Q1 2021 Earnings Call

Company Participants

- Colette Kress, Executive Vice President and Chief Financial Officer
- Jensen Huang, Founder, President and Chief Executive Officer
- Simona Jankowski, Vice President, Investor Relations

Other Participants

- Aaron Rakers, Analyst
- C.J. Muse, Analyst
- Harlan Sur, Analyst
- John William Pitzer, Analyst
- Joseph Moore, Analyst
- Mark Lipacis, Analyst
- Matthew D. Ramsay, Analyst
- Stacy Rasgon, Analyst
- Timothy Arcuri, Analyst
- Toshiya Hari, Analyst
- Vivek Arya, Analyst
- William Stein, Analyst

Presentation

Operator

Good afternoon. My name is Josh, and I will be your conference operator today. At this time, I would like to welcome everyone to NVIDIA's Financial Results Conference Call. All lines have been placed on mute to prevent any background noise. After the speakers' remarks, there will be a question-and-answer session. (Operator Instructions)

Thank you. Simona Jankowski, you may begin your conference.

Simona Jankowski {BIO 7131672 <GO>}

Thank you. Good afternoon, everyone, and welcome to NVIDIA's conference call for the first quarter of fiscal 2021.

With me on the call today from NVIDIA are Jensen Huang, President and Chief Executive Officer; and Colette Kress, Executive Vice President and Chief Financial Officer. I'd like to remind you that our call is being webcast live on NVIDIA's Investor Relations website. The

webcast will be available for replay until the conference call to discuss our financial results for the second quarter of fiscal 2021. The content of today's call is NVIDIA's property. It can't be reproduced or transcribed without our prior written consent.

During this call, we may make forward-looking statements based on current expectations. These are subject to a number of significant risks and uncertainties, and our actual results may vary materially. For a discussion of factors that could affect our future financial results and business, please refer to the disclosure in today's earnings release, our most recent Form 10-K and 10-Q, and the reports that we may file on Form 8-K with the Securities and Exchange Commission. All our statements are made as of today, May 21, 2020, based on information currently available to us. Except as required by law, we assume no obligation to update any such statements. During this call, we will discuss non-GAAP financial measures. You can find a reconciliation of these non-GAAP financial measures to GAAP financial measures in our CFO commentary, which is posted on our website.

With that, let me turn the call over to Jensen.

Jensen Huang {BIO 1782546 <GO>}

Thanks, Simona.

Before Colette describes our quarterly results, I'd like to thank those who are on the front lines of this crisis; first responders, health care workers, service providers, who inspire us every day with their bravery and selflessness. I also want to acknowledge the incredible efforts of our colleagues here at NVIDIA. Despite many challenges, they have barely broken strides during one of the busiest periods in our history. Our efforts related to the virus are focused in three areas.

First, we're taking care of our families and communities. We've pulled in raises by six months to put more money in our employees' hands, and NVIDIA and our people have donated thus far more than $10 million. Second, we're using NVIDIA's unique capabilities to fight the virus, a great deal of science being done on COVID-19 uses NVIDIA technology for acceleration.

Some of the many examples, including sequencing the virus, analyze new drug candidates, imaging the virus and molecular resolution with cryo-electron microscopy and identifying elevated body temperature with AI cameras. And third, because COVID-19 won't be the last killer virus, we need to be ready for the next outbreak. NVIDIA technology is essential for this end-to-end computational defense system. A system that can detect early, accelerate the development of a vaccine, contain the spread of disease and continuously test and monitor.

We are racing to deploy the NVIDIA CLARA computational healthcare platforms. There are Parabricks, an accelerated genome analysis from days to minutes. For imaging, we'll continue to partner with leading research institutes to develop state-of-the-art AI models to detect infections. And Clara Guardian will connect AI to cameras and microphones in hospitals to help overloaded staff watch over patients.

We completed the acquisition of Mellanox on April 27. Mellanox is now NVIDIA's networking brand and business unit, and will be reported as part of our Data Center market platform. And Israel is now one of NVIDIA's major technology centers. The new NVIDIA has a much larger footprint in data center, computing end-to-end and full stack expertise and data center architectures and tremendous scale to accelerate innovation. NVIDIA-Mellanox are a perfect combination and position us for the major forces shaping the IT industry today, Data Center scale computing and AI.

For microservice cloud application, the machine learning and AI, accelerated computing and high performance networking are critical to modern data centers. Previously, a CPU compute node was the unit of computing. Going forward, the new unit of computing is an entire data center. The basic computing elements are now storage servers, CPU servers and GPU servers, and are composed and orchestrated by hyperscale applications that are serving millions of users simultaneously.

Connecting these computing elements together is the high-performance Mellanox networking. This is the era of data center scale computing. And together, NVIDIA and Mellanox can architect end to end. Mellanox is an extraordinary company, and I'm thrilled that we're now one force to invent the future together.

Now, let me turn the call over to Colette.

Colette Kress {BIO 18297352 <GO>}

Thanks, Jensen. Against the backdrop of the extraordinary events unfolding around the globe, we had a very strong quarter. Q1 revenue was $3.08 billion, up 39% year-on-year, down 1% sequentially and slightly ahead of our outlook, reflecting upside in our data center and gaming platforms.

Starting with gaming, revenue of $1.34 billion was up 27% year-on-year and down 10% sequentially. We are pleased with these results, which exceeded expectations in the quarter, marked by the unprecedented challenge of the COVID-19. Let me give you some color.

Early in Q1, as the epidemic unfolded, demand in China was impacted, with iCafes closing for an extended period. As the virus spread globally, much of the world started working and learning from home and gameplay surged. Globally, we have seen 50% rise in gaming hours played on our GeForce platform, driven both by more people playing and more gameplay per user.

With many retail outlets closed, demand for our products has shifted quite efficiently to e-tail channels globally. Gaming laptops revenue accelerated to its fastest year-on-year growth in six quarters. We are working with our OEMs, channel partners to meet the growing needs of the professionals and students engaged in working, learning and playing at home.

In early April, our global OEM partners announced a record new 100 NVIDIA GeForce powered laptops, with availability starting in Q1 and the most to ship in Q2. These laptops are the first to use our high-end GeForce RTX 2080 SUPER and 2070 SUPER GPUs, which have been available for desktop since last summer.

In addition, OEMs are bringing to market, laptops based on the RTX 2060 GPU at just $999, a price point that enables a larger audience to take advantage of the power and features of RTX, including its unique ray tracing and AI capabilities. These launches are well timed, as mobile and remote computing needs accelerate.

The global rise in gaming also lifted sales of NVIDIA, the Nintendo Switch and our console business, driving strong growth both sequentially and year-over-year. We collaborated with Microsoft and Mojang to bring RTX ray tracing to Minecraft, the world's most popular game with over 100 million gamers monthly and over 100 billion total views on YouTube.

Minecraft with RTX looks astounding, with realistic shadows and reflections, light that reflects, reflex and scatters through surfaces as naturalistic effects like fog. Reviews for it are off the charts.

Ars Technica called it a jaw dropping stutter and PCWorld said it was glorious to behold.

Our RTX technology stands apart, not only with our two-year lead in ray tracing but with its use of AI to speed up and enhance games using the Tensor Core silicone on our RTX class GPUs. We introduced the next version of our AI algorithm called deep learning super sampling in real time DLSS 2.0, can fill the missing bits from every frame, doubling performance. It represents a major step function from the original, and it can be trained on non-gaming specific images, making it universal and easy to implement. The value and momentum of our RTX GPUs continue to grow. We have a significant upgrade opportunity over the next year, with the rising tide of RTX-enabled games, including major blockbusters like Minecraft and Cyberpunk.

Let me also touch on our game streaming service, GFN, which exited beta this quarter. It gives gamers access to more than 650 games, with another 1,500 in line to get on-boarded. These include Epic Game's Fortnite, which is the most played game on GFN and other popular titles such as Control, Destiny 2 and League of Legends, with Cyberpunk joining in the fall.

