NVIDIA Corp Investor Day

Company Participants

- Colette M. Kress, CFO and EVP
- Deepu Talla, VP and General Manager of Mobile Business Unit
- Jeffrey D. Fisher, SVP of Geforce Business Unit
- Jensen Hsun Huang, Co
- Robert Csongor, VP and General Manager of Automotive
- Shankar Trivedi, Unknown
- Shawn Simmons, Unknown
- Unidentified Speaker, Unknown

Other Participants

- Ambrish Srivastava, MD of Semiconductor Research and Senior Research Analyst, BMO Capital Markets Equity Research
- Atif Malik, VP and Semiconductor Capital Equipment and Specialty Semiconductor Analyst, Citigroup Inc, Research Division
- Brett Simpson, Senior Analyst, Arete Research Services LLP
- James Wang, Analyst, Unknown
- Matthew D. Ramsay, Principal and Senior Analyst, Canaccord Genuity Limited, Research Division
- Mitchell Toshiro Steves, Analyst, RBC Capital Markets, LLC, Research Division
- Srinivas Reddy Pajjuri, Senior Analyst, Macquarie Research
- Stacy Aaron Rasgon, Senior Analyst, Sanford C. Bernstein & Co., LLC., Research Division
- Toshiya Hari, MD, Goldman Sachs Group Inc., Research Division
- Trip Chowdhry, MD of Equity Research, Global Equities Research, LLC
- Unidentified Participant, Analyst, Unknown
- Vivek Arya, Director, BofA Merrill Lynch, Research Division
- William Shalom Stein, MD, SunTrust Robinson Humphrey, Inc., Research Division

Presentation

Shawn Simmons {BIO 19500814 <GO>}

All right. Hello, everyone, we're finally going to get started here. So if everyone could go ahead and take a seat. I want to welcome you to NVIDIA's 2017 Annual Investor Day. It's great to see all of you here. We really appreciate you taking the time to be with us here today.

I'm Shawn Simmons from the Investor Relations team, at this time, I'd like to go ahead and quickly walk through the safe harbor before we get to today's agenda.

So we will be making forward-looking statements in today's program regarding our expectations and other future events, which may differ materially from NVIDIA's actual results. And I'd like to refer you to our SEC filings for a description of our businesses and associated risks and other factors, which could cause our results to differ materially from these statements. Also, if we use any non-GAAP financial measures, you'll find the reconciliations to GAAP on our Investor Relations website.

So turning to the agenda, going real quick here, we'll start off with our Founder and CEO, Jensen Huang. Next, we'll move to Jeff Fisher, who will cover Gaming. After that will be Shanker Trivedi, then we'll go into Rob Schlanger, Deepu Chawla and then Colette Kress for the financial. Once we finish the presentations, which will last around 2 hours, all of the executives will join us up here on stage for roughly an hour of Q&A. Following the Q&A session, we will have a reception, starting around 3 pm over in that corner. And just a couple more things, if you do need anything during this time, please do not hesitate to find me, e-mail me. My e-mail is clearly presented up here. We absolutely will help you with whatever you need, whether it's Wi-Fi or anything.

And with that, I do want to -- now please join me in welcoming NVIDIA's Founder and Chief Executive Officer, Jensen Huang.

Jensen Hsun Huang {BIO 1782546 <GO>}

I didn't realize there was clapping in Investor Day. You didn't leave anything for me to read. Who prepared my stuff? I only have one slide, is this it? I just got to click it? I don't want to torture them with this. I don't want to torture you guys with anything. How about we just have that?

Let's see. You know, if there was a way -- if there was a way to talk about what's happening in computing right now, I would characterize it as the perfect storm. On the one hand, a new computing model is taking off. On the other hand, a new computing approach has been found. It's almost like the simultaneous beginning of two different things and simultaneous ending of two others.

We know that general purpose microprocessors won't take us the rest of the way. I think we understand that very well now. And that this computing approach that we've been working on for over a decade called GPU accelerated computing has real promise. And the type of problems that GPU computing solves happens to be exactly the most important problems we want to solve today. And whether it's AI or self-driving cars or robotics or cloud computing or -- it just happens to be -- it happens to be at the perfect place at the perfect time. And I guess it's the perfect storm or serendipity meets destiny. But I think we're feeling all of that right now at the same time.

The thing that we've done, the thing that we have done as a company, that I really feel are good decisions, some of the decisions are this. As a business model, as an architecture of company, the first layer we are horizontal. We advance the computing platform on first principles. We architect it from the processor, to the system, to the system software, to the algorithms. And we work with application providers in each one of the domains that we target. And we target large domains, like for example, our video game industry is $100 billion large. The automotive industry is quite large. And when I say automotive, I mean, autonomous vehicles, is quite large. Healthcare is large. Manufacturing is large. Media and entertainment are large. And so we target these verticals and in these verticals, we go deep into the vertical, we work with all of the majors and we help them incorporate our computing approach into their workflow.

The first time I met Carl Bass was probably, I would say, 10 years ago. And they had 0% GPU accelerated applications. Autodesk, AutoCAD was completely rendered on CPUs. Everything they did was completely CPU based. Not one drop of GPU code. The first time I met Chuck at Adobe, they had no software accelerated code. Adobe had no software accelerated. It was all completely CPU. And they were thinking about moving to SIMDs, the SSC's.

And so -- and for good reasons because at the time, GPU adoption was really low. At the time, GPU adoption was really low. And their point is we develop all the software where will we run it. And so we had to solve this chicken and the egg problem by doing our job first, by being -- by having strong conviction to our architecture, we stuck with it. We're on CUDA 9 right now. Every one on the CUDA is a backward -- can babble with the other CUDAs. And every one of our GPUs on the planet runs with every single CUDA, every single permutation, doesn't matter. You find yourself a GPU, it is architecturally compatible. We have the discipline to stay with it. It's a fairly significant undertaking. And you have to believe in it because you're doing a whole bunch of work that nobody is using for a long time. And we stuck with it. And we stuck with it. And the reason why we stuck with it was because we had that curve in our mind. Just because you're not experiencing the future it doesn't mean the future won't come. And so we believed in it. We believe that Dennard scaling was going to stop because first principal says that it was going to stop. We knew that. We knew that there was only so much instruction level parallelism that you can harvest. We could see that. We all knew that. You multiply the 2 of the things together. And you're going to see the present, we knew that. We knew that. And so if you know something, what do you do about it? That's the hard part. Doing something about something that is really when doing it is really, really painful, fully realizing the future and still doing it in the present where nobody cares was hard. And so we stuck with it. And now, every single application from Adobe, every single application from Autodesk, every single application from every single media, entertainment, design, CAD, you name it, every single one, 100%, completely ported onto CUDA, a 100%. So in 10 years, it went from 0% to basically 100%. And that's our foundational layer, the first foundation.

Then we build teams to go after work with the vertical industries one after another. We work with (Dusso) in manufacturing. We work with all of molecular dynamics companies in healthcare, just one industry after another industry. So that's our horizontal approach, the horizontal approach.

The second thing we do is for some of the select markets, we might decide to go all the way. We might decide to go all the way, like for example, gaming. You're going to hear from Jeff Fisher, we're going to go all the way. We are the world's largest game platform. We're the world's largest game platform. And we serve a lot of gamers. And they know that we will serve them and support them and optimize their system for as long as we shall live. They know that. And we have a facility and a mechanism for doing that, which also shows the self-driving car industry the autonomous vehicle industry and we are going to serve that top to bottom. I drove by our BB8 this morning. And I was driving home and drove by our BB8 going home. And we have cars around world that are driving by themselves. And we created the architecture, the processor, the system, the system software, all the algorithms in between, connected to every single mapping company in the world. And we continuously iterate on that software stack, than entire stack. And the rate at which it advances is incredible once we put it together.

You saw today a few of the curves that I showed, 7x in five years, 7x in five years. The way to think about that is in five years time, in five years time whatever capability I currently have in self driving cars, it will be 7x, think about that for a second. five years from now is only 2022. We're going to be 7x. Whatever functionality we currently have, we'll -- we have improved it by a factor of 7, not 50%, a factor of 7. The power of accelerated computing, the power of thinking from top to bottom, bottom to top and refining and refining and refining the stack. The autonomous vehicle stack.

We decided that for AI computing, we would go all the way. We would go all the way. We would work on every single layer from the architecture to the processors to the systems called DGX, to the system software, all of our deep learning SDKs, the integration with every single application on the planet. And we now containerize it. We created this thing called NVIDIA Docker, NV Docker. We containerize it. We maintain it for every single combination and permutation, meaning you could have somebody in your office using your computer running TensorFlow. And the next person running on the same computer is actually using Caffe, no problem. They're virtualized, they're containerized, that's the benefit of containers. And we'll connect it all the way to our cloud Registry. And this cloud registry now has NVIDIA software stack in it. And this cloud registry can go everywhere. We could download it onto Titan X. We can download it into DGX. We can download it into an HGX in the cloud. And we could download it to any cloud. And so, as a result, deep learning engineers could grab a Titan X and as soon as they outgrow it, they can burst it into the cloud. As soon as they outgrow it, they can burst it into the data center with DGXs. All 1 Registry, 1 account.

And so for deep learning, we decided that we're going to go all the way and we can actually going to be training developing networks, except we're not going to develop generic networks that are horizontal in application. We're not going to do voice recognition networks. There's no point in doing that because all of the cloud companies are going to do a fantastic job doing it. There's no sense us doing it. There's no sense us doing basic visual recognition, computer vision networks. They're going to do a fantastic job doing that. What we are going to do is we going to work with each one of the domains that I've already talked about where we have vertical engagement. And we'll work with them to create networks that are specific to their application, to their domain. We put that up in the cloud. So that everybody who's doing medical imaging can now download a pre-trained network that's 98% smart and then use our software platforms to incrementally train the last 20% with their own domain specific knowledge. We'll do that for healthcare. We'll do that for automotive. We'll do that for the vertical markets we serve.

And so these 3 areas, we're going to go all the way. And so we're part horizontal, part horizontal, computing platform that you can -- the way we monetize it is through a cartridge. That cartridge is called Tesla. But largely is the software around it, largely it's the ecosystem around it. Just has to be monetized in the form of a cartridge. We have a horizontal business model. Take that to the cloud. We take that the OEMs. We take that directly to retail or e-tail. We take that directly through our own systems called DGX, a horizontal computing model and then a vertical, by application, one at a time, gaming, design, the same verticals that we've always been following, gaming, design, self-driving cars, otherwise known as autonomous vehicles and artificial intelligence.

And in these markets, we will offer the entire stack, okay? And so today, you're going to hear us talk sometimes horizontally, sometimes vertically. Does that make sense? And so you just got to sort in your head which one they're in and sometimes they're talking horizontally and they're capturing some customer examples on your behalf to tell you the customer's voice and how they're seeing it and what's the value proposition they see. But that's basically our company. And I think that this entire system that you see is 1 architecture, 1 computing architecture. I actually think that we're the only company in the world at the moment that has an accelerated computing -- a unique computing architecture and it's unified. There's only 1. We use 1 architecture for Tesla, 1 architecture for -- the same architecture for Quadro, the same architecture for GeForce, the same architecture for Tegra, the same architecture for Jetson, the same architecture for Drive, the same architecture for DGX. It's exactly the same one architecture. The reason why that's so valuable to us is, of course, leverage. One tool works across the whole thing. One tool works across the whole thing. You can develop on our platform, deploy it on our platform. So all of our employees inside the company are super delighted by that, that they could have a platform to develop their neural network and then without getting out of their seat, they go from TensorFlow, boom, to TensorRT, boom, it goes into Drive and they're driving the car. No recompile, no move files here and there, it just doesn't. Goes into the cloud and comes back into the car, boom, boom. No problem, one architecture. And so leverage, incredible amounts of productivity, improvement as we go from end to end. Then most importantly, this is the big, big deal that people always forget. The point about a platform is not your investment. The point about the platform is their investment. What makes a great platform is the investment of the ecosystem around you. That's what makes AWS great, right? That's what -- does what makes AWS great. That's what makes Windows great. That what makes these Android great. That's what makes these platforms great. What makes them great is the fact that there are so many developers who have developed on that platform. And here comes the question, why would they develop on your platform if several conditions are not met completely, with complete confidence? One, that this platform adds unique value, otherwise use the current platform they have. Number two, that by supporting this platform, they could have reach. Everybody wants to grow. They're not supporting you for the heck of it. It's a lot of work to support a new computing platform. So it has to have reach. And here comes the tough one, they need to look at you. And they need to know you're committed to it. And if you have 7 architectures in your company, which one are you committed to? All of them? Really? Any of them? Really? Our company is dedicated to one architecture. We are all in.

It's the last thing that you guys want to hear, it's all going to work out or it's going to work out terribly. The work -- however, it works out, we're going together. We're going together. The captain isn't the last person on the ship. They captain is the guy that took the ship down. We're going down together, taking you all with me. So we are in 1 architecture company. We are all in. And because we are all in, everything just works. And everything that -- it because it's 1 architecture, it's everywhere. And we met those 3 simultaneous equations, incredibly, incredibly hard. We add value in a very specialized way. Number two, we are absolutely everywhere. Number three, our commitment is unwavering. The only thing they see when they look at my eyes, crazy eyes, intense crazy eyes, all focused on making sure that this happens. And that's why CUDA was able to lift an entire industry as we go and enjoy life after Moore's Law. I think that that's it. That's how we got here. Crazy amounts of intense commitment because we believed in something so much that we're willing to push it all in.

And that's it. That's my short talk. And so when the team comes up and talk to you about that, just listen to it from that perspective, okay. And I'll come back later on today and answer questions. Thanks for being here today. Thank you.

Ladies and gentlemen, Jeff Fisher, one of NVIDIA's oldest employees in many, many ways, hang on a second. Fish and I grew up together. Here we go. I raised you. He did, he did.

Jeffrey D. Fisher {BIO 2373419 <GO>}

I don't need that, no. I think I'm all set. Thank you, all for joining us today and joining us for keynote. My name is Jeff Fisher. I run the Gaming business. Today, I'm going to talk about PC gaming. I know there's a fair amount of you who grab me have some keen interest in it. I'm going to show you my perspective on where we're at, where the market's going, some of the fundamentals behind PC gaming. I think we're meeting afterwards in case you have any more specific questions. But let me get started.

First of all, it doesn't always go without saying. But it goes without saying that the gaming market is incredibly big and growing. Of the roughly 2 billion people are gaming in the world today. That's about 1/3 of humanity. It's an entire generation. If roughly everybody probably younger than this group are gamers. All of our kids grew up gaming. They started on phones, they moved to tablets and ultimately, they moved to PC or some other platform. The overall gaming market is growing between 8% and 10% a year, about $100 billion, Jensen noted. It's a software revenue within gaming. PC gaming is growing about 6% to 7% a year as a CAGR. Nuzu estimates there's about 400 million core PC gamers in the world scattered in every country in the world. Core gamers are gamers that game roughly once a week or more. They spend money on content. And the core gamers had definitely a lot to do with the growth in PC gaming. And these, in effect, are our target audience.

