Q4 2018 Earnings Call

Company Participants

- Colette M. Kress, Chief Financial Officer & Executive Vice President
- Jen-Hsun Huang, President, Chief Executive Officer & Director
- Simona Jankowski, Vice President, Investor Relations

Other Participants

- Blayne Curtis, Analyst
- C.J. Muse, Analyst
- Christopher Rolland, Analyst
- Craig A. Ellis, Analyst
- Harlan Sur, Analyst
- Joseph Moore, Analyst
- Mark Lipacis, Analyst
- Mitch Steves, Analyst
- Stacy Aaron Rasgon, Analyst
- Toshiya Hari, Analyst
- Vivek Arya, Analyst
- William Stein, Analyst

MANAGEMENT DISCUSSION SECTION

Operator

Good afternoon. My name is Victoria, and I'm your conference operator for today. Welcome to NVIDIA's financial results conference call. The phone lines have been placed on mute to prevent any background noise. After the speakers' remarks, there will be a question-and-answer period.

Thank you. I'll now turn the call over to Simona Jankowski, Vice President of Investor Relations, to begin your conference.

Simona Jankowski {BIO 7131672 <GO>}

Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call for the fourth quarter of fiscal 2018. With me on the call today from NVIDIA are Jensen Huang, President and Chief Executive Officer, and Colette Kress, Executive Vice President and Chief Financial Officer.

I'd like to remind you that our call is being webcast live on NVIDIA's Investor Relations website. It's also being recorded. You can hear a replay by telephone until February 16, 2018. The webcast will be available for replay up until next quarter's conference call to discuss our fiscal first quarter financial results. The content of today's call is NVIDIA's property. It can't be reproduced or transcribed without our prior written consent.

During this call, 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 the disclosure in today's earnings release, 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, February 8, 2018, based on information currently available to us. Except as required by law, we assume no obligation to update any such statements.

During this call, we will discuss non-GAAP financial measures. You can find a reconciliation of these non-GAAP financial measures to GAAP financial measures in our CFO Commentary, which is posted on our website.

With that, I'll turn the call over to Colette.

Colette M. Kress {BIO 18297352 <GO>}

Thanks, Simona.

We had an outstanding quarter and fiscal 2018, led by strong growth in our gaming and data center businesses. Q4 revenue reached $2.91 billion, up 34% year on year, up 10% sequentially, and above our outlook of $2.65 billion. All measures of profitability set records. They also hit important milestones. For the first time, gross margins strongly exceeded 60%. Non-GAAP operating margins exceeded 40%, and net income exceeded $1 billion.

Fiscal 2018 revenue was $9.71 billion, up 41% or $2.8 billion above the previous year. Each of our platforms posted record full-year revenue, with data center growing triple digits. From a reporting segment perspective, Q4 GPU revenue grew 33% from last year to $2.46 billion. Tegra processor revenue rose 75% to $450 million.

Let's start with our gaming business. Q4 revenue was $1.74 billion, up 29% year on year and up 11% sequentially, with growth across all regions. Driving GPU demand were a number of great titles during the holiday season, including Player's Battleground (sic) [PlayerUnknown's Battlegrounds] (3:40), PUBG, Destiny 2, Call of Duty: World War II, Star Wars: Battlefront 2.

PUBG continued its remarkable run, reaching almost 30 million players and recording more than 3 million concurrent players. These games deliver stunning visual effects that require strong graphics performance, which has driven a shift toward the higher end of our gaming portfolio and adoption of our Pascal architecture.

E-sports continues to grow, expanding the overall industry and our business. In one sign of their popularity, Activision's Overwatch League launched in January and reached 10 million viewers globally in its first week.

We had a busy start to the year with a number of announcements at the annual Consumer Electronics Show in Las Vegas. We introduced NVIDIA's BFGD, Big Format Gaming Displays, in a partnership with Acer, Asus, and HP. These high-end 65-inch 4K displays enable ultra-low latency gaming and integrate our SHIELD streaming device, offering popular apps such as Netflix, Gaming Video, YouTube, and Hulu. The BFGD won nine Best of Show awards from various publications.

We expanded the free beta of GeForce NOW beyond Macs to Windows-based PCs, and we enhanced GeForce experience with new features, including NVIDIA Freestyle for customizing game play with various filters, an updated NVIDIA Ansel photo mode, and support for new titles with ShadowPlay Highlights for capturing gaming achievements. Additionally, the Nintendo Switch gaming console contributed to our growth, as it became the fastest selling console of all time in the U.S.

Strong demand in the cryptocurrency market exceeded our expectations. We met some of this demand with a dedicated board in our OEM business, and some was met with our gaming GPUs. This contributed to lower than historical channel inventory levels of our gaming GPUs throughout the quarter. While the overall contribution of cryptocurrency to our business remains difficult to quantify, we believe it was a higher percentage of revenue than the prior quarter. That said, our main focus remains on our core gaming market, as cryptocurrency trends will likely remain volatile.

