Investor Day

Company Participants

- Colette Kress, NVIDIA Corporation
- Jeffrey Fisher, NVIDIA Corporation
- Jen-Hsun Huang, NVIDIA Corporation
- Simona Jankowski, NVIDIA Corporation

Other Participants

- Aaron Rakers, Wells Fargo Securities, LLC, Research Division, MD of IT Hardware & Networking Equipment and Senior Equity Analyst
- Atif Malik, Citigroup Inc., Research Division, Director & Semiconductor Capital Equipment and Specialty Semiconductor Analyst
- Blayne Curtis, Barclays Bank PLC, Research Division, Director & Senior Research Analyst
- Brett Simpson, Arete Research Services LLP, Senior Analyst
- Christopher Muse, Evercore ISI Institutional Equities, Research Division, Senior MD, Head of Global Semiconductor Research & Senior Equity Research Analyst
- Harlan Sur, JPMorgan Chase & Co, Research Division, Senior Analyst
- John Pitzer, Credit Suisse AG, Research Division, MD, Global Technology Strategist and Global Technology Sector Head
- Matthew Ramsay, Cowen and Company, LLC, Research Division, MD & Senior Technology Analyst
- Stacy Rasgon, Sanford C. Bernstein & Co., LLC, Research Division, Senior Analyst
- Timothy Arcuri, UBS Investment Bank, Research Division, MD and Head of Semiconductors & Semiconductor Equipment
- Unidentified Participant, Unknown Company, Analyst
- Vivek Arya, BofA Securities, Research Division, Director
- William Stein, Truist Securities, Inc., Research Division, MD

Presentation

Simona Jankowski {BIO 7131672 <GO>}

Hi, everyone. Welcome to NVIDIA's 2021 Investor Day. I'm Simona Jankowski from Investor Relations.

I'd like to welcome all of you joining us on the webcast and hope to see many of you in person for future events. Let me quickly walk through the safe harbor before getting to today's agenda.

We will be making forward-looking statements regarding our expectations and other future events, which may differ materially from NVIDIA's actual results. Please refer to our SEC filings for a description of our businesses and associated risks and other factors, which could cause our results to differ materially from these statements.

All our statements are made as of today, April 12, 2021, based on information currently available to us. Except as required by law, we assume no obligation to update any of these statements.

Also, if we use any non-GAAP financial measures, you'll find the reconciliations to GAAP on our IR website. Today's agenda includes presentations by Jensen Huang, Jeff Fisher and Colette Kress.

Once we finish the presentations, you will have an opportunity for Q&A with Jensen and Colette. Now please join me in welcoming NVIDIA's Founder and Chief Executive Officer, Jensen Huang.

Jen-Hsun Huang {BIO 1782546 <GO>}

Good morning. I hope you're enjoying GTC 2021. It is the best ever. Over 180,000 registered attendees, 1,600 from touring award winners, Gordon Bell award winners, CAGL Data Science, Grandmasters. (inaudible) in AI researchers, Yoshua Bengio, Jeff Henton, Jan Lekan, JŸrgen Schiever, so many others. 1,600 talks highlighting research on accelerated computing.

AI, 5G, quantum computing, natural language understanding, recommender systems, selling cars, health care, cybersecurity, robotics, 5G, Edge iOT, incredibly diverse fields. There are many vertical industry tracks, so be sure to tell your colleagues covering health care or transportation or retail, so they know to listen to the talks.

We announced some important things. Let me describe how they fit into our strategy and highlight some key points. Omniverse, a platform to create and simulate shared virtual 3D worlds. It connects to other worlds using USD, Universal Seen Description, if you will, the HTML of 3D worlds. And it simulates and physically in the physically and photorealistic way.

With Omniverse, you can connect designers using different tools into 1 world, a shared world to create a scene or a game. With Omniverse, you can also connect robots, AI characters, performing various tasks into 1 world, a shared world like the BMW simulator factory. Omniverse is the foundation platform of our AR and VR strategies.

Our design and remote collaboration strategies, our med averse virtual world strategies and our robotics and autonomous machine AI strategies. You're going to see a lot more of Omnibus. It's a really important platform.

We announced the new DGX station, the new DGX Superpod, the world's first cloud-native supercomputer. It is now the world's first cloud-native supercomputer.

We announced 3 important new. MegaTron to train giant transformer language models. And never the #1 internet company in Korea is an excellent example of why domain and region-specific language model development on Superpods is necessary and a trend. NVIDIA Clara for computational drug discovery announced new algorithms and partnerships with Oxford Nanopore, and recursion. There are so many more.

We announced Co-Quantum, our quantum Circuit simulation platform, running on DGX, simulate Quantum circuits in days that otherwise takes decades. Researchers in industry, national labs and academia around the world are investing hundreds of billions in Quantum Research.

Quantum, Co-quantum will benefit designers of Quantum computers and those designing hybrid GPU and quantum architectures, researchers inventing new quantum algorithms like a cryptographic algorithm for the post Quantum world to scientists who are simulating science with Quantum physics.

We announced Bluefield 3, our next-generation data center infrastructure platform, isolating control from application planes to enhance security, offloading and accelerating virtualization, networking, storage and security.

1/3 of today's processing and software-defined data centers is already in the infrastructure software and will grow substantially with Zero Trust security models. The time for Bluefield has come.

We announced DOCA 1.0, the software stack for data center infrastructure computing. DOCA will be for Bluefield, what CUDA has been for our GPUs. We announced Grace, our First Data center CPU, designed especially for a giant data scale, AI and HPC. Grace is 30x the system and memory bandwidth of our DGX, the fastest computer in the world today, just to put that in perspective.

We'd expect to see an order of magnitude LEAP for giant AI models like Megatron transformers. We will sample next year and ship 2023 in DGX and the giant Swiss supercomputer called ELPS, which is going to be 10x the AI performance of the fastest supercomputer in the world today.

One of the most important announcements is NVIDIA EGX for enterprise with aerial for 5G industrial edge. And with aerial for 5G industrial edge. We believe enterprise industrial Edge will be where AI makes the biggest impact. In health care, warehouse logistics, manufacturing, retail, agriculture, transportation, the world's largest industries.

We announced new customers of drive Orin and a new chip called Atlan and a reference AAV car computer system called Hyperion 8. The announcements feature the most powerful dynamics that are shaping the industries today.

Accelerated computing is the path forward. And not only has it kept computing performance from plateauing, we've set computing on a new supercharged curve.

AI is software that writes software, no human can. It has profound is how software is developed. And open new opportunities to automate tasks never before possible. The data center is the new unit of computing. The software as a service trend has caused software to be refactored into disaggregated microservices that can e-scale out and run across the entire data center, like it's 1 computer.

AI and 5G are the 2 critical technologies that enterprises of the world's largest industries need to deploy new products, services and business models. The era of robotics is here. The first significant example is self-driving cars.

Thousands of companies around the world are building robotic services. One of the most important missing technologies is finally here. Omniverse, a simulated virtual world so accurately that robots can learn to be robots. Grace is our first data center CPU is designed especially for giant scale, AI and HPC.

In addition to chips from industry partners, we now have the Essential 3 chips needed to innovate in a diverse range of computing from cloud data centers to 5G industrial edge. Our road map rhythm will combine CPU, GPU and DPU to deliver a big boost each year.

We will support both X86 and arm, and many CPU designs optimized for diverse segments of computing from Intel, AMD, Amazon, Marvell, Ampere Computing, Mediatek and others.

As you see in my talks and announcements, I'm a big fan of arm, their CPU architecture and their open licensing business model. We believe there are great many new market opportunities for arm. To engage it, we will have to develop new platforms, new ecosystems and new markets. These are things that NVIDIA excels.

We announced several partnerships to expand Arm's ecosystem in opportunities. We announced that AWS and NVIDIA are building cloud computing platforms with Graviton2 and NVIDIA GPU. We announced that M Ampere Computing and NVIDIA are building reference platforms for AI, scientific computing and cloud.

We announced that Marvell and NVIDIA are building reference platforms for 5G and cloud and 5G edge in cloud. And we announced that MediaTek and NVIDIA building reference platforms for PC and mobile devices.

Our combination will accelerate and expand Arm's opportunities. Arm and NVIDIA are great businesses separately. Together, we will create new growth platforms for our partners and build a premier computing company for the age of AI.

We're working with regulators in the U.S., Europe and Asia to explain our vision for Arm and obtain the necessary approvals. Discussions with regulators are expected and constructive. We continue to expect closing in 2022.

AI, the automation of intelligence that can operate and scale out at the speed of light is the most powerful technology force of our time. We're expanding the reach of AI in 4 waves. First is to reinvent computing and software for AI. This led to the invention of NVIDIA DGX, the Tensor core GPU and the NVIDIA AI platform.

The second wave is cloud and CSP adoption. The drive to maximize cloud throughput flexibility utilization led to the invention of the Ampere Universal GPU for train and inference, TensorRT optimizing compiler, Triton inference server and MIG multi instance GPU, Rapids data processing, and integrated cloud graphics and a global team of accelerated computing software experts working with CSPs.