The fundamentals of PC gaming, as I'll talk about today, are also strong. What's driving PC gaming, eSports, competitive gaming, AAA gaming, notebook gaming, all those fundamentals remain strong.

So let's talk a bit about GeForce. GeForce grew about 44% year-over-year, FY '16 to FY '17. We saw growth in both our desktop and more importantly, or even higher growth in our notebook market. Gaming notebook. Q1, we also saw year-over-year growth, both in the desktop and notebook segments. The CAGR over five years revenue has been increasing about 25%. Our ASPs have been increasing about 12%, I'll talk a bit more about that. And our units are increasing about 11% year over the last five years, I'll talk a bit about that.

We've seen growth in all of the key regions. Developed in China, the largest market is up about 43% year-over-year revenue. Last year when we met, we talked about emerging markets, accelerated increase in broadband penetration in emerging markets. As a result, more PCs and more PC gaming. PC gaming has come along with broadband penetration in the emerging markets. Roughly 60% revenue growth, off of a small base. But continues to grow and be a very important part of our business. If you look within our installed base, last year, we launched Pascal, our biggest GPU launch. It's our 10th generation architecture. When I say it's our biggest GPU launch, we launched the entire stack of Pascal in a period of about four months. That's one of the fastest transitions we've had for a new architecture. Yet in the desktop segment, which is where the gamers go first, about 17% of our installed base have moved to Pascal. Still got a lot of work ahead of us.

Success of the GeForce business has really been rooted in a transformation of our business from selling GPUs to selling a platform. And those of you who have been following us a long time, as Jensen noted, including myself, being a part of this a long time, we started -- the company really started as a GPU supplier. We built a chip, we had reference drivers. And we had a referenced board designed. Companies like Diamond Multimedia, CB, Creative Labs would take that and build their own GPUs. From that early beginning, that humble beginning, we moved on to build the entire software stack, to build an entire production worthy board. And in some cases, sell our own board, as we do in the very high end. And we migrated from that to provide a complete -- to really support the complete gaming ecosystem. Tons of engineers working with developers is a key part of that. But the heart of it really has turned into our gaming client, GeForce experience. GeForce Experience was launched in 2013. And it's really designed to be an essential client, the hub of any gaming PC, certainly, a GeForce gaming PC. Our GeForce Experience client now is installed in over 90 million GeForce-based PCs worldwide. It's about a 15% year-over-year increase. 27 billion hours of gameplay have taken place on clients with GeForce Experiences as the gaming hub, 20% year-over-year. So within our gamers, GeForce Experience gamers are more engaged or gaming more than the general population in terms of year-over-year growth.

The first feature of GeForce Experience, to really capture a gamer's attention is this capability to optimize your PC and keep your PC always updated. And anybody that I talk to who was a new PC gamer that has GSE installed, really loves of fact, the elegance, the simplicity gaming on a PC with GeForce experience. The reason is that every game has a multitude of settings. And every PC is different. GeForce Experience has been designed to make sure in a single click, the game settings will be set up exactly for your PC for the best possible experience. 1 click, console-like experience, if you will. On top of that, we added this capability to update your PC for the latest drivers. We developed the capability inside NVIDIA called game ready drivers, with every game, you want your PC optimized to play that game best. So we -- prior to every game launch. And there were 63 titles last year that we had -- we did about a year of QA, created a game ready driver and updated the 90 million PCs with the latest driver prior to that game launch. So gamers don't have to worry about not having the best possible experience. 63 games were game ready driver -- game ready driver for last year. And that's about 3x over the year before. And as we go forward every major game, we want to make sure our gamers have the best possible experience.

As we added features and there's a lot more, every month we're adding new features to the GeForce Experience. About a year ago, we added a feature we called Share. As you know, gaming is very social. We want gamers to be able to record their gameplay and share with friends or stream their gameplay up to YouTube Live, up to Twitch, up to Facebook, wherever there's a live streaming capability so their friends can watch them game, watch their kills. We had over 300 million gameplay recordings occurring on GeForce Experience past year. And that's about a 3x engagement year-over-year as well.

One of the more recent features we added to the GeForce Experience is rewards. We want to become even more engaged with our gamers. So we're offering promotions and discounts and free in game items and free games in some cases, in for gamers who are using GeForce Experience. While we just kicked this off several months ago, we've given out about $1 million in rewards to our installed base.

So that's GeForce Experience. Really the heart of our gaming platform.

Let's talk about some growth drivers now. I mentioned eSports. More and more gamers are coming to the PC platform for competitive gameplay. And when I say eSports, what may come to mind is the big tournaments are the pros and that's really not what I mean by eSports specifically. When you think about football, you don't just think about the Super Bowl or the NFL. You think about all the kids that are playing football, the entire culture, the lifestyle of football. And that's really what we refer to when we talk about eSports. Kids that liked to get into gaming, the competitive gaming, the social gaming nature of the PC have moved into the PC in great numbers. And eSports competitive gaming, social gaming is really what I would call a TAM driver. This segment of the market is what's bringing units in into the PC. And eSports, like any sport, is developing like any sport. If you look at the key attributes of a sport, viewership is increasing, 2x from 2016 to 2020, the audience of enthusiast people who are watching eSports or game content every month. As a result, the advertising is coming along. People are advertising in competitive events. It's about 3x from 2016 to 2020. Attendance is way up at eSports events, as is -- as you would expect, betting. Now there's a whole industry being built on betting on eSports and prizes. Prize money's up to about $100 million a year, almost 2x year-over-year. All the things that are developing around developing eSports, competitive gaming as a real game. You can -- if you pay attention to this space, as I encourage you to do, there's new news coming out all the time. Like in the last couple of weeks, Overwatch, one of the latest entries from Blizzard has a competitive gaming title, announced they have 30 million gamers now on playing Overwatch. That's from a launch of about a year ago. Incredible growth. There's a governing body that's been founded to organize competitive gaming on college campuses. You might consider it the NCAA of eSports in the U.S. So far, 34 universities have signed on with eSports programs within their colleges to start intercollegiate competition and start to develop official and official eSports competition at the university level.

Also, you may have read in the last couple of weeks, eSports is in route to a Olympic sport, believe it or not. The Asian games have announced in 2022, eSports will be a medal sport. And ultimately, that, with continued momentum, will evolve into an international Olympic sport. New gamers are joining the market. This is mobile gamers DSC has projected about a 30% growth in mobile gamers. This is one category of eSports League of Legends, Dota 2 type of a category, from 2016 to 2020. If I look within our installed base, those that are buying GPUs that are motivated by eSports. And that is someone who's buying a GPU and right away, they start playing eSports class titles, competitive gaming class titles, 75% of new purchases in emerging markets in China are motivated by eSports. In developed markets, it's about 55%. The combination of new gamers coming into eSports, the motivation to buy GPUs is really a key driver of our business.

Tournaments as well are getting behind GeForce. You may or may not know it, I suspect you don't watch us too carefully. But there is a season to eSports, just like any other sport. And it is kicking off right now. Qualifiers around the world are underway to take the best teams to the finals with any game. The first one to go live is CS Go, Counterstrike Go. It's usually a popular game in Asia and Europe and soon-to-be released in China. The CS Go finals are taking place in Poland in July. So between now and July, it will be a worldwide tournament qualification. The best teams will go to Poland for that championship. Dota 2 is now live. Dota 2 finals, you may have heard about these finals, it's an epic event in Seattle, in August. Last year's prize pool was $20 million for Dota 2. This year will easily exceed that. Qualifiers are underway. And ultimately, the best teams will end up in Seattle. And finally, League of Legends. League of Legends is the most popular eSports title in the world. Qualifiers start soon. And that's finals happen in November in Beijing. Throughout this entire competitive season, you will see -- we see increased engagements, new gamers coming to play and amplification of interest in the titles.

We also -- if you look at that top 10 tournaments around the world driving hundreds of millions of views, hundreds of thousands of attendees, millions of dollars of sponsorship, all of those (inaudible) count on GeForce to power their GPUs throughout the tournament. And as a result, it's the #1 GPU choice of pros as well. That's eSports. Driving the TAM, incredible momentum, especially true in APAC and China.

Okay, let me talk about AAA content, give you an update here. And what I want to show you is the momentum of AAA content and how it relates to GPUs and gamers basically mixing up and buying up. So this chart, let's see, well, on this side of the stage, is a plot. And I'll break this down for you. It's a plot of all the top franchises that GPU load, the performance of a GPU necessary to deliver 1080p, full HD, 60 frames per second. Over time, the performance of a GPU required to deliver 60 frames per second at 1080p has increased about 2.5 to 3x, from 2013 to today.

The solid line is the average -- the performance of the average GPU purchased. If I look within my entire sales end of gaming GPUs, the average performance delivered by my GPUs falls along that line. That's where my average buyer is. And my ASP, the ASP for that average GPU, equivalent street price. So you can understand what gamers are paying for a board, has increased from $140 to $180. The story here is, is as production value of games increases, gamers need to upgrade their PC, buy up of a high-performance PC, either on a replacement rate -- increasing as part of our replacement of their PC. And the value the GPU has increased to the gamer over time. One cannot argue that the value of a CPU over time has not provided much value to a gamer. The value of the GPU has provided a great deal of value to the gamer. And as a result, they continue to buy up. And even still at $180, if you upgrade your PC, it's far less than you would pay for any console, you're getting an amazing experience and I think as well an amazing value.

Now one point I'll make about this slide as well is that this is 1080p. There's a vast number of gamers who are playing up at 2K and 4K, they play at higher settings. And this is an average who want a better experience even within the franchises. So we can see continued momentum of gamers buying up to get the best possible experience as production value increases.

If I look within my installed base, what % of my installed base needs to upgrade? I think we've talked about this last year as well. But it's an important metric for us. And I guess it's important also, what it is they want to do with the PC. About 60% of my installed base. And when I say my installed base, please keep in mind, I'm not -- I don't have a view into AMD's installed base. I don't have a view into Intel's IGP installed base, which we know, just looking at Steam. And I'm sure you all do, that there are gamers who play on very low end PCs. So this is just my installed base, which is probably one of the highest performance bases. About 60% of my installed base is below 60 frames per second for playing Overwatch. Now I mentioned Overwatch, it's as an eSports title, which is indeed is. But it is the most graphically demanding eSports title.

Average load, if I look at the midpoint, say roughly 2.5 on my load scale, about 80% of my installed base is below my average delivered performance today, by my buyers today. Then Scorpio, we've talked about console in the past and how console is part of that content ecosystem. Game developers want to write for console and PC together. Microsoft announced Scorpio. It's a high-end version of Xbox. It's coming at the end of the year. Scorpio is roughly just under a 4.0 on this scale. That's going to drive PC content even further. And roughly 95% of my installed base is below Scorpio class. Continue to drive opportunity and drive production value and ultimately, keep pushing the GPU business.

More tactically, short term, if I look at the fall, with several blockbuster titles are on their way out, Destiny 2, maybe some of you are console games, I don't know. But Destiny was one of the most important console games that come out in the last two years. For the first time, Destiny is coming out on PCs in September. Expected to be a huge blockbuster for PC, Destiny 2. And we'll be marketing this title heavily. And also coming out this fall, there's a sequel to Battlefront, the Star Wars title Battlefront 2, Call of Duty World War 2 should be very big. And in addition, a sequel to Shadows of Mordor called Shadows of War. It's coming in the September timeframe, which we expect should be also be very big for PC.

Last, I wanted to talk about virtual reality. I believe virtual reality is poised for growth. I think last year, it had a very good year. It was probably not as large as what maybe some of you had expected, maybe others. But we're very pleased with the results. When we survey gamers, 81% say that virtual reality is still very important to them. About 65% say virtual reality is an important metric in their buying decision of GeForce cards. And when we look at our installed base in terms of the GPU that gamers buy to attach to HMDs, they're very high end. They want the best experience. 70, 80 class GPUs get attached to virtual-reality HMDs.

And finally and something I'm very excited about, Microsoft has a new virtual reality product they call Holographic. Holographic is effectively a reference design that they've created. Holographic is a virtual reality HMD. This is not the augmented reality, the AR helmet that you've seen in the past. They have licenses to 6 to 8 OEMs, HP, Dell, Asus, Acer, Lenovo are all going to come out with this HMD in the fall. The price they're targeting is $299. It's incredible value. The resolution is higher than rift in (inaudible). It will take a decent PC to drive it. It will be tied into Microsoft store. And they're underway pushing content to enable and get some traction behind Holographic. I think this is very exciting for VR and could be really an impetus for this market, really start to gain more traction.

Okay, let's see. Last thing I want to talk about is notebooks. I mentioned that our notebook gaming market is growing faster than desktop in terms of % growth. And I want to explain why that is and how exciting it is, this market it is for us in the future. You're going to hear more a lot about notebook from us. First of all, back in the day, when we designed our prior architectures, Tesla, GPU architecture, Fermi, our main mentality was to build a product that was fast. We weren't paying a lot of attention to power. The fastest possible GPU. The problem was there is a notebook market started consumer -- the notebook market started getting going, they would have the force fit a gaming capable GPU into a notebook. And we've shown them to you before. These are big, beefy, effectively desktop replacements back in the Fermi days. In Kepler, it got a little better. But still no gaming notebooks, true gaming notebooks were not transportable. We sold a lot of GPUs in the notebook. But they were entry OEM class GPUs, not gaming capable. And if I overlay this with my gaming chart previously, you'd see that in Fermi and Kepler, we are well below the delivered performance on notebook that was needed to play games at 60 FPS.

Come Maxwell, we totally shifted our strategy. Performance, of course, is incredibly important. But we really got a per watt (inaudible) sensitivity into Maxwell. So with Maxwell, it was the first time we could really deliver a notebook that was portable, that was gaming capable. Now let me frame this a bit, portable. What probably the most important metric in my mind is what would a student take to school, take to university, something that they can use for productivity and they can use for gaming and has a good gaming experience. Maybe we can call that backpack worthy. That's one of my favorite terms. Something you put in a backpack and carry around all day, take it home, pop it open. And game on it. This is a notebook that's about 22 millimeters, less than about an inch thick, under 5 pounds, 15-inch or smaller, this is a constrained notebook. You only have so much room to put a lot of power into. And until we got to Maxwell, we could not deliver gaming performance in a backpack worthy notebook.

Now, considering there's about 150 million college students around the world, expected to grow to about 250 million by 2025, I mean, this is a class of customer that, as we said, are all growing up gaming, that need a notebook that can be used for productivity and gaming. So it's a huge opportunity for us.

With Pascal, we are able to deliver an entire stack from the 60s on up to the 80s in a backpack worthy student type notebook. I'm not trying to create a new market segment called backpack worthy. But I don't have a better name for it. So I'm going to stick with that for now. And as a result, our notebook business has grown about 10x -- starting with Maxwell, about 10x. And relative to the overall consumer market and this is a consumer trend with gaming taking out, which is the small part of it. But the total consumer notebook has been in decline, whereas the gaming market has been growing rapidly. And this is a new class of notebook. It's an important market segment. And I think you're going to hear a lot more from us over the course of the year in terms of notebook gaming -- gaming notebooks.

