# Rosenblatt 3rd Annual Technology Summit

## Company Participants

- Ian Buck, Vice President and General Manager, Hyperscale and HPC
- Simona Jankowski, Vice President and Head of Investor Relations

## Other Participants

- Hans Mosesmann, Analyst, Rosenblatt Securities

## Presentation

### Hans Mosesmann {BIO 1522582 <GO>}

Good morning. Good afternoon, everybody. Thank you for joining us, for the NVIDIA Fireside. Before we get started, Simona Jankowski is going to read us disclosures and then we'll jump in.

Simona?

### Simona Jankowski {BIO 7131672 <GO>}

Yes. Good morning, and thank you very much for hosting us. I just wanted to quickly remind the audience that our comments today may contain forward-looking statements, and investors are advised to read our reports filed with the SEC for information that relates to the risks and uncertainties facing our business.

Back over to you.

### Hans Mosesmann {BIO 1522582 <GO>}

Simona, thanks. Well, we're delighted to have Ian Buck. It's been a couple of years. We couldn't talk last year, there were some conflicts. Ian is a veteran. He runs everything that has to do with accelerated computing at NVIDIA, which is what the interest level is all about these days as it relates to AI. It includes all hardware, all the software, all the third-party enablement and marketing activities.

Ian is known for being basically I'll call him the father of CUDA which has led to having a formidable moat around NVIDIA's business in terms of compiler technology, acceleration libraries, framework optimizations and so on. So there couldn't be a better person I think to talk about what's happening in the world of AI, be and how NVIDIA is playing its part in all of this.

So, Ian, welcome. How are you?

### Ian Buck {BIO 18454865 <GO>}

Pretty busy, as you can imagine. So, yeah, it's been an amazing journey and an amazing last few years since we've talked actually in last -- actually last six months, even more exciting, definitely riding that exponential. So we're cooking here. Yeah.

## Questions And Answers

### Q - Hans Mosesmann {BIO 1522582 <GO>}

Okay, great. How about -- if we just start. You've been at NVIDIA for 20 years -- almost 20 years, and it's gotten even more intense in terms of what's going on. What's the state of AI today? I know that there's lots of discussions on ChatGPT, generative AI. Just briefly, what's the state of AI in terms of NVIDIA's view, and how you guys are participating? And it's an open-ended question, but I think we started off with that and we can go from there.

### A - Ian Buck {BIO 18454865 <GO>}

Yeah. Like you said, it's been a 20-year journey so far in accelerated computing as a whole, and the -- it started with making our GPUs more programmable, and launching CUDA in 2006, and investing in that ecosystem since that. Internally within NVIDIA, building up a software foundation platform for accelerated computing, as well as working with everybody in the ecosystem -- everybody in the ecosystem to enable GPUs as a computing platform.

That of course was a broad goal. There were certain markets in high-performance computing, simulation and others that adopted first and that's been broadening since then, of course, since 2012, AI. We didn't invent AI. [ph] AI found us, but because of the activation and making it everywhere, putting it into the GPU, those researchers have been at Canada were able to find and realize that this thing they were working on neural networks with (inaudible) and others.

Turns out the math was a pretty good fit for what we're doing in -- for CUDA. So it took off from there. There's been a couple of inflection points along the way for AI. That first one which is basically the initial work by (inaudible) in Canada was all the back in 2012, and initially enriching that competition. And that was the starting point and really that was AI for basic image recognition. What is this a picture of, a beach ball stop sign and a birthday party. And that image recognition expanded into other forms of what is this -- what is this statement of would there be images and sentiment, so taking their view on a web page or tweet and understanding its positive or negative sentiment to understanding content. And that described the initial -- that concept of AI, the use of AI was the initial production ramp that we saw. And we probably all remember the -- Jeff Dean talking about finding cats and videos as an example.

### Q - Hans Mosesmann {BIO 1522582 <GO>}

All right.

## A - Ian Buck {BIO 18454865 <GO>}

And obviously, for the hyperscalers and the cloud providers and social media and the internet needed to understand their content. It's the first place where they can really turn their data and use AI to understand what this -- what people were posting, reviews, products, et cetera. And the shift and next shift in AI, those became more and more capable, and then along the way, AI shifted from a recognition problem to a generative one, being able to not just understand the content, be it text, speech, video, whatever, but be able to generate meaningful content, to create content and to create a product description that would want users to click on a link or. And that started small and really hit -- and in fact, with set -- with actually BERT. If you remember the original BERT model, which was the first transformer-based model, it was its ability to not just understand textbook, produce simple text statements and general text.

