Q1 2019 Earnings Call

Company Participants

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

Other Participants

- Atif Malik, Analyst
- Blayne Curtis, Analyst
- Chris Caso, Analyst
- Christopher Rolland, Analyst
- Craig A. Ellis, Analyst
- Joseph Moore, Analyst
- Mark Lipacis, Analyst
- Mitch Steves, Analyst
- Sajal Dogra, Analyst
- Stacy Aaron Rasgon, Analyst
- Timothy Arcuri, Analyst
- Toshiya Hari, Analyst
- Vivek Arya, Analyst
- William Stein, Analyst

MANAGEMENT DISCUSSION SECTION

Operator

Good afternoon. My name is Kelsey and I am your conference operator for today. Welcome to NVIDIA's financial results conference call. All lines have been placed on mute. 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 first quarter of fiscal 2019. 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 May 16, 2018. The webcast will be available for replay until the conference call to discuss our financial results for the second quarter of fiscal 2019. 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 the reports that we may file on Form 8-K with the Securities and Exchange Commission.

All our statements are made as of today, May 10, 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, let me turn the call over to Colette.

Colette M. Kress {BIO 18297352 <GO>}

Thanks, Simona. We had an excellent quarter with growth across all our platforms led by gaming and datacenter. Q1 revenue reached a record $3.21 billion, up 66% year-over-year, up 10% sequentially and above our outlook of $2.9 billion. Once again, all measures of profitability set records, with GAAP gross margins at 64.5%, operating margins at 40.4% and net income at $1.24 billion.

From a reporting segment perspective, Q1 GPU revenue grew 77% from last year to $2.77 billion. Tegra Processor revenue rose 33% to $442 million.

Let's start with our gaming business. Revenue was $1.72 billion, up 68% year-on-year and down 1% sequentially. Demand was strong and broad-based across regions and products. The gaming market remains robust and the popular Battle Royale genre is attracting a new wave of gamers to the GeForce platform.

We also continue to see demand from upgrades with about 35% of our installed base currently on our Pascal architecture. The launch of popular titles, like Far Cry 5 and Final Fantasy XV continued to drive excitement in the quarter.

Gamers are increasingly engaging in social gameplay and gaming is rapidly becoming a spectator sport, while the production value of games continues to increase. This dynamic

is fueling a virtuous cycle that expands the universe of gamers and drives a mix shift to higher end GPUs.

At the recent Game Developers Conference, we announced our real-time ray tracing technology, NVIDIA RTX. Ray tracing is movie quality rendering technique that delivers lifelike lighting, reflections and shadows. This has long been considered the holy grail of graphics, and we've been working on it for over 10 years.

We look forward to seeing amazing, cinematic games that take advantage of this technology come to the market later this year, with the pipeline building into next year and beyond. And we expect RTX, as well as other new technologies like 4K and virtual reality, to continue driving gamers' requirements for higher GPU performance.

While supply was tight earlier in the quarter, the situation is now easing. As a result, we were pleased to see that channel prices for our GPUs are beginning to normalize, allowing gamers who had been priced out of the market last quarter to get their hands on the new GeForce GTX at a reasonable price.

Cryptocurrency demand was again stronger than expected, but we were able to fulfill most of it with crypto-specific GPUs, which are included in our OEM business at $289 million. As a result, we could protect the vast majority of our limited gaming GPU supply for use by gamers. Looking into Q2, we expect crypto-specific revenue to be about one-third of its Q1 level.

Gaming notebooks also grew well, driven by an increasing number of thin and light notebooks based on our Max-Q design. And Nintendo Switch contributed strongly to year-on-year growth, reflecting that platform's continued success.

Moving to datacenter, we had another phenomenal quarter with revenue of $701 million, up 71% year-on-year, up 16% sequentially. Demand was strong in all market segments and customers increasingly embraced our GPUs and CUDA platform for high-performance computing and AI. Adoption of our Volta architecture remained strong across a wide range of verticals and customers.

In the public cloud segment, Microsoft Azure announced general availability of Tesla V100 instances joining Amazon, IBM and Oracle. And Google Cloud announced that the V100 is now publicly available in beta. Many other hyperscale and consumer Internet companies also continued their ramp of Volta, which delivers five times the deep learning performance of its predecessor, Pascal.

Volta has been chosen by every major cloud provider and server maker, reinforcing our leadership in AI deep learning.

In high-performance computing, strength from the broad enterprise vertical more than offset the ramp down of major supercomputing projects such as the U.S. Department of Energy's Summit system. We see a strong pipeline across a number of vertical industries,

from manufacturing to oil and gas, which should help sustain the trajectory of high-performance computing next quarter and beyond.

Traction is also increasing in AI inference. Inference GPU shipments to cloud service providers more than doubled from last quarter. And our pipeline is growing into next quarter. We dramatically increased our inference capabilities with the announcement of the TensorRT 4 AI inference accelerator software at our recent GPU Technology Conference in San Jose.

TensorRT 4 accelerates deep learning inference up to 190 times faster than CPUs for common applications, such as computer vision, neural machine translation, automatic speech recognition, speech synthesis and recommendation systems. It also dramatically expands the use cases prepared with the prior version. With TensorRT 4, NVIDIA's market reach has expanded to approximately 30 million hyperscale servers worldwide.