Okay, the last thing I want to say real quick, Jensen mentioned it on the call and we announced it in CES, our new gaming service called GeForce now, basically a GeForce PC in the cloud. This will allow us to reach customers that are otherwise unreachable for PC gaming like initially on the Mac. We've got several hundred internal employees, including myself, that are playing it every day and really love it. We plan to roll out a external beta, an early beta by the end of the quarter. We've got a waiting list of tens of thousands of people that want to try it. And this is a service, as Jensen has mentioned, that we will be building over time. We're super excited about the promise. But we've got a lot of work to do. But the world will start to be able to experience, I hope you try it sometime later this quarter. That's my update on PC gaming.

Next is Shanker, Mr. Datacenter, who's going to give you an update on Datacenter. Thank you.

Shankar Trivedi

Thank you, Fish. Thanks. Hi, I'm Shanker. Thank you very much for having me here again. I was just reflecting on Fish's 150 million students. And as an Asian parent, I thought this is the most important reason that my college going kid needs to have a gaming notebook because if he or she has one, they have an opportunity to become or a job -- in the hottest job you can get right now, which is to be a deep learning data scientist, right. All those gaming notebooks are the deep learning data science capable today. Okay. As I said, I'm Shankar, Shankar Trivedi. I'm going to share with you around our enterprise business and I'm good share with you our Datacenter opportunity. I'm going to share with you our Datacenter strategy. I'm going to go into three of our large markets and explain why this is a big opportunity, what value we deliver to the customers, why customers keep coming back. And I'll illustrate to you how operates in our industry verticals. And of course, that will tee up very well for Rob Schlanger whose coming up and going to talk to you about automotive.

But let me start by summarizing, we really had a great year in Datacenter. Our business tripled compared to a year ago. We are now $400 million a quarter. And the reason -- there are many reasons why we had a great year. First and foremost, we -- for the first time, we launched our own AI supercomputer, called DGX. DGX-1 was the first generation. Today, Jensen introduced the second generation. And we had our first DGX user group at this conference, for the first time ever, we actually had a user group, focusing on all the goodness of the software that goes with that supercomputing AI appliance. And many of our customers are presenting their DGX story. I think at about 4:30 one of -- I think it's Facebook is presenting what they're doing with their DGX systems in Facebook AI research, it's about 4:30, which is really good talk you should go to.

So we also launched the Tesla Pascal platform. And Tesla Pascal had a fantastic year. I just looked up -- we have a tracker of all the OEMs and ODM servers, right now, we have 506 datacenters servers available from 43 OEMs and ODMs worldwide, all supporting the Tesla accelerated Datacenter platform in their servers. And it's amazing, just five years ago, we could count on -- we had barely a handful of OEM and ODM partners that used to do these accelerated computing servers. The main reason for the drive and growth is we have more and more applications, which are accelerated. So today, our accelerated applications catalog is about 450 entries. And what you'll see is as you go around this conference is how many enterprise ISVs are actually showing and talking about their GPU accelerated applications. And these applications are what drive demand. Then of course, as more and more people consider, should I go on-prem or on to the cloud, GPUs have now become available in every single public cloud service, whether it's a small public cloud like Olympics that's running a boutique for high-performance computing, or the largest cloud service providers, such as Amazon, AWS, (Macwood) was with Jensen on stage a short while ago.

And what I'm seeing more and more is enterprises are adopting AI accelerated computing. You can't go into a meeting now with any customer, any large corporation, any large institution, any government department and not talk about AI. I think this week's Economist had a really nice shot, showing the number of corporations that are talking about AI in their earnings releases. It's something like 400 companies are talking -- I don't know what they talk about. But they apparently mention AI is a really important part of their strategy. So AI enterprise adoption is really driving. And what I'm sharing with you on the right side for the first time is the growth in our number of customers. So when we started accelerated computing, only a handful of large supercomputing centers and research laboratories, the bleeding edge innovators, the people who have such a big problem that they have to have the bestest, fastest most flops in their computer, only those people bought a GPU accelerated server, a GPU accelerated system. But now, we have 8 times as many customers. And these are by the way the large customers that we can track because we can track all of our customers. So these are the larger ones. And what's fascinating to see is the breadth and depth of the kind of customers who are buying both our appliance and buying servers from our OEMs and our ODM partners. So you can see over there, about 8x growth over the last few years. And most of them are in large enterprises. So that's pretty interesting for us.

So we had a great year. And I want to summarize why though, why, what's the reason people are adopting this accelerated datacenter for their datacenter requirements. And so, it all starts with the application. It's all about, as Jensen said, you have to accelerate an application. You can't -- and then you can't subtract value from the CPU. The worst thing you can do is let it just -- it all works fine on your existing CPU, whatever that might be. But what we do is we accelerate the application. And today's GDC that you see all the few thousands of developers out there are creating that next generation application. Then they in turn, go and join corporations and cloud services companies and those companies come out with software and cloud services that customers use, right? So what I'm showing you is how it started with the top, the high-performance computing applications. These are the ones that are solving the weather and earthquakes and tsunamis and genomics and molecules and chemistry and quantum physics. All of those applications started to get accelerated. And today, all 10 of the top 10 high-performance computing applications are all accelerated. And what you're seeing that's happening in the area of AI. And then AI is kind of like data analytics because you're looking at lots and lots of pieces of data. And so the data analytics is and the AI framework, the data analytics companies are all embracing this accelerated computing. What you saw on the keynote, SAP's brand analytics. It's analyzing huge amounts of data, training our neural network and then allowing customers to infer how much at their advertising was worth for that particular proposition. Okay. So SAP is out there. Then it's interesting how many other enterprise companies are out there. So these developers are solving difficult problems, whether they are simulation problems or data analytics problems. Then they apply these problems to industry solutions. So we're right now in 2017. One of our first customers, an oil company, came to us and said we have this algorithm called reverse time migration. And according to their calculations, this was in 2009, they said that in this year, 2017, they would deploy reverse time migration to search for oil under the sea, under the salt, that's way down, 7 miles under the surface of the sea. And we were able to work with them by optimizing their application. And we delivered it into their production datacenter in 2013. So that's an example if you take a tough problem and you find a solution working with the developers and then you implement it in an application today, it's called Schlumberger Omega, which is the application for their on-premise, on the cloud, working everywhere and it's delivered to their customers in the oil and gas industry. And so we go into this approach with this horizontal platform into industry verticals with solutions, solutions for governments, whether it's in defense or security or cyber security. And so on, solutions in the Internet services companies, whether they're the largest companies or smaller companies like Pinterest or Dropbox or (ChingDong) which is China's second-largest retailer. So these are the verticals that we go into. And once you make that solution, once you have an industry solution, then you have to make sure that it's available everywhere. And today's people want a choice. They want to either buy a data center system or they want to buy a service in a public cloud. And as I said, it's available everywhere.

So that's kind of a strategy. That's what we do. strategy is an action. And let me show you, how it works in these 3 major categories of accelerated Datacenter. So the first is the classic one, which is accelerated, high-performance computing. And the best way to think about this market is the top 500. And today, there are about 1.2x exaflops installed in the top 500 systems. These are the Oak Ridge National Laboratory, The Riken National Laboratory, China's largest supercomputing center. And so on and so forth. There's 500 of these. Because of CORAL, which is going to be our nation, the United States' fastest supercomputer, we are predicting that by next year, there'll be 1.7x exaflops, 1.7x exaflops and so 1.2x in this top 500. So that's one interesting data point. And if you then think about all of the nations that said, this is -- we are all going to exascale. Every continent, every country, every subsegment of high-performance computing, whether it's weather, earthquakes, climate and so on, all want to need to do genomics, needs to reach exascale. So we postulate that maybe the first exascale computer will be somewhere in 2020 to 2022. Then a few years later, there'll be more exascale computers. You can work out that somewhere in the -- by the year 2020, there will be about 8 exaflops of computing in the top 500. Then, of course, you have to calculate what will be the number of processes in there and what will be the price paid for exaflop of computing. And it turns out, my best calculation at this point of time is about $4 billion. Last year, I said to you, that this market was probably an opportunity of $3 billion for NVIDIA. I'm now convinced, based on more applications, better value being delivered, more reach and indeed, when you look at it, the problems are getting bigger, the problems are getting bigger. And so I think that our opportunity is about $4 billion. And why do people buy an accelerated system? It's molecules convened to be modeled faster. Now there's 273 nanoseconds of computing that's going to be done on an Amber supercomputer. Whereas previously, it used to be 7 nanoseconds. So this is a protein interacting so we know how that protein is going to interact with your body. This is a classic JAC benchmark. And you can see that on a CPU-only, non-accelerated system, 19 racks of computing can be done with 1 rack of our Pascal P100. In other words, the savings are $13 million per instance. And these savings are just huge for our customers. The reason there's not a single high-performance computing RFP, request for proposal, out there that does not require some number of accelerators or GPUs. Many of them now specify 100% acceleration because they can see that in the future, more and more applications are accelerated. And so if you want to buy something that's going to last you for 3 to five years for your research community, it makes sense to go all-in on acceleration.

I'll give you an example. The National -- our National Center for Supercomputing Applications, NCSA, which is hosted in the UIUC, University of Illinois at Urbana-Champaign has the Blue Waters. At that time, which was 2012, they decided, out of that 27,000 nodes, they would have 5,000 nodes accelerated with 1 GPU, 1 GPU per node.

And sort of now -- well, it turns out that we have saved them 20 million hours of computing that would otherwise have been done on a GPU.

Now here's the interesting thing. Had they, at that point of time, decided to go for a richer number of GPUs in NCSA Blue Waters, it turns out, instead of saving 20 million hours of computing, they would have saved 50 million hours of computing. And so that's the amazing -- even now, there are some applications that run on CPU only on that system. So obviously they need CPU-only. But the beauty is, over time, the value of the supercomputer actually increases because more and more applications are accelerated.

So what's interesting in this market right now is every single high-performance computing center is embracing AI. RIKEN, Japan's National Laboratory, it hosts the third-fastest computer in the world, it's called the K computer, 24 racks of DGX-1s to run an AI supercomputer. It's the fastest currently available facility in Japan for the AI research community. And Japan has made a huge commitment to improving their AI research base. TiTech, which runs the TSUBAME supercomputer, just announced their third generation. So they've always been with NVIDIA GPUs. That third generation will be Japan's fastest AI supercomputer and probably the fastest AI supercomputer in the world when it's installed very, very shortly.

And the reason I put up SATURNV, which is our own supercomputer, it's #28 in the top 500 right now. But the most important thing about SATURNV, besides being a really, really good AI deep learning machine, is it's #1 on the Green 500. And it came within a whisker of the magic number of 10 gigaflops of computing per watt of energy. It is simply the most efficient, energy-efficient computer on the planet. And that's really an amazing result.

Then last but not least, an example of what's happening in high-performance computing and why AI is everywhere in high-performance computing is CANDLE, which is the cancer deep learning environment set up by our national laboratories with the NCI, the National Cancer Institute, dataset. So this dataset, together with a deep learning framework, is being made available to all of the NIH researchers to help us find a cure for cancer faster than the traditional way. So deep learning can help cure cancer. And so that's the reason why all of this segment, which is a very important segment for us as a scientific research community, is embracing deep learning and artificial intelligence.

So now let me talk about the opportunity for deep learning training. And so deep learning training is everywhere. Alexa, which is Amazon's assistant, is trained with NVIDIA GPUs. Cortana is trained with NVIDIA GPUs. Google Voice is accelerated by NVIDIA GPUs. And so all of these neural networks, which have now all gone superhuman -- Microsoft ResNet took image recognition on ImageNet to superhuman accuracy. Baidu Deep Speech took speech recognition to more than superhuman accuracy. And Google's NMT, neural machine translation, has taken simultaneous machine translation to superhuman accuracy. All of these neural networks are trained using NVIDIA GPUs at the heart. And then one of the frameworks that people talk about, whether its Cognitive Toolkit or Caffe2 or MXNet and so on.

And the way I think about this is it's all about training data. And the data happens to be in the new way we think about data as images and of video as sort of 30 images per second. And of course it's in motion and all sorts of interesting things happen as a result of things that are going on in a video. Then there's speech and speech to text. Then of course speech becomes with phrases into language and language has context and meaning. And of course, when a -- in medicine, every doctor is trained to dictate their notes. And doctors' languages -- I mean, both of my parents were doctors. The doctor's language is dramatically different to the way you and I speak to each other. And so the neural network for medicine has to be trained in a different way. And that's why people say, "Hi. isn't all this training just one-off?" Training is all the time. Training is domain-specific. Training is going to go on forever as the datasets get larger and larger. So the question is, how big is this opportunity?

And by the way, as Jensen already explained, we have taken training times down in this instance from sort of -- we always say it's from -- it used to be from months to days. And now it's from weeks to days. And it's going to be weeks into hours. Less than a shift to train a humongously complicated neural network is now completely within reach.

And so the training value proposition ultimately, in this example I worked out, if you were using a CPU-only system, assuming it could even do it, right, it would take you about 500 hours more to do that training than doing it on a Pascal-accelerated GPU server. And so you work out, okay, what's your -- what are your data scientists worth? Well they're worth a lot. They're the gladiators. They're the people who are going to get the best jobs. And so 500 hours of their time is worth a lot of money. And it's a huge improvement in their productivity and helping enterprises get these trained neural networks to market in whatever shape they're looking for.

So when I look at sizing this market, I sort of said, well, right now I can see here are all my customers, here's the kind of -- they're buying about 1.4 exaflops of training computing from us right now over the past few years. And based on the growth curves that I see, in terms of how people are using these neural networks and how often they train, we can predict something like 55 exaflops of training will happen by the time we get to 2020.

And I can size that market in a number of ways because I have to assume how much -- how many flops will be in the processor, how many processors per system, et cetera, et cetera, what the average price would be. And you can all make your own assumptions. I could have sized this at $16 billion. I could have sized it at $13 billion. But I was conservative. And I thought, okay, I think at this point of time, our best forecast on where this is growing is about $11 billion. And this is based on real data, real adoption rates of -- and of course you have to assume that some people are early adopters, some people are opinion leaders, some people are early majority. And then some people are late majority. So there's some sort of adoption cycle that goes into this model. And so we think it's about $11 billion of market opportunity for NVIDIA.

Then the new market that we are now seeing is the inferences that are happening as a result of the trained neural networks. So every time you say, "Okay, Google," you want an answer really quickly. Every time -- Amazon has Alexa. JD has DingDong. All of them have these bots. And you want the AI to give you an answer in a relatively short time. You don't want to be waiting for hours. So the new thing is all about real-time inferencing, how -- and of course, the real-timers have to assume some network latency. Once you say something into your phone or you type something in your pad or your digital assistant or whatever, it has to go up, go to the data center, come back. Sometimes these things get done on the edge. So it's pretty complicated.