Moving to data center, revenue of $606 million was up 105% year on year and up 20% sequentially. This excellent performance reflected strong adoption of Tesla V100 GPUs based on our Volta architecture, which began shipping in Q2 and continued to ramp in Q3 and Q4. V100s are available through every major computer maker and have been chosen by every major cloud provider to deliver AI and high-performance computing. Hyperscale and cloud customers adopting the V100 include Alibaba, Amazon Web Services, Baidu, Google, IBM, Microsoft Azure, Oracle, and Tencent.

We continued our leadership in AI training markets, where our GPUs remain the platform of choice for training deep learning networks. During the quarter, Japan's Preferred Networks trained the ResNet-50 neural network for image classification in a record of 15 minutes by using 1,024 Tesla P100 GPUs. Our newer generation V100s deliver even higher performance, with the Volta architecture offering 10 times the deep learning performance of Pascal.

We also saw growing traction in the AI inference market, where NVIDIA's platform can improve performance and efficiency by orders of magnitude over CPUs. We continue to view AI inference as a significant new opportunity for our data center GPUs. Hyperscale inference applications that run on GPUs include speech recognition, image and video analytics, recommender systems, translation, search, and natural language processing.

The data center business also benefited from strong growth in high-performance computing. The HPC community has increasingly moved to accelerated computing in recent years, as Moore's Law has begun to level off. Indeed, more than 500 HPC applications are now GPU-accelerated, including all of the top 15.

NVIDIA added a record 34 new GPU accelerated systems to the latest Top500 supercomputer list, bringing our total to 87 systems. We increased our total petaflops on the list by 28%, and we captured 14 of the top 20 spots on the Green500 list of the world's most energy-efficient supercomputers.

During the quarter, we continued to support the build-out of major next-generation supercomputers. Among them is the U.S. Department of Energy's Summit system, expected to be the world's most powerful supercomputer when it comes online later this year. We also announced new wins such as Japan's fastest AI supercomputer, the ABCI system, which leverages more than 4,000 Tesla V100 GPUs.

Importantly, we are starting to see the convergence of HPC and AI, as scientists embrace AI to solve problems faster. Modern supercomputers will need to support multi-precision computation for applying deep learning together with simulation and testing. By combining AI with HPC, supercomputers can deliver increased performance that is orders of magnitudes greater in computations ranging from particle physics to drug discovery to astrophysics.

We are also seeing traction for AI in a growing number of vertical industries, such as transportation, energy, manufacturing, smart cities, and healthcare. We announced engagements with GE Health and Nuance in medical imaging, Baker Hughes, a GE Company, in oil and gas, and Japan's Komatsu in construction and mining.

Moving to professional visualization, fourth quarter revenue grew to a record $254 million, up 13% from a year ago, up 6% sequentially, driven by demand for real-time rendering as well as emerging applications like AI and VR. These emerging applications now represent approximately 30% of pro visualization sales.

We saw strength across several key industries, including defense, manufacturing, energy, healthcare, and Internet service providers. Among key customers, high-end Quadro products are being used by GlaxoSmithKline for AI and by PEMEX oil and gas for seismic processing and visualization.

Turning to automotive, in automotive for the fourth quarter, revenue grew 3% year on year to $132 million and was down 8% sequentially. The sequential decline reflects our transition from infotainment, which is becoming commoditized, to next-generation AI

cockpit systems and complete top-to-bottom self-driving vehicle platforms built on NVIDIA hardware and software.

At CES, we demonstrated our leadership position in autonomous vehicles, with several key milestones and new partnerships that point to AI self-driving cars moving from deployment to production.

In a standing room-only keynote that drew nearly 8,000 attendees, Jensen announced that DRIVE Xavier, the world's first autonomous machine processor, will be available to customers this quarter. With more than 9 billion transistors, DRIVE Xavier is the most complex system-on-a-chip ever created. We also announced that NVIDIA DRIVE is the world's first functionally safe AI self-driving platform, enabling automakers to create autonomous vehicles that can operate safely, a necessary ingredient for going to market.

Additionally, we announced a number of collaborations at CES, including with Uber, which has been using NVIDIA technology for the AI computing system in its fleets of self-driving cars and freight trucks. We announced that ZF and Baidu are using NVIDIA DRIVE self-driving technologies to create a production-ready AI autonomous vehicle platform for China, the world's largest automotive market. Production vehicles utilizing this technology, including those from Chery, are expected on the road by 2020. We also announced a partnership with Aurora, which is working to create a modular, scalable, Level 4 and Level 5 self-driving hardware platform incorporating the NVIDIA DRIVE Xavier processor.

Jensen was joined on stage by Volkswagen CEO Herbert Diess. They announced the new generation of intelligent VW vehicles will use the NVIDIA DRIVE Intelligent Experience or DRIVE IX platform to create the new AI-infused cockpit experiences and improved safety.

Later at CES, Mercedes Benz announced that MBUX, its new AI-based smart cockpit, uses NVIDIA's graphics and AI technologies. The MBUX user experience, which includes beautiful touchscreen displays and a new voice-activated assistant, debuted last week in a Mercedes-Benz A-Class compact car and will ship this spring. And earlier this week, we announced a partnership with Continental to build AI self-driving vehicle systems from enhanced Level 2 to Level 5 for production in 2021.