At GTC, we also announced other major advancements in our deep learning platform. We doubled the memory of Tesla V100 to 32 GB VRAM, which is a key enabler for customers building virtual networks for larger data sets. And we announced a new GPU interconnect fabric called NVIDIA NVSwitch, which joins up to 16 V100 GPUs at a speed of 2.4 terabytes per second or five times faster than the best PCIe switch.

We also announced our DGX-2 system, which leverages these new technologies and is updated, fully optimized software stack to deliver a 10x performance boost beyond last year's DGX. DGX-2 is the first single-server capable of delivering 2 petaflops of computational power. We are seeing strong interest from both hyperscale and enterprise customers and we look forward to bringing this technology to cloud customers later this year.

At our Investor Day in March, we updated our forecast for the datacenter addressable market. We see the datacenter opportunity as very large, fueled by growing demand for accelerated computing in applications ranging from AI to high performance computing across multiple market segments and vertical industries. We estimate the TAM at $50 billion by 2023, which extends our previous forecast of $30 billion by 2020.

We see strong momentum in the adoption of our accelerated computing platform and the expansion of our development ecosystem to serve this rapidly growing market. About 8,500 attendees registered for GTC, up 18% from last year. CUDA downloads have continued to grow, setting a fresh record in the quarter. And our total number of developers is well over 850,000, up 72% from last year.

Moving to pro visualization, revenue grew to $251 million, up 22% from a year ago and accelerating from last quarter, driven by demand for real-time rendering, as well as emerging applications like AI and VR. Strength extended across several key industries, including public sector, healthcare and retail.

Key wins in the quarter included Columbia University, using high-end Quadro GPUs for AI, and Siemens, using them for CT and ultrasound solutions. At GTC, we announced the

Quadro GV100 GPU with NVIDIA RTX technology, capable of delivering real-time ray tracing to the more than 25 million artists and designers throughout the world.

RTX makes computational intensive ray tracing possible in real time, when running professional design and content creation applications. This allows media and entertainment professionals to see and interact with their creations with correct light and shadows and do complex renders up to 10 times faster than a CPU alone.

And the NVIDIA OptiX AI denoiser built into RTX delivers almost 100 times the performance of CPUs for real-time noise-free rendering. This enables customers to replace racks of servers in traditional render farms with GPU servers at one-fifth the cost, one-seventh the space and one-seventh the power.

Lastly, automotive. Revenue grew 4% year-on-year to a record $145 million. This reflects the ongoing transition from our infotainment business to our growing autonomous vehicle development and production opportunities around the globe. At GTC and Investor Day, we made key product announcements on the advancement of autonomous vehicles and established a total addressable market opportunity of $60 billion by 2035.

We believe that every vehicle will be autonomous one day. By 2035, this will encompass 100 million autonomous passenger vehicles and 10 million robo taxis.

We also introduced NVIDIA DRIVE Constellation, a platform that will help car companies, carmakers, Tier 1 suppliers and others developing autonomous vehicle test and validate their systems in a virtual world across a wide range of scenarios before deploying on the road.

Each year, 10 trillion miles are driven around the world. Even if test cars can eventually cover millions of miles, that's an insignificant fraction of all the scenarios that require testing to create a safe and reliable autonomous vehicle.

DRIVE Constellation addresses this challenge by enabling cars to safely drive billions of miles in virtual reality.

The platform has two different servers. The first is loaded with GPUs and simulates the environment that the car is driving in, as in a hyper real video game.

The second contains the NVIDIA DRIVE Pegasus autonomous vehicle computer, which possesses the simulated data, as if it were coming from the sensors of a car driving on the road. Real-time driving command from the DRIVE Pegasus are fed back to the simulation for true hardware-in-the-loop verification.

Constellation will enable autonomous vehicle industry for safety test and validate their AI self-driving systems in ways that are not practical or possible with on-road testing.

We also extended our product roadmap to include our next-generation DRIVE autonomous vehicle computer.

We have created a scalable AI car platform that spans the entire range of autonomous driving, from traffic jams, pilots, to level 5 robo taxis. More than 370 companies and research institutions are now using NVIDIA's automotive platform. With this growing momentum, we remain excited about the intermediate and long-term opportunities for our autonomous driving business.

Now moving to the rest of the P&L, Q1 GAAP gross margins were 64.5% and non-GAAP was 64.7%, records that reflect continued growth in our value added platforms. GAAP operating expenses were $773 million. Non-GAAP operating expenses were $648 million, up 25% year-on-year.

We continue to invest in key platforms driving our long-term growth, including gaming, AI and automotive. GAAP net income was a record $1.24 billion and EPS was $1.98, up 45% (sic) [145%] (15:33) and 151% respectively from a year earlier. Some of the upside was driven by a tax rate of 5% compared to our guidance of 12%. Non-GAAP net income was $1.29 billion and EPS was $2.05, both up 141% from a year ago, reflecting the revenue strength as well as gross margins and operating margin expansion and slightly lower tax.

Our quarterly cash flow from operations reached record levels at $1.45 billion. Capital expenditures were $118 million.

With that, let me turn to the outlook for the second quarter of fiscal 2019. We expect revenue to be $3.1 billion plus or minus 2%. GAAP and non-GAAP gross margins are expected to be 63.6% (sic) [63.3%] (16:26) and 63.5%, respectively, plus or minus 50 basis points.