But as I -- the simplest way to think about it, right now inferencing is all happening on conventional servers, non-accelerated servers. Some of the people are using FPGAs. But the vast majority of it is in non-accelerated servers. But as we saw with TensorRT, the value proposition of TensorRT, together with Pascal or, indeed, Volta, is so much more powerful that it's, I think, starting to get disruptive.

So Jensen shared 15 to 16 racks replaced with 1 rack. With Pascal, 12 racks replaced with 1 rack, okay, to do inferencing within 7 milliseconds for a trained neural network. And so when you look at this thing and you say, "Hi. I could save you $2 million the next time you're looking at inferencing servers in your data center," that's a very, very strong value proposition, for people starting to adopt inferencing using accelerated computing.

And the capabilities are not just the traditional voice, image, textual search, or speaking. I already gave you the example of sort of medical dictation or lawyers dictating their briefs to -- for their customers. There's also a huge play in video analytics. And Deepu will share more about this when he talks about the implementation in smart cities for safety and security and so on. So this is a very, very large opportunity, traditional hyperscale servers.

Oh, the dataset. So here's the way to think about it. 3.6 trillion searches on Google. 3.6 trillion searches. Andrew Ng says 25% of those searches are going to be voice and image searches. And that's going to increase over time, right? So you then say, okay. So how many searches are there going to be? How many inferences are there going to be in the future? And we all know. You've all seen the explosion in Facebook Live. You've seen the explosion in terms of number of photos uploaded, the size of photos, the quality of photos. The datasets, I put it to you, are getting bigger and bigger. By the time we get to genomic datasets for personalized medicine or targeted therapies, these datasets are going to become way, way large.

So I did a calculation. I said, okay, right now, it seems to me that something like 50 exa-inferences -- 56 trillion inferences per second based on Google search is 3.6 trillion. And I scale it out to other companies and so on. And so 56 trillion inferences per second. And we think, based on the datasets, the sizes that are growing, it's probably going to go up about 10x that, to around 500 trillion, 450 trillion is what our best estimate was at this point of time.

Then once again, you need to work out, well, how does that equate to in terms of racks? Obviously we're going to have way fewer racks. These are going to be inferencing at way faster speeds. The latency requirement is probably going to go lower and lower because of the network build-out. And so I conservatively size this at about $15 billion. Fine.

So we think inferencing is a very, very large, new opportunity for us in the data center. Our first customers are already deployed. Our Pascal P4 for inferencing, it's a great product. Yesterday, I did an announcement with iFly Tech. They are doing the speech inferencing, the Chinese-to-English translator, their little speech assistant. And they found that it's 50x better on a P4 than their current CPU-based inferencing service that they have. And of course, iFly Tech has 1 billion users of their service. And their service is embedded in the self-driving cars that people are putting in. It's embedded in DingDong, which is JD's AI assistant, JD's Alexa and so on. So they're kind of like an Alexa platform for speech recognition. And they've already deployed P4 into their data center.

So where could this take us in terms of industries? And as I already shared with you, we have people in all of these industries. Their job is to work with the key influencers to discover the most difficult problems and apply accelerated computing to them.

So I wanted to share with you where we might be in terms of health care and what that opportunity is or could be. And so right -- the first thing to think about is detection. All of us have MRIs or CTs or ultrasounds. And the world of medicine is full of images. And there are so many start-ups and so many companies that are looking -- whether it's an ophthalmic image or an ultrasound, they're looking at how do they solve -- apply deep learning to solve this imaging problem. And the use case is detection. All we're looking for is the traditional thing everyone talks about in deep learning, which is anomaly detection. You'll hear this word everywhere: anomaly detection. In the case of imaging, it's detecting where is the issue and assisting the radiologist or the doctor or the clinician to identify and show them the possible anomaly.

And GE Ultrasound has been a great customer of ours. The cardiac group which does the EK -- electrocardiograms, this guy called Eric Steen, he's been at our (GTUCs) for the last nine years. He's like a CUDA Fellow. And he decided to apply this technique, the deep learning technique to identify for their imaging that are coming off these echocardiograms to allow the technician to get a -- what's the best image to present to the doctor. Because when you're on the treadmill and you're doing this echocardiogram, there's a lot of noise in these images. And they've got to identify what's the best one to show to your cardiologist as part of this test. So this is an example of detection.

An example of diagnosis is Massachusetts General Hospital. They are working to basically figure out how do we better diagnose prostate cancer? 1 in 8 men will get prostate cancer, are likely to get prostate cancer. That's a huge image scanning problem. How do we figure out -- and there's just not enough radiologists to do this job. In the United States there's, I think, 1 radiologist for every 100,000 people. So 6 for 1 million. I think that's the number. I may -- but in countries like India and China, there's way fewer radiologists per person. And we need to scan everybody. And the radiologist needs to be able to know how -- what's the likelihood that this person might need to be tested for prostate cancer. It's way cheaper to do imaging than to do biopsies and so on. So we want to do that. And MGH is working on this dataset and they're trying to solve this important problem of diagnosing prostate cancer better.

Then finally into treatment. Deep Genomics, which is a really interesting start-up, is taking this whole problem of the gene -- a gene is 3 billion base pairs. It's a lot of data. And many countries, many hospitals are now fully sequencing the genome so we get a better dataset. And so Deep Genomics is applying this -- their knowledge of deep learning to these types of complex datasets. And the outcome will be a targeted treatment.

Yesterday, GlaxoSmithKline, which is one of the world's largest pharmaceutical companies, they were talking about their (ATAM) program. So they have taken years and years of historical data that Glaxo has had, that all the clinical trials, the times they've tested all these molecules on animals and people and so on and documented it in a recorded dataset of 2 million case studies that they fully documented and curated and now making it available through the national laboratories. And the whole game here is to reduce the time it takes from a target molecule to a candidate from 5.5 years to 12 months. So that's what they are trying to do. And we think -- there's $1.7 billion that's already been invested in health care start-ups. PathAI is working with Philips Healthcare to take digital pathology to market. So start-ups are working with large corporations. Start-ups create the innovation, go into large corporations, they industrialize it. So we think health care is a really, really big market for us in the future. And it's just starting.

So lastly, where are we in terms of momentum? And as I said, it all starts with the developers. You've seen -- our developer base has grown from about 50,000 to over 500,000 people. The number of applications, when I first -- when we first, Ian and I, we kind of created this GPU-accelerated catalog, we decided -- it had 100 entries when we first published this catalog. That happened in 2012. five years later, we had nearly 500 applications. We'll probably cross the 500 pretty soon. So 5x as many applications. And what's interesting is, a year ago, we didn't have a start-up program. But we were already engaged with about 300 of these AI start-ups. And now we have 1,300 of these start-ups enrolled in our start-up program. We're working with them, taking -- on obviously accelerating their applications. But more importantly, taking them into the enterprise ecosystem so that they can monetize their offerings in a better way.

Then lastly, our number of go-to-market partners, we call it NPN, the NVIDIA Partner Network, has grown from less than 300 to over 2,100 in just a year. So we've really got strong, strong momentum in our go-to-market program. So everyone wants to be part of NVIDIA's go-to-market activity.

So to sum up, yes, we had a great year. But I think our opportunity has become bigger. It's become bigger because HPC has converged with AI. It's become bigger because the AI training opportunity has grown because more and more datasets need to be trained. Every single enterprise will be training multiple sets of data. Some of them will run it on the cloud. Some of them will run it on-premise. I don't yet know. Right now, market analysts are saying 50-50. I think more will go to the cloud than stay on-premise for this type of activity.

Then inferencing is a very large new opportunity.

And so overall I think, compared to a year ago, our opportunity has grown to about $30 billion for 2020. And our momentum, I believe, with the new Volta, the new DGX Volta, the new TensorRT, our GPU cloud service now with the secure containers, allowing these developers to package their applications in a nice, secure container gives them an even bigger opportunity. And so we're seeing all these enterprises -- large enterprises, small enterprises -- adopting AI. We have a really, really good optimized stack which is available on every cloud provider through all of the data center server OEMs and ODMs. I think we are in a very, very strong position for the future.

Thank you very much. Oh. And please welcome Rob Csongor, our head of -- SVP of our automotive business.

Robert Csongor {BIO 3210739 <GO>}

Thanks, Shanker. Hi, everyone. This morning you saw in the keynote we talked about the fact that we're developing a complete computing solution stack for automotive. And we outlined some of the opportunities for us. We outlined some of the partners that we're partnering with. But one of the things that I want to point out to you, one of the reasons we have our Investor Day here at GTC is that you have an opportunity to go out and actually see the grounds up or see what is actually happening beyond slides or beyond things that we say.

So let me just give you a sense of what's happening here at GTC in terms of how important is this solution that we're developing. Not only are we developing it as a full stack. But it's open. So how important is it for other people and what kind of activity is happening on top of NVIDIA's automotive solution?

So just to give you a sense, there are 245 separate companies that registered themselves at GTC as automotive companies who are here specifically because of the work that NVIDIA is doing. And they are all using the computing solution, the AI stack and the fact that it's open to build on top of this to solve a myriad number of problems.

Just to give you a sense, there's 35 vehicle makers here. I call them vehicle makers because it's not just carmakers. There's 20 carmakers. But John Deere is here. They're looking at how to develop autonomous vehicles that can use AI to harvest a field more efficiently. Harley-Davidson is here. There's a whole bunch of start-ups that are here. The number of vehicles in the final mile and autonomous vehicles within a factory -- trucks, Paccar, DAF -- all of those are looking for autonomous solutions to solve a particular problem and to do it more efficiently and more safely.

There's 24 Tier 1s here. There are 37 start-ups. 10 HD mapping companies. And all of this in 42 presentations. If you actually went and sat on a chair, you just would not be able to see all the automotive presentations. So this kind of gives you a feel of the importance of the work that we're doing and how does it affect the outside world and the value of what we're building to the outside world in building on top of it.

If you take a look at some data, beyond the revenue that we accomplished in this past year, which is still historical, of course, mostly in nature, it's mostly infotainment, the significant data on the right is that, in this past year, we launched our DRIVE PX 2 platform. And you can see the growth of the platform. And fundamentally, when people started using DRIVE PX, a lot of people were working with PCs in a trunk. They developed a CUDA. Then they take that software and they port it to a DRIVE PX 2.

But of particular note, you can see in the last chart, in Q4 of FY '17, we introduced the TensorRT AI compiler to the automotive industry. And what you see is the number of companies developing with AI almost tripled, went from 60 to 170 different companies, people who were basically looking at, "I need a tremendous amount of computation. And I need an AI solution that I can develop on and start doing serious work on." And what you're seeing is the explosion that occurred as soon as we released that technology.

And with DRIVE PX 2, as you know and we've talked about before, we have 4 processors on that board: 2 Parkers, 2 GP106. Powerful processors. All of that, at the end of this year, is replaced by one single Xavier SoC. So the platform that we build and the work that's going on is just amazing. It gives you an indication of the type of things that matter.

And I think what matters is, I think everyone has universally recognized that autonomous vehicles, different from ADAS, autonomous vehicles require several things. First, conventional code is not sufficient. It requires a completely new computing model to achieve the goals of an autonomous vehicle. And for that, you need deep learning.

Secondly, the reason that there are so many HD mapping companies here at GTC is because the process of creating an HD map is very difficult, requires a tremendous amount of point cloud processing, requires a tremendous amount of AI. And as a result, generating the HD map is something that NVIDIA uniquely can do in the automotive industry. And the HD map is essential for a self-driving car.

Then finally, the reason DRIVE PX 2 exists, the reason why we are building Xavier, the reason why we need Tensor cores and the architectural breakthroughs that Jensen introduced this morning in a small footprint that's available for everybody is that level of processing is required to take care of all of the computation and all the tasks and algorithms that are required for a self-driving car. Okay?

The reason, when you break it down, is that a self-driving car just is extremely hard to do. Starting with perception, the amount of objects that you have to detect, whether it's signs, lanes, pedestrians, landmarks. The fact that you have to process all of that data. You have to reason. You have to do something with the data. You have to path-plan. You have to drive. You have to localize yourself against an HD map. Then in all of this, you have to do a lot of inferencing and AI computing. So AI, we believe, is a requirement to get to the level of an autonomous vehicle.

Now if you look, one of the most common questions we get is, "What is the difference between what you do and what solutions are out there today?" Fundamentally, what exists today on the road is ADAS, assist -- advanced driver assistance systems. But the next step beyond that, the first step of an autonomous vehicle would be a Level 2-plus or Level 3 vehicle, which fundamentally goes much further beyond detection. Aside from detecting, you have to localize your (inaudible). You have to create an occupancy grid. In other words, you have to take all of the objects that you can detect and then create a 3D map around the car so that you know what you have to do next. And that is, of course, navigate your way or path-plan or drive your way through there.

You have to take into account vehicle dynamics. The computation of how you drive is extremely complex. You have to take into account the weight of the car, the egomotion. A Prius will handle very differently than an Audi. So all of these factors have to be taken into consideration. And of course, this vehicle, it will just simply not be possible to take advantage of the benefits of AI to be able to gather data and to learn without the capability of having over-the-air update. And as you saw previously, one of the advantages of NVIDIA is, because we are engaged in so many different industries, we can take the best of what we learn in those industries and leverage them into automotive. So as a result, the fact that we maintain an over-the-air update system to 100 million gamers in GeForce means we understand how that system needs to work and we can apply it into automotive, just like we do inferencing, learning, training, all of those things.

Beyond Level 3, Level 4 introduces an entirely new level of complexity. It introduces the concept of fail operations. If something happens to the computer, steering, braking, all of these things have to have a backup, a fail operational solution. It requires a tremendous amount of autonomy. Basically the driver is not required to be paying attention. And that means that the computer has to be able to drive and maintain autonomy. It has to be able to handle situational awareness, like construction scenes. That's why at CES, when you saw the demonstration of our technology to navigate through cones or construction signs that are pulled in front of the path of the car, these are all things that AI and deep learning and different kind of algorithm redundancy is of benefit and necessary to implement in the car. Then just imagine the complexity of a typical urban environment, the amount of objects you have to detect, the kind of complex path-planning you have to do.

And all of that requires a tremendous increase in the amount of processing required. So it is the reason why we invested and why we started investing many years ago in the Xavier chip. The Xavier chip, which uses the Volta GPU core that Jensen talked about this morning, is required, we believe, to deliver the kind of processing. And when we're talking about this kind of driving, we're talking about processing on the order of 30 trillion operations per second to achieve this level.

And beyond self-driving, our vision is that if the car is self-driving or if you are driving, the AI is always on. So the idea that you have an AI car, a car with a copilot or guardian angel that is running a number of applications within the car to monitor the driver and to make sure that the driver doesn't go into any harm. So applications like face tracking, head detection, lipreading, gaze tracking, all of these things are algorithms that we believe will be running in the AI car.

The end result is we believe the requirement for a tremendous amount of DNNs, deep neural networks, all running together with a tremendous amount of processing to make the whole thing happen.