Since launching in February, GFN has added two million users around the world, with both sign-ups and hours of game playing boosted by stay-at-home measures. GFN expands our market reach to the billions of gamers with underpowered devices. It is the most publisher-friendly, developer-friendly game streaming service with the greatest number of games and the only one that supports ray tracing.

Moving to Pro visualization. Revenue was $307 million, up 15% year-on-year and down 7% sequentially. Year-on-year revenue growth accelerated in Q1, driven by laptop workstations and Turing adoption. We are seeing continued momentum in our ecosystem

for RTX ray tracing. We now have RTX support for all major rendering visualization and design software packages, including Autodesk Maya, Dassault's CATIA, Pixar's RenderMan, Chaos Group's V-Ray and many others.

Autodesk has announced that the latest release of VRED, it's automotive 3D visualization software supports NVIDIA RTX GPUs. This enables designers to take advantage of RTX to produce more like life designs in a fraction of the time versus CPU based systems. Over 45 leading creative and design applications now take advantage of RTX, driving a sustained upgrade opportunity for Quadro powered systems, while also expanding their reach.

We see strong demand in verticals including healthcare, media and entertainment and higher education among others. Healthcare demand was fueled, in part, by COVID-19-related research at Siemens, Oxford and Caption Health. Caption Health received FDA clearance for an update to its AI guided ultrasound, making it easier to perform diagnostics-quality cardiac ultrasounds. And in media and entertainment, demand increased as companies like Disney deployed remote workforce initiatives.

Turning to automotive and robotic autonomous machines. Automotive revenue was $155 million, down 7% year-on-year and down 5% sequentially. The automotive industry is seeing a significant impact from the pandemic, and we expect that to affect our revenue in the second quarter as well, likely declining about 40% from Q1. Despite the near-term challenges, our important work continues. We believe that every machine that moves some day will have autonomous capabilities.

During the quarter, Xpeng introduced the P7, an all-electric sports sedan with innovative level 3 automated driving features powered by the NVIDIA DRIVE AGX Xavier AI compute platform. Our open programmable software defined platform enables Xpeng to run its proprietary software, while also delivering over-the-air updates for new driving features and capabilities. Production deliveries of the P7 with NVIDIA DRIVE begin next month.

Our Ampere architecture will power our next-generation NVIDIA DRIVE platform called Orin, delivering more than 6x the performance of Xavier solutions and 4x better power efficiency. With Ampere scalability, the DRIVE platform will extend from driverless robotaxis, all the way down to in-windshield driver assistance systems, sipping just a few watts of power.

Customers appreciate the top to bottom platform, all based on a single architecture, letting them build one software defined platform for every vehicle in their fleet.

Lastly, in the area of robotics, we announced that BMW Group has selected the new NVIDIA Isaac robotics platforms to automate their factories, utilizing logistic robots built on advanced AI computing and visualization technologies.

Turning to data center. Quarterly revenue was a record $1.14 billion, up 80% year-on-year and up 18% sequentially, crossing the $1 billion mark for the first time. Announced last

week, the A100 is the first Ampere architecture GPU. Although just announced, A100 is in full production, contributed meaningful to Q1 revenue and demand is strong.

Overall, data center demand was solid throughout the quarter. It was also broad-based across hyperscale and vertical industry customers as well as across workloads, including training, inference and high-performance computing. We continue to have solid visibility into Q2. The A100 offers the largest leap in performance to date over our eight generations of GPUs, boosting performance by up to 20x over its predecessor. It is exceptionally versatile, serving as a universal accelerator for the most important high performance workloads, including AI training and inference, as well as data analytics, scientific computing and cloud graphics.

Beyond its leap performance and versatility, the A100 introduces new elastic computing technologies that make it possible to bring rightsized computing power to every job. A multi instance GPU capability allows each A100 to be partitioned into as many as seven smaller GPU instances. Conversely, multiple A100s interconnected by our third generation, NVLink, can operate as one giant GPU for ever larger training tasks.

This makes the A100 ideal for both training and for inference. The A100 will be deployed by the world's leading cloud service providers and system builders, including Alibaba Cloud, Amazon Web Services, Baidu Cloud, Dell Technologies, Google Cloud platform, HPE and Microsoft Azure among others.

It is also getting adopted by several supercomputing centers, including the National Energy Research Scientific Computing Center and the JUWELS Supercomputing Center in Germany and Argonne National Laboratory. We launched and shipped the DGX A100, our third generation DGX and the most advanced AI system in the world. The DGX A100 is configurable from 1 to 56 independent GPUs to deliver elastic software defined data center infrastructure for the most demanding workloads, from AI training and inference to data analytics.

We announced two products for edge AI, the EGX A100 for larger commercial off-the-shelf servers and EGX Jetson Xavier NX for micro-edge servers. Supported by full AI, optimized cloud, native and secure software, the EGX platform is built for AI computing at the edge. With the EGX, hospitals, retail stores, farms and factories can securely carry out real-time processing of the massive amounts of data streaming from trillions of edge sensors.

NVIDIA EGX makes it possible to securely deploy and manage and update fleets of servers remotely. EGX is also ideal for the massive computational challenge of 5G networks, which we are working on with our partners like Ericsson and Mavenir. Additionally, we announced CUDA 11, another important software harnessing the A100's performance and universality to accelerate three of the most complex and fast-growing workloads; recommendation systems, conversational AI and data science.

First, NVIDIA Merlin is a deep recommend data application framework that enables developers to quickly build state-of-the-art recommendation systems, leveraging our pre-

trained models, with billions of users and trillions of items on the Internet. Deep recommend-daters [ph] are the critical engine powering virtually every Internet service.

Second, NVIDIA Jarvis is a GPU accelerated application framework that makes it easy for developers to create, deploy and run end-to-end real-time conversational AI applications that understand terminology unique to each company and its customers, using both vision and speech. Demand for these applications are surging and are the shift to working from home, telemedicine and remote learning.

And third, in the field of data science and data analytics, we announced that we are bringing end-to-end GPU acceleration to Apache Spark, an analytics engine for big data processing that uses more than 500,000 data scientists worldwide. Native GPU acceleration for the entire Spark pipeline from extracting, transforming and loading the data to training to inference, delivers the performance and the scale needed to finally connect the potential of big data with the power of AI.

Adobe has achieved a 7x performance improvement and a 90% cost savings in an initial test using GPU-accelerated data analytics with Spark. Our accelerated computing platform continues to gain momentum, underscored by the tremendous success of GTC digital, our annual GPU Technology Conference, which shifted this spring to an online format.

More than 55,000 online developers in AI research registered for the online event, which includes hundreds of hours of free content from AI practitioners and industry experts who leverage NVIDIA's platforms. Our ecosystem is now 1.8 million developer strong. Times like this truly test a computing platform's mettle and the utility it brings to scientists racing for solutions. Researchers around the world are deploying our GPU computing platform in the fight against COVID-19.

Scientists are combining AI simulation to detect changes in the pneumonia cases, sequence the virus and seek effective biomolecular compound for a vaccine or treatment. The first breakthrough came from researchers at the University of Texas at Austin and National Institute of Health, who used the GPU accelerated application to create the first 3D atomic scale map of the virus using NVIDIA GPUs.

This was followed by researchers at Oak Ridge National Laboratory who screened 8,000 compounds to identify 77 promising drug targets using the world fastest supercomputer, Summit, which is powered by more than 27,000 NVIDIA GPUs. The V100 GPUs at Oak Ridge are in high demand, as they can analyze 17 million compound protein combinations in a day. To help understand the virus spread pattern, The University of California, San Diego, researchers ported the microbiomic analysis software to GPUs in the San Diego supercomputing cluster of 500x analysis speed up from what some people are more susceptible to the virus.