GAAP and non-GAAP operating expenses are expected to be approximately $810 million and $685 million, respectively. GAAP and non-GAAP OI&E are both expected to be income of approximately $15 million. GAAP and non-GAAP tax rates are both expected to be 11%, plus or minus 1%, excluding discrete items. Capital expenditures are expected to be approximately $130 million to $150 million. Further financial details are included in the CFO Commentary and other information available on our IR website.

In closing, I'd like to highlight a few upcoming events for the financial community. We'll be presenting at the JPMorgan Technology Conference next week on May 15 and at the Bank of America Global Technology Conference on June 5. We will also hold our Annual Meeting of Stockholders online on May 16.

We will now open the call for questions. Simona and I are here in Santa Clara and Jensen is dialing in from the road. Operator, would you please poll for questions? Thank you.

Q&A

Operator

Your first question is from Stacy Rasgon with Bernstein Research.

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

Hi guys, thanks for taking my questions. First, I had a question on gaming seasonality. It's usually down pretty decently in Q1. It was obviously flat this time as you were trying to fill up the channel. Now that's done. I was just wondering what the supply demand dynamics as well as like any thoughts on crypto might mean for typical – the seasonality into Q2 versus what would be typical where it would usually be down – or usually be up pretty decently. How are you looking at it? And this is a question for Colette.

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

Jensen, why don't you start on the question for Stacy and I'll follow-up afterwards after you speak.

A - Jen-Hsun Huang

Okay. Hi Stacy, so let's see. Q1, as you probably know, Fortnite and PUBG are global phenomenons (sic) [phenomena] (18:51). The success of Fortnite and PUBG are just beyond comprehension, really. Those two games are a combination of Hunger Games and Survivor has just captured imaginations of gamers all over the world. And we saw the uptick and we saw the demand on GPUs from all over the world.

Surely, there was scarcity as you know. Crypto miners bought a lot of our GPUs during the quarter and it drove prices up. And I think that a lot of the gamers weren't able to buy into the new GeForce as a result. And so we're starting to see the prices come down. We monitor spot pricing every single day around the world. And the prices are starting to normalize. It's still higher than where they should be. And so obviously, the demand is still quite strong out there.

But my sense is that there's a fair amount of pent-up demand still. Fortnite is still growing in popularity. PUBG is doing great. And then we've got some amazing titles coming out. And so my sense is that the overall gaming market is just really, is super healthy. And our job is to make sure that we work as hard as we can to get supply out into the marketplace. And hopefully, by doing that, the pricing will normalize and the gamers can buy into their favorite graphics card at a price that we hope they can get it at. And so I think there's a fair – I mean the simple answer to your question is Fortnite and PUBG. And the demand is just really great. They did a great job.

Operator

Your next question is from Joe Moore with Morgan Stanley.

Q - Joseph Moore {BIO 17644779 <GO>}

I wonder – Colette had talked about the inference doubling in sales quarter-over-quarter with cloud. Can you just talk about where you're seeing the early applications for inference? Is that sort of as-a-service business or are you looking at internal cloud workloads? And just any color you can give us on where you guys are sitting in the inference space. Thank you.

A - Jen-Hsun Huang

Sure hi Joe. So as you know, there are 30 million servers around the world. And they were put in place during the time when the world didn't have deep learning. And now with deep learning and with machine learning approaches, the accuracy of prediction, the accuracy of recommendation has jumped so much that just about every Internet service provider in the world that has a lot of different customers and consumers are jumping onto this new software approach. And in order to take this neural network – and the software that's written by deep learning in these frameworks are massive software. The way to think about these deep neural nets is it has millions and millions and millions of parameters in it and these networks are getting larger every year. And they're enormously complex. And the output of these neural nets have to be optimized for the computing platform that it targets.

How you would optimize the neural network for a CPU or a GPU is very, very different. And how you optimize for different neural networks, whether it's image recognition, speech recognition, natural language translation, recommendation systems, all of these networks have different architectures and an optimizing compiler that's necessary to make the neural network run smoothly and fast is incredibly complex.

And so that's why we created TensorRT. That's what TensorRT is. TensorRT is an optimizing graph neural network compiler. And it optimizes for each one of our platforms. And even each one of our platforms has very different architectures. For example, we invented recently – reinvented the GPU and it's called the Tensor Core GPU, and the first of its kind is called Volta. And so TensorRT 4.0 now supports, in addition to image recognition, all of the different types of neural network models.

The answer to your question is internal consumption. Internal consumption is going to be the first users. Video recognition, detecting for inappropriate video for example all over the world, making recommendations from the videos that you search or the images that you're uploading. All of these types of applications are going to require enormous amount of computation.

Operator

Next question is from 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, I have two questions about the datacenter, one, from a growth and the second from a competition perspective. So from the growth side, you guys are doing about, say, $3 billion or so annualized but you have outlined a market that

could be $50 billion. What needs to happen for the next inflection? Is it something in the market that needs to change? Is it something in the product set that needs to? How do you grow and address that $50 billion market, right, because you have only a few percent penetrated today in that large market. So what needs to change for the next inflection point?

And then on the competition side, as you are looking at that big market, how should we think about competition that is coming from some of your cloud customers, like a Google announcing a TPU 3.0 or perhaps others looking at other competing technologies? So any color on both sort of how you look at growth and competition would be very helpful. Thank you.

