Q2 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

- Aaron Rakers, Analyst
- Atif Malik, Analyst
- Blayne Curtis, Analyst
- C. J. Muse, Analyst
- Harlan Sur, Analyst
- Joseph Moore, Analyst
- Mark Lipacis, Analyst
- Matthew D. Ramsay, Managing Director & Senior Research Analyst
- Timothy Arcuri, Analyst
- Toshiya Hari, Analyst
- Vivek Arya, Analyst

MANAGEMENT DISCUSSION SECTION

Operator

Good afternoon. My name is Kelsey and I'm 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 Second 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 August 23,

2018. The webcast will be available for replay until the conference call to discuss our financial results for the third 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 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, August 16, 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. This is a big week for NVIDIA. We just announced the biggest leap in GPU architecture in over a decade. We can't wait to tell you more about it, but first, let's talk about the quarter.

We had another strong quarter, led by datacenter and gaming. Q2 revenue reached $3.12 billion, up 40% from a year earlier. Each market platform; gaming, datacenter, pro visualization and automotive hit record levels with strong growth both sequentially and year-on-year. These platforms collectively grew more than 50% year-on-year.

Our revenue outlook had anticipated cryptocurrency-specific products declining to approximately $100 million, while actual crypto-specific product revenue was $18 million, and we now expect a negligible contribution going forward.

Gross margins grew nearly 500 basis points year-on-year, while both GAAP and non-GAAP net income exceeded $1 billion for the third consecutive quarter. Profit nearly doubled. From a reporting segment perspective; GPU revenue grew 40% from last year to $2.66 billion; Tegra Processor revenue grew 40% to $467 million.

Let's start with our gaming business. Revenue of $1.8 billion was up 52% year-on-year and up 5% sequentially. Growth was driven by all segments of the business, with desktop, notebook and gaming consoles up all strong double digit percentages year-on-year. Notebooks were a standout this quarter with strong demands for thin and light form factors based on our Max-Q technology.

Max-Q enabled gaming PC OEMs to pack a high-performance GPU into a slim notebook that is just 20 millimeters thick or less. All major notebook OEMs and ODMs have adopted Max-Q for their top of the line gaming notebooks, just in time for back to school. And we expect to see 26 models based on Max-Q in stores for the holidays.

The gaming industry remains vibrant. The eSports audience now approaches 400 million, up 18% over the past year. The unprecedented success of Fortnite and PUBG has popularized this new battle royale genre and expanded the gaming market. In fact, the battle royale mode is coming to games like the much anticipated Battlefield 5. We are thrilled to partner with EA to make GeForce the best PC gaming platform for the release of Battlefield 5 in October.

We have also partnered with Square Enix to make GeForce the best platform for its upcoming Shadow of the Tomb Raider. Monster Hunter: World arrived on PCs earlier this month and it was an instant hit. And many more titles are lined up for what promises to be a big holiday season.

It's not just new titles that are building anticipation. The gaming community is excited over the Turing architecture announced earlier this week at SIGGRAPH. Turing is our most important innovation since the invention of the CUDA GPU over a decade ago. The architecture includes new dedicated ray tracing processors, or RT Cores, and new Tensor Cores for AI inferencing, which together will make real-time ray tracing possible for the first time. We will enable the cinematic-quality gaming, amazing new effects powered by neural networks and fluid interactivity on highly complex models.

Turing will reset the look of video games and open up the $250 billion visual effects industries to GPUs. Turing is the result of more than 10,000 engineering years of effort. It delivers up to 6x performance increase over Pascal for ray-traced graphics and up to 10x boost for peak inference FLOPs. This new architecture will be the foundation of a new portfolio of products across our platforms going forward.

Moving to datacenter, we had another strong quarter with revenue of $760 million accelerating to 83% year-on-year growth and up 8% sequentially. This performance was driven by hyperscale demand, as Internet services used daily by billions of people increasingly leverage AI.

Our GPUs power real-time services such as search, voice recognition, voice synthesis, translation, recommender engines, fraud detection, and retail applications. We also saw a growing adoption of our AI and high performance computing solutions by vertical industries, representing one of the most fastest areas of growth in our business. Companies in sectors ranging from oil and gas to financial services through transportation are harnessing the power of AI and are accelerating computing platform to turn data into actionable insights.

Our flagship Tensor Core GPU, the Tesla V100, based on Volta architecture, continue to ramp for both AI and high-performance computing applications. Volta has been adopted by every major cloud provider and hyperscale datacenter operator around the world.

Customers have quickly moved to qualify the new version of V100, which doubled the on-chip DRAM to 32-gig to support much larger data sets and neural networks. Major server OEMs, HP Enterprise, IBM, Lenovo, Cray and Super Micro also brought the V100 32-gig to market in the quarter.