Now one of our strategies. And as we talked about this morning, is that in order to do a system like this and one of the advantages of NVIDIA's approach is that we are delivering an end-to-end platform. In fact, everything that Shanker talked about on the training side is something that we supply. So for example, the typical flow for developing a car, self-driving car would start with survey cars that have a DRIVE PX in it which are collecting data. And you are taking that data, bringing it into a data factory, where you have to label the data, you have to annotate it and you have to do it properly. There's a lot of know-how and expertise on how to train a DNN efficiently. And from that, you're going to use it to train a deep neural network system. There's opportunities, of course. And you're already seeing automotive companies now buying DDX systems to do that training. Also the mapping companies who are now doing map processing to create those HD maps that self-driving cars are using. And at this point, I think we've announced with every single mapping company a partnership where we are working with them to solve this problem.

All that information and the result of the training have to be OTA'd through an OTA server. Then the final result goes into a car where you have a computer and a full software stack which is doing the things I mentioned earlier: detection, localization, path-planning, driving and the AI.

This complete end-to-end platform is something we have to develop and we have to use to develop our own stack. But it is something that the ecosystem and the world also needs in order to do it. And we uniquely have the expertise and the products to do it.

Aside -- once you're in the car. And as we've talked about, not only are we developing that full and open platform, not only are we developing that full platform. But it is open. So we offer a variety of ways that you can build on the platform. At the lowest level, you can interact with APIs and write your own. You can develop algorithms directly to CUDA (inaudible) and then use TensorRT. Aside from that, NVIDIA has developed computer vision libraries which are optimized for our hardware. And you can build on top of that, what we call the DriveWorks SDK. So imagine the entry point to the way to build on the platform is the APIs, the DriveWorks SDK. And then you develop AI applications on top. And there you're doing applications like localization, path-planning. You're doing detectors, lane detection, those kinds of things.

Our road map includes, of course, the Level 4, Level 5 with Xavier, DRIVE PX 3. But also DRIVE PX 2 with the Parker solution, which is our Level 2-plus and Level 3 solution. The 2 of them together are a Level 4/5 solution. DRIVE PX 2 with Parker available now. And then at the end of this year DRIVE PX 3 coming. And the DRIVE PX 2 development system is the development solution for that.

We announced a number of partnerships this morning. You heard about, of course, Toyota. I think one of the significance of the different announcements that we've made is that you are seeing adoption of AI and these kinds of systems by companies that regard quality and safety as a fundamental thing. These are not companies who are given to speculative risks. And a lot of these companies, like Bosch and (Zetef) and Toyota, are companies that have the capability to take the technology and the capabilities that we have and have the scale to make it available anywhere in the world. And we believe that this is -- these are powerful indicators of our future success and the growth and the valuableness and the requirement of the technology platform that we've built.

We've also announced, as I mentioned, with just about every HD mapping partner out there, who also understand the requirement for what we do in order to deliver the HD maps.

So all of this, if you look at the opportunity, there's a number of different ways you can slice it. But fundamentally, we view our opportunity as the segments of the automotive market where artificial intelligence and high-performance computing are required for that market to exist. So within the segments that I talked about, ABI

Research numbers, targeting numbers at about 5 million for Level 4/5 and 15 million cars for Level 2-plus and Level 3.

Basically, Level 4 and Level 5, just to paint it for you, that type of a car today is a car with 7 PCs in the trunk. There is no room in the trunk for anything, okay? The idea of having something with fail operational capability, a tremendous amount of processing, requires that you have multiple processors and an enormous amount of software capability that's operating in that vehicle. The Level 2-plus and Level 3 is something that is capable right now with the Parker solution. And those are types of designs that we are capable of delivering to right now.

The AI car capability right now, conservatively we assign it as roughly the same level as the Level 4 or 5. But ultimately we believe that AI car capability, those capabilities that we demonstrated in our demos this morning, are the kinds of capabilities that we believe will have grand -- lots of application. If you look currently today and if you drive a car that has advanced driving capability, traffic jam assist, it is an extremely compelling solution. The ability to have the car drive you, even if it's just in traffic, is extremely valuable and compelling. And the attach rates of those type of systems within the manufacturers are very high. And we believe that will be something that will matter a lot in the attach rate and the overall opportunity. Okay?

So the key takeaways, I believe, for the automotive opportunity. First of all, autonomous vehicles, we believe, require AI. NVIDIA's position, I believe, in all of these, as I think you understand, is we are at the center of AI. And as a result, if you look around GTC and you look at all the work that companies are doing, that's why NVIDIA is at the center of the work that they are doing. Our strategies are to build and deliver this end-to-end system. And you're already seeing that, auto companies building and buying on DGX, building their deep learning systems; mapping companies using NVIDIA to do the HD mapping processing; and fundamentally, companies coming to events like GTC to take advantage of the know-how and the expertise that we have been able to deliver to them so they can solve the problems they need to solve.

We are building a full and open software stack on top of a powerful processing platform. That is the solution. And that is the product that we're delivering to the market. The advantage of it, it's one architecture; one software interface; scalable from a Level 2-plus, Level 3, all the way up to a Level 4, Level 5.

Then it's easy to say but very, very hard to do, developing an ecosystem, developing SDKs, working with all the companies you see here at GTC, it's easy to say but very, very hard to do, Took us a long time to get to the point that we have now. But if you look around. And you can get a feel of the type of energy and development work that's going on top of our work, I think you can't do anything but conclude that what's happening and what NVIDIA is doing in automotive is something very important to a lot of companies' efforts to develop these kind of autonomous vehicles. Okay? Thank you very much.

Oh, I'd like to introduce Deepu Talla, who is going to talk to you about AI cities.

(presentation)

Deepu Talla {BIO 18035782 <GO>}

Okay. Good afternoon, everyone. So video today is, by far, the largest generative data. Approximately 75% of Internet traffic is video. Across many cities in the world, there's hundreds of millions of cameras, traditionally for surveillance purposes, that are capturing data and storing them 24/7. These cameras are deployed in government buildings, traffic intersections, transportation hubs like airports and train stations, malls, offices, schools, you name it.

Now many of these cameras are deployed for safety reasons first, traditional public safety applications like video surveillance, law enforcement, forensics. But we're also seeing them being increasingly used for efficiency or smart city reasons, or what people call smart cities, which essentially is using the same video cameras but use them for things like traffic management, retail analytics, how is the foot traffic in a mall or in an intersection. And also optimizing the resources.

Now an AI city is both a safe city and a smart city. Using the power of deep learning, we can actually take these billions of video frames, anonymize it and then transform that into meaningful insight. And that is what an AI city is about.

Now as I mentioned, there are several hundred million dollars cameras that are already deployed today. Another half a billion cameras are going to be coming online in the next 3 to four years. So we end up with approximately 1 billion camera streams that are 24/7 by 2020. Today, what's happening with those camera streams basically is that they're going to a video recorder. And a very small fraction of those video streams are being watched by human beings in extreme sensitive safety areas. But otherwise, they're basically going into a recorder. And if and when something bad happens, somebody goes in and reviews it after the fact.

Now if you want to solve this problem using humans, you can imagine, 30 billion frames a day, it basically takes 3 billion 8-hour work shifts. So it's not really a practical problem. Now in the past 10 years or so, video analytics was a solution proposed to solve this. But the problem with video analytics in the past has been that, because of challenging real-world conditions, the accuracy was never really good enough. It's so much inferior to human beings that a false alarm is actually a bigger issue than really trying to solve the problem.

And now we power up deep learning and AI, as you've seen in many, many industries and applications. We can use deep learning now to solve this problem. And so achieving an AI city, both a safe city and a smart city through the power of deep learning.

Now let's look at a -- what an architecture of an AI city looks like. It turns out that we actually need an edge to cloud architecture to solve this problem. Now here I show you in this picture -- first of all, there's cameras installed everywhere, right? So you could have a body camera on a police officer. Or this could be a camera that's deployed in a parking entrance, a garage that's basically doing your building, looking at your license plate. So it would be nice to have some of these cameras actually be AI-enabled.

Now then take it one level higher. Imagine an airport. You've got hundreds of cameras, maybe even 1,000 cameras. And there, you'd want to actually capture all of the data together. And in these sort of applications, what happens across cameras matters, right? For example, imagine you're looking at it from this viewpoint in a camera. And then what's happening 100 feet away from you is probably relevant to what's happening in this viewpoint. For example, a lost child, you want to track them, it has to be done across cameras.

Also equally important is across time. What happened 5 minutes ago is probably relevant to what's happening right now. So because of that, you actually have these powerful, on-premise servers or appliances that are going to be taking these tens and hundreds of cameras and doing an analysis. And that's an on-premise server.

Now when you take it to a city scale all the way. And basically now you're talking about a cloud infrastructure, because you're getting this data from across multiple precincts or blocks. And if you want to do a full public safety, like a city center, looking at all the different areas. And then if you want optimize resources across a city, what the traffic patterns look like or something -- if there is a sort of safety reason, something bad is happening, you would be able to look at alternate routes.

So to really solve this problem, you need AI, not only at the edge. But also in the cloud and also in an on-premise server. So we've announced earlier this week a platform for this. It's called NVIDIA Metropolis, wherein we apply the different NVIDIA products that we have and across the board, all the way from edge to the cloud.

Now you've heard about this earlier in the talks and also you obviously know by now, there's 2 pieces to AI and deep learning. First piece is training or creating the neural network. And that's happening typically in the cloud. We have DGX-1. And once you train the neural network, you have to deploy that network. And that's called inferencing. And that deployment can happen even in the cloud or in an on-premise server all the way to the edge (access) of the camera.

So we created a platform using the same software stack starting with our CUDA. But then on top of that, we have unified our different software development kit for AI to create a singular platform for doing intelligent video analytics to bring AI to cities. And that's the Metropolis platform.

So typical installations have cameras, they have a video recorder or a server, an appliance on-premise. And we have -- we're using our Tesla platform for these high-end servers and Jetson platform for a very high-end camera or a small recorder that's typically taking 4 channels or 8 channels. Actually, if you do have the time, in the Expo floor, we have a classic example of a law enforcement, where the police car, what the future of police cars would look like, wherein you have police cars that have cameras in the front and the sides and body cams on the police officer. And if there's an Amber Alert, today what happens is basically the police officers are looking out for something. Then if they capture something, they're going to go back and dump all the data into a server after their shift. Instead of that, what we're showing in the Expo area is using the power of AI to process all these camera streams real-time. And whether it's detecting cars, license plates, faces from the body cam. So you can actually prevent, hopefully, prevent something bad from happening as opposed to doing it after the fact.

So we started on this journey a couple of years ago, started working with a few partners in the space. And last year, for the first time, AI-enabled servers or even some high-end cameras started as proof of concepts for this AI city, both smart and safety applications. We did approximately $10 million of revenue last year. Now we are projecting that next year we will would be doing more than $100 million in revenue. And this is all an inferencing opportunity on those either on-premise servers, appliances or high-end cameras.

And even with that growth, the 10x growth that we expect from last year to next year, bulk of the cameras are still not going to be AI-enabled. If you think about it, as I mentioned, there's 500 million camera streams out there right now. And almost none of them are AI-enabled. And many more cameras are going to be deployed this year and next year. Many of them will still not be AI-enabled. So we're still very early in this phase of transforming all these humans watching video streams or just going to the recorder to transforming it to deep learning, doing real-time video analytics. So even next year, we think that less than 5% of the total streams, cumulative streams would be AI-enabled.

Then going into calendar year 2020, we think the NVIDIA opportunity is $2 billion. And basically, even then, the number of streams would increase in terms of how much AI is going to be enabled. But increasingly, the networks and the capability of AI is also going to increase. And where we are focused on is primarily on-premise servers, like things that go into -- that process tens, hundreds, thousands of channels in an on-premise environment like an airport or a train station or a building like this. So that's the server. Then we're also focused on things like these small recorders and also these high-end cameras. So we think that it's going to be a $2 billion opportunity for us, growing fairly fast, coming from proof of concept last year.

So the AI city platform, Metropolis, we have already over 50 platform partners that we've announced -- or they have announced products or engagements with us. And many of them actually are here at GTC showing their products. And many of them have actually started shipping these products as well. And these are some of the few examples that you see. Before, they've been working on traditional video analytics for almost a decade. And just by using deep learning AI, within a year they could increase accuracy significantly, going from 70%, 80% to over 90%. Then soon we expect that to increase significantly. Then traditionally those video analytics were being done on, obviously, traditional servers. And they're seeing significant speed-ups, like double-digit order of magnitude, double-digit, 10 to 30x type of speed-up by using NVIDIA GPUs in these appliances.

So these are some of the list of the 50-plus partners. And the reason I show you these partners, many of them you might not recognize the names. This industry is a pretty fragmented industry. There are many, many verticals in this space. And there are many, many geographies that require different solutions. So in order to address this market, we decided to go in a platform strategy. We have many small customers that make up this ecosystem. So we are doing an end-to-end platform, all the way from training to inferencing, from edge to cloud, providing SDKs and then working with these different partners to launch the partners. Okay. So that's basically what I had to tell you about AI city.

With that, it's my pleasure to invite Colette, CFO of NVIDIA, to give you the next presentation.

Colette M. Kress {BIO 18297352 <GO>}

Okay. Good afternoon. For some of you, I feel like I just finished talking to you. Oh, that's right, I did. That was right. All last night we got the opportunity to speak to you after our earnings. And I really appreciate that you're here today for the keynote this morning as well as for Investor Day.

What I'm going to try and do is kind of go through what you hopefully heard today, was to answer some of the questions. We speak often. You have key questions in trying to understand key parts of our business. Always what I hear is about what's the size, what's the opportunity going forward. Hopefully, you got that from the presentations today. But I'll just kind of do an overview of some of the key highlights.

So last year, a really, really great year for us. We were able to grow nearly $2 billion in terms of revenue to $6.9 billion and growing 38%. Overall, our overall gross margins increase over that period of time of one year, also a 240 basis point increase to a record level as well. Then additionally, on the operating income, taking a look in terms of our value-added platforms, thinking through our overall investment in the business and surely our growth across every single platform, we were able to double our overall operating income from where we were just a year ago.

If we looked back over the last three years in terms of the growth of the overall business, I think several years ago I came here and talking about the transformation that we were on, the transformation of thinking differently, of not a chip company and moving to our overall platform approach. You're now starting to see what that has evolved into in terms of our movement to a platform.

We started off just three years ago at about a $4 billion company, doubling quite nicely in several of those businesses. But if you overall look in terms of where we are now, the growth platforms just represented about 60% of our business just three years ago. And we still had our OEM and IP business nearly 40% of our overall business. Right now, when you look at what we finished in terms of this last year, our Gaming business is about the size our total company was three years ago. Then we added on more platforms on top of that. You have our Pro Vis, you have our Data Center business. Those 2 combined are now nearly 25% of our overall business. And our Automotive continuing to grow through there, to reach the 6.9%. And over those three years, an overall CAGR of 20% and those growth platforms growing more than 35%.