There are now more than 320 companies and research institutions using the NVIDIA DRIVE platform. That's up 50% from a year ago and encompasses virtually every carmaker, truck maker, robo-taxi company, mapping company, sensor manufacturer, and software startup in the autonomous vehicle ecosystem. With this growing momentum, we remain excited about the intermediate to long-term opportunities for autonomous driving.

Now turning to the rest of the P&L, Q4 GAAP gross margins were 61.9% and non-GAAP was 62.1%, records that reflect continued growth in our value-added platforms. GAAP operating expenses were $728 million and non-GAAP operating expenses were $607 million, up 28% and 22% year on year respectively. We continue to invest in the key platforms driving our long-term growth, including gaming, AI, and automotive.

GAAP EPS was $1.78, up 80% from a year earlier. Some of the upside was driven by a lower than expected tax rate as a result of U.S. tax reform and excess tax benefits related to stock-based compensation. Our fourth quarter GAAP effective tax rate was a benefit of 3.7% compared with our expectation of a tax rate of 17.5%.

Non-GAAP EPS was $1.72, up 52% from a year ago, reflecting a quarterly tax rate of 10.5% compared with our expectation of 17.5%.

We returned $1.25 billion to shareholders in the fiscal year through a combination of quarterly dividends and share repurchases. Our quarterly cash flow from operations reached record levels at $1.36 billion, bringing our fiscal year total to a record $3.5 billion. Capital expenditures were $416 million for the fourth quarter, inclusive of $335 million associated with the purchase of our previously financed Santa Clara campus building.

Let me take a moment to provide a bit more detail on the impact of U.S. corporate tax reform on the quarter and our go-forward financials. In Q4, we recorded a GAAP-only one-time net tax benefit of $133 million or $0.21 per diluted share. This is primarily related to provisional tax amounts for the transition tax on accumulated foreign earnings and remeasurement of certain deferred tax assets and liabilities associated with the Tax Cuts and Jobs Act. We previously accrued for taxes on a portion of forward earnings in excess of the provisional tax amount recorded for the transition tax, hence the one-time benefit.

For fiscal 2019, we expect our GAAP and non-GAAP tax rates to be around 12%, which is down from approximately 17% previously. This does not take into effect the excess tax benefit from stock-based compensation which, depending on stock price and vesting schedule, could increase or decrease our tax rate and GAAP in a given quarter.

In terms of our capital allocation priorities, we continue to focus first and foremost on investing in our business, as we see significant opportunity ahead. Our lower tax rate strengthens our ability to invest in both OpEx, such as adding engineering talent, as well as CapEx, such as investing in supercomputers for internal AI development. In addition, we remain committed to returning cash to shareholders, with our plan remaining at $1.25 billion for fiscal 2019.

With that, let me turn to the outlook for the first quarter of fiscal 2019. We expect revenue to be $2.9 billion, plus or minus 2%. GAAP and non-GAAP gross margins are expected to be 62.7% and 63% respectively, plus or minus 50 basis points. GAAP and non-GAAP operating expenses are expected to be approximately $770 million and $645 million respectively.

GAAP and non-GAAP OI&E are both expected to be nominal. GAAP and non-GAAP tax rates are both expected to be 12%, plus or minus 1%, excluding discrete items. For the full fiscal year 2019, we expect our operating expenses to grow at a similar pace as in Q1. Further financial details are included in the CFO Commentary and other information available at our IR website.

In closing, I'd like to highlight a few upcoming events for the financial community. We'll be presenting at the Goldman Sachs Technology and Internet Conference on February 13 and at the Morgan Stanley Technology, Media & Telecom Conference on February 26. We will also be hosting our annual Investor Day on March 27 in San Jose on the sidelines of our annual GPU Technology Conference, which we are very excited about.

We will now open the call for questions. Operator, will you poll for questions, please?

Q&A

Operator

Absolutely. Your first question comes from the line of C.J. Muse from Evercore.

Q - C.J. Muse

Good afternoon, thank you for taking my question. I guess first question, when I think about normal seasonality for gaming, that would imply data center potentially north of $700 million-plus into the coming quarter. And so I'm curious if I'm thinking about that right or whether crypto is being modeled more conservatively by you guys. And so I would love getting your thoughts there.

A - Jen-Hsun Huang

Which way is more conservatively, C.J.?

Q - C.J. Muse

Yes?

A - Jen-Hsun Huang

When you say conservatively, which direction were you saying it was – are you implying up or down?

Q - C.J. Muse

I was just curious to hear your thoughts there.

A - Jen-Hsun Huang

We modeled crypto approximately flat.

Q - C.J. Muse

Okay. And then I guess as part of the larger question, how are you thinking about seasonality for gaming into the April quarter?

A - Jen-Hsun Huang

There are a lot of dynamics going on in gaming. One dynamic of course is that there's fairly sizable pent-up demand going into this quarter. But I think the larger dynamics that are happening relate to just the really amazing games that are out right now. PUBG is just doing incredibly well, as you might have known, and it's become a global phenomenon. And whether it's here in the United States or in Europe, or in China, in Asia, PUBG is just doing incredibly well. And we expect other developers to come up with similar genres like PUBG that are going to be coming out in the near future, and I'm super excited about these games.