In fact, NVIDIA was one of the -- we had a BERT day, if you remember. In fact, it was got noticed even in the markets to identify this idea of -- this new kind of neural network called the transformer that could understand text. Prior to that, most applications in AI were convolution-based. They were basically looking at neighborhoods of information and building up understanding from localized data. This makes sense in image recognition to recognize the face. You first recognize individual shapes in certain positions in certain places. My face has two circles, a nose and a line from mouth, and you build up the notion of a face which is localized.

Language is different. Language has all sorts of interactions. What I'm speaking about right now is filled with pronounce and context. It's only known for other parts of the text or speech that are far away from what I'm saying right now to understand it. Transformers were based around this idea of attention, figuring out that those distant relationships and incorporating them into the neural network. It started with BERT and mentioned by Google, and then it took off from there and we offer GPT, and the T in GPT is transformer accessing idea.

And NVIDIA, that was obviously, swarming convolutions, image recognition, CNNs, and now we focus -- no, we still do that and it's still a growing used-case and important one. But now transformers have taken over, including some video to understand just relationships in the primary use case for that in speech and human understanding.

Speech and language are hard a problem. If you think about computer vision like dogs, cats, even bugs can do basic computer vision. And from a brain perspective, it's a -- you can find highly tune in to do a reasonable job. And only really humans have the gift of language and it's built upon the -- a deep understanding of knowledge. That was the first -- that was a next inflection point, transformers, which is sort of understanding the knowledge and being able to connect knowledge with language. So -- I mean, most of the understanding and a little bit of generation.

Today, we're in the era of generative AI. It started with -- in two areas. It's two areas that kick it off. One is obviously image generation. Being able to describe it like a picture of a teddy bear swimming (inaudible) and it generates a picture of that. And generative AI of ChatGPT, being able to have a conversation. I have understanding what I'm saying and repeating back and extracting information. There is no database back there. It's one large neural network. And with generative AI, we not only can make -- we've moved from an era of recognition, recognition only to -- which is important. I can pick that data and understand what my content and make decisions based on that informed -- what AI is informing me. But now AI itself can provide the content, provide a review, provide mid text, engage with customers, generally, which is help artists, optimize business, build new applications and build new kinds of services.

What's interesting also though, unlike other revelations of like in the PC space or mobile space, you guys all have seen and the new kinds of applications, new kinds of platforms, new kinds of software, and this one, generative AI is actually making the old stuff more interesting. We look at Office 365, and I think could be -- I should -- probably doesn't like me saying this, but my stuff listening more to it, Excel isn't that interesting and Words wasn't that interesting, but with generative AI, wow, it's way interesting, yeah.

So in this revolution, we're seeing generative AI opt to create new startups, new kinds of services, but it's also making all the old stuff, super interesting again too, which is a fun double exponential. So that's where we're at. I think we're really at the -- that cusp of that begins at generative AI. Everyone sees the opportunities. You guys do. The market does and the VC community -- the investor community does, and you see the amazing startups there are being created. And it's -- that's what makes this real super fun right now is seeing all the different applications of the new services and old that are getting amplified and changed with AI.

## Q - Hans Mosesmann {BIO 1522582 <GO>}

Hey, Ian, the compute, I think a lot of people in the industry or observers talk about AI and parameter complexity doubling every three, four, five months. What does that? How much longer do we half through that? And what are the compute implications from -- for NVIDIA and for the industry currently use or custom ASICs or even FI? [ph]

## A - Ian Buck {BIO 18454865 <GO>}

Yeah. That's a great question, one I get asked a lot. So first off, to do generative AI, you have to -- the AI has to have knowledge, and not like access to knowledge, certainly, having access to knowledge like a database something that courier or pro found providers is important and most do. But then the -- the reason why GPT is so big or mega -- the megatrend $530 billion model that we trained on our supercomputer is so large. It has to capture knowledge at some level and/or be a starting point for a generative model. So that drives the model size up.

The other thing that drives up, but it's not just model size. Well, first of all, model size tends to limit the -- you know, bigger the model, you want it extreme, in fact. [ph]

And people build models to limit that's practical. So they don't want to wait. And it's not just one training job. To build a model at that scale, you are constantly iterating on that model, on the data, on the tuning of the parameters to give to converge to a level of intelligence. There is a lot of AI that training drops at don't complete but informe to inform the next one actually. So it takes many months or some years to build a truly intelligent model. The final change is the potential convergence of that effort.