Breaking this down in terms of each of those market platforms and looking at the CAGR over those three years. You've heard Fish come up here and talk about our overall Gaming business, the overall success that we have seen. And Gaming is still a very, very large growth driver for us going forward. We've talked about what drives that overall growth. The entire ecosystem around gaming is tremendously healthy. You've seen us fuel that overall ecosystem in terms of what we do with GeForce Experience, what we do with our game developers out there, as well as how we see just a tremendous platform for them to do what they want to do, is play the best games. The production value of the games have continued to rise over this period. And we've been able to benefit from producing better and better technology for them to do their gaming.

So what we have here, just in this last year, growing 44%. Yes. We launched our Pascal architecture last year. You saw that in our Q3 and Q4 results. And you're now even seeing the continued growth, as we just finished this last quarter also with very strong growth in terms of gaming. And we do expect that to continue to go forward.

If you think about the lineup that we have right now, tremendously fresh, a lineup full, everything from a 1050 all the way up to 1080 Ti and then our TITAN in order to dazzle the overall gamers out there. And you're coming into a time in terms of thinking about the back-to-school as well as also the new games that will be coming out in the fall. So we're very excited in terms of what we've seen in terms of our gaming history as well as what we see going forward.

Our Pro Vis business, a very well established, doing very nicely. You've seen it actually grow for all of last year as well as what we saw in this just last quarter in terms of Q1. A mature business. But we're continuing to see the mobility needs of workstations. But the mobility needs still would be high-performance. And we're just very, very right for that. We're a leader in this industry. We have probably the highest market share out there. And you've seen our performance do quite nicely.

Data Center. I think today and everything that you've heard today is really focused a lot about our developers and really the transformation that we're seeing in the overall computing environment. The computing environment, that it changes to GPU computing and what does that mean? Remember, our Data Center business is not just one single total business. There's probably about 5 different businesses that are within there. You have our deep learning training, you have our inference, you have our high-performance computing. We sell full systems, DGX supercomputers, as we also have our grid business in order to do virtualization in the overall cloud.

There's many different ways that we are bringing this to market as well. We're bringing this to market definitely selling our Tesla overall platforms. But we have also the ability to put together full supercomputers for them and also enable it through the cloud, which has been a very, very big part of the growth that we've seen over the last year.

Moving to Automotive. Rob's talking about our great opportunities that we have going forward with autonomous. And yes, our Automotive business today, the lion's share of that is still in terms of our infotainment systems. And a small piece of that is moving in terms of the development process, the overall development work that we are doing with many of the car manufacturers out there, start-ups and others. And the partnerships that we're doing to build the autonomous driving. But that's still not the majority of our revenue. And we're still benefiting for the high-end systems that we are producing to put in the overall cars.

So as you can see, our overall Gaming business, a CAGR of 40%. Our Pro Vis also doing nicely, growing a little bit in terms of the mid-low digits. And what we have in our Data Center, growing 60% over the last three years and our Automotive business growing 70%. Quite a unique position for a company to have that many strategic growth platforms all continuing to execute one on top of each other. Okay? And they're all really based on the same underlying architecture that has enabled us.

So growth margin expansion. I get a lot of questions. Is there still room? Is there still room for gross margin expansion, given what you've seen? Just three years ago, down at 55%. We finished this last year at a record, over 59%. And we actually just finished Q1 at 59.6%. Strong. Really thinking about what is driving this is overall our value-added platforms. That's a different statement than saying it is based on our higher-priced platforms. It's about the value that we're delivering inside of the platforms and where the overall investment is in terms of what we've done to build those. So much of our platform approach has been the overall end-to-end systems, the overall software development that we have included in terms of so many of the systems and the ecosystems around what we're doing. That investment is in OpEx. That investment is not in gross margin. So it's overall enabling our gross margins to continue to grow. And it's probably one of our biggest drivers as we go forward.

As you can see from the slide, our continued mix change of moving to our growth platforms and away from our OEM and IP businesses has continued to help our overall gross margins also lift up in the right direction. So yes, we do still see a continuation of this going forward.

As you can see from the slide, our continued mix change of moving to our growth platforms and away from our OEM and IP businesses has continued to help our

overall gross margins also lift up in the right direction. So yes, we do still see a continuation of this going forward.

Our operating expenses and operating expense investments. We have what we called a unified architecture approach, the same architecture across all of our different platforms. That is quite unique that we can serve up a platform to gamers, at the same time, running mission-critical application for AI in our Datacenter, all with the exact same underlying technology. Yes. There's a lot of work in terms of building out the full systems and the softwares on top of that. But we have just a great overall platform underneath.

We've done very solid in terms of looking at those investments and continuing to find efficiencies in terms of how we do that work. We had grown just about 8% in this last year in fiscal year '17. But you saw in the latter half of last year as well as going into this year, we started to increase that. We've talked about that we want to, for several quarters, in the high-teens in terms of an overall growth rate. And you're seeing us do that. It's probably one of your biggest asks in terms of, are you continuing to invest back into the business? The answer is yes. We need to think about how we continue to expand and reach some of these large markets that we're doing. And that is really from the investment. Most of that is headcount. Most of that is investing in the key engineering functions that we have that are building out these overall platforms. We have, over the last 10 years, invested more than $11 billion in overall R&D. Over the last 10 years, software has quadrupled in terms of our investment size. And now software, we actually spend more than actually the underlying hardware in terms of engineering. So yes, you will continue to see us invest and invest back into the business.

Operating margin expansion. So thinking about this overall leveraged approach, you're seeing right now our overall operating margins, starting back in 2014 at 16%. And we just finished the whole fiscal year '17 at 32%. And this last quarter at 33%. So we've doubled. We've doubled our margins over three years. But yes, we're still investing and that's going to be our #1 focus to make sure that we have the right investment. But we have a great model, a great model, that ability to grow in terms of on the top line enable higher gross margins, a unified approach in terms of OpEx, we have the ability to expand our margins. Probably one of the better metric to think about, as a bottom line, as so much as key of what we are building is in our OpEx.

Focusing on cash and cash flow. Used to get a lot of questions in terms of our cash levels, our U.S. cash levels. So let me try and highlight some of the key things. Over this period of time, over three years, more than doubling in terms of our overall annual cash flow. And what are we thinking about when we think about that cash flow? We do make sure that we want to be able to benefit from that cash flow, as we have also increased our cash. But as you see in our overall net cash position is just slightly up and is nowhere near in terms of that overall cash flow. And that is from some of the key returns that we have done in terms of our cash that we have collected. Additionally, our U.S. cash, although we've had a change in terms of where we receive our U.S. cash, is still about the same levels, even where it was about three years ago. So we're continuing to find new sources of U.S. cash. As you know, we are a high global company and we receive a significant amount of our profit from overseas as well.

Capital return. The key component in terms of our shareholder value message. And we have continued. Since our restart back in 2013, we have returned more than $4 billion to this date to overall shareholders. That represents more than 85% of free cash flow returned to shareholders. So we're doing the best that we can to make sure: One, we've invested properly in the business. And that may be investment in terms of OpEx, investment in terms of the capital that we need to grow this business. We'll think about other investments in terms of tuck-ins that may overall help our overall business. And that will be #1 priority. After that, we choose to return about as much cash flow as we can to our overall shareholders. And you'll see us continue to do that.

Driving shareholder value. This is my last piece before I actually bring everybody back on stage. But when you actually think over the last three years and you look at our overall invested capital that we did in the business in terms of our debt and overall equity. And that overall profit that we've gained through there, we've invested about in terms of where you see the NASDAQ 100 probably a little bit faster. But now when you look in terms of our overall return on that capital, in terms of '16 and our overall improvement that we have done in terms of since '13 of continuing to return to the bottom line from our overall leverage model, we've seen a clear outperform on that. Then lastly, you can see that return to you in terms of total shareholder return over this period and over 800%, okay? So this is still a very big component in terms of what we're here to deliver to you. We still have a lot of work to do. There's a lot of opportunity out there for the grab, driving shareholder value is probably one of our key things. Okay?

So with that, I thank you all for coming. We're going to open up. We're running a little bit behind. But we still have plenty of time for overall Q&A. We're going to bring up our overall speakers that you heard today, Jeff Fisher, Schankar, Rob, Deepu. And I'm going to see Jenson as well. I'm going to bring and open up the lights in terms of the group. I will be able to answer any questions. We could, of course, talk about yesterday. But let me know if we can try and gear our discussion in terms of what we heard today from the keynote or anything else that might be on your mind. Okay?

Questions And Answers

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

Hi. ladies and gentlemen, what do you think about my management team? pretty awesome, huh, guys. Wow. I hardly had to do anything.

Q - William Shalom Stein {BIO 15106707 <GO>}

Will Stein from SunTrust here. Thanks so much for hosting the Analyst Day. It's very educational and helpful. And in particular, sizing your opportunity in Datacenter, I think that's something that was very helpful. But it's a big number. In fact, it's bigger than Intel's DCG today. So I think what we'd really love is maybe an idea for what you think the future holds for the architecture of the Datacenter of the future relative to traditional Intel architecture servers versus GPUs and your new architectures that you talked about today?

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

Yes. Thanks for that. First of all, you guys know that the Datacenter architecture has evolved over the course of the last -- well gosh, in my time in the industry in the last 30 years, the architecture had gone from vectors in supercomputing, vectors, to massively parallel processors. And each one of these generations, you guys heard some names that went along with them. And its not necessary to bring back the past. But the vectors massively parallels. Then we saw some SIMDs, some measured multiprocessing SIMD machines. And that was kind of the generation of Silicon Valley's computer companies. Prior to that, it was kind of East Coast. Then prior to that, it was kind of, as you know, Midwest. Then after that, the microprocessor showed up. And one of the most important things that ever happened in the computer industry was the announcement of Pentium. The fact that it could do 64 bits, the fact that it was a workstation class computer, the floating-point mathematics was quite powerful. And the Pentium microprocessor made it possible for a single chip to really be as powerful as almost a custom server consisting of multiple ASICs. And the advantage that a single chip had was that every single year, it would pump it out. And it just rode that Moore's Law as you guys know. And the thing that it created was this idea called COTS, okay, commodity off-the-shelf datacenters. The idea that you would build custom servers and custom-oriented chips for mainframes and supercomputers was completely crushed by it. And the economics made sense. The economics made sense. Intel's R&D budget was just so much more than any $200 million or $400 million or $500 million high performance computing or supercomputing company could possibly afford. And it simply crushed them and wiped them out. And today, if you look at supercomputing architectures, what you'll find is that the vast majority of them are COTS. And if you look at datacenters, hyperscale datacenters, the vast majority of them are COTS. The problem with COTS, the problem with COTS, is you look at a datacenter, it's pretty much just empty space. It's just air. Air separated by motherboards and sheet metal and it's connected by a whole bunch of cables. And if you look at the cabling alone, the cabling alone in a datacenter, it's really quite amazing. I mean, if you had a high-performance computer and you have 2 cables coming out of the back and they're both InfiniBand, let's say, -- even the fiber optic -- between the neck and the fiber optic channel cable, it's a couple of thousand dollars. And suppose you had 300,000 of them. 300,000 times $2,000, it's a fair amount. And so the idea, the idea of scaling computers like that, at some level, doesn't make sense if you had an alternative, if you had an alternative. Because all we want to do is, as you know, we want to create computers that are denser, not more vacuous. Computers separated by air is not a good way to design computers. And so we want to compress them. And so this is -- leads me to the answer, which is the architecture of the future are going to be denser. Quite frankly, if you just had a couple racks of GPU-accelerated servers, as I demonstrated today, you'd replace 500 servers. And you would. You would take an entire row of computers and you replace them by 2 racks. And so you would take an entire datacenter, replace it by 4 racks. And so that's kind of the future now. Of course, on the -- in the other hand, on the other hand, the amount of AI that we're going to start to do is going to start growing. So these 2 -- so there's a dynamic that's going to create more density, there's another dynamic that going to create more expansion. But my sense is that, you're going to create density faster than you expand for some time. Then -- and then, we're going to see expansion again. But long-term, how could the world not need a lot more computing? It's not even logical. The question isn't whether the -- the industry, the world needs more computing. The real question is in what architecture would it become? That's really the greater question. I think for a long time, as you know, we believe in accelerated computing. So I personally believe in competent CPUs. And if we have competent CPUs that we connect them with competent GPUs, the result is actually pretty remarkable. And so I prefer that architecture. So I think we're going to, for some time, continue to see opportunities for both processors. But in this new computing architecture.

Q - Vivek Arya {BIO 6781604 <GO>}

Vivek Arya from Bank of America. Thanks for a very informative Analyst Day. I actually had one clarification. I think, Colette, you started talking about operating margins. I just wanted to make sure that you're still committed to expanding operating margins? I understand the need for investment. But I just wanted to understand how you think about margin expansion. But my real question is on, again, the datacenter. I think you convinced us that the training market is actually bigger than we think it is, that it's not just a one-off, that it has a lot more legs to it. But inference is still going to be a bigger market. The fact that you are leading in training, does that give you an advantage of some kind when somebody trains them more but when they moved to the inference stage, are they more likely to use the GPU? Or do you think it still is as easy for them to move to an alternative architecture. Do you have those sustainable advantages when it comes to inference also?

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

You said something earlier that -- in terms of a commitment, let me just answer that question. We're investing into this opportunity. And we're committed to this investment. The reason why we believe, you can't commit to -- we believe in margin expansion. And the reason why we believe in margin expansion is we believe our top line has the opportunity to grow faster than our investment. We're committed to the investment. We believe we will see margin expansion. One of them is the game, the other one is the score. We can't play the score. And so I just want to clarify that. We believe we will enjoy market -- margin expansion, enjoy margin expansion. Because the top line -- our revenues is growing so fast. So the -- you made the statement earlier that you believe to be true. But it's not, okay? And so the statement is inferencing is a larger market. It's not. And the reason for that is it's only larger in units, it's not larger in computing. And the reason for that is because it takes trillions of cycles to train these billions of parameters otherwise known as neurons or weights. It takes trillions of cycles to train billions of weights. Trillions is a very large number. Trillions is a very large number, as you know. And so inferencing inferences on the billion. Notice there's a trillion factor on it on training. Until we get to a trillion, billion, the inferencing is only a billion, right? And so you got to inference it either millions of people simultaneously hit the Internet to infer something or we have that network on a trillion devices. Until that happens, it is just not true. And so if you look at the total inferencing market today, I believe it's approximately small. Okay? And that's why there's no inferencing companies that you know of. We talk about a lot because it's a sexy conversation. But the fact of the matter is there's no evidence of it. Now we are going to lean into it, we think that it's about time that people are going to start taking these networks and they're going to start deploying into their datacenters. And that's why we decided to make Volta a pretty killer inferencing machine. I mean, it's off the charts. And if somebody would like to build something better, on first principles it's pretty hard. It is really, really hard. And so the question then becomes who has $2.5 billion R&D budget or pick your number, $0.5 billion R&D budget to go build a processor that has approximately no market at the moment? Well it turns out we can. And the reason we can is because our processor has a day job. You guys understand? It's waiting around for inferencing as a night shift. But at the moment, the day job is game away, train away, HPC away, workstation away. Then, when the inferencing market comes, we're going to be right there for it. That's why the magic works, that's why this equation works. If I were an inferencing start-up company today, I'm going to be chasing that dollar for a long time. Pretty rough. I'm not suggesting people shouldn't start companies. I think people ought to start companies. Lots of them.

