 without a DPU. So the first is creating an opportunity for our GPUs.

Second is, every one of the VMware stacks goes with a virtualization stack of our GPUs and that hypervisor -- there is the VMware hypervisor and then there is essentially the NVIDIA hypervisor for virtualizing all of our GPUs. The virtualization of GPUs is really complex and it opens up CUDA, opens up our graphics RTX, it opens of cuDNN, it opens up all of the capabilities that is buried underneath the hypervisor otherwise. And so the second thing is, it opens up the virtualization stack, we call it the GPU for all of our data centers. Okay?

And so the third and this is really the biggest one, which is the ability for enterprises to be able to accommodate all three domains of computing from scale out virtualized to micro services. All three of these harmoniously and basically transparently in their data center using VMware. And so the ability for enterprise IT to easily adapt and really accelerated AI is now you don't have to think about it, it was just like before. VMware has never fully transparently integrated NVIDIA GPUs or any accelerators aside from the CPU until now. This transition is a big, big deal and it opens up great opportunities for VMware into all of the worlds of AI, opens up great opportunities for us to be able -- to transparently, seamlessly, easily integrate NVIDIA AI into all world's data centers. So three ways.

## Q - Stacy Rasgon {BIO 16423886 <GO>}

Got it. Thank you, guys.

## A - Jensen Huang {BIO 1782546 <GO>}

Yeah, thanks a lot, Stacy.

## Operator

And your next question will come from John Pitzer from Credit Suisse. Your line is open.

## Q - John Pitzer {BIO 1541792 <GO>}

Yeah. Good morning, guys. Thanks for the presentation. Two questions here. Jensen, both in your presentation and I believe Paresh's presentation, you talked about software writing software and sort of this iPhone moment, I'm wondering if you could just help us elaborate on that a little bit. Is the analogy here that you're sort of the iOS and you'll be collecting a recurring revenue stream on some of these sort of AI apps the same way Apple does? And if so, should we expect to hear about other Mercedes likes deals coming down the pipeline and as you think about monetizing software is that part of your $100 billion incremental TAM that Colette talked about this morning or is that above and beyond?

And then secondly for Colette, you talked about kind of Ampere being extremely strong and you guys clearly guided for that in October. You also guided gross margins to be under a little bit of pressure as you ramp to new product. I'm just

curious, is the strength that you're seeing just that demand is outstripping supply or you're having some supply issues as well, how should we think about some of your gross margin comments? Thanks.

## A - Jensen Huang {BIO 1782546 <GO>}

Our $100 billion TAM does not include many of the things or almost all of the software things that we've talked to you guys before. It didn't even include what I mentioned just now to Stacy about our virtual compute. It doesn't include GeForce NOW, it doesn't include DRIVE, it doesn't include Fleet Command, it doesn't include our software stacks for enterprise. And so we'll have plenty of opportunities to talk to you guys about that in the future. Today, we want to just stay very, very focused, keep it nice and conservative and we have plenty to talk about. Our NVIDIA is a full stack accelerated computing company as you know and our opportunities and we're an open computing platform.

We work with our network partners in any way they would like. And in some -- for some they would like to hold stack and reason for that is because ours is just so good, we're so good at it. And for some, we'd like to develop some parts of it and use hours. And so we work with them to figure out which part of it they would like to use on their own and which part of it they would like to use ours. And some people would like to build all of their own and use a lot of our open source tools or our libraries, like CUDA-X or our RAPIDS open source stacks to build their data analytics services. And so we're an open computing platform that works across the multiple layers from the technology to the system, SDKs to the application frameworks. We want to be able to work with the entire world's industries and democratize AI and bring accelerated computing to as many places as we can. And -- but what we've only captured so far in our TAM is the hardware stuff, the hardware stuff which is pretty good as you know.

## Q - John Pitzer {BIO 1541792 <GO>}

Just on the Mercedes deal, as you run numbers, it's not hard to envision kind of a software recurring revenue stream, which is as large as the hardware revenue stream, is that how we should think about the potential for service and software recurring revenue relative to the $100 billion TAM?

## A - Jensen Huang {BIO 1782546 <GO>}

Yeah, that's a great way to think about it, if not more. And the reason for that is this -- and I don't mean for you guys to go make any changes in your models, okay. But just listen to the strategy, listen to the strategy and think about the implications and think about what we're going at and building it all in public and all these pieces are being talked about at GTC and so you guys know where we're going.