Okay. Moving to the rest of the P&L. Q1 GAAP gross margins were 65.1% and non-GAAP was 65.8%, up sequentially and year-on-year, primarily driven by GeForce GPU product mix and higher data center sales. Q1 GAAP operating expenses were $1.03 billion and

non-GAAP operating expenses were a headwind [ph], $21 million, up 10% and 9% year-on-year, respectively. Q1 GAAP EPS was $1.47, up 130% from a year earlier and non-GAAP EPS was $1.80, up 105% from a year ago.

Q1 cash flow from operations was $909 million. Before I turn to the outlook, let me make a few comments on our Mellanox acquisition. Beyond the strong, strategic and cultural fit that Jensen has discussed, Mellanox has exceptionally strong financial profile. The company reported revenue of $429 million in its March quarter, accelerating to 40% year-on-year growth, with GAAP and non-GAAP gross margins in the mid-to-high 60% range.

We expect the acquisition to be immediately accretive to non-GAAP gross margins, non-GAAP earnings per share and free cash flow. We aim to retain the full Mellanox team and accelerate investments in our combined roadmap, as we jointly innovate on our shared vision for the future of accelerated computing.

With that, let me turn to the outlook of the second quarter of fiscal 2021, which includes a full-quarter contribution from Mellanox. We have assumed in our outlook the potential ongoing impact from COVID-19. We expect our automotive platform sales to be down 40% on a sequential basis and pro viz to decline sequentially.

In gaming, while we will likely see ongoing impact from the partial operations or closures of iCafes and retail stores, we expect that to be largely offset by shift to Etail channels. Overall, the precise magnitude of the impact is difficult to predict, given uncertainties around the reopening of the economy.

Overall, we expect second quarter revenue to be $3.65 billion, plus or minus 2%. The contribution of Mellanox revenue is likely to be in the low-teens percentage range of our total Q2 revenue. We are providing this breakout to help with comparability between Q1 and Q2. But going forward, it will become an integrated part of our data center market platform. GAAP and non-GAAP gross margins are expected to be 58.6% and 66%, respectively, plus or minus 50 basis points. The sequential decline in GAAP gross margins primarily reflects an increase in acquisition-related costs, most of which are non-reoccurring.

GAAP and non-GAAP operating expenses are expected to be approximately $1.52 billion and $1.04 billion, respectively. The sequential change in GAAP operating expenses reflects an increase in stock-based compensation and acquisition related costs. GAAP and non-GAAP operating expenses for the full year are expected to be approximately $5.7 billion and $4.1 billion, respectively. For the full year, stock-based compensation and acquisition-related costs also influence.

GAAP and non-GAAP OI&E are both expected to be an increase of approximately $50 million and $45 million, respectively. GAAP and non-GAAP tax rates are both expected to be 9%, plus or minus 1%, excluding discrete items. Capital expenditures are expected to be approximately $225 million to $250 million.

Further, financial details are included in the CFO commentary and other information available on our IR website. New this quarter, we have also posted an Investor Presentation, summarizing our results and key highlights.

In closing, let me highlight upcoming events for the financial community. Next, Thursday, May 28, we will webcast a presentation and Q&A with Jensen on our recent product announcement, moderated by Evercore. We will also be at Cowen's TMT Conference on May 27, Morgan Stanley's Cloud Secular Winners Conference on June 1st, BofA's Technology Conference on June 2nd, Needham's Fourth Automotive Technology Conference on June 3rd, and NASDAQ investor conference on June 16th.

Operator, we will now open for questions. Can you please poll for questions, please?

Questions And Answers

Operator

Certainly. (Operator Instructions) And your first question comes from Aaron Rakers with Wells Fargo. Please go ahead.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yeah. Thanks and congratulations on the solid quarter. Colette, I'm curious of your commentary around visibility in the data center side. That's the comments over the last couple of quarters. How would you characterize your visibility today relative to maybe what it was last quarter? And how do we think about the visibility in the context of trends maybe into the back half of the calendar year? Thank you.

A - Colette Kress {BIO 18297352 <GO>}

Thanks for the question. You're correct, we have in the quarters ago that we were starting to see improved visibility of the digestion period in the prior overall fiscal year. As we move into Q2, we still have visibility and solid visibility into our overall data centers. So at this time, I'd say, they are relatively about the same of we had seen going into the Q1 period. And we think that is a true indication about our platform and most particularly, our excitement regarding A100 and that's launched.

Now regarding the second half of the year, seeing broad-based growth in both the hyperscale and the vertical record levels in our terms of inferencing continuing to grow as well, as well as we're also expanding in terms of edge AI. Our strong demand of A100 products, including the delta board, but also in terms of our DGXs are just starting a initial ramp. However, we do (Technical Difficulty) so it's still a little bit too early for us to give a true certainty in terms of the macro situation that's in front of us. But again, we feel very good about the demand for A100.

Operator

Stacy Rasgon with Bernstein Research. Please go ahead.

Q - Stacy Rasgon {BIO 16423886 <GO>}

Hi, guys. Thanks for taking my questions. I first wanted to follow up on your gaming commentary. You sort of mentioned a couple of offsets. COVID potentially still a headwind, Etail a tailwind and maybe offsetting each other. Were you trying to suggesting that (Technical Difficulty) gaming was kind of flattish into Q2, but I know it has a typical seasonal pattern, which is (Technical Difficulty) what are the kinds of things we should be thinking about when it comes to seasonality, Colette, in the Q2 around that business segment?

A - Colette Kress {BIO 18297352 <GO>}

Let me start and I'll see if Jensen also wants to add on to it. I think you're talking about (Technical Difficulty) our sequential between Q1, right. Some of the pieces that we had seen related to COVID-19 in Q1 may carry (Technical Difficulty) COVID-19, in fact, had an impact in terms of our retail iCafes. However, as we discussed, we efficiently moved to overall etail. We have normally been seasonally down in desktop between Q1 and Q2. So, we do see the strength in terms of laptops and overall consoles as we move for Q1 to Q2. (Technical Difficulty) grow sequentially (Technical Difficulty) for our overall gaming business.

And I'll turn it over to Jensen to see if he has additional commentary.

A - Jensen Huang {BIO 1782546 <GO>}

No, that was great. That was fantastic.

Q - Stacy Rasgon {BIO 16423886 <GO>}

Yeah. I guess just a follow up (Technical Difficulty). In prior years, we've seen it grow like very strong double digits. Obviously, the mix of the business was different back then. But do you think that the kind of -- I mean, we're thinking kind of is up some what. You don't -- is there any chance that it could be up like on, so what we've seen in terms of like typical levels in the past? Like can you give us any sense of magnitude that would be really helpful?

A - Colette Kress {BIO 18297352 <GO>}

When we think about that sequential growth, we will probably be in the low, moving up to probably the mid-single digits in terms of -- that's what our guidance is right now and we'll just have to see how the quarter goes.

Q - Stacy Rasgon {BIO 16423886 <GO>}

Yeah. That's very helpful.

A - Jensen Huang {BIO 1782546 <GO>}

Stacy, the things that I would add is, I would say, I think the guidance is exactly what Colette mentioned. But if you look at -- look at the big picture, there's a few dynamics that are working really well in our favor. First, the ray tracing is just homerun. Minecraft was

phenomenal. We have (Technical Difficulty) we're shipping just about every game developers signed onto RTX and ray tracing. And I think it's a forgone conclusion that this is the next generation. This is the way computer graphics is going to be in the future. So, I think RTX is a homerun.

The second, the notebooks that we created with RTX and Max-Q is just doing great. We got 100 notebooks and gamebooks designed for either mobile workstations or what we call, NVIDIA Studio, for designers and creators. And the timing was just perfect. With everybody needing to stay at home, mobile gaming platform and a mobile workstation, it was just perfect timing.

And then, of course, you guys know quite well that our Nintendo Switch is doing fantastic. There are -- there are three -- the top 3 games in the world. The top games in the world today are Fortnite, Minecraft and Animal Crossing, are all three games on NVIDIA platforms. And so I think we have a -- we have all the dynamics working in our favor and then we've just got to see how it turns out.