We continue to gain traction with AI inference solution, which helped expand our addressable market in the datacenter. During the quarter, we released our TensorRT 4 AI inference accelerator software for general availability. While prior versions of the TensorRT optimized image and video-related workloads, TensorRT 4 expands the aperture to include more use cases, such as speech recognition, speech synthesis, translation and recommendation systems. This means we can now address a much larger portion of deep learning inference workloads delivering up to 190x performance speed-up relative to CPUs.

NVIDIA and Google engineers have integrated TensorRT into the TensorFlow deep learning framework, making it easier to run AI inference on our GPUs. And Google Cloud announced that NVIDIA Tesla P4 GPU, our small form factor GPU for AI inference and graphic virtualization, is available on Google Cloud Platform.

Datacenter growth was also driven by DGX, our fully optimized AI server, which incorporates V100 GPUs, our proprietary high-speed interconnect and our fully optimized software stack. The annual run rate for DGX is in the hundreds of millions of dollars. DGX-2, announced in March at our GPU Technology Conference, is being qualified by customers and is on track to ramp in the third quarter.

At GTC Taiwan in June, we announced that we are bringing DGX-2 technology to our HGX-2 server platform. We make HGX-2 available to OEM and ODM partners, so they can quickly deploy our newest innovations in their own server designs.

In recent weeks, we announced partnerships with NetApp and Pure Storage to help customers speed AI deployment from months to days or even hours with highly integrated optimized solutions that combine DGX with the company's all flash storage offerings and third-party networking.

At GTC Taiwan, we also revealed that we set five speed records for AI training and inference. Key to our strategy is our software stack, from CUDA to our training and inference SDKs as well as our work with developers to accelerate their applications, it is the reason we can achieve such dramatic performance gains in such a short period of time. And our developer ecosystem is getting stronger.

In fact, we just passed 1 million members in our Developer Program, up 70% from one year ago. One of our proudest moments this quarter was the launch of the Summit AI Supercomputer in Oak Ridge National Laboratory. Summit is powered by over 27,000 Volta Tensor Core GPUs and helped the U.S. reclaim the number one spot on the TOP500 Supercomputer list for the first time in five years.

Other NVIDIA power systems joined the TOP500 list were Sierra at Lawrence Livermore National Laboratory in the third spot, and the ABCI Japan's fastest supercomputer in the fifth spot. NVIDIA now powers five of the world's seven fastest supercomputers reflecting the broad shift in supercomputing to GPUs.

Indeed, the majority of the computing performance added to the latest TOP500 list comes from NVIDIA GPUs and more than 550 HPC applications are now GPU accelerated. With our Tensor Core GPUs, supercomputers can now combine simulation with the power of AI to advance many scientific applications from molecular dynamics, to seismic processing, to genomics and materials science.

Moving to pro visualization. Revenue grew to $281 million, up 20% year-on-year and 12% sequentially, driven by demand for real-time rendering and mobile workstations, as well as emerging applications like AI and VR. These are emerging applications now represent approximately 35% of pro visualization sales. Strength extended across several key industries, including healthcare, oil and gas, and media and entertainment. Key wins in the quarter include Raytheon, Lockheed, GE, Siemens and Philips Healthcare.

In announcing the Turing architecture at SIGGRAPH, we also introduced the first Turing-based processors, the Quadro RTX 8000, 6000 and 5000 GPUs, bringing interactive ray tracing to the world years before it has been predicted. We also announced that the NVIDIA RTX Server, a full ray tracing global illumination rendering server that will give a giant boost to the world's render farms as Moore's Law ends.

Turing is set to revolutionize the work of 50 million designers and artists, enabling them to render photorealistic scenes in real-time and add new AI-based capabilities to their work flows. Proto GPUs (13:51) based on the Turing will be available in the fourth quarter.

Dozens of leading software providers, developers and OEMs have already expressed support for Turing. Our ProViz partners view it as a game-changer for professionals in the media and entertainment, architecture and manufacturing industries.

Finally turning to automotive; revenue was a record $161 million, up 13% year-on-year and up 11% sequentially. This reflects growth in our autonomous vehicle production and development engagements around the globe as well as the ramp of next-generation AI-based smart cockpit infotainment solutions. We continue to make progress on our autonomous vehicle platform with key milestones and partnerships announced this quarter.

In July, Daimler and Bosch selected DRIVE Pegasus as the AI brand for their Level 4 and Level 5 autonomous fleets. Pilot testing will begin next year in Silicon Valley. This collaboration brings together NVIDIA's leadership in AI and self-driving platforms, Bosch's hardware and systems expertise as the world's largest Tier 1 automotive supplier and Daimler's vehicle expertise and global brand synonymous with safety and quality.