There is no question that in the case of autonomous vehicles, the business model that we created with Mercedes is really quite extraordinary, it's potentially the best one. And the reason why is because, first of all it's incredibly hard to be able to create a computing platform that integrates into the most safety concerned, the most safety conscious companies and industries with accommodation for all of the heritage and all of the existing technologies to be able to integrate into that, infuse

into that harmoniously is a great challenge. And so that's the beginning. That took us 10 years to learn. Now that we're inside the car and we're the computing platform, then we can create the applications that sit on top of it. Because unlike it -- unlike a smartphone, you're not going to be able to create an application in the cloud and download that works in every single Android phones, it's just not going to be like that and the reason for that is because of safety. These applications are safety concerned first and so each one of the application opportunities belong to the car company.

And so this is one of the reasons why such an extraordinary thing. If Daimler can figure out a way to make all of their car software defined and turn them into application platforms, they'll grow that application platform 2.5 million cars a year, they'll grow instantaneously. Over the course of a decade, you could just imagine the economic opportunity that they're going to create. Our business model with them is to share that. And so we're doing a lot of the hard work of course as well and that becomes one that's a significant opportunity.

For us, now we would like to do this -- we're an open platform company and we would like to do this with other car companies and there will be other opportunities. And the reason for that is very simple. The number of companies in the world that could create an end to end self-driving car stack that is world-class, and then you can deliver on real streets is very, very few, that integrates into the existing car industry is very, very few. And so, I think this is a great opportunity and we're going to continue to scale it. But you're going to see other examples like that. You can see other examples like that. We will always have free community versions, that's just one of the policies we have. The community versions will be free, the developer versions could be free, and there always be free versions.

But for some companies, they want to make sure that we're on the hook on it and they want to make sure that we have some kind of an enterprise agreement and enterprise business model that allows them to get in front of open source or allows them to get in front of developers and others so that they can get the software in hands towards software debug.

## A - Colette Kress {BIO 18297352 <GO>}

And let me see if I can touch based on your second question that was regarding overall Ampere demand and the impacts in terms of our gross margins. So far, the gaming demand for our RTX 30 series has just been off the charts. We had expected a really great holiday season. We knew that our overall platform that we will bring in was the best generation to generation performance ever. We've got a great release of fall games that are coming out. And the work from home is even bringing more and more to the gaming and the entertainment and social arena. So we are racing to catch up to that demand, but the ramp is going well. The yields are very good. So all of that is intact.

Now when we think about our gross margin that remind what we did in terms of our gross margin outlook for the Q3. Our outlook for Q3 and as usual with most quarters mix is our driver of overall gross margin. We expect a very strong sequential increase

for our gaming. And with that gaming piece of our business being so strong, we took a slight sequential dip in terms of our guidance for gross margin. Everything seems to be in place, intact, so no change in terms of our gross margin in terms of what we're seeing.

### Q - John Pitzer {BIO 1541792 <GO>}

Very helpful. Thanks guys.

### A - Jensen Huang {BIO 1782546 <GO>}

Thank you.

### Operator

And your next question will come from Mark Lipacis from Jefferies. Your line is open.

### Q - Mark Lipacis {BIO 2380059 <GO>}

Hi. Thanks for the presentation today. The question, I think this one is for Jensen, to provide an integrated data center scale architecture, I'm trying to make sure I understand how far across the data center value chain and how deep in the value chain that NVIDIA -- you feel NVIDIA has to go in order to deliver that. I think it's pretty clear, NVIDIA is not a chip company but a platform company, but maybe if you could compare what you feel you need to deliver across that value chain today to today's data center value chain.

This is NVIDIA effectively becoming the equivalent of companies that are selling processors in the data center today, companies that are selling servers in the software, the networking companies on the software, from the OS to application side, how far up the software stack you're going? And you could provide maybe if you had an analogue in today's data center versus the data center scale architecture you're delivering in the future? Maybe that would be helpful to level set. Thank you.

### A - Jensen Huang {BIO 1782546 <GO>}

Sure. The first -- the first -- there are three questions I'll hit right away, and then I'll explain. One, we're not like a company that exist today, because AI is a problem and an opportunity, a challenge and an opportunity unlike any software that's ever been written before, otherwise why could we do all the things that we're doing right now, number one.