And then of course there's Call of Duty, there's Star Wars. There are just so many great games that are out in the marketplace today, Overwatch and League of Legends still doing well. There are just a countless number of great franchises that are out in the marketplace, and the gaming market is growing and production value is going up. And that's driving increased unit sales of GPUs as well as ASPs of GPUs. And so I think those are – that's probably the larger dynamic of gaming.

Operator

Your next question comes from the line of Mark Lipacis with Jefferies.

Q - Mark Lipacis {BIO 2380059 <GO>}

Hi, thanks for taking my question. The first question, the checks we've done indicate that the Tensor Cores you put into Volta give it a huge advantage in neural network applications and the data center, and I'm wondering whether the Tensor Cores might also have a similar kind of utility in the gaming market.

A - Jen-Hsun Huang

First of all, I appreciate you asking a Tensor Core question. It is probably the single biggest innovation we had last year in data centers. Our GPUs, the equivalent performance to one of our GPUs, one of our Volta GPUs would take something along the lines of 20-plus CPUs or 10-plus nodes, and so one GPU alone would do deep learning so fast that it would take 10-plus CPU-powered server nodes to keep up with.

And then Tensor Core comes along last year, and we increased the throughput of deep learning, increased the computational throughput of deep learning by another factor of eight. And so Tensor Core really illustrates the power of GPUs. It's very unlike a CPU where the instruction set remains locked for a long time, and it's hard – it's difficult to advance. In the case of our GPUs and with CUDA, that's one of its fundamental advantages. We can continue to – year-in and year-out continue to add new capabilities to it. And so Tensor Core's boost of the original great performance of our GPU has really raised the bar last year.

And as Colette said earlier, our Volta GPU has now been adopted all over the world, whether it's in China with Alibaba, Tencent, and Baidu, iFLYTEK too. Here in the United States, Amazon and Facebook and Google and Microsoft and IBM and Oracle, and in Europe, in Japan, the number of cloud service providers that have adopted Volta has been terrific, and I think everybody really appreciates the work that we did with Tensor

Core. And all of the updates that are now coming out from the frameworks, Tensor Core is a new instruction set, it's a new architecture, and the deep learning developers have really jumped on it. And almost every deep learning framework is being optimized to take advantage of Tensor Core.

On the inference side, and that's where it would play a role in video games, you could use deep learning now to synthesize and to generate new art. And we've been demonstrating some of that at GTC, if you've seen some of that. Whether it's improve the quality of textures, generating artificial characters, animating characters, whether it's facial animation for speech or body animation, the type of work that you can do with deep learning for video games is growing, and that's where Tensor Core could be a real advantage.

You can take a look at the computational capability now we have in Tensor Core, compare that to a non-optimized GPU or even a CPU, it's now two-plus orders of magnitude greater computational throughput. And that allows us to do things like synthesize images in real time, synthesize virtual worlds, animate characters, animate faces, bring a new level of virtual reality and artificial intelligence to these video games.

Operator

Your next question comes from the line of Vivek Arya with Bank of America.

Q - Vivek Arya {BIO 6781604 <GO>}

Thank you for taking my question and congratulations on the strong growth and the consistent execution; Jensen, just a near and longer-term question on the data center. Near term, you had a number of strong quarters in data center. How is the utilization of these GPUs, and how do you measure whether you're over or under from a supply perspective?

And then longer term, there seems to be a lot of money going into startups developing silicon for deep learning. Is there any advantage they have in taking a clean sheet approach, or is GPU the most optimal? Like if you were starting a new company looking at AI today, would you make another GPU or would you make another ASIC or some other format? Just any color would be helpful.

A - Jen-Hsun Huang

Sure. In the near term, the best way to measure customers that are already using our GPUs for deep learning is repeat customers. When they come back another quarter, another quarter, and they continue to buy GPUs, that would suggest that their workload is continuing to increase. With existing customers that already have very deep penetration, another opportunity for us would be using our GPUs for inference, and that's an untapped growth opportunity for our company that's really, really exciting, and we're seeing traction there.

For companies that are not at the forefront, the absolute forefront of deep learning, which, with the exception of one or two or three hyperscalers, almost everybody else I would put in this category, their deployment, their adoption of deep learning, applying deep learning to all of their applications is still ongoing. And so I think the second wave of customers is just showing up.

And then there's the third wave of customers, which is they're not hyperscalers. They're Internet service applications, Internet applications for consumers. They have enormous customer bases that they could apply artificial intelligence to, but they run their application in hyperscale clouds. That third phase of growth is now really spiking, and I'm excited about that.

And so that's kind of the way to think about it. There are the pioneers, the first phase. Are they returning customers? Then there's the second phase that's now ramping, the third phase that's now ramping. And then for everybody we have an opportunity to apply our GPUs for inference.