Q - Atif Malik {BIO 15866921 <GO>}

Atif Malik, Citigroup. I have a question on Gaming. I'm a little bit surprised at the Pascal refresh or upgrade rate at 17%. And if you look at the Gaming decline that you guys saw in the last reported quarter, about 24%, a little bit kind of below seasonal in regard to your peers. So just trying to better understand what you guys think of the back half of this year in terms of Gaming demand? Then longer-term, you showed a 6% to 7% compound annual growth rate, I believe, from a third party. How should we think about that rate versus some of your customers like MSI in Taiwan talking about 20% year-over-year Gaming sales growth this year?

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

Yes. So I guess, I'll -- the numbers that I showed in terms of the overall market, the $100 billion is the software revenue. So it's kind of the strength, the ecosystem of overall Gaming. Gaming hardware has been growing faster than that. And especially notebook. You mentioned MSI is a big notebook customer of ours. They're one of the early, I guess, leaders in terms of pushing form factors and performance in the notebook space. So I think, Pascal really kind of turbocharged the hardware, Gaming hardware market by providing such a step up in performance. And I think this year should be a good year for Gaming hardware. We'll see seasonality as we do. But the end markets appear fairly robust to us looking forward. And the fundamentals that I covered, eSports, AAA gaming, those types of things all look to have continued momentum through the year.

Q - Srinivas Reddy Pajjuri

Srini Pajjuri from Macquarie. Thanks for the presentations again. Jensen, a question on competition. Obviously, you're the only game town in the Datacenter right now. But if you look out, you got AMD with graphics, Intel with a variety of solutions and you got internal solutions as well as FPGAs. Can you maybe talk about puts and takes when you look at the competitors? And also who you're most concerned about, longer-term?

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

First of all, before you can answer the competition question, before you could figure out what is -- who's your competitor in a game, you got to figure out what is the game you're playing. And we used to be a very different sports franchise. We were in the NASCAR business in the past. You race around for 490 laps and then the last 10 laps is chaos. And I think that's how it works, right? You eat pizza for the first 490 laps. And so -- but our game is really very different today. This is much more of a marathon. It's really -- you really have to play -- you had to play a long game. And the long game that we've done here is we selected some markets that are going to take at least 10 years to happen. The market that we were in before existed. PC graphics chips that existed. And we competed rigorously for it. I'm really happy for it. I really felt that the refinement of the execution system in our company, the speed of our company, the intense competitive nature of our company is really honed out of that. You just can't be successful in NASCAR. And there are people in this audience that have tracked us for a long time. I chose NASCAR because it kind of started like that. There were -- I don't know how many cars are in NASCAR that start. But there were 150 other companies with us in our lap. And it was chaos. It was chaos. But we remained. And it was because of the precision of the company, the ability to really hone its art to a very high level. But today, we're playing a very long game. We selected markets that are very, very large. But they're far out. And these far out markets are software intensive. It has nothing to do with Windows, it's got nothing to do with industry standard, it's got nothing to do with components. It's really a software problem. And we invested over the course of the last 10 years, just a massive amount of software. We're a software company today. We're a software company today. And that's why we could, on the one hand, solve this AI computing problem like nobody can. On the other hand, we can create a self-driving car. On the other hand, we could train robots in an alternative universe. Because we're a software company. You can't do that with a -- graphics chips don't do that. If I gave you a graphic chip, we could put it on the stage and say, "Train a robot." It won't do it. Just sit there like a graphics chip. And so -- it will boot Windows. And so I think the company is just a very different company today. We compete with different people in different markets. I believe there are really only 2 platforms today in self-driving cars, us and Intel. I believe that -- and we have different approaches. We have different approaches. But I don't -- I think long-term in this $10 trillion market, there are going to be 2 platforms that people can choose. And ours is architected very differently. In the case of AI, the question is where do people want to take the battle? AI is going to be huge. This is going to be the future of software. It's going to be the future of computing. If you were a strategist, I wouldn't take on this game. I would take on a different part of AI, go and endeavor a new frontier that we don't focus on. And so our competitor for AI is really the expansion of the market. If the market doesn't expand fast enough, our investment level becomes too high, right? I mean, we're investing very high -- very, very, very hard into this. And so we would like the market to expand. Fortunately, the market has expanded quite nicely. But that's an imperative for us. And that's one of the reasons why we do so much in this thing called Deep Learning Institute. We teach deep learning all over the world. And we partner with Microsoft, we partner with Google, we partner with all of these companies that are coming to us because we help educate the world's deep learning engineers. We're going to educate 100,000 engineers this year. And so we would like to create future engineers that can take advantage of this new art form, this new computing way. Then we also want to make sure that on the end points, on the end points, that we completely democratize the ability to do inferencing, to the point where we open-source it. We open-source our design. We've got the world's best engineers building the world's best CPU basically, we open-sourced it. And the reason for that is because we want to proliferate as quickly as possible the consumption of networks on the other side. And so the sooner that somebody puts a DLA in a lamppost, fantastic. Somebody puts it in speakers, fantastic. And so we would like to expand the market. Our competitor in the core AI market is consumption, which is not unlike a lot of companies who are building long-term. And at some point, of course, it takes off and we're starting to experience that. So it's different for different markets.

Q - Mitchell Toshiro Steves {BIO 3255357 <GO>}

Mitch Steves, RBC. So 2 questions. So first on the Datacenter side. So it seems like the implied message that AI is actually probably going to be the faster growing piece. So if that's the case, does that mean that essentially it's growing faster than the 186% you saw in the latest quarter? Then secondly on the Automotive side. The one case I'm having a hard time understanding is once you've trained the car, what prevents you from moving to something like a DSP, a lower powered chip, now that it knows how to drive versus using a GPU?

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

Welll, first of all, DSP is way less efficient than a GPU. I can't imagine doing deep learning on a DSP. So why? Do you see what I'm saying? I mean, is there something about a DSP that you particularly think of? They're just 2 different 3-letter acronyms. The GPU at least has the benefit of having tensor instructions, a DSP doesn't, right? DSP is actually good at time sequence, digital signal processing, that's why they use it for audio. Deep learning is a massively parallel problem. And so all of those arrays and all of those matrix multiplies, you want to really happen almost simultaneously. And so I think GPU is actually quite uniquely wonderful. The architecture is really quite uniquely wonderful at doing it. We just had to add deep learning finesse to it. We had to add deep learning data formats, deep learning specific instructions. And now we have actually included deep learning computational arrays right into CUDA. CUDA is on CUDA 9. CUDA is one of those computing architectures that's quite extensible. We've been extending it since the very beginning. Every single year, we add new things to it. This is one of those architectural innovations that started at the beginning of our company that's really carried on and I'm very proud that we have invested in it. And so I think the way to think about it is Xavier was designed to do deep learning inferencing or AI computing for the car from scratch. It's purpose-built. If people don't use Xavier for cars, we're not likely to use it for coffee machines. And we're not going to -- if Xavier doesn't work out for cars, we're not going to use them for cell phones, for example. And so if that's helpful, it's dedicated to do one thing, run the entire software stack of self-driving cars.

Q - Toshiya Hari {BIO 6770302 <GO>}

Toshiya Hari, Goldman Sachs. Thank you very much for the opportunity today. I think you provided long-term TAM forecast for all your businesses but Gaming. So Jeff, I think you ran through all the growth drivers and collectively as a group, we appreciate all of those. But if you can help us translate some of those growth drivers into actual numbers for the next, say, 3 to five years, that would be much appreciated.

A - Jeffrey D. Fisher {BIO 2373419 <GO>}

Well I don't have the specific TAM for graphics. I think that if you look at the total PC market. And the areas we're investing in, in our platform, it's a little hard to put an upper bound on it. GeForce now, for example, can get us outside of our -- it gets us outside of our traditional market. We're adding value in moving up the value chain in terms of what we're providing to gamers. The total software TAM is $100 billion -- Jon Peddie says the hardware TAM is about $30 billion. We're in monitors now with G-SYNC. So I guess it's hard for me to put a specific TAM that I can tie back to our business today. I don't know how you would -- how you -- what you want to add to that...

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

Three billion people. That's what I would have done. Three billion people. I believe, during my tenure, for -- in my -- in the arc of my career here, which is another 50 years, I believe I'm healthy, I enjoy my job. There's plenty to do. In that 50 years, I believe there's no question in my mind, every human will become a gamer. I think that's just -- it's logical. It's reasonable. Every human watches television. In my parent's generation, only 10% did probably. Our generation, give me an example of one who doesn't. I mean, it's hard to find. And so it's just another form of entertainment. And the only difference, of course, is that this form of entertainment is computationally intensive. And therefore, the delivery of the vehicle, the delivery of the medium is hard, which is the reason why our cartridges look like supercomputers. GeForce cartridges look like supercomputers. But step-by-step-by-step we're solving the problem of being able to provision that service to you in a virtualized way. It's going to still take us several years. But as Fish mentioned, GeForce now was created so that we can reach everyone. Long-term, the question is how many gamers and who are the world's top -- who are the world's top fundamental gaming platform provider? I can almost rule out who they're not. And -- but I can almost say for absolute certainty, we will be one of them. And I think everybody in this room would say, yes, pretty -- yes, I would go along with that. Pretty darn sure they're going to be one of the platform makers of gaming someday. Kind of hard to imagine they don't. And so 3 billion -- 3 billion, how many of them do we serve? When do we serve them? Not sure. Who is absolutely going to be at the table? Absolutely sure. That's how strategies are made. Okay, everyone, this will be our last question.

A - Unidentified Speaker

There's like -- we're having so much fun.

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

We are hosting a reception with our management team right out here after this last question.

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

It's Matt Ramsay from Canaccord. I don't know if this is for Shankar or Jensen or how you guys want to take it. But it's been a great thing for the company that you guys have one common GPU platform across all of the businesses and we've all seen the benefits of that. But you're currently able to generate significantly higher ASPs in pieces of your business, I think, particularly in Datacenter. I think that's justified by the benefits that your customers get from that. I just want to ask a little bit, as we get scale into some of the TAMs that you guys laid out, how do you feel about pricing in the datacenter space and your ability to segment your product portfolio as we move forward. And just some perspectives on that going forward?

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

First of all, the idea that our pricing is somehow higher in certain segments is just not right. And the reason for that is this: we think about cost wrong. We think about cost wrong. NVIDIA's cost is not our chips. That's not our chip, that's the least of my cost. If no more -- our chips is not our cost, no more than CD ROMs is Microsoft's cost. We are a software company. We're an information company. And so our cost are human costs. And they're not variable. Chips are variable. Sometimes I have more inventory, sometimes I have less inventory. But our engineers are employed everyday. And so that's really our cost. When you think about it from that perspective, the Datacenter business is pretty high cost. The cost is much, much higher per unit than it is in GeForce, for example. Because GeForce sells tens of millions, Datacenters don't. And yet we're investing real numbers into advancing Datacenters. So the cost is very different. I just want to let you -- in each one of our businesses, we have to think about it from the first principles of the business model. And what makes that business work and that's why we price it accordingly. But as far as I'm concerned, I frankly think that -- well, the operating margins of GeForce, obviously, are quite good.

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

Let's see if we can do some more questions. We have food and drinks out there. But -- I don't know. A little overrated food and drinks, we can just keep here on some questions. Sounds good?

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

Shawn's like an investor police.

Q - Ambrish Srivastava {BIO 4109276 <GO>}

Ambrish from BMO. Thank you, Jensen. You laid out a vision for AI which is pretty compelling, I've never doubted that. But I wanted to move on to a more mundane topic like AR and inventory and free cash flow.

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

Augmented reality? That's not mundane. Come on. Let's talk about it.

Q - Ambrish Srivastava {BIO 4109276 <GO>}

We can touch on that as well. So Colette, accounts receivable and just going back to last night, because I didn't get the chance to ask a question. Accounts receivable, up 18% q-over-q on a down 11%. Even on the year-over-year basis, pretty large increase. Inventory has been going up as well. Can you just help us understand what we should be expecting both on AR as well as in inventory as we go forward?

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

So let me highlight on AR. AR is always influenced in terms of the timing of our new products. Whether or not they're in the beginning of the quarter, whether they're in the middle of the quarter, whether or not they're in the end of the quarter. There's never an issue in terms of our collection. And my team here doesn't really look at the end of the month like I do. They shift when they think they need to on that. So not necessarily any concerned other than just timing in terms of our AR. The inventory. We've talked about a couple of things this morning that are coming up that is very important in terms of our part of our inventory. When I look and we all understand in terms of our inventory, it's probably about 1/3 of our quarter of inventory. And we have a significant amount of businesses in front of us. Businesses in terms of Datacenter, many different architectures that we are continuing to sell, as well as our Gaming business, as well as our Automotive business. So the reality is the inventory level for the size of our company is actually quite reasonable, is actually low. Will you have enough in terms of that? So again, we all feel very comfortable with what we have. But probably the key things in that is just timing within there. And I'm not concerned.

Q - Brett Simpson {BIO 3279126 <GO>}

This is Brett Simpson at Arete Research. I had a couple of 2-part question on Datacenter. First of all, Jensen, I wanted to get your perspective on CPU. Because you're accelerating a lot of workloads, you've always been hosted by Xeons, particularly in deep learning training. So the interface between the CPU and the GPU is obviously holding back system performance. And so my question is, why not circumvent the Xeon, particularly with the work you've done with NVLink and Power9, why don't you develop your own CPU or license Power9 instead? I've got a follow-up.

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

Well Power9 is really a fantastic CPU, there's just no question about it. And the work that they've done on Power9 is, well, for a single-threaded processor, it is unparalleled. And it's a great processor. On the other hand, on the other hand the x86 ecosystem is really rich. I mean, there's a lot of -- the body of work on top of x86 and the Intel architecture is quite rich. And we use it for our own products. We use it to develop software on. We run -- NVIDIA runs on x86, runs on Intel. I'm not ashamed of that. And it's a great processor. We select the good ones. In terms of the datacenters, I think datacenters use Intel because their energy efficiency is so good. I mean, the fact of the matter is, the craftsmanship of Intel Xeons are so great that practically most people -- and I don't know of one that is going to violate this rule that I'm about to say, okay? I don't know anybody who could actually give their CPUs away for free to companies and still be lower total cost of ownership than Intel. That says something about their craftsmanship. And until single-threaded software, single-threaded code is completely gone out of the world, which is not likely, which is not likely, Intel Xeons are fantastic. And so that's why we use it. Now we circumvent then PCI Express through all kinds of techniques, through all kinds of techniques. And so -- and Intel's also advancing PCI Express to Gen 4 and so on. And so I'm comfortable with the connection with them.