Number two, we are not doing integrated data center, we're not doing that. Number three, we innovate as much as we need to -- as much as we need to and as little as we can -- we innovate as much as we need to as little as we can which is a guiding principle of NVIDIA to do as little as we can. We're not trying to do everything, we are only trying to do the things that we have to, we're only doing the things that the world relies on us to do, we're only doing the things that we do if we don't do it, the world just they can have it.

It is absolutely the case that if we don't do what we do today, the world doesn't have it. It is absolutely the case that when Gelsinger and I worked on the VMware partnership with us, if the two of us don't do it, it just doesn't get done. It just won't get done. And so, we have to go and do the things that we do because on what we do what the world doesn't have it. So I answered those three questions very quickly. Let me now give back off a little bit and explain.

One, the way that AI is written, the way that AI software, it's a computer that are learning from data, it's learning from data that we collect and we asked, we coach it, we influence it on the type of neural network architecture and the type of data that we presented, and the way that by which we wanted to learn, we coach it, we're like a coach, we're like a teacher, and then it goes off, you know runs for days and weeks, and it does it over and over and over again on junk [ph] amounts of data and writes the software. When it's done writing that software, we've made -- we can't read it. It's unreadable. It's like a neural -- it's like a brain dump of somebody's brain and it's not readable and it requires a new type of computer to run it.

And so, from the way that the software is written, the methodology by which is written, the infrastructure pressure it creates and during the POD [ph], there were some (technical difficulty) -- talking about it, the four speakers were fantastic today, I really appreciate their work. And you could hear in their phrases, the tip of the iceberg of the challenges in computing that we're solving. And so, the software is the different, the way it's written is different, the pools [ph] are different, the pressure on the infrastructure is different and the coordinates now that is different.

In no time in history did we see that ALBUS 1 [ph] the data center now, the enterprise data center advance to the hybrid cloud as the enterprise data center has to manage three computing environment. That's never happened before. One computing environment is bare metal (inaudible) computing like a supercomputer.

Number two, it's virtualized multi-tenant. Virtualized, easily manageable, easily scalable, easily secured, multi-tenant. And, third, continuized micro services. The core is far out of the Edge, you'll never visit it again. You drop the server, you connect it to your network, and hopefully you never go back to that warehouse and you never go back to that store room ever again. And you manage it from one thin glass very, very far away. These three types of computing domain has never happened in one company before. We've got to go make it happen and it's never done for AI and it's never done with the GPUs.

And so, we have to go create the necessary GPU. When we are done creating it -- when we're done creating it and this is a very, very important. When we're done creating it, we open it up like SDK, we open it up like SDKs, system SDKs, ADPS [ph], EGX, AGX, they're all SDK, they are hardware components that OEMs can integrate. Then we put on top of it our entire software SDK, and then we put on libraries on top of that and then for the application developers when they create tools like application frameworks which are basically AI skills that we pre-train. All of this stuff is put into the cloud.

All of this stuff is put into the cloud, and it's all certified, it's all optimized, it runs in data centers, it runs in cloud. We, as a result, can connect up a network of partners of software developers, system makers, solution makers, cloud service providers as they want the partners all over the world. And they can all run into the AI, they can all run NVIDIA that's already computing, they can all run NVIDIA Clara, they can all run NVIDIA RAPIDS, they can all run NVIDIA Isaac, Jarvis, Mervin, all of the SDKs that we created. So I hope that's helpful. I know we look very different. However, accelerated computing needs to look different, because Moore's Law is finished.

And number two, AI is different, because it's not written by humans, it written by machine. So the world has changed, that's why new type of company needs to be created.

## Q - Mark Lipacis {BIO 2380059 <GO>}

Very helpful. Thank you.

## A - Jensen Huang {BIO 1782546 <GO>}

All right. Thanks a lot.

## Operator

Your next question will come from Matt Ramsay from Cowen. Your line is open.

## Q - Matt Ramsay {BIO 17978411 <GO>}

Yes. Thank you very much. Good afternoon and good morning. Jensen, I had a couple of questions for you. On the first one, I noticed in the BlueField DPU roadmap eventually you guys integrate sort of the full AI engine into the DPU. And so I wondered if you could talk a little bit about on the acceleration side what AI opportunities and low-hanging fruit might be available in -- particularly in the security domain? And then the second question is, you now have AI acceleration, you have a DPU acceleration, for an integrated stack, I mean whether you buy Arm or don't buy Arm, it seems like you could make a CPU and maybe you could talk a little bit about the pros and cons of going after that one piece of the TAM that you're not addressing today? Thanks.