If I had all the money in the world and I had, for example, billions and billions of dollars of R&D, I would give it to NVIDIA's GPU team, which is exactly what I do. And the reason for that is because the GPU was already inherently the world's best high-throughput computational processor. A high-throughput processor is a lot more complicated than linear algebra done that you would substantiate from a synopsis tool. It's not quite that easy. The computation throughput, keeping everything moving through your new chip with supreme levels of energy efficiency, with all of the software that's needed to keep the data flowing, with all of the optimizations that you do with each and every one of the frameworks, the amount of complexity there is just really enormous.

The networks are changing all the time. It started out with just basically CNNs and then all kinds of versions of CNNs now. It started out with RNNs and simple RNNs, and now there are all kinds of LSTMs and gated RNNs, all kinds of interesting networks that are growing. It started out with just 8 layers and now it's 152 layers, going to 1,000 layers. It started with mostly recognition and now it's moving to synthesis with GANs, and there are so many versions of GANs. And so all of these different types of networks are really, really hard to nail down, and we're still at the beginning of AI.

So the ability for our GPUs to be programmable to all of these different architectures and networks is just an enormous advantage. You don't ever have to guess whether NVIDIA GPUs could be used for one particular network or another. And so you could buy our GPUs at will and know that every single GPU that you buy gives you an opportunity to reduce the number of servers in your data center by 22 nodes, by 10 nodes, 22 CPUs. And so the more GPUs you buy, the more money you save. And so I think that capability is really quite unique.

And then if I could just give you one example from last year or from the previous year. We introduced 16-bit mix precision. We introduced 8-bit integer. We introduced NVLink the year before this last year. This year, this last year we introduced Tensor Core which increased it by another factor of nearly 10. Meanwhile, our GPUs get more complex.

Energy efficiency gets better and better every single year, and the software richness gets more amazing.

And so it's a much harder problem than just a multiply accumulator. Artificial intelligence is the single most complex mode of software that the world has ever known. That's the reason why it's taken us so long to get here. And these high-performance supercomputers is an essential ingredient, an essential instrument in advancing AI. And so I don't think it's nearly as simple as linear algebra. But if I had all the money in the world, I would certainly invest it in the team that we have.

Operator

Your next question comes from the line of Stacy Rasgon with Bernstein Research.

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

Hi, guys, thanks for taking my questions. I have a question for Colette. So if I correct for the switch revenue growth in the quarter, it means the gaming business ex-switch was up – I don't know, maybe $140 million, $150 million. In your Q3 commentary you did not call out crypto as a driver. You are calling it out in Q4. Is it fair to say that incremental growth is all crypto?

And I guess going forward, you mentioned pent-up demand. Normally your seasonality for gaming would be down probably double digits. Do you think that pent-up demand is enough to reverse that normal seasonal pattern or normally down? And frankly, do you think gamers can even find GPUs at retail at this point to buy in order to satisfy that pent-up demand?

A - Colette M. Kress {BIO 18297352 <GO>}

So let me comment on the first one. We did talk about our overall crypto business last quarter as well. We indicated how much we had in OEM boards, and we also indicated that there was definitely some also in our GTX business. Keep in mind that's very difficult for us to quantify down to the end customer it is. But yes, there is also some in our Q3, and we did comment on it. So here we are commenting in terms of what we saw in terms of Q4. It's up a bit from what we saw in Q3, and we do again expect probably it going forward. I'll let Jensen answer regarding the demand for gamers as we move forward.

A - Jen-Hsun Huang

So one way to think about the pent-up demand is we typically have somewhere between six to eight weeks of inventory in the channel. And I think you would ascertain that globally right now the channel is relatively lean. We're working really hard to get GPUs out into the marketplace for the gamers, and we're doing everything we can to advise e-tailers and system builders to serve the gamers. And so we're doing everything we can. But I think the most important thing is we've just got to catch up with supply.

Operator

Your next question comes from the line of Mitch Steves with RBC.

Q - Mitch Steves {BIO 19155169 <GO>}

Hey, guys, thanks for taking my question. I actually want to circle back on the autos since I was at CES. It's still on track for calendar – towards calendar year 2019, at the end of that where we see the autonomous ASP uplift. And just to clarify, the expected ASP uplift is somewhere around $1,000. Is that about right?

A - Jen-Hsun Huang

Yes, it just depends on mix. I think for autonomous vehicles that still have drivers, passenger cars, branded cars, ASPs anywhere from $500 to $1,000 make sense. For robot taxis where they're driverless, they're not autonomous vehicles, they're actually driverless vehicles, the ASP will be several thousand dollars.

And in terms of timing, I think that you're going to see larger and larger deployments starting this year and going through next year for sure, especially with robot taxis. And then with autonomous vehicles, cars that have autonomous driving capability, automatic driving capability starts late 2019. You could see a lot more in 2020. And just almost every premium car by 2022 will have autonomous – automatic driving capabilities.

Operator

Your next question comes from the line of Toshiya Hari with Goldman Sachs.

Q - Toshiya Hari {BIO 6770302 <GO>}

Great, thanks very much for taking the question. Jensen, I was hoping to ask a little bit about inferencing. How big was inferencing within data center in Q4 or fiscal 2018? And more importantly, how do you expect it to trend over the next 12 to 18 months? Thank you.