Q - Stacy Rasgon {BIO 16423886 <GO>}

Got it. That's helpful. Thank you, guys.

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Thanks.

Operator

Your next question comes from Joe Moore with Morgan Stanley. Please go ahead.

Q - Joseph Moore {BIO 17644779 <GO>}

Great. Thank you. I wanted to ask about the rollout of Ampere. How quickly does that roll into the various segments between hyperscale, as well as on the DGX side, as well as on the HPC side? And is it a smooth transition? Is there a -- I remember when you launched Volta, there was a (Technical Difficulty) can you tell us how you see that ramping up with the different customer segments?

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Thanks a lot, Joe. So first of all, taking a step back, accelerated computing is now common sense in data centers. It wasn't the case when we first launched Volta. If you went back to Volta, Volta was the first generation that did deep learning training in a really serious way, and it was really focused on training. It was focused on training and high-performance computing. We didn't come until later with the inference version called T4. But over the course of the last five years we've been accelerating workloads that are now diversifying in data centers.

If you take a look at most of the hyperscalers machine learning is now pervasive, deep learning is now pervasive. The notion of accelerating deep learning and machine learning

using our GPUs is now common sense. It didn't use to be. People still saw it as something, something esoteric, but today data centers all over the world expect a very significant part of their data center being accelerated with GPUs.

The number of workloads that we've accelerated since, in the last 5 years has expanded tremendously, whether it's imaging or video or conversational AI or deep recommendation, deep recommender systems that -- probably unquestionably at this point, the most important machine learning model in the world. And so the number of applications we now accelerate is quite diverse. And so that's really -- that's contributed greatly to the ramp of Ampere.

When we started to introduce Ampere to the data centers, it was very common -- sensible to them that they would adopt it. They have a large amount of workload that's already accelerated by the NVIDIA GPUs. And as you know our GPUs are architecturally compatible from generation to generation. We're fully compatible, we're backwards compatible, everything that runs on T4 runs on A100, everything that runs on V100 runs on A100. And so I think the transition is going to be really, really smooth.

On the other hand, because V100 and T4, which by the way, V100 and T4 had a great quarter. It was sequentially up. And then on top of that we grew with the A100 shipment. A100 that was going to be V100 and T4 are now quite broadly adopted in hyperscalers for their AI services, in cloud computing, in vertical industries, as Colette mentioned earlier, which is almost -- roughly half of our overall HPC business, all the way out to the edge, which had a great quarter. And then a much smaller part of course, Supercomputing is important, but it's a very small part of high-performance computing. But that's also -- we also shipped A100 to supercomputing centers.

So I think the general sense of it, the summary of it is that the number of workloads for accelerated computing has continued to grow. The adoption of machine learning and AI and all the clouds and hyperscalers has grown. The common sense of using acceleration is now a foregone conclusion, and so I think we're ramping into a very receptive market with a really fantastic product.

Q - Joseph Moore {BIO 17644779 <GO>}

Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Thanks a lot, Joe.

Operator

Your next question comes from Vivek Arya with Bank of America. Please go ahead.

Q - Vivek Arya {BIO 6781604 <GO>}

Thanks for taking my question, and congratulations on the strong growth and execution. Just a quick clarification, Colette, is 66% kind of the new baseline for gross margin? And then the question Jensen for you is, give us a sense for how much inference as a workload and Ampere as a product are expected to contribute. I'm just curious where you are in terms of growing in the inference and Edge AI market and where are we kind of in the journey of Ampere penetration. Thank you.

A - Colette Kress {BIO 18297352 <GO>}

Let me start on the first question regarding the gross margin, and our gross margin as we look into Q2. We are guiding Q2 non-GAAP gross margins at 66%. This is -- would be another record gross margin quarter. Just as we finished a overall record level, and even as we are continuing right now to ramp our overall Ampere architecture within that.

Q2 also incorporates Mellanox. Mellanox is -- had very similar overall margins to our overall data center margins as well. But we see this new baseline as a great transition and likely to see some changes as we go forward. However, it's still a little early to see where these gross margins will go, but we're very pleased with the overall guidance right now at 66% for Q2.

A - Jensen Huang {BIO 1782546 <GO>}

Accelerated computing is just at the beginning of its journey. If you look at -- I would characterize it as several segments. First, is hyperscaler AI micro services, which is all the services that we enjoy today, that has AI. Whenever you shop on the web it recommends a product. When you're watching a movie, it recommends a movie, or it recommends a song.

All of those -- or recommends news or recommends a friend or recommends a website, the first 10 websites that they recommend, all of these recommenders that are powering the Internet are all based on machine learning today. It's the reason why they're collecting so much data. The more data they can collect the more they could predict your preference.

And that predicting your preference is the core to a personalized Internet. It used to be largely based on CPU approaches. But going forward, it's all based on deep learning approaches. The results are much more superior and a few percentage change in preference prediction accuracy could result in tens of billions of dollars of economics. And so this is very, very big deal and the shift towards deep learning in hyperscale micro services, or AI micro services is still ramping.

Second is cloud. And as you know, cloud is a $100 billion market segment of IT today, growing about 40% into a trillion dollar opportunity. This cloud computing is the single largest IT industry transformation that we ever received. The two powers that is really the force -- the two forces that is really driving our data center business is AI and cloud computing. We're perfectly positioned to benefit from these two powerful forces.

So the second is cloud computing and that journey is -- has a long ways to go. Then the third is industrial edge. In the future -- today it's not -- it's not the case today, but the combination of IoT, 5G, industrial 5G and artificial intelligence, it's going to be -- is going to turn every single industry into a tech industry. And whether it's logistics or warehousing or manufacturing or farming, construction, industrial every single industry will become a tech industry. And there'll be trillions of sensors. And they will be connected to little micro data centers. And those data centers will be in the millions, will be distributed all over the edge. And that journey has just barely started.

We announced three very important partners in three domains, and they are the lead partners that we felt that people would know, but we have several hundred partners that are working with us on edge AI. We announced Walmart for smart retail, we announced the US Postal Service, the world's largest mail sorting service and logistic service. And then we announced this last quarter, BMW, who is working with us to transform their factory into a robotics automated factory of the future.

And so these three applications are great examples of the next phase of artificial intelligence and where Ampere is going to ramp into. And that is just really at its early stages. And so I think it's fair to say that we're really well positioned in the two fundamental forces of IT today, data center, scale computing and artificial intelligence, in the segments that it's going to -- it's going to make a real impact are all gigantic markets.

Hyperscale, AI, cloud and edge AI.

Q - Vivek Arya {BIO 6781604 <GO>}

Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Thanks a lot, Vivek.

Operator

C.J. Muse with Evercore. Please go ahead.

Q - C.J. Muse

Yeah. Good afternoon, and thank you for taking the question. I guess I'm going to ask two. Colette, can you help us with what you think the growth rate for Mellanox could look like in calendar '20? And then, Jensen, a bigger picture question for you. And really not specific to healthcare, more broad based. But how do you think about the long-lasting impact of COVID on worldwide demand for AI? Thank you.

A - Colette Kress {BIO 18297352 <GO>}

C.J., can you help me? You cut out in the middle of your sentence to me. Can you repeat the first part of it for me? Thank you.

Q - C.J. Muse

Sorry about that. I'm curious if you could provide a little hand-holding on what we should think about for growth for Mellanox in calendar '20?

A - Colette Kress {BIO 18297352 <GO>}

At this time, it's a little early for us and as you know, we generally just give a one quarter out. And we're excited to bring the Mellanox team on board, so we can start beginning the future of building products together for the overall market. Again, you've seen their overall performance over the last couple quarters. They had a great last year. They had a great March quarter as well. And we're just going to have to stay tuned to see equally with them what the second half of the year looks for them. Okay.