We've created so much to enable AI in the cloud. The the next waves of AI are big, the enterprise industrial edge and robotics. EGX is our AI platform for enterprise and industrial edge. AGX is our robotics computing -- is our robotics computing platform for autonomous machines.

NVIDIA EGX is the AI platform for enterprise and industrial 5G edge. Every layer of the computing stack required massive engineering and a large ecosystem of partners were brought together so EGX can seamlessly integrate into the world's enterprise IT infrastructure.

From the bottom up, 14 global computer makers are offering 55 new high-volume servers designed for the world's enterprise, integrating our new A10, A30 and aerial A100 GPUs optimized for enterprise data center environments. A10 is optimized for AI as well as graphics. A30 is optimized for AI and computing.

Ariel A100 combines a 100, Bluefield 2 and our aerial 5G VRAN stack. All of these servers are NVIDIA certified to run Nvidia AI and NVIDIA Omniverse, and NVIDIA optimized VMware vsphere.

Together with Vmware, we did some great computer science. We engineered Vsphere to incorporate NVIDIA Tetracore GPUs and NVIDIA Bluefield technologies so that NVIDIA AI can achieve bare metal performance in a VMware virtualized environment.

This is important because 80% of the world's enterprise runs Vmware. We're offering NVIDIA AI enterprise, a major software product with enterprise-grade service levels for mission-critical AI operations. And to run on top of the NVIDIA AIOS, if you will,

we have the NVIDIA suite of NVIDIA pretrained models, state-of-the-art AI models, designed for production and performance optimized to deploy.

There are many personas of AIs that we offer, but think of it as site, speech, language, animation, understanding and recommendation AIs. If NVIDIA AI is the OS, then NVIDIA pretrained models are the application suites.

Let me highlight one of the new pretrain AIs, Jarvis. Jarvis is an interactive conversational AI. State-of-the-art deep learning model end-to-end, 100 milliseconds blink of an eye response time. World-class results in speech recognition and translation.

Jarvis supports 5 languages today, English, Spanish, French, German, Japanese and Russian. Pretrain models are a company with some great tools and NVIDIA to customize Jarvis to the language lingo of your domain, health care, technology, insurance, retail, customer service of some kind.

And NVIDIA fleet command to securely deploy and manage your AI to your fleet of EGX computers, end-to-end, top-to-bottom complete. And this is most important. Jarvis and NVIDIA pretrained models runs in any cloud and now on-prem and to the 5G edge on NVIDIA EGX.

If you deploy to the 5G edge, and aerial A100 GPU. And with our ariel 5G VRAN stack turns the EGX server into a 5G base station that runs AI services like the cloud. We brought the power of the cloud to the edge.

I'm delighted to see the IT industry join us to bring EGX AI platform to the world's industries. Dell, ATAS, HP, Vmware, Red Hat, Google Cloud, Splunk, Cloudera, Omni side, Check Point, Fortinet. And in 5G, Ericsson, Maviner, Fujitsu and so many more I didn't list.

I appreciate all of your support. We may have noticed -- you may have noticed that the leading cybersecurity companies are also joining our platform. NVIDIA Morpheus, real-time, all packet inspection platform is really exciting and really essential in this future, 0 trust security model.

Autonomous driving is one of the first mass market robotics applications. It is also one of the most intense machine learning applications that requires decade-only -- decade-long investment. NVIDIA strategy has 3 central pillars: first, build an end-to-end and full stack drive service that we will operate with automakers.

Second, build an open platform for the drive building blocks for the entire transportation industry to build AVs. And third, generalize the learnings from drive to create other machine learning applications like robotics and edge AI.

On the left is the pipeline. It's an incredible thing that drive -- autonomous driving requires. Hyperion 8 AV system in BBA test car collects data that are data factory processes into label training data. Label data and Omniverse synthetically generated data trains our models. Then drive AV application runs an Omniverse drive simulation on real car computer, but in our data center, these computers are called constellation.

And the software is OTA into the car with fleet command and tested in BBA. Effectively, a massive AI computer and a fleet of AV computers that are running continuously in a loop improving continuously.

This is the machine learning loop. All of this work will go into the 2024 Mercedes-Benz EQS fleet. On the right is our AV computer road map, the growth in computing is staggering and needed. We announced Atlan, our car computer for the 2025 generation.

It's 1,000 tops in a chip and fuses our most advanced GPU and AI technologies, the Bluefield 4 data center technologies. The most advanced arm server class CPUs, years of functional safety and security expertise and tens of thousands of engineering years of software.

All software developed on Xavier runs on Orin and then on Atlan. One architecture. Our car computing platform that offers carmakers 1 consistent and programmable installed base that they can leverage software investments across and a growing installed base that they can deliver value to for the life of the car.

We announced Hyperion 8 a fully production able, functional, safe and secure car computing platform, including sensor, network, computer and software for data collection, testing and production.

Hyperion 8 is compatible with NVIDIA's entire AV stack. Hyperion 8 is a standard reference AV car system that can fit into most vehicles. Hyperion 8 is like the PC ATX motherboard, a standard reference that invigorated and accelerated the PC industry.

We announced more Orin customers. The transportation industry is quickly becoming a technology industry. Makers now realize something very important. They now realize that the car is more than the vehicle. But an installed base, an installed base platform that they own. And if the computer is powerful, it will host value software services they invent for years to come.

This is a giant new observation and is transforming the industry. A few million cars installed based on the road -- a few million cars installed base on the road for 15 to 20 years, even a few hundred dollars of services each year represent an immense value creation opportunity.

Putting an Orin a few into the car is the best way to build a valuable installed base. NVIDIA is an accelerated computing company and innovates across the entire stack. NVIDIA is built as an open platform with 3 layers: chips and systems, software platforms and application frameworks.

With the global NVIDIA partner network connected to the layers of our platform, we serve customers and markets from the layer best for them. Each higher layer offers an order of magnitude more opportunity for our company. Our chips and systems market opportunity is profoundly greater with NVIDIA AI and NVIDIA Omnivere.

Our markets -- our company's market opportunity is profoundly greater by offering drive as a service to a 10 trillion miles a year industry and Jarvis to automate hundreds of billions of hours of spoken language each year. All of this is built on 1 architecture, each layer, leveraging and enhancing the layer below.

We had a great GTC and lots of new products, new partners and new markets. These are the points that I highlight. Bluefield and Grace make NVIDIA chip company to do data center scale computing. NVIDIA EGX is our new AI platform for enterprise and industrial 5G edge.

And this NVIDIA AI enterprise is our new software product. Others are NVIDIA GPU. NVIDIA Omniverse, base command and fleet command. And I'm looking forward to NVIDIA pretrained models like Drive and Jarvis going into production services.

NVIDIA is a computing platform company, innovating across 3 chips and a full stack 3 layers. Jeff Fisher will be up next to talk to you about gaming. So with that, I'll hand it over to Jeff. Fish?

Jeffrey Fisher {BIO 2373419 <GO>}

Thanks, Jensen. Welcome, everyone, to Analyst Day, 2021, excited to give you guys an update on our gaming business.

Every person born today will be a gamer. That's 140 million more gamers this year alone. As gaming has become a past time for millions more people, it is transformed into one of the largest and fastest-growing forms of entertainment. Gaming is no longer about playing a game.

Gaming is an immersive social network connecting like-minded people building lifetime friendships. Discord which began with a mission that connect gamers has grown into a social network and more than doubled its active users to 140 million in 2018 -- since 2018.

Gaming is a sport. Esports have fueled a new generation of e-athletes who compete for bragging rights and has span professional careers. 75% of Gforce gamers play esports. Esports influence is growing, and the world is tuning in.

In the past two years, the esports audience has grown by 75 million, totaling 436 million viewers. Gaming has offered -- gaming also offers interactive storytelling. While watch a movie, when you could play a movie that's a cinematic and rich as anything produced in Hollywood.

And gamers look to share their passion for gaming, live streaming their gameplay and connecting content for others to watch. In 2020, a 100,000,000 hours of gaming content was watched on YouTube. That's twice of 2018.

The growth of gaming is also evident on steam. One of the top destinations for gamers, where 120 million people played every month in 2020. With 25 million peak concurrent gamers that's up 1.5x from 2018. In that same time, the Epic Game store has grown from its launch in December of 2018 to 160 million PC gamers.

Gamers are demanding much more from their hardware as well. So our GeForce gaming platform continues to deliver more value. For example, this past year, we launched DLSS 2.0. And DLSS uses AI to deliver a significant performance increase in games. Running on RTX tensor cores, DLSS can increase frame rates by up to 100%. Press and gamers have widely recognized DLSS as a must-have feature for gaming.

We also delivered a powerful weapon to Esports Gamers NVIDIA Reflex. System latency is a killer literally for our competitive gamers. Shots need to be spot on, not a few frames behind the opponent.

In the case of Blizzard's Overwatch, which recently integrated NVIDIA Reflex, reflects reduces system latency by 50%. And now 8 of the top 10 competitive shooters have integrated NVIDIA Reflex. And last, we brought AI to streamers and video content creators with NVIDIA Broadcast. NVIDIA Broadcast can turn any room into a broadcast studio, like my office here.