A - Jen-Hsun Huang

Thanks, Vivek. First of all, at its core, this is something we all know now, that CPU scaling has really slowed. And if you think about the several hundred billion dollars worth of computer equipment that's been installed in the cloud, in datacenters all over the world, and as these applications for machine learning and high-performance computing approaches come along, the world needs a solution. CPU scaling has slowed.

And so here is the approach that we pioneered a decade and a half ago called GPU computing. And we've been determined to continue to advance it during this time because we saw this day coming and we really believe that it was going to end. I mean, you can't deny physics. And so we find ourselves in a great position today.

And as Colette already mentioned, we have something close to 1 million developers on this platform now. It is incredibly fast, speeding up CPUs by 10, 20, 50, 100 times, 200 times sometimes depending on the algorithm. It's everywhere. The software ecosystem is just super rich. And as Colette mentioned, there's already almost 1 million developers around the world that's grown 70% year-over-year. And so I think at the core, it's about the fact that the world needs a computing approach going forward.

With respect to our ability to address the TAM, there are three major segments. There's more than that, but there's three major segments. One is of course training for deep learning. The other is inferencing and TRT 4 is intended to do just that to expand our ability to address all of the different types of algorithms, machine learning algorithms that are running in the datacenters.

The third is high-performance computing and that's molecular dynamics, to medical imaging, to earth sciences, to energy sciences. The type of algorithms that are being run in supercomputers all over the world is expanding. And we're doing more and more of our product designs in virtual reality. We want to simulate our products and simulate its capabilities in simulation in this computer rather than build it in the beginning. And then the last category would be graphics virtualization.

We've taken with GRID and our Quadro Virtual Workstation and now with NVIDIA RTX, we turned the datacenter into a powerful graphics supercomputer. And so these are the

various applications and segments of datacenter that we see. I think in the case of training, we're limited by the number of deep learning experts in the world.

And that's growing very quickly. The frameworks are making it easier. There's a lot more open source and open documentation on sharing of knowledge. And so the number of AI engineers around the world is growing super fast. The second is inference. And I've already talked about that. It's really limited by our optimizing compilers and how we can target these neural network models to run our processors. And if we could do so, we're going to save our customers enormous amounts of money.

We speed up applications. We speed up these neural network models 50 times, 100 times, 200 times over a CPU. And so the more GPUs they buy, the more they're going to save. And high-performance computing, the way to think about that is, I think, at this point, it's very clear that going forward, supercomputers are going to get built with accelerators in them. And because of our long-term dedication to CUDA and our GPUs for acceleration of all these codes and the nurturing of the ecosystem, I think that we're going to do super well in the supercomputing world. And so these are the different verticals.

With respect to competition, it all starts with the core. And the core is that the CPU scaling has slowed. And so the world needs another approach going forward. And surely because of our focus on it, we find ourselves in a great position. Google announced TPU 3.0 and it's still behind our Tensor Core GPU. Our Volta is our first generation of a newly reinvented approach of doing GPUs. It's called Tensor Core GPUs. And we're far ahead of the competition and – but more than that, it's programmable. It's not one function. It's programmable.

Not only is it faster, it's also more flexible. And as a result of the flexibility, developers could use it in all kinds of applications, whether it's medical imaging or weather simulations or deep learning or computer graphics. And as a result, our GPUs are available in every cloud and every datacenter, everywhere on the planet and which developers need so that – accessibility, so that they could develop their software.

And so I think that on the one hand, it's too simplistic to compare a TPU to just one of the many features that's in our Tensor Core GPU. But even if you did, we're faster. We support more frameworks. We support all neural networks.

And as a result, if you look at GitHub, there are some 60,000 different neural network research papers that are posted that runs on NVIDIA GPUs. And it's just a handful for the second alternative. And so I'd just kind of gave you a sense of the reach and the capabilities of our GPUs.

Operator

Your next question comes from Toshiya Hari with Goldman Sachs.

Q - Toshiya Hari {BIO 6770302 <GO>}

Great. Thank you so much. Jensen, I had a question regarding your decision to pull the plug on your GeForce Partner Program. I think most of us read your blog from last Friday. I think it was, so we understand the basic background. But if you can describe what led to this decision and perhaps talk a little bit about the potential implications, if any, in terms of your ability to compete or gain share. That will be really helpful. Thank you so much.

A - Jen-Hsun Huang

Yeah. Thanks for your question, Toshiya. At the core, the program was about making sure that gamers who buy graphics cards knows exactly the GPU brand that's inside. And the reason for that is because, we want gamers to – the gaming experience of a graphics card depends so much on the GPU that is chosen.

And we felt that using one gaming brand, a graphics card brand, and interchanging the GPU underneath causes it to be less – causes it to be more opaque and less transparent for gamers to choose the GPU brand that they wanted. And most of the ecosystem loved it. And some of the people really disliked it.

And so instead of all that distraction, we're doing so well. And we're going to continue to help the gamers choose the graphics cards, like we always have, and things will sort out. And so we decided to pull the plug because the distraction was unnecessary and we have too much good stuff to go do.

Operator

Next question is from C.J. Muse with Evercore ISI.