A - Jensen Huang {BIO 1782546 <GO>}

Yeah, C.J., thanks for the question. The -- this pandemic is really quite tragic and it's reshaping industries and markets. And I think it's going to be structural. I think it's going to remain. And I think your question is really good because now is a good time to think about where to double down. There is a few areas that I believe are going to be structurally changed. And I think that once I say it, it will be very sensible. The first is that, that the world's enterprise digital transformation and moving to the cloud, that's going to accelerate. Every single company can't afford to rely just on on-prem IT. They have to -- they have to be much more resilient. And having a hybrid cloud computing infrastructure is going to provide them the resilience they need. And so that's one. And when the world moves and accelerates into this $1 trillion IT infrastructure transformation, which is now $100 billion into that journey, it's growing 40% a year, I wouldn't be surprised to see that accelerate. And so cloud computing AI is going to accelerate because of that.

The second is the importance of creating a computational defense system. The defense systems of most nations today are based on radar. And yet in the future, our defense systems are going to detect things that are unseeable. It's going to be infectious disease. And I think every nation and government and scientific lab are now gearing up to think about what does it take to create a niche country that is based on computational methods. And NVIDIA is an accelerated computing company. We take something that, that otherwise would take a year, in the case of -- in the case of Oak Ridge and they filter a billion compounds in a day.

And that's what you need to do. You need to find a way to have an accelerating computational defense system that allows you to find insight, detect early warning asap. And then, of course, the computational system has to go through the entire range from mitigation to containment, to living within a monitoring. And so scientific labs are going to be gearing up, national labs are going to be gearing up.

The third part is that AI and robotics, we are going to -- we're going to have to have the ability to be able to do our work remotely. NVIDIA has a lot of robots that are helping us in our labs. And without those robots helping us in our labs, we have a hard time getting our work done. And so we need to have remote autonomous capability for -- to handle all of these, either dangerous circumstances to disinfect environments, to fumigate

environments autonomously, to clean environments, to be able to interact with people where as little as possible in the event of an outbreak. All kinds of robotics applications have been trained [ph] up right now to help society forward in the case of another outbreak.

And then lastly, I think more and more people are going to work permanently from home. There is a strong movement of those companies that are going to support a larger percentage of people working from home. And when people work from home, it's going to clearly increase the single best home entertainment, which is video games. I think video games is going to represent a much larger segment of the overall entertainment budget of society. And so these are some of the trends I would say. I would say cloud computing AI, I would say national labs, a computational defense system, robotics and working from home are structural changes that are going to be here to stay and these dynamics are really good for us.

Operator

Your next question comes from Toshiya Hari with Goldman Sachs. Please go ahead.

Q - Toshiya Hari {BIO 6770302 <GO>}

Hi, guys. Good afternoon, and thank you very much for taking the question. I had one for Colette and then one for Jensen as well, if I may. Colette, I wanted to come back to the gross margin question. You're guiding July essentially flat sequentially, despite what I'm guessing is better mix with Mellanox coming in and automotive guided down 40% sequentially. But I guess the question is what are some of the offsets that are pulling down gross margins in the current quarter? And sort of related to that, how should we be thinking about the cadence in OpEx going forward, given the six month pulling [ph] that you guys talked about on the compensation side?

And then one quick one for Jensen. I was hoping you could comment on the current trade landscape between the US and China. I feel like you guys shouldn't be impacted in a material way directly or not indirectly. But at the same time, given the critical role you play in scientific computing, I can sort of see a scenario where some people may claim that you guys contribute to efforts outside of the US. So if you can kind of speak on that, speak to that, that will be helpful. Thank you.

A - Colette Kress {BIO 18297352 <GO>}

Thanks, Toshiya, for your question. So regarding our gross margins in the second quarter, our second quarter guide at 66% is up sequentially from even a record level in terms of what we had into -- in terms of Q1. This next record that we hope to achieve with our overall guidance is even with including our overall Ampere architecture. So typically when we transition to new architectures, margins can somewhat be a little bit lower on the onset, but tend to kind of move up and trend up over time.

Additionally, as you articulated, our automotive is lower. But also, we're going to see growth in some of our platforms in gaming such as consoles, which may offset those two.

But overall, there's nothing structural to really highlight, other than our mix in business and the ramp of Ampere and its transition.

A - Jensen Huang {BIO 1782546 <GO>}

Let's see the -- the trade tension. We've been living in this environment for some time, Toshiya. And as you know, the trade tension has been in the background for -- coming up on a year, probably quite longer. And China's high performance computing systems are largely based on Chinese electronics anyhow. And so -- so that's I think -- I think our condition won't materially change going forward.

A - Colette Kress {BIO 18297352 <GO>}

So, Toshiya, let me respond to your second question that you had for me, which was regarding to our OpEx and our decision to pull forward our overall (inaudible) into Q2. This is something that we've normally done later in the year. We felt it was prudent during the current COVID-19. Although our employees are quite safe, we just wanted to make sure that their family members also were safe and had the opportunity to have cash upfront. It is about a couple months, about four months earlier than normal. And it is incorporated in our guidance for Q2.

Operator

Your next question comes from Mark Lipacis with Jefferies. Please go ahead.

Q - Mark Lipacis {BIO 2380059 <GO>}

Hi, thanks for taking my question. Question coming back to the A100 and trying to understand how this of kind of fits into the evolution of your solutions set over time and the evolution of the demand for the applications. is -- I mean -- I guess if I think about it going back, you had a solution, which is largely training based and then you kind of introduced solutions that were targeted more inferencing, and now you have a solution, it sounds to my understanding that is -- it's solves both inferencing and training efficiently.

And so I guess I'm wondering is 3 years, 5 years, 10 years down the line Is this part of the kind of general-purpose computing or acceleration framework that you had talked about in the past, Jensen, where Ampere is kind of like an Ampere class product, or is this -- would you still -- should we still expect to see inferencing specific solutions in the market and then training specific solutions, and then an Ampere solution for a different class application, if you could provide a framework for thinking about Ampere and those context, I think that would be helpful. Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Thanks a lot, Mark. Good question. I think the -- if you take a step back, currently in our data centers, the current set up in data centers, starting from probably all the way back 6, 7 years ago, but really accelerating in the last five years and then really accelerating in the last couple of years, we learned our way into it. There are three classes of workloads and they kind of came into acceleration over time.

The first class of workload that we discovered was -- the major workload was deep learning training. And the ideal set up for that today prior to Ampere, or yesterday prior to Ampere is the V100 SXM with NVLink, 8 GPUs on one board, and that architecture is called scale-up. It's like a supercomputer architectures -- it's a weather simulation architecture. It's -- if you were trying to build the largest possible computing node you can for one operating system, called scale-up.

And the second thing that we learned along the way was then cloud computing, started to grow because researchers around the world needed to get access to a accelerated platform for developing their machine learning algorithms, and because they have a different degree of budget and they want to get into it a little bit more lightly and have the ability to scale up to larger notes, the perfect model for that was actually a B100 PCI Express, not SXM, but PCI Express that allows you to offer one GPU all the way up to many GPUs.

And so that versatility, V100 PCI Express, not as scalable in performance as the B100 SXMs but it was much more flexible for renters, cloud renting was really quite ideal. And then we started to get into inference and we're now in our 7th generation of TensorRT, TensorRT 7. Along the way we've been able to accelerate more and more, and today we largely accelerate every deep learning inference computational graph that's out there.

And the ideal GPU for that was something that has the reduced precision, which is called 8-bit integer, reduced precision, not with electronics that is focused more for inference, and because inference is a scale-out application where you have millions of queries and each one of the queries are quite small, versus scale-up where you have one training job, and that one training job is running for days. It could be running for days and sometimes even weeks. And so scale-up application is for one user that uses it for a long period of time on a very large machine.