2020 was an exciting year for our gaming business. two years ago, we introduced a breakthrough in graphics, real-time ray tracing, and AI-based DLSS. We called it RTX. In 2020, we doubled down. We introduced the Ampere architecture. Ampere featured a new Shader design a second-generation RT core for ray tracing and a third-generation tensor core for AI.

It was our biggest generational leap ever. And gamers who are waiting to upgrade to RTX jumped in. Combined with strong gaming market fundamentals, the increasing production value of AAA games, global Esports growth and an increasing number of creators and streamers, we delivered a record year. Our 5-year gaming GPU CAGR is 21%, with growth in units shipped and ASP.

Looking forward, RTX represents a huge reset of the installed base. RTX was featured in the biggest games of last year, including Cyberpunk and WatchDogs Legion and in massive hit titles like Fortnite, World of Warcraft, and the best-selling game of all time, Minecraft.

And if you haven't seen some of the beautiful RTX worlds that Minecraft gamers are creating, do a quick Google search of RTX Minecraft.

Trust me, you will be amazed. With every major operating system, game engine and console supporting ray tracing and a strong pipeline of games on the horizon, RTX is the new standard.

Looking just at our installed base of 140 million GeForce gaming GPUs, 85% were designed to run traditional games. Turn on rate tracing and like this example of the hit game control, the games become unplayable on these GPUs. RTX with RT Cores and AI-based DLSS delivers beautiful ray trace games at terrific performance and.

And the upgrade opportunity extends well beyond this. Considering the hundreds of millions of gamers playing on a wide range of hardware, just look at steam for a peak at this hardware profile. We believe RTX is at the front end of a major upgrade cycle and we are off to a great start.

The excitement was really high prior to the Ampere launch. If you recall, rumors were all over the Internet. Google Search for NVIDIA RTX was up 6x compared to that around the turning launch. Since the RTX 3000 was pulled from last September, sales have been off the charts.

With new buyers and those upgrading sales have outpaced prior generations by 2x. These are end market sales. While demand continues to outstrip our supply, gamers are getting their hands on Ampere. Amper share on steam is twice that of turnings at the same time after launch.

More important, as the GeForce gaming platform continues to add more value to gamers, gamers are buying up each generation. This chart on the right shows the end market ASP of our desktop stack.

This was calculated for the six months after each architecture launch. In effect, today's desktop card ASP is $360. That's based on MSRP. That's 20% higher than the same time after turing launch.

And I believe there's plenty of ASP headroom, especially compared to what gamers are paying for consoles, which start at $500. Now let's talk about the fastest-growing game platform, gaming laptops.

This past January, at CES, we launched the RTX3000 series for laptops. With 70 new models from every OEM starting at just $999. This was our best laptop launch ever. Laptops represent a major growth opportunity. As new buyers are choosing laptops, choosing gaming laptops to learn, play and create.

Fueling this growth is Max-Q. Max-Q is a system design approach that delivers high-performance in thin and light gaming laptops. It has fundamentally changed how

laptops are built.

Our third-generation Max-Q was introduced with the Ampere architecture. It includes dynamic boost 2.0. For the first time, dynamic Boost uses AI to dynamically shift power between the GPU, CPU and system memory, depending on where it is needed most.

And most important, it shifts power away from where it isn't needed. Max-Q also features resizable bar, which enables more efficient memory access to boost performance while using no more power and an updated whisper mode, which maintains high system performance while minimizing acoustics.

Over twice the number of Max-Q models shipped this year -- or shipping this year. 2021 will offer the thinnest, highest performance gaming laptops ever. And I've got mine right here. This has got an RTX3070, high-performance gaming laptop, thin and light. This follows a strong -- this follows years of strong growth at over 20% CAGR.

Gaming laptops are outpacing the consumer laptop market and out selling the most popular game console. New students, gamers and creators are clearly choosing gaming laptops for their next PC. And our gaming business is more than just playing. The creation of digital content is exploding.

For the 45 million creators and growing, we launched NVIDIA Studio. NVIDIA Studio is our accelerated platform for creators, speeding up ray tracing and AI in over 60 creative and design applications, including Adobe Photoshop, DaVinci Resolve and Blender.

Studio includes monthly dedicated studio drivers that we release. They offer new features, faster performance and enhanced stability for creative applications. Since the launch of Studio in 2019, there have been over 100 purpose-built studio laptops and desktops from every major OEM.

Our Ampere architecture takes NVIDIA Studio to the next level. Rendering is up to 5x faster than pascal. Video editors can now use AI to simplify workflows.

Video incode times are reduced by up to 75%. RTX will change the way creators work. We estimate there are over 30 million streamers globally. On Twitch alone, the number of streamers more than doubled over the past year.

In China, streamers are becoming a new e-commerce channel, selling gaming hardware direct to their followers, including GeForce RTX GPUs. Sharing your gameplay while streaming requires high-performance hardware. Many enthusiasts use 2 PCs, 1 to play and the other to broadcast.

NVIDIA broadcast paired with GeForce RTX solves that. Broadcast uses AI to eliminate background noise and to green screen your background. And the powerful video encoder in RTX is capable of streaming video without impacting your gameplay.

NVIDIA Broadcast works seamlessly with all the popular streaming and video conference apps, it turns any room into a broadcast studio, and all you need is an RTX GPU.

VR has long been viewed as the next thing for gaming. With the availability of lighter, more capable headsets like Quest 2 Valve's index and HP's River G2, along with compelling new games, like Microsoft Flight SIM and VR and Valves half-life, Alex, which drove a 71% increase in VR game sales on steam. VR is coming into its own.

It's reported that Facebook has over 10,000 people working on augmented and virtual reality. And as you heard today, NVIDIA is building the omniverse. Our portal in a VR portal into a metaverse.

VR headsets have twice the resolution of a desktop gaming monitor and demand very high, smooth frame rates. It's very unforgiving from a performance perspective. This requires a very high-performance GPU and provides more motivation for gamers to upgrade.

There are expected to be over 30 million PC capable VR headsets sold in the next five years. And from what you've heard today, that could be conservative. This past year, we officially launched GeForce Now, 10 years in the making. GeForce Now leverages our GeForce PC platform into a cloud gaming service.

Gamers on underpowered PCs, Chromebooks and mobile devices can effectively subscribe to a virtual GeForce PC in the cloud. And the PC ecosystem is coming along with us.

Gaming stores and publishers like Valve, Epic and Ubisoft, see the opportunity that we see to reach billions more gamers, and we are not building it out alone. Our strategy is to team up with ISPs and telcos or GFN Alliance partners around the world to offer GeForce Now to their subscribers.

Alliance Partners manage the infrastructure. We operate the service, and we share the revenue. This will become increasingly more powerful as 5G blankets the world. With a high resolution, low latency requirements of gaming, cloud gaming is the killer app for 5G.

Today, GFN has passed 10 million registered users. And offers 1,000 instantly playable games with much more coming online every Thursday in our GFN Thursdays, 2-minute and watch for that.

GFN is offered in 27 countries. And today, I'm excited to sit to announce that we are adding South America to the list for 2021, including our most requested country and 1 I'm most excited about Brazil. Brazil is traditionally a low end hardware gaming market, and they are going to love GFN.

Brazil has 95 million gamers alone. Over time, I see GFN extending beyond gaming to all kinds of interactive experiences. This past Sundance Film Festival, GFN hosted Disney's interactive short AMAG streams as part of their new frontiers showcase.

To wrap up gaming, we see a long runway for growth. RTX resets everything and will drive a major upgrade cycle. The entire PC ecosystem needs to upgrade, including 85% of our gaming installed base. As the value of our platform grows, gamers are buying up.

Laptop continues to offer a strong growth opportunity. Gaming laptops are outpacing the consumer laptop market. Max-Q makes laptops thinner and lighter and new buyers are choosing gaming laptops.

RTX and RTX laptops will power the 100 million and growing number of e athletes, creator streamers and virtual reality adopters. Last, GeForce Now continues to scale up. The PC ecosystem is coming along with us. GeForce Now gives us the opportunity to extend our PC platform to billions more gamers.

I want to thank you all for joining me and joining us on our Analyst Day. And with that, let me hand it over to Colette Kress.

Colette Kress {BIO 18297352 <GO>}

Good Morning, everybody. Fiscal year 2021 was a record-breaking year for NVIDIA. We achieved record revenue and EPS, launched our Ampere Architecture for both gaming and data center into incredible demand.

Completed the acquisition of Mellanox and announced our transformative acquisition of Arm. Let's first look at some high-level highlights of our P&L. Our fiscal year 2021 revenue increased 53% year-on-year to $16.7 billion, fueled primarily by the tremendous ramp of the Ampere Architecture across our data center and gaming platforms.

We grew our non-GAAP gross margins by 310 basis points as data center increased as a percentage of revenue. We also demonstrated strong operating leverage. Our non-GAAP operating income increased 82% year-on-year, and our non-GAAP EPS increased 73% year-on-year to $10 a share.