## A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Excellent question. Number one, I'll give you example, intrusion detection. Intrusion detection will be distributed not at the perimeter, the data center will be protected at every single transaction at every single node, at every single application, it will not be protected only at the doors. And the reason for that is, don't forget, all of the intruders are largely inside the building already. In the future, you also have public clouds. The entire data center is opened to the world, you can't allowed to have East-West intrusion. The moment an intruder goes inside a data center, goes sideways, East-West, into the data center imagine the damage they could do.

Security is incredible. Every single node will become a super firewall. Firewall technology today, intrusion detection technology today based on AI. We've got to put AI processing right at the network, number one -- number two.

Network shaping, network traffic shaping, that's an AI problem, it's optimization problem that cannot be done with a simple set of equations and it's a heuristics problem, it's a dynamic problem. It's one of those computer science problem that goes, well, it depends; what's the solution? Well, it depends. And so the, well, it depends requires intelligence as we want to put intelligence right at the network.

What computer sciences and what we had described in the past as in network computing, these are two examples of in-network computing. Okay. So very, very big deal. We're super excited about doing that and this is one of the reasons why -- one of the reasons. And if you go back to the very early days when I talked about the acquisition of Mellanox, I talked about normal computing, I talked about how the network itself will become a fabric when we do a lot of computational, a lot of AI, this is the first step.

There are several things that we can't do. There are many things we can do with Arm, there are many things we can do with Arm, can do with Arm, and we can build a CPU. However, there will never be another CPU built like Arm ever and it's not a computer science problem anymore. It started out as a computer science problem 30 years ago, an energy-efficient architecture that's designed like no other. That was visionary and the reason for that is because if you are energy efficient and the world hits the wall because of the end of Moore's Law, you've got more runway. There is more runway in Arm than there is in x86, there's just more, I mean that's just the bottom line.

The architecture is more energy efficient, therefore they got more runway. However, Moore's Law end is for them as well. And so we need to bring accelerated computing to Arm. Arm number one is just the genius of an architecture, but it's also a genius of a business model and the reason for that is because they wanted Arm to be the most popular CPU in the world, they want it to be used in all kinds of things from everything to everything, cars to phones to televisions to you name it. And so that required a business model that allow them to license their IP in soft form, soft flexible form that fits into other people's chips, because many computers in the future are full -- are just the whole data center is on one chip, the whole computer is on one chip, the phone is on one chip, TVs on one chip. There is no such thing as two chips. Just on one chip.

And so the second thing that they have because of the business model created the third thing, which is ultimately the most valuable thing today, which is their vast ecosystem. The execution machine of Arm that knows how to build soft IP, their IP for smartphones, for embedded systems, microcontrollers to the data centers and now increasingly PCs, there are number of CPU cores, the engine they have behind it that creates the soft IP, productizing it and delivering it to customers, helping them integrated into their chips, that's phenomenal, that's phenomenal. And, as a result, they created this ecosystem of thousands of chip companies, they ship 22 billion

chips last year, NVIDIA shipped 0.1 billion. So the difference between NVIDIA and Arm is 22 billion. So as you can put it in perspective, the reach of their ecosystem, that's the value to us.

You can't do that by building another CPU, it doesn't matter what another CPU is, I don't care what it is. I just don't think the ecosystem will ever be as rich as this one, it's a 30 years to build that took enormous character enormous vision to built it. The team that have built it is phenomenal. They love Cambridge, they work in Cambridge, it's a great computer science team, I love to work that they've done, that ultimately is the asset that we're buying, the combination of all that and the ecosystem that they've built up that just simply won't get replicated again.

## Operator

And your next question will come from Harlan Sur from JPMorgan. Your line is open.

## Q - Harlan Sur {BIO 6539622 <GO>}

Good morning, and thank you for taking my question. One of the powerful dynamics that the team is creating for itself is leveraging the entire portfolio to target vertical markets and so question for Kimberly for the vertical market focus like healthcare, the drug discovery opportunity that you talked about is primarily focused on high performance computing platforms like EGX but moving across the portfolio how is the team leveraging the edge and inferencing portfolio like EGX and Jetson product families, within your healthcare franchise and how involved is your team in hoping to define next generation hardware and system platforms?