Scale-out is for millions of users, each one of them have a very small query and that query could last hundreds of milliseconds, where ideally you'd like to get it done in hundreds of milliseconds. And so notice, I've got three different architecture in the data center today. Most data centers today has a storage server, has CPU servers and it has scale up acceleration servers with Volta, has scale-out servers with GeForce and then has scale cloud computing flexible servers, based on A100.

And so the ability to predict workload is so hard and therefore the utilization of these systems will be spiky, and so we created an architecture that allows for three things. So things -- the three characteristics of Ampere are number one, it is the greatest generational leap in history. I mean, I don't remember a generation where we increased throughput for training and inference by 20x. It's just a gigantic -- for training and for influence, it is a gigantic GeForce [ph]. The second, it's the first architecture that is unified.

We could use this computational -- the computation engine of Ampere accelerates the moment the data comes into the data center from data processing, it's called ETL, the engine, which many of you probably know is the single most important computational engine in the world today for big data. It used to be Hadoop and now it's Spark. Spark is

used all over the world, 16,000 [ph] customers. We finally have the ability to accelerate that. And then it's -- Ampere is also good for training, deep learning machine learning, XGBoost as well as deep learning, all the way out to inference.

And so we now have a unified acceleration platform for the entire workload. And then the third thing is it's the first GPU ever, the first acceleration platform ever that's elastic. You can reconfigure it. You could configure it for either scale-up or you can configure it for scale-out. When you configure for scale-up, you're getting a whole bunch of GPUs together using NVLink and it creates this one gigantic GPU.

When you want to scale it out that same computation node becomes 56 small GPUs. Each one of those 56 partitions, each one is more powerful than Volta. I mean it's really quite extraordinary. And so Ampere is a breakthrough on all of these fronts, for-performance or the fact that it unifies the workload and you can now have one acceleration cluster and then number 3, it's elastic. You could use it in the cloud, you could use for inference, you can use of training. And so the versatility of Ampere is the thing that I'm most excited about. And now you could have one acceleration cluster that serves all of your needs.

Q - Mark Lipacis {BIO 2380059 <GO>}

Thank you. It's very helpful.

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Thanks a lot, Mark.

Operator

Your next question comes from Timothy Arcuri with UBS. Please go ahead.

Q - Timothy Arcuri {BIO 3824613 <GO>}

Thanks a lot. Actually I had two, I guess Jensen first for you. Just on the data center business, things have been very strong recently. Obviously there's always concerns that customer are pulling in CapEx. But it sounds like you have pretty good visibility into July. But I guess last time (technical difficulty) kind of (technical difficulty) really was so low that you would be immune to any digestion, but that wasn't the case. I guess I'm wondering, things are different (technical difficulty) but my question is, how do you handicap your ability to this time maybe get through any digestion on the CapEx side?

And then I guess second question, Colette stock comp has been running like 220 a quarter and the guidance implies that it goes like 460 a quarter. So it goes up a lot. Is that all executive retention and is that sort of the right level as you look into 2021. Thanks.

A - Jensen Huang {BIO 1782546 <GO>}

Colette, did you want to handle that first and then I'll do the --?

A - Colette Kress {BIO 18297352 <GO>}

Sure. So let me help you on the overall GAAP adjustments. So the delta between our GAAP OpEx and our non-GAAP OpEx. If you look at it for the full year and what we guided, we probably have about $1.55 billion associated with GAAP level expenses. Keep in mind, there is more in there then just our stock-based compensation. We have also incorporated the accounting that we will do for the overall Mellanox. And a really good portion of those costs are associated with the amortization of intangibles, and also in terms of intangibles and also in terms of acquisition-related costs and deals [ph]. So, our stock-based compensation includes what we need for NVIDIA and also the onboarding of Mellanox. There is some retention with the overall onboarding of Mellanox. But for the most part, it is just working them into the year for three quarters, which is influencing the stock-based compensation.

A - Jensen Huang {BIO 1782546 <GO>}

Tim, there are several differences between our condition then and our condition today. So the first -- the first difference is the diversity of workload we now accelerate. Back then, we were early in our inference. We're still early in our inference and most of the data center acceleration was used for deep learning. And so, today, the versatility spans from data processing to deep learning and the number of different types of AI models that's been trained for deep learning is growing tremendously, from detecting, from training video, from training a model to detecting unsafe video. The natural language understanding of the conversational AI to now a gigantic movement towards deep recommender systems.

And so the number of different models that are being trained is growing. The size of the models are gigantic. Recommendation systems are gigantic. They are training on models that are hundreds -- the data size is hundreds of terabytes, hundreds of terabytes. And it would take 10s of -- 100s of servers to hold all of the data that is needed to train these recommender systems.

And so the diversity from data analytics to training all the different models to the inference of all the different models, we didn't inference recurrent elements at the time, which is probably the most important model today, language models, speech models (inaudible). And so those models were early for us at the time. So number one is the diversity of workloads.

The second is the acceleration to cloud computing. I think that accelerated cloud computing is a movement that is going to be a multi-year, if not a decade-long transition from where we are today. It's only $100 billion industry segment of the IT industry. It's going to be a trillion dollars someday and that movement is just starting. We're also much more diversified out of the clouds. At the time, it was largely where our acceleration went for a deep learning. And today, hyperscale only represents about half. And so we've diversified significantly out of cloud -- not out of cloud, but including vertical industries. And a lot of that has to do with edge AI and inference. And as I mentioned earlier, we're working with Walmart and BMW and USPS. And that's just the tip of the iceberg.

And so, I think the conditions are a little different. And then what I would say lastly is Ampere. I mean, we ramped a few weeks. Even though it was quite significant, it was a

great ramp. The demand is fantastic. It is the best ramp we've ever had. The demand is the strongest we've ever had in data centers. And we're starting to ramp a multi-year ramp. And so those are some of the differences. I think the conditions are very different.

Q - Timothy Arcuri {BIO 3824613 <GO>}

Thank you, Jensen.

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Thanks a lot, Tim.

Operator

Your next question comes from Harlan Sur with J.P. Morgan. Please go ahead.

Q - Harlan Sur {BIO 6539622 <GO>}

Good afternoon. Thanks for taking my question. Jensen, the team [ph] has shown the importance of networking fabric and the Mellanox acquisition. For example, when you guys moved from both the DGX-1 to both the DGX-2, you guys didn't change the GPU chipset. But by adding a custom networking fabric chip and more Mellanox network interface cards, among other things, you guys drove a pretty significant improvement in performance for GPU.

But now when we think about scaling out compute acceleration to data center skilled implementation, how does Mellanox's Ethernet switching platforms differ from those provided by other large networking OEMs, some of whom have been your long-term partners and then how does the Cumulus acquisition fit into the switching and networking strategy as well?

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Great. Thanks a lot, Harlan. Appreciate the question. So DGX, this is our third-generation DGX, and it's really successful. People love it. It's the most advanced AI instrument in the world. If you're a serious AI researcher, this is your instrument. And in the DGX [ph], there are eight A100s and there are nine Mellanox mix, the highest-speed mix they have. And so we have a great appreciation for HyperPoint's networking. HyperPoint's networking and HyperPoint's computing go hand in hand.

And the reason for that is because the problems we're trying to solve no longer fit in one computer. No matter how big it is. And so it has to be distributed. And when you distribute a computational workload of such intense scale, the communications overhead becomes one of its greatest bottlenecks, which is the reason why Mellanox is so valuable. There is a good reason why this company is so precious and really a jewel and one of a kind. And so it's -- and it's not just about the link speed. It's not mostly. I mean, we just have a deep appreciation for software. It's a combination of architecture and software and electronics design, chip design. And that combination, Mellanox is just world-class. And

that's the reason why they're in 60% of the world's supercomputers. That's why they are in 100% of the AI supercomputers.