This quarter, we started shipping development systems for DRIVE Pegasus, an AI supercomputer designed specifically for autonomous vehicles. Pegasus delivers 320 trillion operations per second to handle diverse and redundant algorithms and is architected for safety as well as performance. This automotive grade, functionally safe production solution, uses two NVIDIA Xavier SoCs and two next-generation GPUs designed for AI and visual processing, delivering more than 10x greater performance and 10x higher data bandwidth compared to the previous generation. With co-designed hardware and software, the platform is created to achieve ASIL-D ISO 26262 the industry's highest level of automotive functional safety.

We have created a scalable AI car platform that spans the entire range of automated and autonomous driving from traffic jam pilots to Level 5 robotaxis. More than 370 companies and research institutions are using NVIDIA's automotive platform. With this growing momentum and accelerating revenue growth, we remain excited about the intermediate and long-term opportunities for autonomous driving business.

This quarter, we also introduced our Xavier platform for Jetson for the autonomous machine market. With more than 9 billion transistors, it delivers over 30 trillion operations per seconds, more processing capability than a powerful workstation, while using one-third the energy of a lightbulb. Jetson Xavier establishes customers to deliver AI computing at the edge, powering autonomous machines like robots or drones with applications in manufacturing, logistics, retail, agricultural, healthcare and more.

Lastly, in our OEM segment, revenue declined by 54% year-on-year and 70% sequentially. This was primarily driven by the sharp decline of cryptocurrency revenues to fairly minimal levels.

Moving to the rest of the P&L. Q2 GAAP gross margin was 63.3% and non-GAAP was 63.5%, in line with our outlook. GAAP operating expenses were $818 million. Non-GAAP operating expenses were $692 million, up 30% year-on-year. We can continue to invest in the key platforms driving our long-term growth, including gaming, AI and automotive.

GAAP net income was $1.1 billion and EPS was $1.76, up 89% and 91% respectively, from a year earlier. Some of the upside was driven by a tax rate near 7% compared to our outlook of 11%. Non-GAAP net income was $1.21 billion and EPS was $1.94, up 90% and 92%, respectively, from a year ago, reflecting revenue strength as well as gross and operating margin expansion and lower taxes. Quarterly cash flow from operations was $913 million. Capital expenditures were $128 million.

With that, let me turn to the outlook for the third quarter of fiscal 2019. We're including no contributions from crypto in our outlook. We expect revenue to be $3.25 billion, plus or minus 2%. GAAP and non-GAAP gross margins are expected to be 62.6% and 62.8%, respectively, plus or minus 50 basis points.

GAAP and non-GAAP operating expenses are expected to be approximately $870 million and $730 million, respectively. GAAP and non-GAAP OI&E are both expected to be income of $20 million. GAAP and non-GAAP tax rates are both expected to be 9%, plus or minus 1% excluding discrete items. Capital expenditures are expected to be approximately $125 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 some of the upcoming events for the financial community. We'll be presenting at the Citi Global Technology Conference on September 6 and meeting with the financial community at our GPU Technology Conferences in Tokyo on September 13 and Munich on October 10. And our next earnings calls to discuss our financial results is in the third quarter of 2019 will take place on November 15.

We will now open the call for questions. If you could limit your questions to one or two and operator, would you please poll for questions? Thank you.

Q&A

Operator

Your first question comes from Mark Lipacis with Jefferies.

Q - Mark Lipacis {BIO 2380059 <GO>}

Hi. Thanks for taking my question. The question is on ray tracing, to what extent is this creating new market's versus enabling greater capabilities in your existing markets? Thanks.

A - Jen-Hsun Huang

Yeah, Mark. So, first of all, Turing, as you know, is the world's first ray tracing GPU. And it completes our new computer graphics platform, which is going to reinvent computer graphics altogether. It unites four different computing modes: rasterization, accelerated ray tracing, computing with CUDA, and artificial intelligence. It uses these four basic methods to create imagery for the future.

There's two major ways that we'll experience the benefits right away. The first is, for the markets of visualization today, they require photorealistic images. Whether it's a IKEA catalog, or a movie, or architectural engineering, or a product design, car design, all of these types of markets require photorealistic images. And the only way to really achieve that is to use ray tracing with physically-based materials and lighting. The technology is rather complicated. It's been computing-intensive for very long time. And it wasn't until now that we've been able to achieve it in a productive way.

And so Turing has the ability to do ray tracing, accelerated ray tracing, and it also has the ability to combine very large frame buffers because these datasets are extremely large. And so that marketplace is quite large and it's never been served by GPUs before. Until now, all of that has been run on CPU render farms, gigantic render farms in all these movie studios and service centers and so on and so forth. The second area where you're going to see the benefits of ray tracing, we haven't announced.

Operator

Your next question...

Q - Mark Lipacis {BIO 2380059 <GO>}

Okay. If I could have a follow-up, on the gaming side where do you think the industry is on creating content that leverages that kind of capability? Thank you.