Let's turn to the performance of our market platforms. Our gaming business grew 41% year-on-year to record $7.8 billion, in fiscal year '21, with broad based strength, driven by growth in our desktop, notebook and console businesses.

We reinvented the graphics with the launch of our RTX 30 Series GPUs. And and the demand has been off the charts, resulting in the fastest launch in the company's history, twice as fast as Turing.

Gaming revenue has seen a 4-year compounded annual growth rate of 18%, driven by a combination of GPU unit and blended ASP growth. Unit growth has been driven by the expanding universe of gamers, our phenomenal success in growing gaming laptops and strong demand in our console business.

Blended ASP growth has been driven by the gamers buying up the as they adopt new features and capabilities like NVIDIA ray tracing and deal less us. As the production value of games continues to increase.

Turning to pro visualization. Our revenue declined 13% in fiscal year '21 to $1.1 billion as we believe enterprises deferred purchases due to the pandemic. Despite these headwinds, we grew revenue at a 4-year compounded annual growth rate of 6% as we benefited from the continued growth of a number of the GPU accelerated applications.

And as RTX technology gains broad industry adoption. Automotive revenue declined 23% year-on-year in fiscal 21 to $536 million due to lower global production volumes and the expected decline in infotainment revenue. We grew revenue to a 4-year compounded annual growth rate of 2% as strong growth in our autonomous and AI cockpit solutions were largely offset by the declines in infotainment.

Autonomous solution and AI cockpit are approaching 2/3 of our automotive revenue. And we expect mix to continue to shift to these businesses over time as substantial wins in these areas ramp in the coming years.

Data center had a record year with revenue increasing 124% year-on-year to $6.7 billion, including almost 70% growth for data center compute. The A100 based on our Ampere architecture delivers up to 20x performance increase versus the prior generation, our largest generation we've ever.

It delivers high utility driven by its unified architecture, allowing you to process number numerous workloads, including training, inference, data analytics and graphics. MIG technology allows the A100 to efficiently scale up for demanding training workloads and scale out for high-volume inference use cases.

These advances of the A100 combined with the forces of AI and cloud computing drove strong demand in fiscal '21 across hyperscales and enterprise customers.

From a workload perspective, we saw strength across training and inference as exponential increase in AI model complexity and compute requirements drove the demand for NVIDIA accelerated compute and networking products.

Finally, Mellanox had an outstanding year with growth stemming from hyperscale supercomputing and AI customers. From a product perspective, Mellanox saw strength across Ethernet and infiniband offerings. In total, we have seen tremendous growth in data set with a 4-year compounded annual growth rate of 69%.

AI is the most powerful technology force of our time, and we see a long runway of growth ahead of us. As you know, NVIDIA is a full stack computing platform company, spending across silicon, systems and software.

So far, our software has largely been offered as part of the platform and not directly monetized on a stand-alone basis. Jensen earlier discussed our 3-layer model of customer engagement, which conveys our go-to-market strategy as well as our growing revenue opportunity as we move up the stack to offer software commercially.

This helps unlock large new market opportunities and will add reoccurring revenue to our P&L over time. Let me now highlight 1 such opportunity. We recently announced NVIDIA AI enterprise, a comprehensive suite of enterprise-grade AI software that speeds development and deployment of AI workloads and simplifies management of enterprise's AI infrastructure.

Through our partnership with Vmware, hundreds of thousands of VSphere customers will be able to purchase NVIDIA AI enterprise with the same familiar pricing model that IT managers use to procure, VMware infrastructure software. NVIDIA AI enterprise software is offered in a perpetual license per CPU socket with annual maintenance.

We also offer the software suite as a subscription. We believe the NVIDIA AI enterprise software suite represents a multibillion-dollar opportunity. NVIDIA AI enterprise, combined with our EGX enterprise platform is democratizing AI and helping to bring NVIDIA AI and accelerated computing to the world's largest industries.

Our EGX platform is gaining rapid adoption as enterprise customers such as Lockheed Martin and Mass General bring in. Additionally, we are supporting these systems with powerful processor road map, as shown with the launch of our A30 and A10 GPUs, our aerial A100 platform and our Bluefield DPU road map.

This powerful enterprise and edge computing platform positions NVIDIA for the next wave of AI adoption and will be driven by the vertical industries. Let me shift gears to our automotive opportunity. The NVIDIA drive platform is seeing wide adoption across the transportation industry, which will create a significant software component.

Today, vocal cars announced that it will build next-generation vehicles on NVIDIA drive alone. This further extends our partnership with Volvo to now include more software-defined vehicles in its lineup beginning in 2023.

Volvo joins other established OEMs, such as Mercedes, Audi and Hyundai, who were all developing on investment (technical difficulty) driving including many trucking. The auto we are harnessing the new season GPUs, but also their incredible energy efficiency.

Last year, we announced a landmark partnership with Mercedes-Benz, which will adopt NVIDIA's full stack drive platform to enable their entire fleet of vehicles to software-defined and perpetually upgradable.

This deal was transformational in that in addition to the hardware, it includes a revenue share component for the software sales that Mercedes to make on their fleet of connected vehicles, such as AutoPart.

Vehicle owners will be able to purchase over-the-air software and service offerings to enhance the capabilities of their vehicles and increase the joy of driving. With software content potentially in thousands of dollars per vehicle, this could be a multibillion revenue opportunity for both Mercedes and NVIDIA.

Not only is this deal transformational to NVIDIA's business model, but for the auto industry is large. There are 100 million cars sold each year globally. And over time, we believe all vehicles will be autonomous, software-defined and upgradable.

We see the potential of similar deals with other automakers. This large and growing list of wins across the transportation industry is set to ramp in the coming years and will drive a major inflection in our revenue.

We have over 8 billion in automotive design win pipeline through fiscal year 2027, with a good amount of this revenue expected to ramp in the latter part of this time frame.

I talked earlier about the revenue growth that our data center and gaming market platforms experienced over the last few years. Over the past four years, our data center business has grown nearly 3x the rate of the company's 25% compounded annual growth rate. This mix shift has led to a favorable expansion of our gross margins, as our fastest-growing business is also our highest margin business.

Given the secular forces of AI and cloud computing, combined with this rapid adoption of our computing and networking platforms, we believe this trend can persist.

Additionally, the gross margin profile of our gaming GPU business has increased over time as gamers have bought up our stack, resulting in rising blended ASPs. We expect this trend to continue over the coming years as gamers upgrade to Ampere RTX continues in its rapid pace of adoption and the overall production value of games continues to increase.

Overall, our gross margin has increased from 59.2% in fiscal year 2017 to 65.6% in fiscal year '21. We continue to see uplift to our gross margin profile as our mix shifts to higher-value platforms. Furthermore, software is a significant opportunity for NVIDIA.

This revenue as it scales, will provide an additional tailwind to gross margins. As discussed, our non-GAAP gross margins have increased 640 basis points over the past four years, and our non-GAAP operating margins have also increased at even a faster rate, growing from a 32.1% in fiscal year 2017 to 40.8% in fiscal year '21.

We delivered significant operating leverage, even as we invested heavily across the market platforms to support the growth we have demonstrated over the years.

And as we look to take advantage of the material opportunities that lie ahead of us. We have a single architecture across all of our platforms that forms the basis for our product offerings and innovations.

Our 1 architecture approach provides us leverage and core IP to innovate across the entire technology stack from silicon, systems and software, to compute, networking and storage technologies. We are able to rapidly develop new products and technologies, move quickly into new markets and nurture a large and growing ecosystem of partners and developers innovating on our platform.

This, in turn, fosters increased adoption of our platform and helps to drive revenue growth and our ability to invest. This flywheel of innovation is accelerating, as seen by our incredible breadth and depth of products and technology as found into many new announcements discussed today.

NVIDIA's unique business model allows us to innovate like no other company while driving attractive economic returns. Going forward, we believe we can continue to drive strong revenue growth, while delivering earnings growth and healthy cash flow generation.

On revenue and earnings growth over the past years, it has also resulted in material increase to our cash flow generation with cash flow from operations growing from $1.7 billion in fiscal year 2017 to $5.8 billion in fiscal year 2021, a 37% compounded annual growth rate.

This increase shows our ability to invest for growth. Our capital expenditures increased from approximately $200 million in fiscal year 2017 to $1.1 billion in fiscal year 2021.

It has also allowed us to engage in transformational M&A as seen with our purchase of Mellanox and the announced acquisition of Arm. We maintain a disciplined capital return possibly with $5 billion returned to shareholders in the form of dividends and share repurchases since fiscal 2017.

We remain committed to paying a dividend. We employ a conservative financial quality and have a healthy balance sheet with $11.6 billion in cash and marketable securities and $7 billion in debt.

Overall, our business is highly cash generative, and we expect continued growth in our cash flow in the coming years. This will help fund our investments as we seek to take advantage of the growth opportunities across our business.

We start our fiscal year with great momentum across our business, while fiscal Q1 is not yet complete, our Q1 total revenue is tracking above for $5.3 billion outlook provided during the fiscal Q4 earnings call.