And the understanding of large scale distributed computing is second to none. Now in the world -- and I just talked about scale out. And you are absolutely right. Now the question is why scale-out? And the reason for that is this. This is the reason why they are doing so well. The movement towards disaggregated microservice applications where containers -- microservice containers are distributed all over the data center and orchestrated, so that the workload could be distributed across a very large hyperscale data center. That architecture and you probably know the three most important application in my estimation in the world today, number one would be (Technical Difficulty) and PyTorch. Number two would be Spark and number three would be Kubernetes. You could rank it, however, your desire. And these three applications in the case of Kubernetes. It's a brand new type of application where the application is broken up with a small pieces and orchestrated across an entire data center. And because it's broken up into small pieces and orchestrated across the entire data center, the networking between that compute nodes becomes the bottleneck again.

And that's the reason why they're doing so well, by increasing the network performance, by offloading the communications after CPUs. You increase the throughput of a data center tremendously. And so it's the reason why they had a record quarter last quarter. It's the reason why they've been going 27% per year. And their stock was that, their integration into the hyperscale cloud companies. There are low latency, their incredibly low latency of their link makes them really unique, even -- whether it's Ethernet or InfiniBand in both cases. And so it's a really fantastic stack.

And then lastly, Cumulus, we would like to innovate in this world, where the world is moving away from just a CPU as a compute node. The new computing unit, a software developer is writing a piece of software that runs on the entire data center. In the future going forward, the fundamental computing unit is an entire datacenter. It's so incredible. It is utterly incredible. You run an application, one human can run an application and it would literally activate an entire data center. And in that world, we would like to be able to innovate from end-to-end, from networking storage, security, everything has to be secured in the future, so that we can reduce the attack surface down to practically nothing.

And so, networking storage, security are all completely offloaded, all incredibly low latency, all incredibly high performance and all the way to compute, all the way through the switch. And then, the second thing is we'd like to be able to innovate across the entire stack. You know that NVIDIA is -- it's just supremely obsessed about software stacks and the reason for that is because software creates markets, you can't create new markets like we're talking about, whether it's computational health care or autonomous driving or robotics or conversational AI or recommender systems or edge AI, all of that requires software staks, it takes software to create markets. And so, our obsession about software and creating open platforms for the ecosystem in all of our developer partners, Cumulus plays perfectly into that.

They are -- they pioneered the open networking stack. And they pioneered in a lot of way, software defined data centers and so, we are super, super excited about the team and now we have the ability to innovate in a data center scale world from end to end and then from top to bottom the entire stack. Okay?

Q - Harlan Sur {BIO 6539622 <GO>}

Yes. Thank you, Jensen.

A - Jensen Huang {BIO 1782546 <GO>}

Hey, thanks a lot.

Operator

Your next question comes from William Stein with SunTrust. Please go ahead.

Q - William Stein {BIO 15106707 <GO>}

Great, thank you for taking my question. Jensen, I'd like to focus on something you said, I think it was in one of your earlier responses, you said something about a very significant part of data centers are now accelerated with GPUs. I'm sort of curious how to interpret that. If we think about sort of the evolution of compute architecture going from almost entirely, let's say racks and racks of CPUs to some future day where we have many more accelerators and maybe a much smaller number of CPUs relative to those. Maybe you can talk to us about where we are in terms of that architectural shift and where you think it goes sort of longer term, where we are in the position of that.

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. I appreciate the question. And this for computer architecture geeks and people who follow history, you know well that in the entire history of time, there are only two computing architectures that has made it so far, which is -- one of them is x86, the other one is ARM in any reasonable way. And if you get an ARM computer, you get an x86 computer, you can program it.

And in fact, there is no such thing as an accelerated computing platform until we came along. And today we're the only computing -- accelerated computing platform that you could really largely address, we're in every cloud, we're in every computer company, we're in every country, we are at every single size, we're -- and we accelerate applications from computer graphics, to video games to scientific computing, to workstations, to machine learning, to robotics. This journey to 20 some-odd years, inside our company, it took 20 some-odd years and we've been focused on accelerated computing since the beginning of our company.

And we made a general purpose, we made it general purpose really starting an endeavor CG, C for graphics and then it became CUDA. And we have been working on accelerated computing for quite a long time and I think at this point, it's a forgone conclusion that accelerated computing has reached the tipping point as well beyond it. The number of

developers this year that support -- that we supported was almost 2 million developers around the world and it's growing what appears to be exponentially. And so I think accelerated computing is now a well-established -- NVIDIA accelerated computing is well established, it's common sense and people who are designing data centers expect to put accelerated computing in it.

The question is how much? How much accelerated computing do you use and what part of the date in your pipeline do you do it? And the big -- the gigantic breakthrough of course we know well now and NVIDIA is recognized as one of the three pillars that ignited the modern AI, the big bang of modern AI and the other two pillars are of course is deep learning algorithm and the abundance of data. And so the three -- these three ingredients came together and it was, people use NVIDIA accelerated computing largely for training, but over time, we expanded training to have a lot more models and as I mentioned earlier, the single most important model of machine learning today is the recommender system.

It's the most important model because it's the only way that you and I could use the Internet in any reasonable way. It's the only way that you and I could use a shopping website or a video web or a video app or a music app or a book or news or anything. And so it is the engine of the Internet from the consumers' perspective and the company's perspective, it is the engine of commerce, without the recommender system, there is no way they could possibly make money. And so, their accuracy in predicting user preferences is core to everything they do, you just go up and down the list of every company. And that engine is gigantic, it is a -- just a gigantic engine. And from the data processing part of it, which is the reason why we went and spent three years on Spark and RAPIDs, which made Spark possible and all the work that we did on NVLink and all that stuff was really focused on big data analytics.

The second is, all of the training of the deep learning models and influence. So the number of applications, the footprint of accelerated computing has grown tremendously and its importance has grown tremendously, because of the applications are the most important applications of these companies. And so, I think when I mentioned, when I said that, that acceleration was -- is still growing, it is, but the major workloads, the most important workloads of the world's most important companies are now solidly require acceleration. And so, I'm looking forward to a really exciting ramp for Ampere for all the reasons that I've just mentioned.

Operator

Your next question comes from John Pitzer with Credit Suisse. Please go ahead.

Q - John William Pitzer {BIO 1541792 <GO>}

Yeah, hi guys. Thanks for letting me ask the question. Just two quick ones, Colette, I hate to ask something as mundane as OpEx, but just given the full-year guide, there is sort of a lot to unpack and you talked about some of it like the raises. I mean, I think you also probably have some COVID plus or minuses in that, I think there is an extra week this year

as well. And then of course, there is Mellanox and how you're thinking about investing in that asset.

I guess I'm just kind of curious, when we look at the full year guide, is there something structural going on, on OpEx as you try to take advantage of all these opportunities or it can be used as a sort of a guide post to how you're thinking about revenue for the back half of the year as well? How do I understand that?And then Jensen, just a quick one for you, kind of makes sense to me that COVID is accelerating activity in sort of HPC and Hyperscale and maybe even in certain verticals like healthcare, but in the other verticals, has the sort of shelter in place kind of hurt engagement and could we actually come out of COIVD with some pent-up demand in those vertical markets?

A - Colette Kress {BIO 18297352 <GO>}

Okay. Thanks, John for the question. Let's start from the first perspective on the overall OpEx for the year. We've guided the non-GAAP at approximately $4.1 billion for the year. Yes, that incorporates three full quarters of Mellanox. Mellanox and its employees we have about close to 3,000 Mellanox employees coming on board. You are correct, we have a 53rd week in this quarter, excuse me, not this quarter, this year and that is -- has been outlined in the SEC filings that you should expect that as well. We pulled forward a little bit our focal by several months in order to take care of our employees.

And then lastly, we are investing in our business, we see some great opportunities. We've seen some great results from our investment and there's more to do. We are hiring and investing in those businesses. So there is nothing different structurally, but just this onset of Mellanox and are investing together, I think will produce long-term great results.