Q - Sajal Dogra {BIO 19959997 <GO>}

Hi. This is Sajal Dogra calling in for C.J. Muse. Thank you for taking my question. So I had a question on HPC. TSMC, on their recent call, raised their accelerator attach rate forecast in HPC to 50% from mid-teens. So I would love to get further details on what exactly NVIDIA is doing to software services, et cetera, that's kind of creating this competitive positioning in HPC and AI basically. And then, if I could ask a follow-up basically in benchmarks. So, there's been some news on AI benchmarks whether it's Stanford DAWNBench et cetera. So I would love to get your thoughts on A, the current state of benchmarks for AI workloads and B, your relative positioning of ASICs versus GPUs especially as we move towards newer neural networks like RN and GAN, et cetera. Thank you.

A - Jen-Hsun Huang

Yeah, thanks for the question. Well, HPC. First of all, at the core, CPU scaling has stalled and it's reached the limits of physics. And the world needs another approach to go forward. We created the GPU computing approach a decade and half ago. And I think at this point, with the number developers jumping on, the number of applications that's emerging, it's really clear that the future of HPC has accelerated. And our GPU approach, because of its flexibility, because of its performance, because of the value that we create

that as a result of the throughput of a datacenter, we save people so much money just in cables alone, often times, more than pays for the GPUs that they buy. And the reason for that is because the number of servers reduced dramatically. And so I think the future of HPC is about acceleration and the NVIDIA CUDA GPUs are really in a great position to serve this vacuum that's been created.

With respect to benchmarks, you might have seen that earlier this week, we released three speed records: the fastest single GPU, the fastest single computer node, a definition of a computer node is something that fits in a box that runs one operating system, one node; and one instance, one cloud instance. We now have the fastest speed record for one GPU, one node, and one instance.

And so we love benchmarks. Nothing is more joyful than having a benchmark to demonstrate your leadership position. And in the world of deep learning, the number of networks is just growing so fast, because the number of different applications that deep learning is able to solve is really huge. And so you need a lot of software capability and the versatility of your platform needs to be great.

We also have a lot of expertise in the company in software. I mean, NVIDIA is really a full stack computing company, from architecture, to system software, to algorithms, to applications. We have a great deal of expertise across the entire stack. And so we love these complicated benchmarking – benchmarks that are out there. And I think this is a great way for us to simplify our leadership position.

I think long-term, the number of networks that are going to emerge will continue to grow. And so the flexibility of ASICs is going to be its greatest downfall. And if someone were to create a general-purpose parallel accelerating processor like ours and had it designed to be incredibly good at deep learning, like recently what we did with our Tensor Core GPU, which is a reinvented GPU, and Volta is the first one, it's going to be hard. It's going to be expensive. And we've been doing it a very long time. And so I think this is a – it's a great time for us.

Operator

Next question is from Blayne Curtis with Barclays.

Q - Blayne Curtis {BIO 15302785 <GO>}

Thanks for taking my question. Jensen, maybe – I wanted to ask on inference side about edge inference and beyond autos when you look at sizing that TAM, what are the other big areas that you think you can penetrate with GPUs in edge inference besides autos?

A - Jen-Hsun Huang

Yeah, Blayne. The largest inference opportunity for us is actually in the cloud and the datacenter. That's the first great opportunity. And the reason for that is there's just an explosion in the number of different types of neural networks that are available. They're

image recognition, there's video sequencing, there's video, there's recommender systems, there's speech recognition, speech synthesis, natural language understanding.

There are just so many different types of neural networks that are being created. And creating one ASIC that can be adapted to all of these different types of networks is just a real challenge. And by the time that you create such a thing, it's called a Tensor Core GPU, which is what we created. And so I think that the first opportunity for us in large-scale opportunity will be in the datacenter and the cloud.

The second will be in vertical markets. The vertical market that you mentioned is self-driving cars. And we see a great opportunity in autonomous vehicles, both in the creation of autonomous vehicles. And I mentioned that before, between now and the time that we ramp our AV computers we call DRIVE, we're going to be selling a whole lot of servers, so that the companies could develop their neural network models for their self-driving cars, as well as simulating in virtual reality their various test drives, as well as testing their neural network and their self-driving car stack against billions and billions of miles of saved up pre-recorded videos.

And so in the vertical markets, we're going to see inference both in the datacenter for developing the self-driving car stack as well as in the self-driving cars themselves. Now, in the self-driving cars, the ASPs for Level 2 could be a few hundred dollars to a Level 5 self-driving car, taxi or driverless taxi being a few thousand dollars.

And I expect that driverless taxis will start going to market about 2019 and self-driving cars probably somewhere between 2020 and 2021.

And I think the size of the market is fairly well modeled. And the simple way to think about that is I believe that every single – everything that moves someday will be autonomous or have autonomous capabilities.

And so the 100 million cars, the countless taxis, all the trucks, all the agriculture equipment, all the pizza delivery vehicles, you name it. Everything is going to be autonomous. And the market opportunity is going to be quite large. And that's the reason why we're so determined to go create that market.

Operator

Your next question is from Tim Arcuri with UBS.

Q - Timothy Arcuri {BIO 3824613 <GO>}

Thank you. I actually wanted to go back to the question about seasonality for gaming in June. Normal seasonal sounds like it's up mid-teens for June in gaming. But obviously, the comps are skewed a little bit because of the channel restock and the crypto stuff. So does the guidance for June assume that gaming is better or worse than that mid-teens normal seasonal? Thank you.