We are experiencing broad-based strength with all of our market platforms, driving upside to our initial outlook. Additionally, we now expect CMP revenue to be approximately $150 million higher than the $50 million included in our fiscal Q1 outlook.

Upside to CMP is not displacing supply from our other platforms. It is incremental. Within data center, we have good visibility, and we expect another strong year. Industries have been increasingly using AI to improve their product and services. We expect this will lead to consumption of our platforms through CSPs, resulting in more purchases as we go through the year.

Our EGX platform has strong momentum, and we expect this will drive increased revenue from enterprise and edge computing deployments in the second half of the year. Overall, demand remains very strong and continues to exceed supply, while our channel inventories remain quite leaning.

We expect demand to continue to exceed supply for much of this year. Our operations team is agile and executing fantastic rate, and we expect our supply will increase as the year progresses. We believe we have sufficient supply to support sequential growth beyond Q1.

Finally, I want to discuss our commitment to ESG. NVIDIA is committed to build one of the world's great companies through people, innovation and energy-efficient technology. This means not only doing what's good for business, but what's good for our employees, our business partners, society at large and the environment.

We have been recognized as one of the best places to work by a number of publications, including Forbes, Fortune and Glassdoor. No time better highlighted our innovations in this past year when our technologies played an important role fighting the pandemic across a wide range of use cases.

Our Clara platform helped with discovery efforts as community races to develop a vaccine and other treatments. Our V100 and T4 GPUs help scientists create the first atomic scale map of the corona '19 virus. And our AGX platform helped keep health

care and frontline workers safe with automated technologies that accomplished house without human invention.

Finally, our RTX and virtual GPU technologies help the world work, learn and play from home. We are also committed to investing in a safer environment. Our GPUs inherent provide more energy-efficient form of computing, up to 42x more efficient for processing AI workloads.

The green 500 list ranks the world's most energy-efficient supercomputers in the world. NVIDIA's own Supercomputer ranks #1, the most energy-efficient supercomputer in the world. And NVIDIA powers 26 of the most energy-efficient systems in the world.

Finally, we are committed to sourcing 65% of our total energy consumption from renewable sources by fiscal year 2025. That wraps up our presentations for today. I want to thank everyone for tuning in to our GTC and presentations, and we would like to now move to Q&A for this portion of our event. So moving it back to you, Simona.

Questions And Answers

A - Simona Jankowski {BIO 7131672 <GO>}

Thank you, very much, Colette. We will now begin the Q&A portion of our event with analysts who have joined us on Zoom. (+++presentation Operator Instructions) Our first question will come from John Pitzer from Credit Suisse.

Q - John Pitzer {BIO 1541792 <GO>}

Yes. Congratulations on the solid DTC at Analyst Day. Jensen, just talk a little bit about the Grace announcement and congratulations. I'm kind of curious how added into the $100 billion data center TAM that you guys discussed a year ago, do you see Grace?

And as part of your presentation, you reiterated your commitment to x86 and Arm and actually mentioned multiple CPU players like Intel and, Marvell, Rabatin. How do you see the co-opetition between your rechip strategy and Grace specifically and the ecosystem that you still want to support?

A - Jen-Hsun Huang {BIO 1782546 <GO>}

Yes, I'll work backwards. The -- first of all, thanks for the question. We -- the world's computing is really diverse. I mean there are so many segments of computing and each one of them are architected slightly differently for good reasons.

Some of them are optimized for a single-threaded performance. Some of it is optimized for many quarters. Some of it is optimized for strong IO performance, some of it is really optimized for large amounts of data.

We designed Grace for -- especially for a particular application that is a set of range of applications, if you will, that is really dedicated towards giant data scale computing. You're working on terabytes and terabytes of data for a very long period of time.

And recommender systems are part of this. Natural language understanding models are part of this in the future. You're going to see multi mobile transformer models that are learning from from video and speech at the same time or images and speech at the same time.

And these are going to be processing just a gigantic amount of data, health care, processing a giant amount of data. And so we're really designing Grace for this particular segment. For -- with respect to co-opetition, we worked very closely with AMD. As you know, they're the CPU and our DGX. It's a fantastic CD.

We work with Intel and enterprise data centers, in notebooks. We build amazing notebooks together. We work with Ampere computing in cloud and cloud gaming and supercomputing, computing. Marvell at the edge MediaTek for PCs and mobile devices.

And so the world of computing is gigantic and the nature of our company is an open platform. We're a platform company. And we use the word platform probably more than any chip company in the world. And we do so purposefully and exactly that this is really about building a platform by which the entire ecosystem can benefit.

There is no sustainable growth that is not inclusive growth. There is no sustainable growth that doesn't include partners and collaborators and developers and and ecosystems and such. And so we're delighted to partner with all of these great companies to build the future. So I appreciate that question. Thank you.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question will come from C.J. Muse from Evicore.

Q - Christopher Muse {BIO 18608702 <GO>}

I guess, Jensen, to follow-up on John's question. I was hoping you could speak a little bit more around our CPU strategy. It looks like Grace into 2023 appears to be high performance, high bandwidth using Nvlink. But said, as you look to 2025 Grace next as well as potentially a couplet strategy. What is your ultimate aim here looking over the next 5, 10 years?

A - Jen-Hsun Huang {BIO 1782546 <GO>}

Yes. Thanks, C.J. Well, first of all, Grace next, I'm going to have to save that that surprise you. I hate to ruin your surprise, and there's so much stuff that we're working on.

Our strategy as a company, our -- if you will, a core philosophy as a company, are really to do things that are uniquely for NVIDIA to do and things that are really, really hard to do. And I prefer things that take a long time to do. And things that, frankly, nobody else in the world is doing.

In every aspect of the conversation that we had in the keynote today, you could see that in everything, whether it's Omniverse or Grace or Bluefield 3 or DOCA in everything, Morpheus, our new cybersecurity platform.

All of these things drive everything, Hyperion 8. The world doesn't have it. And that really needs to be the driving purpose of companies to go solve problems that are incredibly hard that are uniquely specialized for our capabilities and that the world doesn't have.

And so it is the nature of our company to go build CPUs that the world doesn't have. Build CPUS and build products that can somehow expand the envelope, expand the overall size of the marketplace for everybody.

And so, so we're going to continue to work with a whole bunch of CPU partners and many of that I've mentioned many, many more. That are building all kinds of different specialized CPUs, and we'll build ours. And of course, support everybody else with the NVIDIA AI platform, the NVIDIA platform, so that we can bring forward this new method of computing, we pioneered call accelerated computing.

That is, if you will, the highest level bit and everything else is really about expanding more, expanding reach in a way that other people can't.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question will come from Vivek Arya from Bank of America.

Q - Vivek Arya {BIO 6781604 <GO>}

I appreciate the very informative Analyst Day. Jensen, I'm curious to get your strategy on the role that system and software and subscription sales will play at NVIDIA over the next several years because we think of NVIDIA as more a semiconductor company, and obviously, that's the key part of the business.

But then we also see you launch a number of system products, a number of them on subscription-type services, whether it's in the data center or enterprise or gaming. You're talking about subscription sales on the enterprise side and then the automotive side.

I'm curious, what is the strategy here? How big are those businesses for you today? how big can they be over time? And what impact will they have on the financials of the company?

A - Jen-Hsun Huang {BIO 1782546 <GO>}

Yes, Vivek, thanks for the question. The driving -- the driving purpose for the full stack approach is to pioneer accelerated computing. When you create a new form of computing that is brand new, and it doesn't come along very often. And if you take a look at the modern way of doing computing, it's running on CPUs. And then, of course, cloud computing came along.

Largely running on many CPUs, turning an entire data center into a computer. And the approach that we pioneered is accelerated computing. Accelerated computing is very different than accelerators. Our GPUs have video encoders inside, video decoders inside.

Those are accelerators, image processors and accelerated computing as a general-purpose computing platform that is somehow particularly good at a domain of work and accelerated computing platform and a company who is a computing platform company is sensible about architecture compliance, it's sensible about backwards and forward compatibility. It's thoughtful about creating an installed base and developing developers and ecosystems and networks and partners.

And so our primary goal is to pioneer accelerated computing. Well, before you could do that, you really have to build a whole stack because the entire way of computing is refactored. We refactored computing from the application to the algorithms, to the solvers, to the system software, all the way down to the silicon, as you guys know well.

And that's why we say NVIDIA is a full stack company. We're a computing company that's a full stack company. Because in the final analysis, it's essential to pioneer a new way of doing computing.

Now the way -- the thing that we select our approach, our strategy is to develop it in 3 layers, which is the 3 layers of computing. It's the hardware layer, the system software layer, the middleware layer, if you will, operating systems are in that layer. Vmware, for example, in that layer and maybe AI is in that layer, okay?

So those are algorithms and solvers and middleware that connects, if you will, the application on top to the hardware on the bottom. And that transformation is very specific to accelerated computing. Now transformation is very specific to NVIDIA CUDA in our GPU architecture.

And so the third and the third layer, of course, is applications. In the world of AI, in the world of AI, the invention we created underneath was DGX and tensor core GPUs and such. In the middle is this layer called NVIDIA AI. And the upper layer would not be an application, but it would be a skill, it'll be a task. A skill that could perform a task.