A - Jen-Hsun Huang

Yeah, Mark. At GTC this last year in March, GDC and GTC, we announced the brand-new platform called NVIDIA RTX. And this platform has those four computation methods that I described for generating images. We put that platform out with the support of Microsoft. They call it the Microsoft DirectX Raytracing and the major game engine company, Epic, has implemented real-time ray tracing and the RTX into the Epic engine, the Unreal Engine. And at GDC and GTC, we demonstrated for the very first time on four Volta GPUs the ability to do that. And it was the intention to get this platform out to all of the game developers. And we've been working with game developers throughout this time.

This week at SIGGRAPH we announced Quadro, which is the first – the Quadro RTX 8000, 6000 and 5000, the world's first accelerated ray tracing GPUs. And I demonstrated one Quadro running the same application that we demonstrated on four Volta GPUs running in March. And the performance is really spectacular. And so I think the answer to your question is developers all have access to RTX. It's in Microsoft's DirectX. It's in the most popular game engine in the world and you're going to start to see developers use it.

On the workstation side, on the professional visualization side, all of the major ISPs have jumped onto adopt it. And at SIGGRAPH this year, there you could see a whole bunch of developers demonstrating the NVIDIA RTX with accelerated ray tracing, generating photorealistic images. And so, I would say that, in no platform in our history has on day one of announcement, had so many developers jump onto it. And stay tuned, we've got a lot more stories to tell you about RTX.

Operator

Your next question is from Matt Ramsay with Cowen.

Q - Matthew D. Ramsay {BIO 17978411 <GO>}

Thank you very much. Colette, I had a couple of questions about inventory. The first of which is – I understand you've launched a new product set in ProViz and the datacenter business is obviously ramping really strongly. But if you look at the balance sheet, I think the inventory level is up by around mid-30% sequentially and you're guiding revenue up 3% or so. Maybe you could help us sort of walk through the contribution to that inventory and what it might mean for future products. And secondly, if you could talk a little bit about the gaming channel in terms of inventory, how things are looking in the channel as you guys see it during this period of product transition? Thank you.

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

Sure. Thanks for your question. So when you look at our inventory on the balance sheet, I think it's generally consistent with what you have seen over the last several months in terms of what we will be bringing to market. Turing is an extremely important piece of architecture, and as you know, it will be with us for some time. So, I think the inventory balance is getting ready for that. And don't forget our work in terms of datacenter and what we have for Volta is also a very, very complex computer in some cases, in terms of what we have also in terms of there. So just those things together, plus our Pascal architecture is still here, makes up almost all of what we have there in terms of inventory.

A - Jen-Hsun Huang

Matt, on the channel inventory side, we see inventory in the lower ends of our stack. And that inventory is well positioned for back-to-school and the building season that's coming up on Q3. And so I feel pretty good about that. The rest of our product launches and the ramp-up of Turing is going really well. And so I think the rest of the announcements we haven't made, but stay tuned. The RTX family is going to be a real game changer for us and the reinvention of computer graphics altogether has been embraced by so many developers. We're going to see some really exciting stuff this year.

Operator

Next question is from Vivek Arya with Bank of America.

Q - Vivek Arya {BIO 6781604 <GO>}

Thanks for taking my question. Actually just a clarification, and then the question. On the clarification, Colette, if you could also help us understand the gross margin sequencing from Q2 to Q3? And then Jensen, how would you contrast the Pascal cycle with the Turing cycle? Because I think in your remarks, you mentioned Turing is a very strong advancement over what you had before. But when you launched Pascal, you had guided to very strong Q3s and then Q4s. This time, the Q3 outlook, even though it's good on an absolute basis, on a sequential and a relative basis, it's perhaps not as strong. So if you could just help us contrast the Pascal cycle with what we should expect with the Turing cycle?

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

Sure. Thanks, Vivek, for the question. Let me start first with your question regarding gross margins. We have essentially reached, as we move into Q3, a normalization of our gross margins. I believe over the last several quarters, we have seen the impacts of crypto and what that can do to elevate our overall gross margins. We believe we've reached a normal period as we're looking forward to essentially no cryptocurrency as we move forward.

A - Jen-Hsun Huang

Let's see, Pascal was really successful. Pascal, relative to Maxwell, was a leap in fact. And it was a really significant upgrade. The architectures were largely the same. They were both programmable shading. They were both of the same generation of programmable shading. But Pascal was much, much more energy efficient. I think it was something like

30%, 40% more energy-efficient than Maxwell, and that translated to performance benefits to customers. The success of Pascal was fantastic.