A - Jen-Hsun Huang

We're expecting Q2 to be better than seasonality, if I understand your question. We're expecting Q2 to be better than Q1. And we're expecting Q2 to be better than seasonality. Did that answer your question?

Operator

Your next question is from Atif Malik with Citi.

Q - Atif Malik {BIO 15866921 <GO>}

Hi. Thanks for taking my question and good job on the results. And I'd have a question for Colette. Colette, first thank you for breaking out crypto sales in the OEM line and guide for us. I have a question on your gross margins. Your gross margins have been expanding on product mix, despite component pricing headwinds on the DRAM side. When do you expect component pricing to become a tailwind to your gross margins?

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

Thanks so much for the question. When you think about our gross margins, just over this last quarter, as you know, we were working on stabilizing the overall supply that was out there in the market for consumer GPUs.

We benefited from that with a higher gross margin as we filled and completed that. You've seen us absorb a significant amount of the component pricing changes that we have seen, particularly around the memory. We're not here to be able to forecast generally when those pricing of those components will stabilize.

But we believe in terms of the value added that our platforms provide, the components are an important part of finishing that. But I think we have tremendous amount more value that we are adding in terms of the software on top of our platforms, which is enabling our gross margins.

Operator

Your next question is from Chris Caso with Raymond James.

Q - Chris Caso {BIO 4815032 <GO>}

Yes, hi. Thanks for taking the question. My question is the progress on the deployment of Volta into the cloud service providers. You talked in your prepared remarks about the five deployments, including the Google beta. Can you talk about how soon we can expect to see some of those remaining deployments? And of those already launched, how far are they along? I guess, to say proverbially, what inning are we in, in these deployments?

A - Jen-Hsun Huang

Yes. So first of all, Volta is a reinvented GPU. Volta is the world's first GPU that has been designed to be incredibly good at deep learning. We call it the Tensor Core GPU.

It still retained all of the flexibilities of all – everything that CUDA has ever run is backwards compatible, with everything that runs on CUDA. But it has new architectures designed to be incredibly good at deep learning. We call it a Tensor Core GPU. And that's the reason why it has all of the benefits of our GPU but none of the ASICs can catch up to it. And so Volta is really a breakthrough.

We're going to be very successful with Volta. Every cloud will have it. The initial deployment is for internal consumption. Volta has been shipping to the cloud providers, the Internet service companies for the vast majority of last quarter, as you guys know. And they're using it internally. And now they're starting to open up Volta for external consumption of their cloud customers. And they are moving as fast as they can. My expectation is that you're going to see a lot more coming online this quarter.

Operator

Your next question comes from Mark Lipacis with Jefferies.

Q - Mark Lipacis {BIO 2380059 <GO>}

Hi, thanks for taking my question. I had a question about the DGX family of products. Our own fieldwork is indicating very positive reception for DGX. And I was wondering, Jensen, if can you help us understand, the high-growth we've seen in the datacenter business, to what extent is that being driven by the DGX. And when DGX-2 starts to ramp in the back half of the year, is this something that kind of layers on top of DGX – does DGX-2 layer on top of DGX, are they going after different segments and you're kind of segmenting the market with these two different products? Any color on how to think about those two products would be helpful. Thank you

A - Jen-Hsun Huang

Hey Colette, could you give me a brief version of that? It was kind of crackling on my side.

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

So, I'm going to ask the operator if they could ask for the question again because it was also on our side a little crackly.

Operator

Yes, Mark your line is open. Please restate your question.

Q - Mark Lipacis {BIO 2380059 <GO>}

Okay, thanks. Can you hear me better now?

A - Jen-Hsun Huang

Yeah, much better, Mark.

Q - Mark Lipacis {BIO 2380059 <GO>}

Okay, sorry about that, I'm at the airport. So the question was on the DGX family of products. Our own fieldwork indicates a very positive reception. I was wondering Jensen, if you could help us understand the high-growth you've seen in the datacenter market how much is DGX contributing to that. And then when DGX-2 starts to ramp in the second half of the year, how do we think about DGX-1? Does it replace the DGX – the original DGX, or going after different segments, or do they layer on top of one another? Any color on that would be helpful. Thank you.

A - Jen-Hsun Huang

I see. Thank you. DGX-2 and DGX-1 will both be in the market at the same time. And DGX is a few hundred million dollar business. It was introduced last year. So its growth rate is obviously very high. It's designed for enterprises where they don't – they need to have their computers on-premise, but they don't want to build a supercomputer. And they don't have the expertise to do so.

And they would like to pull a supercomputer out of a box, plug it in and start doing supercomputing. And so DGX is really designed for enterprises. It's designed for car companies; it's designed for healthcare companies doing life sciences work or medical imaging work. We recently announced a project called Project Clara, which basically takes medical imaging equipment, virtualizes them, containerizes the software and turns it into a – and most medical imaging equipment today are computational and they – a lot of them run on NVIDIA CUDA anyways.

We can put that into the datacenter, we can virtualize their medical instruments and it gives them the opportunity to upgrade the millions of instruments that are out in the marketplace today. And so DGX is really designed for enterprises and we're seeing great success there. It's really super easy to use and it comes with direct support from HPC and AI researchers at NVIDIA. And the answer to your question at the end is both of them will be in the marketplace at the same time.