A - Jen-Hsun Huang

Thanks a lot, Toshi; first of all, just a comment about inference. The way that it works is you take the output of these frameworks, and the output of these frameworks is a really complex, large computational graph. When you think about these neural networks, and they have millions of parameters, millions of parameters, millions of anything is very complex. And these parameters are waves and activation layers and activation functions, and there are millions of them. And it's millions of them that composes – consists of this computational graph. And this computational graph has all kinds of interesting and complicated layers.

And so you take – this computational graphic comes out of each one of these frameworks, and they're all different. They're in different formats, they're in different styles, and they're different architectures. They're all different. And you take these computational graphs, and you have to find a way to compile it, to optimize this graph, to rationalize all of the things that you could combine and to fold, reduce the amount of conflict across all of the resources that are in your GPUs or in your processor. And these conflicts could be on-chip

memory and register files and data paths. It could be the fabric. It could be the framework for interface. It could be the amount of memory. This computer is really complicated across all these different processors, and the interconnect between GPUs, the network that connects multiple nodes. And so you've got to figure out what all these different conflicts are, resources are, and compile and optimize to take advantage of it to keep it moving all the time.

And so TensorRT is basically a very sophisticated optimizing graph compilation, graph compiler, and it targets each one of our processors. The way it targets Xavier is different than the way it targets Volta, the way it targets our inference, the way it targets for low energy, for different precisions. All of that targeting is different. And so first of all, TensorRT, the software of inference, that's really where the magic is.

Then the second thing that we do, we optimize our GPUs for extremely high throughput to support different precisions because some networks could afford to have an 8-bit integer or even less. Some really can barely get by with a 16-bit floating point, and some you really would like to keep it at a 32-bit floating point so that you don't have to second-guess about any precision that you lost along the way.

And so we created an architecture that consists of this optimizing graph, computational graph compiler, to processors that are very high-throughput, that are mixed precision, so that's kind of the background.

We've been sampling our Tesla P4, which is our data center inference processor, and we're seeing just really exciting response. And this quarter we started shipping. Looking outwards, my sense is that the inference market is probably about as large in the data centers as training. And the wonderful thing is everything that you train on our processor will inference wonderfully on our processors as well.

And the data centers are really awakening to the observation that the more GPUs they buy for offloading, inference, and training, the more money they save. And the amount of money they save is not 20% or 50%. It's factors of 10. The money savings for all these data centers that are become increasingly capital constrained is really quite dramatic.

And then the other inference opportunity for us is autonomous machines, which is self-driving cars. TensorRT also targets Xavier. TensorRT targets our Pegasus robot taxi computer. And they all have to inference incredibly efficiently so that we can sustain real time keeping energy level low and keep the cost low for car companies. So I think inference it is very important work for us. It is very complicated work, and we're making great progress.

Operator

Your next question comes from the line of Blayne Curtis with Barclays.

Q - Blayne Curtis {BIO 15302785 <GO>}

Hey, guys, thanks for taking my question. Just kind of curious as you look at the gaming business, I've lost track of what seasonality is, since you clearly have a big ramp ahead of you. I'm just curious as you think about Pascal versus seasonality ahead of Volta, if you can just extrapolate as you look out into April and maybe July.

A - Jen-Hsun Huang

We haven't announced anything for April or July. And so the best way to think about that is Pascal is the best gaming platform on the planet. It is the most feature-rich, the best software, the most energy-efficient. And from $99 to $1,000, you can buy the world's best GPUs, the most advanced GPUs. And you buy Pascal, you know you got the best.

Seasonality is a good question, and increasingly because gaming is a global market and because people play games every day. It's just part of their life. I don't think there's much seasonality in TV or books or music. People just whenever new titles come out, that's when a new season starts. And so in China, there are iCafes and there's Singles Day, November 11. There's back-to-school in the United States. There's Christmas. There's Chinese New Year. Boy, there are so many seasons that it's hard to imagine what the exact seasonality is anymore. And so hopefully over time, it becomes less of a matter. But the most important thing is that we expect Pascal to continue to be the world's best gaming platform for the foreseeable future.

Operator

Your next question comes from the line of Harlan Sur with JPMorgan.

Q - Harlan Sur {BIO 6539622 <GO>}

Good afternoon and congratulations on the solid results and the execution. I know somebody asked the question about inferencing for the data center markets. But on inferencing for embedded and edge applications, on the software and firmware (44:27) side, you talked about TensorRT framework. On the hardware side, you've got the Jetson TX platform for embedded and edge inferencing applications, things like drones and factory automation and transportation. What else is the team doing in the embedded markets to capture more of the SAM opportunity there going forward?

A - Jen-Hsun Huang

Thanks a lot, Harlan. The NVIDIA TensorRT is really the only optimizing inference compiler in the world today, and it targets all of our platforms. We do inference in the data center that I mentioned earlier. In the embedded world, the first embedded platform we're targeting is self-driving cars.

In order to drive the car, you basically inference or trying to predict or perceive what's around you all the time, and that's a very complicated inference matter. It could be extremely easy like taking the car in front of you and applying the brakes, or it could be incredibly hard, which is trying to figure out whether you should stop at an intersection or not. If you look at most intersections, you can't just look at the lights to determine where

do you stop. There are very few lines. And so using scene understanding and using deep learning, we have the ability to recognize where to stop and whether to stop.