Q - Brett Simpson {BIO 3279126 <GO>}

Okay. And just to follow up on the last question on Datacenter pricing. I mean, it's clear you've got a value-based sell and you talked about the savings you can offer customers from replacing CPU with a GPU for acceleration. But as you scale up the GPU business, particularly with the large hyperscalers who are a concentrated set of buyers going forward, given how big the public clouds are becoming, I guess, they're approaching about $100 million a year on GPU compute, some of these guys with you today. But if you look out over the next 2 or three years, they're going to be buying a lot more product from you guys. How sustainable do you think the current pricing you're getting on GPU is with that specific set of customers?

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

First thing is that we save them money. We save them millions of dollars. And we save them millions of dollars in a way that isn't possible without the ecosystem we've built up and all the applications we've ported. As Shankar said, the top 450 HPC applications on the planet have been ported. And we get more and more ported every day. And so the richness of our ecosystem is what allows them to be able to accelerate those applications and by doing so, reduce their cost. And they reduce their cost by investing a couple of $100,000 to save 10 -- to save $2 million of servers and probably $1 million with the cables, okay? And so that's the math. That's the math they do. Now in the event that we're not able -- in the event they don't run those kind of applications, we are 0% -- we're 0% successful. We basically either help people save millions of dollars, or we don't get considered for the opportunity, not even a little bit. It's an all or nothing. It's an all or nothing. And that's why pricing is largely a non-matter. Now the ecosystem that comes along and the software spec that comes along. And for all of you to hear this now, it's going to be fine. But we've invested as a company close to $10 billion into GPU computing. Long before, long before GPU computing had any value at all, we were carrying CUDA on our chip with shared memory dragging our gross margins for a decade. And so the question is -- that's how we got all of these ecosystems, as I was mentioning earlier, the conviction to go build that ecosystem is what got us here. Then secondarily, all of the investments that we're making today and the richness of the ecosystem and software stack is what allows them to have those applications to decrease their cost. If the applications are not ported and accelerated, you can't benefit from it, okay? So I think the -- there's nothing at all -- I mean the value exchange between the 2 organizations is fantastic. And we are, quite frankly, years from breaking even on the investment, if you will.

Q - Trip Chowdhry {BIO 5306842 <GO>}

Trip Chowdhry with Global Equities Research. Two quick questions I have. First is you know there is a new industry which is evolving and emerging, which is very positive for you, that is a shift from CPU-native databases to GPU-native databases, especially in the columnar database. And there are few companies in your booth who are doing that. My question is, are you driving any SDK or anything else for this ecosystem to grow? That's number one. The second question is...

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

Yes.

Q - Trip Chowdhry {BIO 5306842 <GO>}

Okay. Little slight details, then I'll come to my second question, like, just the SDK you're giving or benchmarks. So what are you giving there?

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

nvGRAPH, cuBLAS, cuDNN, CUDA. If we -- and then a whole bunch of software engineers to help them optimize their software.

Q - Unidentified Participant

When do you think that industry take off? Two years? Three years?

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

I'm not sure. But here's what I know. Graph analytics is one of the most important things we could do. We all have a sea of dark data that we sit on in a company. We just have no -- we have no idea how to turn that unstructured, messy data into something reasonable. If I could take all of that data and just plummet directly into Todd Mostak's MapD, which I'm the process of doing, by the way. Okay? If we could just take all that goop and just ram it into one of our DGXs running MapD, it's going to go figure out, it's going to figure out -- with a little bit more software, it's going to go figure out the relationship between everything and everything else all by itself. And as a result, when it create a corporate graph, we're going to essentially have a corporate Facebook, except it's not people, it's things. And I'll be able to find interaction relationships, trends in that the way that people who own amazing graphs can. And I believe everybody's going to do that. I believe everybody's going to do that. That's not going to say that SQL databases and Hadoop databases, they don't have their place. It's just people who really want to harness the -- take their data out of their underground and really use it and apply graph optimization and analytics to it and apply AI to it, those people are really going to want this. I think companies like Walmart and GE and people who have a lot of data, that they have to find relationships with are going to need it. Not to say that enterprise databases are all going to go away. But this is how high-performance computing is going to come into IT. If you think about IT today, it's really about serving files and maintaining number of employees and number of parts and number of dollars and those kind of things. And ERP systems today in IT, that's IT. But there's a new IT in town and it's HPC IT. And the IT departments of most enterprises don't understand it yet. But they will. HPC is coming in and it has to start with in-memory databases, very high-performance graph analytics, deep learning and AI on top of that, machine learning on top of that, they're coming in. And I think that's going to be the future of IT, really. Yes there, young man? I saw you on TV the other day. Made me proud.

Q - James Wang {BIO 20834257 <GO>}

James Wang, ARK Invest. Jensen, you talked about open-sourcing Xavier. I thought that was the most surprising and radical announcement I've seen out of you guys for a long time. Maybe just talk us through what that thought process is, are you going for an R model, a QUALCOMM model? Am I going to be able to download a blueprint of Xavier? What is it on me? And what's the revenue model. So previously, you've talked about IP as being a line of business. Is this involved? How does that all fit together?

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

Kind of went like this, I think we should open-source Xavier DLA. What do you guys think? And everybody goes, "Damn it, that's right. We should." Then it just took off, just like that. I mean, James, you know how decisions get made. It's we sit around and we talk about it. And the problem we're trying to solve is how do we accelerate the expansion adoption of deep learning? There are a lot of ASIC companies who don't know how deep learning really works right now. They see the mathematics. They see some of it. And they just don't know the end to end of it. They don't know how to design it end to end. There are many markets we're never going to design products for. We're just never going to do it. And there's -- we believe that there's going to be a trillion, a trillion deep learning chips, little tiny TPU things in ASICs that are going to be sprinkled like dust all over the world. And we're not going to go build those things. We're not going to go build refrigerator chips. But somebody will. We're not going to go build coffeemaker chips. But somebody will. We're not going to go build, right, microphone chips. But somebody will. And so we want to make it possible for all of those people who are going to build it, without or without us, to be able to build it using our IP, using our IP that we created that's really superefficient. It's already baked. And they can pull in their deep learning go-to-market by a couple of 2, three years. By a couple of 2, three years. And so we thought that, that expansion of the deep learning market would only help the one conversation we were having earlier about the competitors I worry about. There was a question about competition. And I told you, competition is nonconsumption. Well if competition is nonconsumption, then don't cry about it. Do something about it. And so we did 2 things for nonconsumption. We invested in deep learning institute. We call it DLI. And we invested in deep learning accelerators. Called it DLA. And we're making the 2 of them available fairly affordably. In the case of DLA, it's open-sourced, there's no royalty. No royalty. And just take it and use it and just write home once in a while. But mostly expand the market. Yes?

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

Stacy Rasgon with Bernstein. You talked about customers investing a few hundred thousands to stave $10 million. So I guess in that light, do you view the rise of accelerated compute as helping to increase the overall spend on compute? Or is it a mechanism to better utilize and gain efficiency out of the spend that's already there now? I guess, put another way, do you think about your opportunity -- how should we think about your opportunity as benefiting from a shift in the architecture in terms of where the current spend is actually going versus an ability to -- or a need to actually increase the overall spend on compute that is out there today?

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

Yes, Sean (sic) (Stacy), that's a great question. And in fact, it looks almost exactly like market adoption curves of a new technology that is somehow radically different than the past. And the movie we saw this time looks exactly like that. The first customers who adopt the technology adopted a technology because there were no choice. They needed to simulate something so fast that if they can't get it done, their research grant runs out of money. They're out of a job. One researcher told me that because of my work. And he invited me to his lab to see his supercomputer. He says, "Jensen, because of your work, I'm able to now do my life's work in my lifetime." Now if that is not an imperative, I don't know what is. And so they have no choice. They need a time machine they have no choice. They do need warp speed. The second group are the second generation of people who used our GPUs, where it is possible to do it the other way. But it is just crazy nuts cost prohibitive. Let me give an example. The TITAN, the Oak Ridge Titan was financially not possible without our GPUs. They just had no choice. Their budget was this much money and they had no choice but to use accelerators. And that's why they were the first to take that big chance. Jeff Nichols, who is quoted on Volta, he's here probably. Jeff Nichols took a massive chance to move one of our nation's open science supercomputing centers into accelerated computing when nobody thought it was time. Huge chance. Big courageous move. Soon after that, of course, everybody else discovered the same problem. They don't have enough money. They can't build a datacenter large enough. Datacenters for supercomputers, they get awfully large. And so they just can't build the facility large enough, they can't provision the power, impossible to provision the power. And they simply don't have enough money. And so the reduction in size, the compression that we were talking about earlier, about what the future architectures look like, the reduction in size, the compression of everything and the money they saved allowed them to create a supercomputer that was worthy of being built. Remember, there is no way to go to get a grant for $0.5 billion and say, "Hi, I would like to build a supercomputer that's almost as fast as the 900 top supercomputer in the planet." There's no way to do that. And so you have to build something of meaning and you have to build something that allows to deliver a capability to the nation. And so they're second-generation. The third-generation, simply because they have a new idea that wouldn't have been possible if they didn't have this new technology, deep learning. A new generation of researchers jumped on top of our platform. They have a great idea. It would be practical probably someday. But if it wasn't because of the GPU, it wouldn't be practical today. And so they jumped on the GPU like crazy. Then we're now seeing -- and this is the largest market of all. And this is now, if you will, what people called the tipping point, what people called the mainstream market, if you will. Now as all of this code that has been ported starts going to mainstream, then everybody looks at it and they go, "I need to run this code and this is the list of 10 applications that I have to run. And based on this architecture, which cost this much money, this will be my throughput for these 10 applications. Based on this architecture this is how much it cost. And this is my throughput." And so that's just a spreadsheet. It's simple now. And there, you get 1 of 2 answers. If you have a constant spend, meaning you got a budget. And you just got to spend $10 million. I don't know why you got to do it. But you just got to do it. There, you would like to have the highest throughput. In the event that you're just trying to deliver a throughput to your customer and you're budget conscious. And you would like to figure out a way to spend less money, now you can save money. And those are basically the 2 type of customers that Shankar deals with today. Do you want to increase your throughput with the budget that you have? Or do you want to deliver your throughput at a lower budget? That part of the market is quite large. And that's what we're experiencing today. Then long-term, long-term, this is what we all believe. This is now the next phase. The next phase is the exact same thing that we've enjoyed with PCs, with mobile devices. When we first got -- when they first showed up, everybody goes, "Nah, $700? I think it's going to be a high-end segment. It's really for business executive luxury computing, for example." Do you remember back in the good old days when the PC first came out and they go, "Yes. There's a market probably for, I don't know, 100,000?" And so we're going to see -- we're seeing the same thing now. As it turns out, once you provide a capability, something like deep learning. And we all find value in the power of automation of automation. And we all use it. And we realize how much productivity unleashes into our company, i.e. save money. We're going to buy more and more of it. And so what I really believe is this: the reason why I have so much confidence that the expand -- the market, the market of computing is going to continue to expand, is because that market, that expansion of the computer industry applying machine learning is going to cause us to reduce spend in other areas like you can't believe. I'm just certain of it. Waste of transportation, waste of containers, waste of warehousing, waste of parking lots, waste of fossil fuel, waste of, you name it. We're going to reduce waste. Waste of time, waste of people doing the same thing over and over again. We're going to reduce waste. And the way we're going to do that is we're going to shift that spend to more computing capability in all of our companies. And that's the next phase, expansion of the market. And hopefully, we'll enjoy it differently in each one of the phases. Sean (sic) Stacy , does that answer your question? It's a good question. And I think, we just went through Marketing 101 or business -- Revolutionary Business 101.

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

That's Stacy. But yes, I think that does answer it right there.

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

I'm sorry. Thanks, Stacy.

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

Just get one last one, is that okay, can you get that Shawn?

A - Shawn Simmons {BIO 19500814 <GO>}

We have awards to give.

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

After this, yes.

Q - Unidentified Participant

This question is on autonomous driving. Jensen, you said earlier that we're down to 2 platforms, you and Intel. And I assume you mean also

Mobileye. You also said it's going to be a very big market. So my questions are why do you think your approach might be superior, maybe you can summarize it different approach, the different approaches. What's it going to tell us which approach might get a bigger portion of the market and do you think in 5 or six years whether it's going to be a winner-take-most?

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

Well our two appraches are very different. The -- our approach is based on the computing approach and all of our processors are programmable.

As you know CUDA is the architecture, computing architecture of our GPU. And our DLA has a programmable architecture and everything that we do is from the context of a computer. And that's the reason why 240, 200 companies are able to take our computing platform and add value to it. We also were the first to realize, we realized that deep learning was the right way to go, three years before the rest of the market. And so we've been building our deep learning platform and with Xavier everything is based on deep learning. There is no traditional computer vision hand-coating that goes along with it. I'm not suggesting that it's the only solution that works, it's just as we know it's more robust, it's more diverse, it's able to handle more conditions. It does require more data but data is available. And so all the criticisms of deep learning in the past has largely vaporized either through the maturity of the workflows and the maturity of the engineers, as well as the creation of a processor dedicated just for one thing, called Xavier. Xavier is the world's first processor that is designed for one job. It's optimized

for no other job. And if it doesn't drive we are in trouble, because it doesn't know how to type. And so that sucker is designed for one job and one job only. This is a high risk/high return strategy. Now of course we moderated the risk by having built the entire stack ourselves. So we understand the software so well. We owned our software, we see all the software. So that's how we moderated the risk. But it's pretty high risk. Nobody in a semiconductor company should do it. They wouldn't know all the nuances that went into Xavier. So their approach is a combination of custom ASICs for computer vision. And their traditional computer vision chips. They do have some deep learning capability in the

future. But largely it is still a computer vision chip. Then there are some FPGAs that surround a couple of general purpose CPUs.

And so I think the two approaches, one is based on a AIi computing platform that is rather coherent in my opinion. And the other one is a collection of integrated chips that to cobble up enough capability. And so our tow approaches are very different. As a result of our choice,

our computing platform is programmable, meaning if you were a start-up you could actually get a drive PX2 and make an attempt to program it.

You can't acutally do it on the other one. Because of our acrhitecture, you can actually be a large car company and take our computing stack and add some differentiation on it. You can't do it on the other approach. And so that's just a difference, they're going to find success. And we'll find success.

Q - Unidentified Participant

(Indisceernible) validates your approach?

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

Well 200 companies validated our approach and as you know one of the most conservative companies in the world, not that they're slow, they're careful, has done as detailed of an analysis as you can imagine, Toyota has chosen a drive PX. And ehy had a choice of everything. And that tells you something. And in terms of the ratio, it's 74/26.

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

On that, would you like to close this out?

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

It's always good to have a goal. Okay, well guys, these are extraordinary times, thanks for all the time you've spent with us. And the attention you're paying our company. We are really honored by that. And thanks for all the investors in the room. It's great to have your support and

this is just the beginning. Okay. Thank you.

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 in this transcript constitutes a solicitation of the purchase or sale of securities or commodities. Any opinion expressed in the transcript does not necessarily reflect the views of Bloomberg LP.