## A - Kimberly Powell {BIO 22145194 <GO>}

Yeah. Thanks, Harlan, and thanks for the question. So on drug discovery, you're right, the DGX system as it's a full stack architecture, it does literally cross every computer science domain. Yes, it's very heavy and upcoming in artificial intelligence, but of course it takes advantage of accelerated computing. It's also going to be very instrumental in data science and machine learning and data analytics as we move into these gigantic data sets of genomic data sets or even doing the compound screening, what we do is we generate huge outputs of that analysis that needs analytics to really call out the necessary information. So we literally leverage every corner of NVIDIA's extreme and powerful and world-class computing architectures, whether it'd be accelerated computing, data science, machine learning, deep learning and natural language processing.

And the other areas and Justin touched on it briefly, we have for decades been working in edge devices frankly, revolutionizing the medical instruments that care for us and see inside our bodies and extract our DNA and build out the 3 billion letters that make us up. And so these instruments are one of the edge platforms that in the future just like our phones and our cars, they want to be software defined. You -- the sensor technology that is created in these edge medical instruments is incredibly powerful and we can continue to get amazing insights and new information out of the sensor technology by applying artificial intelligence, but it can't remind the old way of deploying these instruments where you would sell a couple of million dollars

CT scanner and it would refresh every 10, sometimes 15 years. We can't carry on that way anymore. So edge computing is absolutely going to be vital to what is going to be the software defined future of medical instruments.

And then you can imagine not only the new instruments will have -- will leverage everything we've built in our Jetson platform, but it will also leverage what we're building in our EGX platform being able to remotely manage provision and securely operate edge nodes that need to be updated with new AI applications over time is extremely vital.

The example Justin gave was at Northwestern Hospital, where actually there are plenty of sensors that already exist in the healthcare environment, microphones, cameras that can now be coupled with artificial intelligence and conversational AI, so that brand new services can enter the healthcare and hospital environment, just like we would expect, just like we have at our homes, we can talk to smart speakers, and we're unlocking and enabling that environment in the health care system today using what is DeepStream for -- that's used in smart cities, what is Jarvis, that is being used in a lot of our conversational AI platforms. We can leverage all that technology and create a domain specific application framework to over -- literally overnight allow application developers to develop new applications that can be deployed and then leverage the Fleet Command system of EGX to deploy them in the tens of thousands installations and environments that they're going to wireless [ph] at the Edge.

So whether it's new instruments, augmenting existing instruments with compute, coupling all then the amazing sensors that are -- live in our health care environments with AI and then being able to manage securely and deploy applications with EGX. The future is incredibly bright, and we see smart hospitals now and these applications popping up literally out of the woodwork of course as you can imagine to respond to the great demand that the dynamic is putting on the healthcare system.

## Q - Harlan Sur {BIO 6539622 <GO>}

Cool. Thanks, Kimberly.

## Operator

And your next question will come from Raji Gill from Needham & Company. Your line is open.

## Q - Raji Gill

Yes, thank you, and thanks for this excellent presentation. Just when we are thinking about this new cloud class of data center products, that's the DPU, how do we think about kind of the pricing dynamic in terms of the revenue opportunity relative to other class of chips that you're selling. Just trying to understand how this will be kind of integrated and what the go-to-market shares will be for that?

### A - Jensen Huang {BIO 1782546 <GO>}

Yeah, there are two pillars that we could -- we could look towards. One of course is the baseline, which is what a smart NIC goes for, it's a few hundred dollars. And then the other is the modest CPU offload that you will provide such that the application and performance of the server gets a boost. You're effectively going to add a DPU to a server, and my expectation over time is that you'll double the performance of the server, and I kind of put a value on it. So somewhat between those two spacing [ph] and we work it out as we go.

### Q - Raji Gill

Thank you.

### A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Thanks a lot.

### Operator

And your next question will come from Ambrish Srivastava from BMO. Your line is open.

### Q - Ambrish Srivastava {BIO 4109276 <GO>}

Hi. Thank you very much. Jensen, I had a question on inferencing. And then I had one on gaming as well, maybe Colette you could answer that. So on the inferencing, you shared a pretty revealing piece of data in terms of crossing over CPU. And then you gave a projection for a 90% market share. So Intel is incumbent, and so, could you please just help us understand kind of in terms of how much of the training you could translate into NVIDIA inferencing which gives you the confidence that you get to 90% share? And then also, what are you assuming the competition does in that market? And I know Jensen, you know the competition lately?