A - Jensen Huang {BIO 1782546 <GO>}

And as usual, John, you know that we're investing into the IT industry's largest opportunities: cloud computing and AI. And then after these two opportunities is edge AI. And so, we're looking down the fairway with some pretty extraordinary opportunities, but as usual, we're thoughtful about the rate of investment and we're well managed and NVIDIA's leadership team is our excellent managers and you can count on us to continue to do that.

Simona, what was John's question, could you just give me one hint -- I had it at the tip of my mind --

Q - John William Pitzer {BIO 1541792 <GO>}

It's just the idea of engagement levels in verticals just with shelter in place, has that hampered --

A - Jensen Huang {BIO 1782546 <GO>}

Oh, yeah. Right. Right. Yeah, right, a few -- some of the industries have an effected. We already mentioned automotive industry, the automotive industry has been grounded to a halt. Manufacturing has largely stopped and you saw that in our guidance. We expect

automotive to be down 40% quarter-to-quarter. It's not going to remain that way, it's going to come back, and with nobody knows -- nobody knows what level it's going to come back and how long, but it's going to come back and there is no question in my mind that the automotive industry, they're hunkered down right now, but they will absolutely invest in the future of autonomous vehicles, they have to, or they will be extinct.

It's not possible not to have autonomous capability in the future of everything that moves, not so that it could just completely drive without you, that's a nice benefit too. But mostly because of safety and comfort and just a joy of what seems like the car is reading your mind. And of course you are still responsible for driving it and -- but it just seems to be coasting down the road, reading your mind and helping you. And so, I think the future of autonomous vehicles is a certainty. People recognized the incredible economics that the pioneer Tesla is enjoying and the industry is going to go after it. The future car companies are going to be software defined companies and then the technology companies, and they would love to have an economic that allows them to enjoy the installed base of their fleet. And so, they're going to go after it and so, this is -- I am certain that this is going to come back and I have every confidence it's going to come back.

And let's see, the energy sectors are -- have been impacted, the retail sector has been impacted, there is -- those aren't large industries for us, but nonetheless they're impacted. The impact in some of the industries is accelerating their focus in robotics, like for example, on the one hand, BMW has obviously impacted in manufacturing, which is the reason why they're moving so rapidly towards robotics, they have to figure out a way to get robotics into their factories. And same thing with retail, you're going to see a lot more robotic support in retail, you're going to see a lot more robotic support in warehouses, in logistics and so during this time when the market, when the industry is disrupted and impacted, it allows the market leaders to really lean into investing into the future. And so when they come back, they will be coming back stronger than ever.

Q - John William Pitzer {BIO 1541792 <GO>}

Thank you.

Operator

And your next question comes from Matt Ramsay with Cowen. Please go ahead.

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

Thank you very much. Good afternoon. Two different topics Jensen, well, first of all, congrats on Ampere. It's a heck of a product. The first think I want you to --

A - Jensen Huang {BIO 1782546 <GO>}

Thank you, Matt. I'm so proud of it.

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

First question is, it might have been a little bit hard to talk when the deal was pending about this topic, but now that it's closed, maybe you could talk a little bit about opportunities to innovate on and customize the Mellanox stack and the balance of having an industry standard. And the second one is E3 cancelled, Computex moved around, at the same time, there is obviously stay at home gaming demand. Just how are you thinking about gaming product launch logistics and any comments on there would be really helpful. Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Thanks a lot Matt, appreciate the questions. I'll go backwards, because it's kind of cool. On the one hand, I do miss that we can't engage the developers face-to-face, it's just so much fun. GTC is seeing all that work and the hundreds of papers that are presented. I learned so much each time and frankly, I really enjoyed the Analyst Meetings that we have, and so there is all kinds of stuff that I missed about the physical GTC, but here's the amazing thing. We had an almost 58,000 attendees. The GTC kitchen keynote, I did it from my kitchen just right behind me and the kitchen keynote has been viewed almost 4 million times and the video is incredible.

And so I think our reach is -- it could be quite great. And so I'm not too -- we've got amazing marketing team and just we got great people, they're going to find a way to reach our gamers, and whenever we launch something next, you know the gamers are going to be and our customers are going be, our end markets are going to be really excited to see it. And so, I'm very confident that we're going to do just fine. Matt, what was the question before, I should never do backward.

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

Just the industry standard versus customization of Mellanox opportunity.

A - Jensen Huang {BIO 1782546 <GO>}

I see, okay, yeah. There is -- we worked so closely with Mellanox over the years and on the day that we announced GTC, you could see the number of products that we have working together. The product Synergies are really incredible and product the synergies include a lot of software development that went in and a lot of architectural development that went in. DGX comes with nine Mellanox NICs I mentioned. If you look at our data center, we shipped -- before we shipped DGXs to the customers, we shipped it to our own engineers. And the reason for that is because every single product in our company has AI in it. From Jarvis to Metropolis to Merlin to DRIVE to Clara to Isaac to -- right, all of our products has AI in it and we're accelerating frameworks for all of the AI industry. And Ampere comes with a brand new numerical format called Tensor Float 32. And TF32 is just a fantastic new numerical format and the performance is incredible and we had to get it integrated in with the industry standard frameworks. And now, Tensor Float comes standard with Tensor Float with -- NVIDIA's TF32 and PyTorch comes standard with TF32.

And so we need our own large-scale data center and so, the first customer we shipped to was ourselves. And then we started shipping as quickly as we could, to all of the customers. You saw that in our data center, in our supercomputer, we have 170 state-of-

the-art brand new Mellanox switches and almost 1,500 200-gigabit per second Mellanox NICs. And 15 kilometers of cables, fiber optics cables and that is one of the most powerful supercomputers in the world today and it's based on Ampere. And so, we have a great deal of work that we did there together. We announced our first edge computer between us and Mellanox in this new card, we call it the EGX A100. It integrates Ampere and it integrates Mellanox's CX6 Dx which is designed for 5G telcos and edge computing. It's incredible security and has single root of trust and it's virtualized.

And so basically we -- this EGX A100 when you put it into a standard x86 server, turns that server into a cloud computer in a box, the entire capability of a cloud -- of a state-of-the-art cloud, which is cloud native, it's secure, it has incredible AI processing, it's now completely hyperconverged inside one box. The technology that made EGX A100 is really quite remarkable. And so you can see all the different product synergies that we have in working together, we couldn't have done Spark acceleration without the collaboration with Mellanox. They worked on this PC of networking software called ECX. We worked on NCCL together, it made possible the infrastructure for a large scale distributed computing.

I mean it's just the list goes on and on and on. And so we -- the two teams have great chemistry, the culture -- it's a great culture fit, and I love working with them. And right out of the shoot, you saw all of the great product synergies and that it made possible because of the combination.

Operator

That is all the time we have for questions. I will turn the call back to Jensen Huang for closing remarks.

A - Jensen Huang {BIO 1782546 <GO>}

No, it's coming. I'm just. Thank you. We had a great and busy quarter. With our announcements we highlighted several initiatives. First computing is moving to data center scale. Where computing and networking go hand in hand, the acquisition of Mellanox gives us deep expertise and scale to innovate from end-to-end.

Second, AI is the most powerful technology force of our time. Our Ampere generation offer several breakthroughs. It is the largest ever generational leap 20X in training and inference throughput, the first unified acceleration platform for data analytics, machine learning, deep learning, training and influence. And the first elastic accelerator that can be configured for scale up applications like training to scale out applications like inference. Ampere is fast, it's universal and it's elastic. It's going to re-architect the modern data center.

Third, we are opening large new markets with AI software application frameworks such as Clara for healthcare, DRIVE for autonomous vehicles, Isaac for robotics, Jarvis for conversational AI, Metropolis for edge IoT, Ariel for 5G and Merlin for the very important recommender systems. And then finally, we have built up multiple engines of accelerated

computing growth, RTx computer graphics, artificial intelligence and data center scale computing from cloud to edge.

I look forward to updating you on our progress next quarter. Thanks everybody.

Operator

This concludes today's conference call. You may now disconnect.