Operator

Next question is from Mitch Steves with RBC Capital Markets.

Q - Mitch Steves {BIO 19155169 <GO>}

Hey guys. I'm actually going to go to a more nitty-gritty question just on the financial side, just to make sure I'm understanding this right. So the OEM beat was pretty material given a lot of crypto revenue. Is it still the case that OEM is materially lower gross margin than your corporate average at this time?

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

Sure, I'll take that question. Generally, our OEM business can be a little bit volatile. Because remember, OEM business incorporates our mainstream GPUs as well as our Tegra integrated.

So we have development platforms that we sell on some of the Tegra piece of it. But they are slightly below and I think you can go back and refer to our discussion at Investor Day as there's a slide there that talks about those embedded pieces and them being below. So yes, you're correct. Again, a very small part of our business right now.

Operator

Your next question comes from Christopher Rolland with Susquehanna.

Q - Christopher Rolland {BIO 17980513 <GO>}

Hey, guys, thanks for the question. So your competitor thinks that just 10% of their sales were from crypto or like $150 million, $160 million. And you guys did almost $300 million there. And perhaps I think there could actually be some in gaming as well, which would imply that you guys have two-thirds or more of that market?

So I guess what's going on there? Is there a pricing dynamic that's allowing you to have such share there, or do you think it's your competitors that don't know what's actually being sold to miners versus gamers? Why such implied share in that market? Thanks.

A - Jen-Hsun Huang

Well, we try to as transparently review our numbers as best we can. Our strategy is to create a SKU that allows the crypto miners to fulfill their needs and we call it CMP. And to be – as much as possible, fulfill their demand that way.

Sometimes, it's just not possible because the demand is too great but we try to do so. And we try to keep the miners on the CMP SKUs as much as we can. And so I'm not exactly sure how other people do it, but that's the way we do it.

Operator

Your next question is from Craig Ellis with B. Riley.

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

Thanks for sneaking me in and congratulations on all the financial records in the quarter. Jensen, I just wanted to come back to an announcement that you made at GTC with ray tracing. Because the technology looked like it was very high fidelity and I think you noted at that time that it was very computationally intensive. So the question is as we think about the gaming business and the potential for ray tracing to enter that platform group, what does it mean for dynamics that we've seen in the past, for example, the ability to really push the high end of the market with high end capability, 1070 Ti launched late last year, it was very successful. Does this give you further flexibility for those types of launches as you bring exciting and very high end technology to market? Thank you.

A - Jen-Hsun Huang

Yeah, I appreciate it. NVIDIA RTX is the biggest computer graphics invention in the last 15 years. It took us a decade to do. We literally worked on it continuously for one decade. And to put it into perspective, it's basically film rendering, cinematic rendering except it's in real time.

It merges the style of computer graphics, rasterization, and light simulation, what people call ray tracing as well as deep learning and AI, merged it into one unified framework, so that we can achieve cinematic rendering in real time.

What it currently takes is a server about a few hours, depending on the scene, it might take as long as a full day, take a few hours to render one frame.

So it takes a server, one node of a server, several hours to render one frame. And in order to render 30 frames per second, just imagine the number of servers you need. If you take several hours per frame and you need to render 30 frames per second in order to be real-time, it basically takes a high-performance computer, a supercomputer, a render farm, that's why they call it a render farm, it's a full datacenter designed just for rendering.

And now we've created NVIDIA RTX which makes it possible to do in real time. We demonstrated RTX on four Quadro GV100s. It takes four of our latest generation Volta Tensor Core GPUs to be able to render 30 frames per second, the Star Wars cinematic that people enjoyed. And so the amount that we saved, we basically took an entire datacenter and reduced it into one node. And we're now doing it in real time.

And so the amount of money that we can save, people who create movies, people who do commercials, people who use film rendering to create the game content, almost every single game is done that way. There's quite a bit of offline rendering to create the imagery and the textures and the lighting. And then there are of course, architectural design and car design, the number of applications, the number of industries that are built on top of modern computer graphics is really quite large. And I'm certain that NVIDIA RTX is going to impact every single one of them.

And so that's our starting point, is to dramatically reduce the cost of film rendering, dramatically reduce the time that it takes to do it and hopefully, more GPU servers will be purchased. And of course, better content will be created. Long-term, we've also now plotted the path towards doing it in real time. And someday, we will be able to put RTX into a GeForce gaming card and the transformation to the revolution to the gaming industry will be quite extraordinary. So we're super excited about RTX.

Operator

Your next question is from Stacy Rasgon with Bernstein.

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

Hi, guys. Thanks for fitting me in for my follow-up. This is a question for Colette. I want to follow-up again on the seasonality. Understanding the prior comments, normal seasonal

for Q2 for gaming would be up in the double digits. Given your commentary on the crypto declining in Q2, given your commentary on just the general drivers around datacenter and the Volta ramp, I can't bring that together with the idea of gaming being above seasonal within the context of your guidance envelope. So how should I reconcile those things? How are you actually thinking about seasonality for gaming into Q2 within the context of the scenarios that are currently contemplated in your guidance for next quarter?

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

Sure, Stacy. Let me see if I can bridge together, Jensen, and then some comments here. Unfortunately, they're moving quite fast to the next question, so I wasn't able to add-on.