There's just simply no comparison to Turing. Turing is a reinvention of computer graphics. It is the first ray tracing GPU in the world. It's the first GPU that will be able to ray trace light in an environment and create photorealistic shadows and reflections and be able to model things like area lights and global illumination and indirect lighting, that the images are going to be so subtle and so beautiful. When you look at it, it just looks like a movie. And yet it's backwards compatible with everything that we've done. This new hybrid rendering model, which extends what we've built before but added to it two new capabilities, artificial intelligence and accelerated ray tracing, is just fantastic.

So everything of the past will be brought along and benefits, and it's going to create new visuals that were impossible before. We also did a good job on laying the foundations of the development platform for the developers. Now we've partnered with Microsoft to create DXR. VulkanRT is also coming, and we have optics that are used by ProViz renderers and developers all over the world.

And so we have the benefit of laying the foundation stack by stack by stack over the years. And as a result, on the day that Turing comes out, we're going to have a richness of applications that gamers will be able to enjoy.

You mentioned guidance. I actually think that on a year-over-year performance, we're doing terrific. And I'm super excited about the ramp of Turing. It is the case that we benefited in the last several quarters from an unusual lift from crypto.

In the beginning of the year, we thought and we projected that crypto would be a larger contribution through the rest of the year, but at this time we consider it to be immaterial for the second half. And so that makes comparisons on a sequential basis on, I guess, quarterly sequential basis harder. But on a year-to-year basis, I think we're doing terrific. Every single one of our platforms are growing. High-performance computing, of course, datacenters is growing. AI, the adoption continues to sweep from one industry to another industry. The automation that's going to be brought about by AI is going to bring productivity gains to industries like nobody's ever seen before.

And now with Turing we're going to be able to reignite the professional visualization business, open us up to photorealistic rendering for the very first time, render farms and everybody who's designing products that has to visualize a photo realistically to reinventing and resetting graphics for video games. And so I think we're in a great position and I'm looking forward to reporting Q3 when the time comes.

Operator

Your next question is from Atif Malik with Citi.

Q - Atif Malik {BIO 15866921 <GO>}

Hi, thanks for taking my question. Colette, I have a question on datacenter. In your prepared remarks, you talked about AI and high performance computing driving new verticals and some of these verticals are fastest growing. Some of your peers have talked about enterprise spending slowing down in back half of this year, and so unit demand and you guys are not a unit play, but more of an AI adoption. Just curious in terms of your thinking about second half datacenter growth?

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

So we generally give our view on guidance for one quarter out. You are correct that our datacenter results that we see is always a tremendous unique mix every single quarter in terms of what we're seeing. But there's still some underlying points of that that will likely continue. The growth in terms of – used by the hyperscales continued industry by industry coming on board. Essentially just because the needs of accelerated computing for the workloads and for the data that they have is so essential. So we still expect as we go into Q3 for datacenter to grow both sequentially and year-over-year. And we'll see probably a mix of both selling our Tesla V100 Platforms, but also a good contribution from our DGX.

A - Jen-Hsun Huang

Yes. That's right. Atif, let me just add a little bit more to that. I think the, one simple way to think about that is this. In the transportation industry, let's take one particular vertical, there are two dynamics that are happening that are very abundantly clear and that will transform that industry. The first of course is ride-hailing and ridesharing. Those platforms in order to make the recommendation of which taxi to bring to which passenger to which customer is a really large computing problem. It's a machine learning problem. It's an optimization problem a very, very large scale and in each and every one of those instances you need high performance computers to use machine learning to figure out how to make that perfect match or the most optimal match.

The second is self-driving cars. Every single car company that's working on robotaxis or self-driving cars needs to collect data, label data, train the neural network, or train a whole bunch of neural networks and to run those neural networks in cars. And so you just make your list of how many people are actually building self-driving cars. And every single one of them will need even more GPU accelerated servers. And that's just for developing the model. Then the next stage is to simulate the entire software because we know that the industry or the world travels 10 trillion miles per year. And the best we could possibly do is to drive several million normal miles. And what we really want to do is to be able to simulate and stress, stress-test our software stack and the only way to do that is doing virtual reality. And so that's another supercomputer that you have to build for simulating all your software across those billions and billions of virtually created challenging miles.

And then lastly before you OTA the software, you're going to have to re-sim and replay against all of the miles that you've collected over the years to make sure that you have no regressions before you OTA the new models into a fleet of cars. And so transportation is going to be a very large industry.

Healthcare is the same way. From medical imaging that's now using AI just live everywhere to genomics that has discovered deep learning and the benefits of artificial intelligence and in the future pathology, and the list goes on. And so industry after industry after industry we're discovering the benefits of deep learning and the industries could be really, really revolutionized by it.

Operator

Your next question is from C.J. Muse with Evercore.