That task could be driving cars. That task could be recognizing speech and answering a question, a query. That task could be responding to a recommendation.

What movie do you recommend, for example, by click a movie? What's the next 1 you recommend? Soles recommend groceries and you're recommending your cart.

And so these are all skills that are sitting on top. And so if I think about it, and if I answer your question is in the context of AI, then we have the chips on the bottom. We have NVIDIA in the middle, and we have the skills on top. But each one of these layers, each one of the higher layers increases NVIDIA's opportunity by an order of magnitude. Let me give you an example.

There are 100 million cars sold a year. And that's the entire opportunity annually for chips, 100 million cars. However, those cars are driven 10 trillion miles. And so whatever numbers you use, for 100 million cars, if we said $1,000. So just pick a random number for illustration, $100 million $1,000, $100 billion opportunity.

Yet at the driving level, at the task of driving low, 10 trillion miles a year if $1 a mile is $10 trillion. And so that kind of gives you a sense of the economics involved. In the middle, the middleware, you sell a chip one, that's the economics of selling chips. However, in the world of middleware, you have to continuously refine it, enhance it, support to customers as they deliver their their mission-critical applications and services.

And so there's an ongoing support agreement that is in place to respond to customer needs and but fixes and feature enhancements and maintaining long-life, and all of those things are associated with enterprise software licensing. And they tend to live with the GPU for the entire term of the use.

And so we now have an economic model that is about selling, of course, building the most advanced and selling the best chips.

And the customers could very easily and were joy ad they do, use our SDKs and our programming model to develop their entire stack on top or they could use NVIDIA AI stack and NVIDIA Omniverse and NVIDIA a whole bunch of other things, like base command and fleet command and Vgpus, they could develop their own or they can license out with a maintenance fee that -- for the length of their usage of our GPUs.

And then on top of that, it's really skills and tasks, skills. AI skills that perform tasks. And they tend to be per end user or per task or per instance. And I could imagine a day the where and AI is paid by the hour, just like a particular skill, someone who performs a particular task, is paid by the hour, the AI, of course, will be paid by the hour.

And so that's kind of the economic funnel of our company. And my expectation is that over time, the layer that's on top will be the largest of all. But it's built on top of the next 2 layers. It's enabled by the next 2 layers.

And then we'll offer each 1 of the layers open to the industry so that all of our partners could benefit from our knowledge and our skills and developing the capability. And be able to build their own, if they like.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question comes from Stacy Rasgon from Bernstein.

Q - Stacy Rasgon {BIO 16423886 <GO>}

I have a question for Colette. The uplift in outlook for the current quarter, given you suggest that you're still supply constrained, is it proper to take that almost purely as a function of better-than-expected increased supply from the last earnings call until now?

And I guess if that's true, can we take that pace of supply increase is sort of a likely indicator of the likely a continued pace of supply increase as we go through the year.

And then finally, I guess, given that, do you expect supply to improve similar in all the businesses, and hence, you expect all of your business segments to be growing sequentially for the rest of the year.

A - Colette Kress {BIO 18297352 <GO>}

Sure. Let me see if I can answer your question. So yes, we are planning on being higher than where we set out at the very beginning of the quarter with our overall guidance. As we discussed even at earnings last quarter, we will see supply continue to increase throughout this quarter as well as throughout the year.

So it's really a statement of the supply at the right time for the demand that we're seeing today. And we'll continue to work that all year long. Our operations team is very, very focused not only on just what we need to protect, but completing all of our platforms as a whole.

So what we're able to do is seek new supply is going to be coming in a quarter. And that is aiding us as we look in terms of the growth for the rest of the year. As we talked about on our prepared remarks, we talked about really seeing growth even past over Q1 as we do expect supply to be true for the full year.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question is from Matt Ramsay from Cowen.

Q - Matthew Ramsay {BIO 17978411 <GO>}

Yes. Jensen, I wanted to ask a question on your auto business. It kind of dovetails in with the 3-layer strategy that you described a second ago. I think you guys talked about an $8 billion pipeline through fiscal '27.

And I wanted to explore that a little bit. I would think you should be getting visibility at now or sometime soon to hardware sales into some of these OEM platforms. So maybe you could walk me through and break down that $8 billion pipeline a little bit between confirmed hardware sales, maybe software and services sales. And then what you guys are doing with constellation in some of the simulation services? That would be helpful.

A - Jen-Hsun Huang {BIO 1782546 <GO>}

Yes. We cut it off in 2017 for fiscal 2017, I guess, for arbitrary reason. The design wins of cars tend to last a very long time. And the architecture selection will tend to last probably even longer.

The car industry, things and architectures that are thinks about architecture in a decade way. But what is new to the automotive industry is that these computers -- and this is very new.

This is something that is unique to the car industry that is unlike any computing platform in the world. The computers that you launch these computers on wheels, these data center on wheels are going to last into -- in the marketplace for a couple of decades, 15 to 20 years. And electric cars is going to last for a long time.

And so and with the industry starting to think like a technology industry and recognizing that these cars are not just vehicles that they sold, but they're now part of their installed base, that they're part of their fleet, that the installed base, if programmable, could offer them a couple of decades of service offering opportunities.

They're starting to think like cable providers, infrastructure providers, cable set-top box providers. They're starting to think like that phone providers. That these -- and these -- the installed base belongs to them.

They made the effort to create the cars and the installed base is proprietary to them. And they so they have to think along those lines. If you look beyond 2017, the answer, yes, the pipeline goes much deeper than that. But we simply cut it off in 2017. And I think the plan is that we would come back every year and give you guys an update every single year.

And hopefully, that pipeline continues to grow. Because it's growing kind of back-end loaded, you should -- my expectation is that we should see our auto business grow quite fast. Our pipeline grew quite fast. For some time to come. This is really 1 of the largest industries in the world. We've been -- by the time that we should, we would have been investing for a decade.

The arsenal technology that we're bringing to bear to enable this industry, both for our own service, but also creating platforms for the rest of the industry is really quite significant. And leverage is basically all of the might of our company, and we've

been focused on it for -- come out of seven years. And so I think this is a really exciting opportunity for us.

You're right that we can benefit from top to bottom and to end, top to bottom meaning, in the case that we could offer the full-service and operate it with our partners, we'll get the benefit of the service. In the case that a car company would like to provide their own service and they feel that they can develop their own stack, we sell them a great ship.

And these chips are getting more and more powerful because they would like that fleet to be programmable, richly programmable for valued services for decades to come. And so they want to put as much technology into it upfront as possible. And end-to-end in the sense that we have Hyperion 8, which is a reference car computer.

And I think it will prove to be quite profound in its ability to positively impact the ecosystem and make things -- allow people to accelerate their development, just like the PC ATX motherboard did, which was a profound, deeply profound innovation. And then DGX for training models. Constellation for simulation, omni verse for simulation, all the way to, of course, our drive computers that go into the cars.

And we also announced that the same computer, the same architecture could be used for driving AV, but also for in car AI and now integrating 4 computers into one. This is a very big deal.

What used to be multiple ECUs that are sprinkled all over to car really needs to get unified so that the car could become software fine. And it needs to have something that is like a data center. It's virtualized, software-defined, but fully offloaded and accelerated, like the things that we talk about in data centers.

And so it will be architected like a data center, it would be built like a data center. It will be operated like a data center. And as a result, the car be software-defined. And so we have a lot of different ways to engage the car industry we've been talking about the car business for quite a long time because heard me talk about it now for probably about five years.

And/or in the generation coming out or in, it's just -- it's a complete home run. And because we stuck with it, we stuck with what we believe then, and it turned out to be right that the car industry really needs to be a technology industry. That the car is not just a car anymore, but it's a car within an installed base, a fleet.

That fleet is going to be managed like a data center, software-defined. And when you manage that fleet and you grow that installed base, it's going to create a gigantic installed base of future value creation for the carmakers. And I think all of these pieces are really coming together.

A - Simona Jankowski {BIO 7131672 <GO>}

Next question.

Q - Unidentified Participant

More to that. So as discussed, we've been probably more than seven years in the making to get this point. And we're talking about our pulp loan going forward, that six years to 2027, on 2027 out last year that we've provided there. It is providing for the established OEMs, our electrical vehicles or trucking or robotaxi, and we probably expect to see a revenue inflection point somewhere in the time frame of calendar 2023 and 2024.

To your earlier comment regarding the other types of things that we may provide to them in terms of our stock, helping them in the data center, helping them in the oral development process. Again, that's in our data center revenue. So this pipeline out to 2027 is really just about our automotive revenue.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question comes from Tim Arcuri at UBS.

Q - Timothy Arcuri {BIO 3824613 <GO>}

I also had a question on the grace road map. And sort of how or even if you sort of democratize that, when you sort of think about that road map, is it sort of decoupled in any way or monetized separately? I mean, certainly, the propensity is to prioritize for your own road map.

But I'm wondering sort of how you thread that needle, Jensen. And then I guess, Colette is it maybe reasonable to just take the entire $25 billion server TAM, server CPU TAM and just tack that on to the $100 billion by 2024 TAM for the company that you gave last analyst meeting?