And then for Jetson, we have a platform called Metropolis, and Metropolis is used for very large-scale smart cities where cameras are deployed all over to keep cities safe. We've been very successful with smart cities. Just about every major smart city provider and what is called intelligent video analysis company almost all over the world is using NVIDIA's platform to do inference at the edge, AI at the edge.

And then we've announced recently success with FANUC, the largest manufacturing and robotics company in the world, Komatsu, one of the largest construction equipment company in the world to apply AI at the edge for autonomous machines. Drones, we have several industrial drones that are inspecting pipelines, inspecting power lines, flying over large spans of farms to figure out where to spray insecticides more accurately. There are all kinds of applications. So you're absolutely right that inference at the edge or AI at the edge is a very large market opportunity for us, and that's exactly why TensorRT was created.

Operator

Your next question comes from the line of Joe Moore with Morgan Stanley.

Q - Joseph Moore {BIO 17644779 <GO>}

Great, thank you. You had mentioned how lean the channel is in terms of gaming cards. There's been an observable increase in prices at retail, and I'm just curious. Is that a broad-based phenomenon? And is there any economic ramification to you, or is that just retailers bringing prices up in a shortage environment? Thank you.

A - Jen-Hsun Huang

We don't set prices at the end of the market. And the best way for us to solve this problem is work on demand – excuse me, work on supply. The demand is great. And it's very likely that demand will remain great as we look through this quarter. And so we just have to keep working on increasing supply. Our suppliers are the world's best and the largest semiconductor manufacturers in the world. They're responding incredibly, and I am really grateful for everything they're doing. We've just got to catch up to that demand, which is just really great.

Operator

Your next question comes from the line of Chris Rolland with Susquehanna.

Q - Christopher Rolland {BIO 17980513 <GO>}

Hey, guys. Thanks for the question and great quarter. So just to clarify, Jensen, on pent-up demand, one of your GPU competitors basically said that the constraint was memory. I just want to make sure that that was correct.

And then in the CFO Commentary, you mentioned opportunities for professional viz [visualization], like AI and deep learning. Can you talk about that and what kind of applications you would use Quadro versus Volta or GeForce? Thanks.

A - Jen-Hsun Huang

Sure, we're just constrained. Obviously, we're 10 times larger of a GPU supplier than the competition. And so we have a lot more suppliers supporting us and a lot more distributors taking our products to market and a lot more partners distributing our products all over the world. And so I don't know how to explain it aside from the demand is just really great, and so we've just got to keep our nose to it and catch up to the demand.

With respect to Quadro, Quadro is a workstation processor. The entire software stack is designed for all of the applications that the workstation industry uses. And it's used – the quality of the rendering is of course world-class because of NVIDIA, but the entire software stack has been designed so that mission-critical applications or long-life industrial applications and companies that are enormous and gigantic manufacturing and industrial companies in the world could rely on an entire platform which consists of processors and system and software and middleware and all the integrations into all of the CAD tools in the world, to know that the supplier is going to be here and can be trusted for the entire life of the use of that product, which could be several years.

But the data that is generated from it has to be accountable for a couple of decades. You need to be able to pull up an entire design of a plane or a train or a car a couple decades after it was sent to production to make sure that it's still compliant, and if there are any questions about it that it can be pulled up. NVIDIA's entire platform was designed to be professional class, professional grade, long-lived.

Now the thing that's really exciting about artificial intelligence is we now can use AI to improve images. Like for example, you could fix a photograph using AI. You could fill in damaged parts of a photograph or parts of the image that hasn't been rendered yet. You want to use AI to fill in the dots, predict the future, rendering results, which we announced and which we demonstrated at GTC recently.

You could use it to generate designs. You sketch up a few strokes of what you want a car to look like. And based on the inventory, safety, physics, it could – it has learned how to fill in the rest of it, design the rest of the chassis on your behalf. It's called generative design. We're going to see generative design in product design and building design and just about everything. The last, if you will, 90% of the work is after the initial inspiration or the conceptual design is done. That part of it can be highly automated through AI. And so Quadro could be used as a platform that designs as well as generatively designs.

And then lastly, a lot of people are using our workstations to also train their neural networks for these generative designs. And so you could train and develop your own networks and then apply it in the applications.

So AI – think of AI really as, in the final analysis, the future way of developing software. It's a brand new capability where computers can write its own software, and the software that's written is so complex and so capable that no humans could write it ourselves. And so you could teach, you could use data to teach software to figure out how to write the software by itself. And then when you're done developing that software, you could use it to do all kinds of stuff, including design products. And so for workstations, that's how it's used.

Operator

Your next question comes from the line of Craig Ellis with B. Riley.

Q - Craig A. Ellis {BIO 1870408 <GO>}

Thank you for taking the question and congratulations on the very good quarterly execution. A lot of near-term items here on gaming, so I'll switch it to longer term. Jensen, at CES, I think you said that there are now 200 million GeForce users globally. And if my math is correct, then that would be up about 2x over the last three to four years. So the question is, is there anything that you can see that would preclude that kind of growth over a similar period?