Q - C. J. Muse {BIO 6507553 <GO>}

Thank you for taking my question. I guess short-term and a long-term. So for short-term, as you think about your gaming guide, are you embedding any drawdown of channel inventory there? And then longer-term, as you think about Turing Tensor Cores, can you talk a bit about differentiation versus Volta V100, particularly as you think about 8-bit integer and the opportunities there for inferencing? Thank you.

A - Jen-Hsun Huang

We're expecting the channel inventory to work itself out. We are masters of managing our channel and we understand the channel very well. As you know, the way that we go to market is through the channels around the world. We're not concerned about the channel inventory.

As we ramp Turing, whenever we ramp a new architecture, we ramp it from the top down. And so we have plenty of opportunities as we go back to the back-to-school in the gaming cycle to manage the inventory. So we feel pretty good about that. As a result, comparing Volta and Turing, CUDA is compatible. That's one of the benefits of CUDA. CUDA, all of the applications that take advantage of CUDA that are written on top of cuDNN, which is our neural network platform to TensorRT that takes advantage – that takes the output of the frameworks and optimize it for runtime. All of those tools and libraries run on top of Volta and run on top of Turing and run on top of Pascal.

What Turing adds over Pascal is the same Tensor Core that is inside Volta. Of course, Volta is designed for large scale training. Eight GPUs could be connected together. They have the fastest HBM2 memories and it's designed for datacenter applications, has 64-bit double precision ECC, high resilience computing, and all of the software and system software capability and tools that make Volta the perfect high performance computing accelerator.

In the case of Turing, it's really designed for three major applications. The first application is to open-up pro visualization, which is a really large market that has historically used render farms and were really unable to use GPUs until we now have the ability to do full path traced global illumination with very, very large data sets. So that's one market that's brand new as a result of Turing.

The second market is to reinvent computer graphics – real-time computer graphics for video games and other real-time visualization applications. When you see the images created by Turing, you're going to have a really hard time wanting to see the images of the past. It just looks amazing.

And then the third, Turing has a really supercharged Tensor Core. And this Tensor Core is used for image generation. It's also used for high throughput deep learning inferencing for datacenters. And so these applications for Turing, which suggests that there are multiple SKUs of Turing which is one of the reasons why we have such a great engineering team. We could scale one architecture across a whole lot of platforms at one time. And so I hope that answers your question. The Tensor Core inference capability of Turing is going to be off the charts.

Operator

Next question is from Joe Moore with Morgan Stanley.

Q - Joseph Moore {BIO 17644779 <GO>}

Great. Thank you. I wonder if you could talk about cryptocurrency now that the dust has settled. You guys have done a good job of kind of laying out exactly how much of the OEM business has been driven by that. But there's also been, I think, some sense of – some of the GeForce business was being driven by crypto. Looking backwards, can you size that for us? And I guess, if – I'm trying to understand the impact that crypto would have on the guidance for October given that it seems like it was very small in the July quarter? Thank you.

A - Jen-Hsun Huang

Well, I think the second question is easier to answer and the reason – the first one is just, it's ambiguous. It's hard to predict, anyway. It's hard to estimate, no matter what. But the second question, the answer is we're expecting – we're projecting zero basically.

And for the first question, how much of GeForce could have been used for crypto, a lot of gamers at night they could – while they are sleeping they could do some mining. And so, did they buy it from mining or did they buy it for gaming, it's kind of hard to say. And some miners were unable to buy our OEM products and so they jumped onto the market to buy it from retail. And that probably happened a great deal as well. And that all happened in the previous several quarters, probably starting from Q3, Q4, Q1 and very little last quarter and we're projecting no crypto mining going forward.

Operator

Your next question is from Toshiya Hari with Goldman Sachs.

Q - Toshiya Hari {BIO 6770302 <GO>}

Great. Thanks very much. I had one for Jensen and one for Colette. Jensen, I was hoping you could remind us how meaningful your inference business is today within datacenter? And how you would expect growth to come about over the next two years as your success at accounts like Google proliferate across a broader set of customers?

And then for Colette, if you can give directional guidance per each of your platforms? I know you talked about datacenter little bit (45:28) can talk about the other segments and on gaming specifically, if you can talk about whether or not new products are embedded in that guide? Thank you.

A - Jen-Hsun Huang

Thanks, Toshiya. Inference is going to be a very large market for us. It is surely material now in our datacenter business. It's not the largest segment, but I believe it's going to be a very large segment of our datacenter business.

There are 30 million servers around the world, let's kind of estimate, in the cloud and there are a whole lot more in enterprises. I believe that almost every server in the future will be accelerated. And the reason for that is because artificial intelligence and deep learning software and neural net models are going to prediction models, are going to be infused in the software everywhere. And acceleration has proven to be the best approach going forward.