The other thing that drives -- the challenge tends to be training time. Nobody wants to train more than and most a month or two. I think it's, you go past that, the productivity of the just -- it's hard to innovate if you're waiting that long. So the size of the model tends to be a factor of how much capacity they can put in place, and our productivity is at scale, but there's been a general rule like the people deal researchers, the ones really developing the stuff, you don't. They start building foundation models, don't really want to train -- having training jobs that take more than a month because they're just too inpatient.

So as we make faster GPUs, as we figure out how to connect them faster together with InfiniBand to build more optimized infrastructure to do things like Grace Hopper and the new DGX and GH 200, and their productivity increases, what they can train in roughly a month, the model substantially must get bigger because it gets more intelligent.

I will say one other thing though that model size is only one metric, and model size is measured in parameters. $175 billion is typical for GPT. People -- we trained 500. There's trillion perimeter models behind closed doors that they're starting to get a little more secretive and not releasing these huge models that drive us the opportunity and massive intelligence.

The other thing that's driving is model -- the design of the layers, continually tuning the intelligence at each layer, making them more optimized, more clever at each layer, which reduces the complexity per layer. That is noise capturing and parameters because each layer has a bunch of math and calculations in it instead of just being naive connections if you will. And the human brain by the way similar. We have different kinds of neurons for vision processing versus auditory versus memory. So we specialize in the layer and the design. The same happens in AI.

The other one is sequence like. So I don't know if you guys have noticed, but one we play with something like ChatGPT, you can get it to forget the previous conversation and it will drift. And that's a function of sequencing, how much of the information of the confrontation we can keep in its store and its memory is having a conversation.

Sequence length increases compute size significantly in terms -- it's also about the training and inference load which is captioning billions of parameters. It's just how much memory, how much needs to be processed in order to make an informed conversation going forward. And then there is more diversification going on. We

have lots of different models from PaLM to LLaMA to GPT-4, so we see a different specialization happening.

I expect that moving forward, the models will naturally want to get bigger because they can then encapsulate more intelligence. I believe that they -- we are definitely seeing that happen. We're seeing them integrate more deeply with intelligence databases and applying AI into the database and create information itself to better databases is super interesting. I can go -- I can talk forever about that, but those are being tied directly to some of these large models, and now AI is working into their internal systems as well inform them. [ph]

They get -- there's specialization happening. There is -- so we're seeing multiple different models and specialization with different layers and sequence like to keep the conversation more intelligent and keep the AI working memory more adapt which is also significantly increasing compute requirements.

It's a chicken and the egg, and you saw -- which we try to help with. I think it's what's driving every time we launch a new architecture, a new interconnect technology, or do new innovative things in Grace Hopper, DGX, GH 200. And we expand the scope of what these researchers and developers and NVIDIA's own research can do in order to move large language models and generative AI forward.

The next chapter in that probably will be more about reasoning. What's interesting is we're seeing -- right now in generative AI, I can talk more about reasoning in the future, but that's kind of and neither where we're going and that's an even harder blue sky problem.

## Q - Hans Mosesmann {BIO 1522582 <GO>}

Well, okay. So it looks like we're going to be in a growth category for some time. For those that are listening investors or participants, if you like to ask a question, just click on the question or ask the question button on right of your screen and it will come to me and I can read it out for Ian to talk about.

There is a new metric, it's kind of interesting. I was talking to some contacts in Silicon Valley maybe six months ago or so. The price of Hopper and the DGX Hopper was starting to come out and it was really, really expensive. There is some people saying, there is no way we're going to pay that kind of price for this kind of system being so much more expensive have been say and peer, and yet here we are and you're probably hand amount for the better part of this year, which kind of brings to mind that the issue in some and for some of these AI models for training inference have little to do with the upfront price. So it's less relevant and really the TCO aspect or the efficiency aspect that comes into play, how does that determine how you come to market, how you architect your compute GPUs and so on?

## A - Ian Buck {BIO 18454865 <GO>}

Yeah. It's a -- I really appreciate that question, and it's one that gets asked a lot because community this entire community and the world sees pricing and sees

sticker shock. And by the way, they -- usually they don't realize what it takes to build a hyperscale data center and the cost that goes. These are multi-billion dollar investments that are not new. People building datacenters at scale and I get to work with all the hyperscalers about that.

So the productivity and the utility of compute is incredibly important to them in order for them to improve their service, improve what they're doing, optimize their business and increase their revenue. Compute is critical to add generation, putting the right content front of you, keeping those engagement scores high, for keeping the products you want to be -- to provide a service. And nothing is more annoying being getting use those adds but getting ones that actually the things you want and the information you need leads to revenues. It's critical. And while people can see the -- maybe see the cost of a GPU, we create the opportunity for all of them to invest and build those services to make turn AI into the opportunity that it provides for them based on competing on the data.