But let me see if I can add-on here and provide a little bit of clarity in terms of the seasonality. Remember in Q1, we outgrew seasonality significantly. We left Q4 with very low inventory in terms of in the channel. We spent Q1 working on establishing a decent amount of inventory available.

We wanted to concentrate on our miners separately. And then you can see we did that in terms of Q1 by moving that to OEM and moving that to cryptocurrency only boards. So we left Q1 at this point with healthy overall channel inventory levels as far as where we stand.

So that then takes you now to Q2. But if we overshot in terms of seasonality in terms of Q1, we don't have to do those channel fill dynamics again as we get into Q2. But we do have demand out there for our gamers that we can now address very carefully with the overall inventory that we now have available.

So putting together, Q1 and Q2 together, yes, we are within normal seasonality, again, for a guidance. And we'll see how we'll finish in terms of the quarter. But you should be in that range.

So, yes, from a normal seasonality, at a year-to-date inclusive of Q2, yes, we're on that overall seasonality. Always keep in mind, generally, our H2s are usually higher than our overall H1s, and that's what you should think about our overall guidance.

Gaming is still strong. We have to comment that our overall drivers that have taken us to this place over the last three to five years with phenomenal growth and our ability to grow that overall market is still here and all of those things are together. We just had a few quarters in terms of making sure that we get the overall channel correct and put our miners separately.

I hope that clarifies in terms of where we are, in terms of gaming seasonality.

Operator

Your last question comes from Will Stein with SunTrust.

Q - William Stein {BIO 15106707 <GO>}

Hi. Great. Thank you for taking my question and squeezing me in. The question relates to the supply chain challenges that you talked so much about in the gaming end market.

I'm wondering if there's something particular to that end market that is making the shortages concentrated there, or are in fact other end markets in particular, the datacenter end market, also somewhat restricted from what growth they might have achieved if there weren't the shortages that are out there. And maybe talk about the pace of recovery of those. That'd be really helpful. Thanks so much.

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

Let me start off here, and I'll have Jensen finish up on the last part of that question. But overall, our datacenter business did phenomenal. Volta is doing extremely well. And even now with 32-bit, we're seeing tremendous adoption throughout.

Again, remember it's very different than the overall consumer business. You have significant amount of time for qualification and that is moving extremely fast, based on a lot of other industries and their ability to qualify. So no, there is not a supply challenge at all in terms of our datacenter. And our overall growth in datacenter, we're extremely pleased with in terms of how the quarter came out.

I'll turn it over to you, Jensen, and you can answer the rest of the part of it.

A - Jen-Hsun Huang

Yeah. The reason why miners love GeForce is because miners are everywhere in the world. One of the benefits of cryptocurrency is that it's not any sovereign currency. And it's in the digital world, it's distributed. And GeForce is the single largest distributed supercomputing infrastructure on the planet.

Every gamer has a supercomputer in their PC. And GeForce is so broadly distributed, it's available everywhere. And so GeForce is really a good candidate for any new cryptocurrency or any new cryptography algorithm that comes along.

We try the best we can to go directly to the major miners. And they represent the vast majority of the demand. And to the best of our ability, serve their needs directly and we call that C&P (1:02:21) and that's why it's not called GeForce, they're called C&P (1:02:25). And we can serve those miners directly, hopefully to take some of the demand pressure off of the GeForce market.

Because ultimately, what we would like is we would like the market for GeForce pricing to come down, so that the gamers could benefit from the GeForces that we built for them.

And the gaming demand is strong. I mean, the bottom line is Fortnite is a homerun. The bottom line is PUBG is a homerun. And the number of gamers that are enjoying these games is really astronomic as people know very well.

And it's a global phenomenon. These two games are equally fun in Asia as it is in Europe as it is in the United States. And because you team up and this is a Battle Royale, you'd rather play with your friends. So it's incredibly social. It's incredibly sticky.

And more and more – more gamers that play, more of their friends join, and more of their friends join, more gamers that play. And so it's this positive feedback system, and the guys at Epic did a fantastic job creating Fortnite. And it's just a wonderful game genre that people are really enjoying.

And so I think at the core of it, gaming is strong and we are looking forward to inventory normalizing in the channel so that pricing could normalize in the channel, so that gamers can come back to buy the GeForce cards that has now been in short supply for over a quarter. And so the pent-up demand is quite significant. And I'm expecting the gamers to be able to buy new GeForces pretty soon.

Operator

Unfortunately, we ran out of time. I will now turn it back over to Jensen for any closing remarks.

A - Jen-Hsun Huang

Let's see here. Is it my turn again?

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

Yes, it is.

A - Jen-Hsun Huang

Okay. We had another great quarter, record revenue, record margins, record earnings, growth across every platform. Datacenter achieved another record with strong demand for Volta and AI inference. Gaming was strong. We are delighted to see prices normalizing and we can better serve pent-up gamer demand.

At the heart of our opportunity is the incredible growth of computing demand of AI just as traditional computing has slowed. The GPU computing approach that we've pioneered is ideal for filling this vacuum. And our invention of the Tensor Core GPU has further enhanced our strong position to power the AI era. I look forward to giving you another update next quarter. Thank you.

Operator

This concludes today's conference call. Thank you, guys, for joining. You may now disconnect.