And then on the gaming side, Colette, I know in the past such events, you guys have been very kind in giving us the components of the CAGR in terms of units and ASPs. So I was wondering if you could help us out there as well? And then you also said that gaming -- sorry, I'm asking multipart question, but you said this is the fastest ramp ever. So if you could just give a little bit more details on that will be helpful. Thank you.

### A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Yeah, so on the -- it wasn't actually about share, it was just about compute -- aggregate compute in the cloud. For example, in my PC right now, in my PC, I have a decent GPU [ph], and the aggregate -- my computational throughput of my PC is 99% by GPU. In fact, the aggregate DPU versus GPU compute inside the data center, inside a supercomputer or high performance computer is about 99% GPU and the reason for that is because that's its job. Its primary job is in data acceleration, its primary job is to compute.

And that's not to say that the CPU is not useful, that's just not its job. Its job is to manage the application, manage the operating system, manage -- orchestrate the processing of the applications, figuring out who gets priority. Those things are really important, moving things around, managing things. Those are really important. And those single threaded performance, single threaded application code is not operating process, then the single threaded part of your code becomes the critical path, otherwise known as Amdahl's law.

And so, that's all in that. It is that the vast majority of the cloud going forward will be accelerated. In fact, that is a program conclusion at this point, and it's a corollary to -- Moore's Law has ended, and therefore you have to look for another approach to accelerate applications, because Moore's Law has ended, and yet on the other hand, the emergence of this new type of application called AI that require so much computation at exactly the time when CPU performance is not going to double every couple of years, and you can't just wait for it.

And so, the world has to look their code and re-factor it and take advantage of acceleration. It happened in a perfectly good time, and the reason for that because -- is because of the cloud compute. Because of cloud computing, the world had to re-factor its application and disaggregate it. And when you disaggregated it -- whenever you disaggregate into containerized certain modules, certain micro services, you might as well re-tolerate it and you must -- because you're infusing with AI anyway.

And so I think the confluence of both the end of Moore's Law and the beginning of AI and the emergence of this new type of data center and the data center scale computing, we call it, are all working in its favor. Colette?

## A - Colette Kress {BIO 18297352 <GO>}

Yeah, so let me see if I can answer your question regarding our split in gaming between our ASP growth and our unit growth. Both of these are very important to our overall growth, and over this five-year period of time both of them have contributed. One of the key things to note in terms of what is both influencing our units and our ASP growth is the onset of laptops, notebooks, gaming. High-end gaming notebooks for this market have really grown quite well and they have great ASPs force as well. So we continue to uplift overall ASPs as our new gamers coming on board tend to take on the RTX, tend to take on our higher performing overall GPU is just to start off with.

We still provide a slew of different overall price points to attract every single gamer, but you can see our ASPs probably over this period reaching double-digits growth, and there is still a great opportunity as we go forward. We also announced that right now the overall Ampere architecture for gaming is growing quite well. The launch is probably the best launch in history. We are in an opportunity to do it a little bit different than what we did with overall Terrain. Terrain was an opportunity for us to address ray tracing for the very first time, and really start that market in essentially a chicken and the egg type of market. The hardware availability began the beginning of the overall software that was there.

We also had to take a little bit of a pause in terms of some of that launches and really address the whole scope of GPUs over a longer period of time. So we're excited both in terms of the performance improvement that you have with the overall Ampere, as well as the great price points. And so far, the launch of the ramp is growing quite well, and we were just really pleased in terms of how things were going.

## Q - Ambrish Srivastava {BIO 4109276 <GO>}

Okay. Thank you for the detail. Thank you.

## Operator

This will bring us to the end of today's question-and-answer session. I turn the call back over to Jensen for closing remarks.

## A - Jensen Huang {BIO 1782546 <GO>}

This is an amazing time for the computer industry in the world. The age of AI has begun, and NVIDIA is in full throttle to this capability to the world. The breakthroughs of AI can now bring automation to the world's largest industries. This new type of software requires a new type of computer to write the software, validate the software, deploy the software. AI is understandably complex from the chips, systems, software, algorithms to applications. AI requires reinventing every layer of the computing stack. NVIDIA is in full throttle building the full stacks for each computing domain and each computing environment. And from cloud, PC, enterprise, autonomous machine to Edge. NVIDIA is in full throttle building the computing company for the age of AI. Thanks for joining us at GTC.

## Operator

Thank you everyone. This will bring us to the conclusion of today's conference call. You may now disconnect.