Specifically though when we talk about generation and generation, how we should think about introducing new GPU, new technology in the market it is GTU. It is about are we -- how are we revolutionizing, not just the compute capability, but also the TCO analysis of what you can do today with our existing products and more with the next.

Hopper provides six times more compute performance at the transformer level, implementing that transformer layer than Ampere did, six times. And the end it's delivering end-to-end on training. It's delivering three to four times more performance, that's complete training job throughput, and in fronts, even more. Infronts which is -- can be further optimized and then, of course.

So when we -- we think about that. We think -- then we look at more than just the cost of individual one, but what's the throughput of that entire data center is going to be for them based on what they have today and what they're going to be able to do tomorrow. And we save them a ton of money. We save them a ton of money because by transitioning from one generation to next because the opportunity of performance and the economic TCO is hugely in our favor in terms of the throughput of that data center and the productivity of that data center, that one billion dollar investments, the billions that takes to build those data centers all around the world.

That same story plays out in enterprise as well. So by moving workloads from CPU or from previous generous GPUs to new GPUs, the throughput of the system or the rack or the data center at data center scale is measured in X factors often. Certainly, for the model, the transformer-based models, but including, we also look at the breadth of all the different workloads, including the models that are represented in image recognition, benchmarks you see in (inaudible) for example. (inaudible) if you guys haven't heard of it, it is a benchmark that's created by Google to sort of provide a level-playing field, a clean clear benchmark. It was representative of their training workloads and since then Meta has also been contributing their workloads to provide an honest benchmark that changes to the correct level for accuracy or

conversions as a requirement. And we use that to measure our performance to market based on previous generation. And you can see what Hopper has done compared to Ampere.

The other interesting point is that once -- we don't stop after we ship it. We continuously invest in the software and optimization. We suffer a massive part of what we do. I myself who started as a software engineer and manager at NVIDIA doing CUDA have hired thousands of software engineers and others across the Company. And one of the reasons I have this job is because now because of the importance of software and what we do and it's our interface to the rest of the world to people that are consuming our technology, partnering on the frameworks like PyTorch and JAX and tons of flow and everything else, the rest of that ecosystem and the user community.

So NVIDIA at this point has more software engineers than power engineers by good march. And -- so after we do the first round of benchmarking on something like Hopper, we continuously improve it. In fact, the Ampere over its life got, I believe, two and a half, three times faster.

## Q - Hans Mosesmann {BIO 1522582 <GO>}

Okay.

## A - Ian Buck {BIO 18454865 <GO>}

Yeah. From the first time -- if you go and look at the first time we submitted to the (inaudible) benchmark, it's public to where we -- I think we've recently stopped submitting those have shifted over Hopper. You can see 2.5 times X [ph] improvement in some of those models and use cases. So -- I mean, that's kind of what our users experience. I think it's why we have such a loyal community of users, both in the developer community as well as our biggest customers because we're continuously optimizing the whole stack and the platform along with them to improve the TCO.

## Q - Hans Mosesmann {BIO 1522582 <GO>}

Great. Hey, Ian, I did get a question here. It's an interesting one. Can you expand on the current issues of scaling sequence length and how that might be solved? There seems to be a push for new architectures that have more favorable scaling functions. Would this be a risk or opportunity for NVIDIA's advantage with its transformer engine?

## A - Ian Buck {BIO 18454865 <GO>}

Yeah, good question. So let me elaborate a little bit. We want to working with the customers and the users and the community. You can take a relatively small or large model, and larger input sequence length provides more context for the conversation moving forward, as mentioned tuck-ins starting with hundreds now going to thousands and they want to push it up higher. That increases the compute complexity of the inference job and how you want to tune for training.

Scalability is really important there, also capacity is important. It creates a larger working memory with a larger model. (inaudible) there is both ways to address that. One is scaling. Obviously, the throughput, so while there's multiple ways to optimize it. First is, transforming, and you mentioned that. What Hopper did that was so revolutionary and so impactful was it may need some called FPA [ph] fiable. FPA, there is eight bit floating point of presentation, that's basically eight zeros and ones to represent a floating point number.

It's not a lot of information. It's about the number of characters and alphanumeric keyboard, for example times two. So every character you can type on eight per character and roughly double that, that's how balance you can actually move in and bits. [ph] But if you can make training work at FPA, and it's incredibly fast. Obviously, computing on eight bits is faster than computing on 16 bits. Also the memory size is half of what you would have in 16 bit floating point, which is what we had before and which everybody was really. [ph]

The transformer engine specifically designed. You can't just put down and expect to cut the number of information eightfold its exponential -- eightfold in order to make the training successfully. Transformering with [ph] Hopper is actually a combination of both hardware and software to make sure that transformer models can train to convergence with only that if it's information at the corporate community. And it's a ton of work to make that. Actually we consumed a massive amount of our own supercomputing capability to meet that work to understand, tune and figure out how to keep things within the range of those eight bits.