A - Jen-Hsun Huang {BIO 1782546 <GO>}

We're going to offer offer CPUs and technology to our ecosystem partners in the form that best suites them. Today, we put our GPUs into SXM modules better than put into a by 4 or by 8 HCFs carrier board, GPU Board. We also sell our GPUs individually so that they could build their bespoke servers. A lot of -- there are a lot of different configurations of GPU servers that are available now.

In the case of the EGX launch, we worked with 15 of the world's largest computer makers to build 55 different configurations. That will be more coming. And so you could see how many different configurations of servers that 1 use and 2 use and 3 using 4 uses Otherwise -- and HPC versions.

And there's just all kinds of different configurations of servers. And so we'll offer grace as an integrated part of our DGX, we'll offer grace separately to OEMs. And then when the deal closes, we'll continue to license arm to openly to the entire industry because there are so many different versions of CPUs that could get built, and we would love to have very single version built.

And to us, we can add the NVIDIA architecture to it, whether it's RCP or somebody else's CPU, so long as the market is being created, we could add the NVIDIA architecture to it. The GPU, the DPU, CUDA, Doka, NVIDIA AI and NVIDIA Omniverse, all of our AI stacks on top and the AI skills on top.

Look, our economics are just so much so much broader -- so much richer and so much larger when we have more people around the world and the ecosystem supporting our architecture.

So that's -- that I think is the most significant bit and then everything else, we want to create products where the world doesn't have that expands our TAM that expands the market TAM. And then after that, support the customers as best they would like to be supported.

A - Colette Kress {BIO 18297352 <GO>}

Yes. So Tim, regarding your question that you asked regarding how to look at this opportunity going forward. So growth is definitely for our Giant HPC and AI use cases, which is probably a niche of the overall market as a whole.

And when we think about our TAM opportunity going forward, it is breadth and depth growth from a lot of things that you've seen us talk about today. We've talked about incorporating grace inside many of our different systems, what we have. So of course, it's an opportunity, but very hard to overall determine that file at this point.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question will come from Aaron Rakers from Wells Fargo.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yes. Thanks. Great presentation today and congrats on the product announcements. Colette, I wanted to ask you about the commentary around data center. You use the term again, good visibility.

So I'm curious of how you would characterize the visibility today relative to what it was, let's say, a year ago when you started to see kind of the strength demerge in data center. And also on top of that, when do we think about the DPUs, the blue field product starting to become a material incremental driver to the company?

A - Colette Kress {BIO 18297352 <GO>}

Sure. So let's first start with a year ago, again, still a bit different of business at that time versus what we're seeing today. So we have incorporated Mellanox into our overall stack. We continue to work on building products across overall Mellanox as well as full systems in terms of the work that we're doing.

We started off many years ago focused on hyperscales, hyperscales move cloud instances. But as you can see, we're just touching that opportunity for enterprise and enterprise -- what this does is there's just a meaningful amount of opportunities in

many different process for us to grow the overall data soon. So our overall visibility is good.

And we look at it as an opportunity to expand our types of customers with many different views. (inaudible) working with OEMs and enterprises Sepa. Therefore, we consider this to be a vast opportunity in front of us, and we feel good about the growth that we will like receive throughout this fiscal year.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question comes from Brett Simpson from Arete.

Q - Brett Simpson {BIO 3279126 <GO>}

I had a question on the software and services strategy, Jensen. So you laid out the business model for drive with the Mercedes agreement. Can you share with us how NVIDIA might monetize your other AI stacks as they come out of beta, so things like Jarvis and Omnibus.

I think you touched on those. But just curious as to how we should think about the monetization and the time line around that ramping up? And then just on the cloud side of things, do you see NVIDIA more as a competitor to the public cloud players in AI where you host services directly? Or would you expect the hyperscalers to also license these software stacks you're presenting today?

A - Jen-Hsun Huang {BIO 1782546 <GO>}

I'll go backwards. The -- all of our stacks are cloud native. Everything from to of course, now to Clara, to Jarvis, to metropolis to -- they're all cloud-native to -- and they're used by the cloud service providers. We've announced already many, many instances where the cloud service providers are using our libraries to provision their own services to deliver higher throughput. And so we run everything cloud native.

Where we can offer unique value-added is embedded and at the edge and on-prem, where someone needs 10 of the art models, that needs to be -- that are trained incredibly well, that's what they call pre trained. And that is treated like a production model, not like a demo. That will have a company with the dedication to continuously improve it for as long as they shall live. And to provide a suite.

As I mentioned, we provide some fundamental suites from site to language the speech to understanding animation and so on. To tools that allows you to customize it and adapt it to your own domains.

And then run it basically anywhere. You can run in any cloud, you can run an on-prem, you can run it in a robotic system, you can run it in your car. One architecture that spans at all. And so the way that will offer these will largely be embedded software.

And so I think of it much more as intellectual property that will be embedded into systems. And so those are the skills. Now some skills you have to operate. There's no other way to do it about operate it. In those cases, will have a slightly different model. For example, in the case of Mercedes Benz. We'll have to operate that service for basically decades.

And so we'll create a sharing business model like we've described with Mercedes Benz. And in the case of, it will be licensed as a server. Think of it as a server. There's a server and then there's a user component to it. For communities and individual users, it'll be free. It's already in open beta. It's already available and people are using it all over the world doing amazing things.

And so for researchers, for people who are individuals, for people who are playing with it, it's all free. For people who are utilizing it to operate a service, operate, for example, a digital factory, we're operating a large-scale 5G network that's simulating and optimizing in real time. Those will have a server component to it. And then a per user component to it.

And so each 1 of the software licensing, there's no 1 model that kind of cuts across everything. It just depends. And you should always come out of our first principles, what is the best way to deliver the value to the customers. But we care deeply about the software in our company. And we're going to continue to innovate in this area. So the business opportunity is quite large.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question will come from will Stein from Truist.

Q - William Stein {BIO 15106707 <GO>}

Jensen, congratulations on all the exciting announcements today. A few years ago, you surprised with a strong entry into the interface category -- protein category, 1 that you previously haven't played as strong and with the announcement of the T4 chip.

Today, you've discussed this Triton inference server, I was expecting perhaps an announcement bringing T4 up to the current Ampere architecture. Is that essentially what this announcement is about? Or is it much more or different? Any explanation would be helpful.

A - Jen-Hsun Huang {BIO 1782546 <GO>}

Yes. Will, thanks a lot for that. I should have been more clear. The -- so first of all, a comment about inference. People who gets us the most which is what inferencing is guessing, very informed guessing, making predictions.

People who guess the most frequently learn the most. And so I think inferencing has always been very core to our strategy, and we've always believed that to be very hard, and it's proven to be way harder than that.

And so I think inference is just such a great opportunity. And we love it because we like hard things. We -- that when achieved, could make a real impact. Triton is an inference server. It runs a data center scales, and there are so many people using it.

It supports CPUs and GPUs, and it supports every generation of our GPU. And so if you manage a hyperscale data center, it's got all kinds of computer chips inside. And this is the only inference server that's optimized for all of them.

Basically, there's 2 types, and that covers 99%, which is X86 CPUs and and GPUs that goes all the way back to Kepler. And so there are 6 generations of GPUs and all the different versions of them, and we support everything we want and it's optimized for every 1 of them. We'll support them. For as long as we shall live.

And that's the advantage of trite it's also open source so that the CSPs could make bespoke versions and derivatives from the they like. And so we're seeing incredible success with train. The version of T4 that's based on Ampere, we announced at GTC. It's called A30. A30 is our brand-new GPU that is designed for if you will, mainstream use.

And it's designed for enterprise servers, cloud computers. And we're super excited about 830. It's derived from a 100, of course, and it's not as powerful, but it's much more -- it's more -- it fits into a much lower energy or power envelope so that you can deploy it very broadly. Okay. So the inference, Triton and 830 is the mainstream version of.

Operator

Our next question comes from Blayne Curtis from Barclays.

Q - Blayne Curtis {BIO 15302785 <GO>}

So just following back on a question on Bluefield. You announced last GTC, Bluefield 2, just wondering if you provide any clarity or color as to the traction you've gotten over the last year. I don't think you've ever broken out how big that is, but any color on the road map and the traction with Bluefield 2 would be great?

A - Jen-Hsun Huang {BIO 1782546 <GO>}

BlueField 2 is a programmable device and it needs a rich software stack on top of it. Just a our GPUs need a CUDA in order for it to be useful. And on top of CUDA, need a whole bunch of other libraries, we call the CUDA X libraries, acceleration libraries.

All of that has been put together in the case of BlueField called DOCA 1.0, and we just released it. We just released the software. So now DOCA 1.0 and BlueField 2 are ready for production. BlueField 3 is right behind its heels.

And you could tell how excited I am about this area because very simplistically, a data center is becoming software-defined. And you know that well. But what that

means for everybody, is that it takes networking, storage, virtualization and now really, really importantly, cybersecurity, and it puts it in software. It runs on the CPU.