We've been laying the foundations for inferencing for a couple of – two, three years. And as we've described at GTCs, inference is really, really complicated and the reason for that is, you have to take the output of these massive networks that are output of the training frameworks and optimize it. This is probably the largest computational graph optimization problem the world's ever seen. And this is brand new invention territory. There are so many different network architectures from CNNs to R-CNNs to autoencoders to RNNs and LSTMs, there's just so many different species of neural networks these days and it's continuing to grow, and so the compiler technology is really, really complicated.

And this year we announced two things. Earlier this year we announced that we've been successful in taking the Tesla P4 low-profile, high energy efficiency inference accelerator into hyperscale datacenters. And we announced our fourth generation TensorRT optimizing compiler – neural network optimizing compiler. And TRT 4 goes well beyond CNNs and image recognition in the beginning. And now it allows us to support and optimize for voice recognition or speech recognition, natural language understanding, recommendation systems, translation. And all of these applications are really pervasive from Internet services all over the world.

And so now from images to video to voice to recommendation systems, we now have a compiler that can address it. We are actively working with just about every single Internet service provider in the world to incorporate inference acceleration into their stack. And the reason for that is because they need high throughput and very importantly they need low latency. Voice recognition is only useful if it responds in a relatively short period of time. And our platform is just really excellent for that.

And then this week we announced Turing. And I announced that the inference performance of Turing is 10 times the inference performance of Pascal, which is already a couple of hundred times the inference performance of CPUs. And so you take a look at the rate at which we're moving both in the support of new neural networks, the ever increasing optimization and performance output of the compilers and the rate at which we're advancing our processors, I think we're raising the bar pretty high.

Okay. So with that, Colette?

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

Yeah. So when you look at our overall segments, as you've even seen our results in terms of this last Q2, there was growth across every single one of our platforms from a year-over-year standpoint. We probably absolutely see that again in our Q3 guidance, the year-over-year growth across each and every one of those platforms.

Of course, our OEM business will be down likely year-over-year, again just due to the absence of cryptocurrency in our forecast. When we think about sequentially, our hopes is absolutely our datacenter will grow and we'll likely see the growth of our gaming business as well. It's still early. Still we've got many different scenarios on our ProViz and auto, but definitely our gaming and our datacenter are expected to grow sequentially.

Operator

Your next question is from Blayne Curtis with Barclays.

Q - Blayne Curtis {BIO 15302785 <GO>}

Hey. Thanks for taking my question. Two on gross margin. Colette, I just want to make sure I understood July to October gross margins down. I know you've been getting a benefit from crypto, but it's pretty de minimis in July. Is there any other moving pieces? And then kind of a longer picture here, how do you think about the ramp of Turing affecting gross margins? You're obviously enabling a lot of capabilities. You get paid for 12 nanometers fairly stable. Just kind of curious how to think about over the next couple of quarters gross margin with that ramp?

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

Yes. So let me take your first part of the question regarding our gross margins and what we had seen from crypto. Although, crypto revenue may not be large, it still has a derivative impact on our stack in terms of what we are selling and to both replenish the overall channel and such. So over the last several quarters we had (51:30) stabilizing that overall, we did get the great effect of selling just about everything and our margins really been able to benefit from that. Again, when we look at the overall growth year-over-year for Q2 you have 500 basis points in terms of growth. We're excited about what we have now here for Q3 as well, which is also significant growth year-over-year.

Of course, we have our high value-added platforms as we move forward, both those and datacenter those in terms of what we expect the effects of Turing in terms of – on our Quadro piece as well. But that will take some time for that all to partake. So we'll see how that goes. We haven't announced anything further at this time. But yes, we'll see probably over the longer term the effects of what Turing can do.

Operator

Next question is from Aaron Rakers with Wells Fargo.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yeah. Thanks for taking the question. I'm curious as we look at the datacenter business, if you can help us understand the breakdown of demand between hyperscale, the supercomputing piece of the business and the AIPs. And I guess on top of that, I'm just curious, one of the metrics that's pretty remarkable over the last couple of quarters as you've seen significant growth in China? I'm curious if that's related to the datacenter business or what's really driving that as kind of a follow-up question? Thank you.

A - Jen-Hsun Huang

Hi. Yeah, Aaron. I think that you look at the – if you start from first principles here's the simple way to look at it. Demand is continuing to grow at historical levels of 10x computing demand, computing demand. Computing demand is increasing at historical levels of 10x every five years. 10x every five years is approximately Moore's Law.

And computing demand continues to grow at 10x every five years. However, Moore's Law stopped. And so that gap in the world in high performance computing, in medical imaging, in life sciences computing, in artificial intelligence, that gap because those applications demand more computing capability, that gap can only be served in another way. And NVIDIA's GPU-accelerated computing that we pioneered really stands to benefit from that.