And given the recent demand dynamics, I think we've seen that NVIDIA's direct channels have been very good sources for GPUs at the prices that you intend. So as we look ahead, should we expect any change in channel management from the company? Thank you.

A - Jen-Hsun Huang

Thanks a lot, Craig. In the last several years, several dynamics happened at the same time, and all of it were the favorable contributions to today. First of all, gaming became a global market and China became one of the largest gaming markets in the world.

The second, because the market became so big, developers could invest extraordinary amounts into the production value of a video game. They could invest a few hundred million dollars and know that they're going to get the return on it. Back when the video game industry was quite small or when the PC industry – PC gaming was small, it was too risky for a developer to invest that much. And so now an investor, a developer could invest hundreds of millions of dollars and create something that is just completely photorealistic and immersive and just beautiful.

And so the production, when the production value goes up, the GPU technology that's needed to run it well goes up. It's very different than music. It's very different than watching movies. Everything in video games is synthesized in real time. And so when the production value goes up, the ASP or the technology has to go up.

And then lastly, the size of the market, people have wondered how big the video game market is going to be. And I've always believed that the video game market is going to be literally everyone. In 10 years' time, 15 years' time, there's going to be another 1 billion people on earth, and those people are going to be gamers. We're going to see more and

more gamers, and not to mention that, almost every single sport could be a virtual reality sport. So video games is every sport. So e-sport can be any sport and every sport and every type of sport. And so I think when you consider this and put that in your mind, I think the opportunity for video games is going to be quite large, and that's essentially what we're seeing.

Operator

Your next question comes from the line of William Stein with SunTrust.

Q - William Stein {BIO 15106707 <GO>}

Great, thanks for taking my question and congrats on the great results and even better outlook. I'm hoping we can touch on automotive a little bit more. In particular, I think in the past, you've talked about expecting a lull in revenue growth in this market until roughly the 2020 timeframe when autonomous driving kicks-in in a more meaningful way. But of course, you have the AI copilot that seems to be potentially ramping sooner. And you have at least one marquee customer that is ramping now I guess, but volumes aren't quite that large on the autonomous driving side. So any guidance as to when we might see these two factors start to accelerate revenue in that end market? Thanks.

A - Jen-Hsun Huang

Yes, thanks a lot, Will. I wish I had more precision for you. But here are some of the dynamics that I believe in. I believe that autonomous capabilities, autonomous driving is the single greatest dynamic next to EVs in the automotive industry, and transportation is a $10 trillion industry. Between cars and shuttles and buses, delivery vehicles, it's just an extraordinary, extraordinary market. And everything that's going to move in the future will be autonomous, that's for sure. And it will be autonomous fully or it will be autonomous partly. The size of this marketplace is quite large.

In the near term, our path to that future, which I believe starts in 2020, 2019-2020, but starts very strongly in 2022, I believe the path to that in our case has several elements. The first element is that in order for all these companies, whether they're Tier 1s or startups or OEMs or taxi companies or ride-hailing companies or tractor companies or shuttle companies or pizza delivery shuttles, in order to deliver, in order to create their autonomous driving capability, the first thing you have to do is train a neural network. And we created a platform we call the NVIDIA DGX that allows everybody to train their neural networks as quickly as possible. So first is the development of the AI requires GPUs, and we benefit first from that.

The second which will start this year and next year is development platforms for the cars themselves, for the vehicles themselves. And finally, Xavier is here. We have first silicon of Xavier. It's the most complex SOC the world has ever made, and we're super-excited about the state of Xavier, and we're going to be sampling it in Q1. And so now we'll be able to help everybody create development systems. And there will be thousands and tens of thousands of quite expensive development systems based on Xavier and based on Pegasus that the world is going to need, and so that's the second element.

The third element in the near term will be development agreements. Each one of these projects are engineering-intensive, and there's a development agreement that goes along with it.

And so these three elements, these three components are in the near term and then hopefully starting from 2019 going forward and very strongly going from 2022 and beyond, the actual car revenues and economics will show up. I appreciate that question. And I think this is our last question.

We had a record quarter wrapping up a record year. We had strong momentum in our gaming, AI, data center, and self-driving car businesses. It's great to see adoption of NVIDIA's GPU computing platform increasing in so many industries. We accomplished a great deal this last year, and we have big plans for this coming year.

Next month, the brightest minds in AI and the scientific world will come together at our GPU Technology Conference in San Jose. GTC has grown tenfold in the last five years. This year we expect more than 8,000 attendees. GTC is the place to be if you're an AI researcher or doing any field of science where computing is your essential instrument. There will be over 500 hours of talks of recent breakthroughs and discoveries by leaders in the field such as Google, Amazon, Facebook, Microsoft, and many others. Developers from industries ranging from healthcare to transportation to manufacturing and entertainment will come together and share state-of-the-art in AI. This is going to be a great GTC. I hope to see all of you there.

Operator

This concludes today's conference call. You may now disconnect. Thank you for your participation.