The software-defined data center stack is now overloading the CPU, not to mention it's weird to have the control plane, the security plane, agents, all running co-mingled with the application, which could be the intruder. And so the right answer really is to isolate it. That's very, very clear, and everybody agrees with that. To isolate that, offload that and very importantly, accelerate the workload so that you could take it off of the CPUs, which is ultimately what the data centers for. To host applications.

And one of the things that we prove is that we've demonstrated is that BlueField an accelerate an isolate. And it frees back, 1/3 of the CPU course. And it's going to grow. And the reason for that is because of Zero trust security models where every single transaction is going to be monitored.

And so I'm incredibly excited about this new strategy. BlueField 2 is really the world's first version of it. And DOCA 1.0 is the first version of it. And so we're in the process of building that. We have developed out to all the major companies in the world.

You could just imagine, and you saw the IT companies that are all part of the -- part of our launch. And so I think this is going to be a giant business. Here's my prediction, and I predicted this I made a prediction about 20 years ago that every single computer will have GPU in it, and they came true.

And I predict that every single data center will have devices like computing platforms, like BlueField that we call data center infrastructure processing platforms. That every single data center, every single computer will have a blue field like device in the future. And I think it will take less than five years because of the importance of cybersecurity that is top of mind to everybody today.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question comes from Harlan Sur from JPMorgan.

Q - Harlan Sur {BIO 6539622 <GO>}

Good morning, and exciting to be here for another GTC. And it's actually great to see the NVIDIA's accelerated compute architecture and ecosystem move into new opportunities like 5G as we all know, the move to 5G opens up the opportunity for broadband wireless connectivity to more than just smartphones.

So as EGX with the aerial A100 platform, in trial or planning to be trailed by any of these large telco or greenfield networks with the cloud-native open run or architecture soon. And how big is this opportunity for NVIDIA over the next 3 to five years?

A - Jen-Hsun Huang {BIO 1782546 <GO>}

Yes. I'm excited that I'm just about as excited you could imagine about 5G and the reason for that is because particularly private 5G.

I think commercial 5G, consumer 5G is really fantastic, and it will incrementally and over time, increase the capacity of broadband and very importantly, continue to drive down cost of broadband because what goes along with a lot of capacity, is the decrease in cost. And so I think consumer commercial 5G is really fantastic.

But what's brand-new is private 5G, all these spectrums that companies could license to operate a wireless network that is secure and private in farms and entire factories. And as you know, these factories are gigantic. The vast majority of the earth livable space is covered by farms and infrastructure.

And finally, we have the networking available to deliver computation services out to the edge. And the computation service that I think is most exciting is sensors connected to AI applications, AI skills, they're monitoring and predicting, making sense of the world. And 5G makes up possible.

You could install all kinds of cameras now with 5G, little tiny 5G modem that is light and data rate, light low end power, very low cost. And you could just bring up cameras and sensors of all kinds all over the place. As soon as you connect it, you put it up, connected to power, it's on the network.

And so the ability to outfit infrastructures, factories, farms, road, cities is so much easier with 5G. So I think at the foundational layer, 5G is the big catalyst for ending industrial AI to really get out to the edge. The second thing is we need a -- we need it to be software declined.

Just like modern data centers are, you wanted to be software-defined. If you can make a coware defined, it is orchestratable it is -- you can disaggregate it, so you could -- meaning you could put computing power, shift it from place to place across the network to focus the computation on where the workload is.

And all of that is possible if you make software defined and what -- has done with the aerial 5 G&A 100 there are features in it that enables much better 5G. And then with the BlueField, we made it possible to deliver a whole 5G base station inside a standard enterprise server.

And it's a state-of-the-art 5G base stations, 20 gigahertz, it supports up to 9, 100 megahertz, 64/64 massive MIMO radios. And so this is an amazing 5G base station. You saw the announcement of all the partners who are working with us. And we're all racing to get the full stack out there.

Fujitsu is looking with us. is working with us, Ericsson's working with us, of course, as we've announced it. And many, many more will come. And the reason for that is

because they need a programmable software-defined 5G base station that is connected to the AI edge, and that's what EGX is all about.

A - Simona Jankowski {BIO 7131672 <GO>}

And our last question will come from Atif Malik from Citi.

Q - Atif Malik {BIO 15866921 <GO>}

In your updated Q1 guidance, is gaming is still the biggest sequential driver. And then when you spoke about AGX is gaining rapid adoption, when do you expect to recognize software revenues from the NVIDIA AI into enterprise license model?

A - Colette Kress {BIO 18297352 <GO>}

Yes. So thanks for the question regarding our outlook for the quarter. Gaming will still be a material part of our growth from Q4 to Q1. And that is definitely expected.

So from looking additionally on your software opportunities in time, there's many different ways that we've talked about today in terms of how we will monetize the overall software, many different ways, including incorporating in OEMs to our channel, through our enterprise channel through, also looking at it as part of our platform.

And then keep in mind, there's still an opportunity with automotive going forward, working directly with those OEMs and purchasing it overall capable.

So over time, we believe this will continue to grow. Now it is incorporated, of course, in the go today. But looking at it as a separate offering is new and something that we'll see of the opportunities going forward. I'll see if Jensen was to add any more color on that.

A - Jen-Hsun Huang {BIO 1782546 <GO>}

No, that was great CLA. That was great.

A - Simona Jankowski {BIO 7131672 <GO>}

Thank you. This is all the time we have for Q&A today. I will now turn the call back over to Jensen for closing comments.

A - Jen-Hsun Huang {BIO 1782546 <GO>}

Thank you. Thank you for joining us today. I hope you guys all enjoyed GTC. It was -- it's the endeavor of multiple years of development inside the company, some of it, we developed in plain sight. And with a vision to pull it together in just this way today.

It is the hard work and the genius of a couple of 20,000 employees at NVIDIA. And I am so grateful for everything that they've done and work so hard and so diligently

towards this vision.

The strategies that we announced today, the products that we announced and the partnerships that we announced, the new markets that we announced are all targeting and focusing and enabling the forces that are shaping our industry.

And we spoke about accelerated computing and the groundbreaking work that we're doing in health care and even cloud computing. We spoke a lot about AI. AI is not just algorithms, AI is a fundamental way of doing software.

And it's transformed computing completely, and used the self-driving cars as an example. To illustrate how AI and machine learning changes how you should think about developing software. And as a result, even profoundly changes how you think about your products, what used to be a car is now part of an installed base.

And we talked about data center, how the entire data center is going to be programmed like 1 computer, what a miracle it is that technology has reached a point where it's possible for 1 software engineer to write an application that scales across hundreds of thousands of servers. It's a shocking realization and networking technology made it possible.

And it's what's driving Mellanox's, high-speed and low latency networking. It transformed how data centers ought to be architected, software-defined data centers and made possible a brand-new type of computing platform where the computing is done in the fabric, otherwise known as data center infrastructure.

The data center is now a new unit of computing, and it has profound implications on how we architect data centers going forward, not just in the data center proper as we see it, but remember, everything is going to be a data center, 5G base station is going to be a data center.

The car is going to be a data center on wheels. They're going to be architected the same way. And it will be highly secure in the same way. There'll be you'll be able to orchestrate these computers in the same way you use very similar approaches. 5G, incredibly excited about the convergence of 5G and AI. And so now we can deliver AI on 5G.

What a transformative time it is for the industries? We use transportation as a great easy example. And notice how it's possible to apply the combinations of technologies that I've mentioned, to turn your automotive fleet into your installed base. And they think like cable network providers, you could think like smartphone network providers, basically like network providers.

5G and with AI on 5G running on EGX, we would like to do that for the entire industry, for every industry, for health care, for warehouse, for logistics, where companies and enterprises in these industries could reinvent their business models

and their products and services and enjoy their smartphone moment for every industry to become a technology industry from agriculture and warehouse.

And then lastly, the ultimate form of AI, if you will, is autonomous systems. And remember, that autonomous systems in our imagination is not just physical autonomous systems, but there are going to be virtual autonomous systems. There will probably be 1 million times more robots inside virtual worlds.

Than there are in physical world, and there are going to be tons of robots that are autonomous someday. Self-driving cars, of course, is the first example of them. But that's just the first example. And we have the AI technology that are being developed and we have machine learning practices and skills.

And all of the things that we're doing will lead up to be able to deploy robotics into the physical world, but they will have digital twins. And those digital twins will help them learn. And those digital twins will be synchronized to the physical versions and they'll be -- they'll exist in another virtual world that is physically and photo realistically like ours.

It behaves like ours. And so -- but it's completely simulated. And this virtual world and these digital twins will make up some of the industrial that people have been talking about, particularly for consumer applications.

But the physical version, the industrial version of them will be called Digital twins. And it's going to enable robots to learn how to be robots and work with each other and stimulate factories. And the work that we do with BMW hopefully activate some of your imagination there, but there is so much more.

There are just powerful forces shaping and reshaping industries, and we're excited to be part of them. Thanks for joining us today in our keynote and also our analyst meeting to hear us talk about the work that we do that we're so passionate about. And with that, I wish you all well, and stay healthy.