And so at the highest level whether it's supercomputing and this year you heard Colette say earlier that NVIDIA GPUs represented 56% of all the new performance that came into the world's TOP500. The TOP500 is called the TOP500 because it reflects the future computing. And my expectation is that more and more from one vertical industry after another, and I mentioned transportation, I mentioned healthcare, the vertical industries go on and on, that as computing demand continues at a factor of 10x every five years, developers are rational and logical to have jumped on NVIDIA's GPU computing to boost their demand.

I think that's probably the best way to answer it.

Operator

Your next question is from Harlan Sur with JPMorgan.

Q - Harlan Sur {BIO 6539622 <GO>}

Good afternoon. Thanks for taking my question. When we think about cloud and hyperscale, we tend to think about the top guys right? They're designing their own platform using your Tesla based products or sometimes even designing their own chips for AI and deep learning. But there's a larger base of medium to smaller cloud and hyperscale customers out there who don't have the R&D scale. And I think that's where your HGX platform seems to be focused on. So Jensen, can you just give us an update on the uptake of your first generation HGX-1 reference platform and the initial interest on the HGX-2? Thanks.

A - Jen-Hsun Huang

HGX-1 was, I guess, kind of the prototype of HGX-2. HGX-2 is doing incredibly well for all the reasons that you mentioned. It is and even the largest hyperscale datacenters can afford to create these really complicated motherboards at the scale that we're talking about. And so we created HGX-2 and it was immediately adopted by several most important hyperscalers in the world. And we were at GTC Taiwan and we announced basically all of the leading server OEMs and ODMs supporting HGX-2 and are ready to take it to market. So we're in the process of finishing HGX-2 and ramping into production. And so I think HGX-2 is a huge success for exactly the reasons that you mentioned. We could use it for essentially a standard motherboard like the ATX motherboard for PCs that could be used for hyperscalers, it could be used for HPC, it could be used for datacenters. And it's really – it's a really fantastic design. It just allows people to adopt this really complicated and high performance and really high speed interconnect motherboard in a really easy way.

Operator

Your next question is from Tim Arcuri with UBS.

Q - Timothy Arcuri {BIO 3824613 <GO>}

Thank you. Actually I had two questions, Jensen both for you. First, now that crypto has fallen off, I'm curious what you think the potential is that maybe we see a slug of cards that get resold on eBay or some other channel and that could cannibalize new Pascal sales. Is that something that keeps you up at night? Number one.

And number two, obviously the stories about gaming and datacenter and I know that you don't typically talk about customers, but since Tesla did talk about you on their call, I'm curious what your comments are about the development for Hardware 3 and their own efforts to move away from your DRIVE Platform? Thank you.

A - Jen-Hsun Huang

Sure. Well the crypto mining market is very different today than it was three years ago. And even though new cards – at the current prices, it doesn't make much sense for new cards to be sold into the mining market. The existing capacity is still being used and you could see that the hash rates continue. And so my sense is that the installed base of miners will continue to use their cards. And then probably the more important factor though is that we're in the process of announcing a brand new way of doing computer graphics. And with Turing and the RTX platform, computer graphics will never be the same. And so I think this, our new generation of new GPUs is really going to do great.

I also think that I appreciate Elon's comments about our company and I also think Tesla makes great cars and I drive them very happily. And with respect to the next generation, it is the case that when we first started working on autonomous vehicles they needed our help. And we used the three year old Pascal GPU for the current generation of autopilot computers. And it is very clear now that in order to have a safe autopilot system, we need a lot more computing horsepower.

In order to have safe computing, in order to have safe driving, the algorithms have to be rich, it has to be able to handle corner conditions in a lot of diverse situations, and every time that there's more and more corner conditions or more subtle things that you have to do or you have to drive more smoothly or be able to take turns more quickly, all of those requirements require greater computing capability. And that's exactly the reason why we built Xavier.

Xavier is in production now. We're seeing great success and customers are super excited about Xavier. And that's exactly the reason why we built it. And I think it's super hard to build Xavier and all the software stack on top of it. And if it doesn't turn out for whatever reason, it doesn't turn out for them, you can give me a call and I'd be more than happy to help.

Operator

Unfortunately, we have run out of time. I will now turn the call back over to Jen-Hsun for any closing remarks.

A - Jen-Hsun Huang

We had a great quarter. Our core platforms exceeded expectations even as crypto largely disappeared. Each of our platforms, AI, gaming, ProViz and self-driving cars continue to enjoy great adoption. These markets we're enabling are some of the most impactful to the world today. We launched Turing this week. It was 10 years in the making and completes the NVIDIA RTX platform. NVIDIA RTX with Turing is the greatest advance since CUDA nearly a decade ago. I'm incredibly proud of our company for tackling this incredible challenge, reinvesting the entire graphic stack and giving the industry a surge of excitement as we reinvent computer graphics. Stay tuned as we unfold the exciting RTX story. See you guys next time.

Operator

Thank you for joining. You may now disconnect.

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