I mentioned that for sequencing because by doing so, we reduce the size of the model, the size of the working set. They can fit more in 96 or 94 or 80 gig GPU depending on your flavor with available at Hopper, and of course, it keeps the response time how fast it can respond to a question with the range of usability. After that, we scale. So we scale. If we need more GPU computing in order to expand further, we scale with AMD Link. So we have technology called AMD Link, which allows it's -- on Hopper, it's 900 gigabytes a second, which is a lot. It's roughly seven times I think more than what you get with like PCIe if we try to use standard PCIe and connect new devices and start the system.

So basically combine two GPUs together into one. So we'll split the model and actually execute the model in parallel across two GPUs. We need that much bandwidth between the GPUs in order to keep things going, to keep things -- to make -- allow both GPUs to operate as one and start to model and keep the latency and response time to low. If you need more, you can go from two GPUs with NVL with H100 NVL product, which is actually two PCIe carved to the bridge to eight-way. So we have (inaudible) system that can go across eight, and then beyond that, we can use their tricks to use InfiniBand, or we can go all the way to ranger as our DGX GH 200, which is our 256 GPUs all connected to AMD Link. We just announced that in Computex two weeks ago.

The other thing is size of model. So how can we do bigger models, even if we don't need the latency or do some smaller models with longer sequence links, but could

be served with a single GPU, a single Hopper in terms of performance. For that, we have Grace Hopper. Grace Hopper is our -- we've announced that. We've been talking about that in our GTC. If you haven't seen our GTC conferences, you just check it out.

Grace Hopper basically is the 600 gigabyte GPU. So we combine GPU which has upwards of 96 gigabytes of HBM memory and then with our own CPU gloom together with it and we are linking it, so that the GPU can take advantage of all the CPU memory, which operates it upwards at 500 gigabytes, 600 gigabytes a second. So we can now have effectively a 600 gigabyte GPU and then also helps with these -- doing larger sequence links.

There are lots of ways to -- and actually, I pitty the -- your community to do all this analysis. It's becoming a complex matrix of model size, latency requirements and sequence length, and we're blanketed in this space. So we have -- and that's why we're creating-- you see us creating so many different brands and different products of Hopper in a PCA form factor, Hopper and AMD Link DGX form factor, two PCAs bridge together and now Grace Hopper as all of which can be used to deploy inference at a scale.

## Q - Hans Mosesmann {BIO 1522582 <GO>}

That's -- it's a good answer.

## A - Ian Buck {BIO 18454865 <GO>}

Yeah. I apologize. This is what I do every day, and making sure that working with each of those hyperscalers and those startups and everyone else to dial in and create new products to address it.

## Q - Hans Mosesmann {BIO 1522582 <GO>}

And so it looks like because you're blanketing the market with various types of products, it can counter some of the different proprietary or new architectures that have emerged out there that are being composed. Is that kind of like what you're saying?

## A - Ian Buck {BIO 18454865 <GO>}

Yeah. The others multiple -- there's not one click anymore in NVIDIA's roadmap. I think that's kind of how it used to be. Here is our Pascal GPU, and three years later, it's Volta, and here three-year later, there is Ampere and (inaudible)

What NVIDIA has been -- we've been working on is diversifying the ways in which we can add value. And instead of bringing them now, we build CPUs, GPUs BPUs. We work on (Technical Difficulty) InfiniBand and Ethernet, and make both of those platforms AI-capable for different paths. And then we can play with and optimize how we can connect all these things together and build different products within -- even within one GPU traditionally generation and meet the demand wherever it wants to go based on where you guys are going.

So that agility is really important in AI, and things are being invented all-time. And NVIDIA being the -- sort of one AI company that works with every AI company, that's why you're seeing this product is because we can -- we are meeting different -- see different aspects of what we believe to be able to dial in and bring-to-market. Perhaps in parallel of our partners who have this now are trying to meet this demand and to meet -- to optimize that workload.

The other thing I'll say is that we've also accelerated GPU roadmap. So we used to do GPU -- 100 class GPUs every three years. We're now down to two years, in some case 18-month cycle. Jensen has talked about Hopper-Next, and that timeline and in addition, its time to -- and would have Grace next, its time with quantum next for interconnect, and we've accelerated that now. So we're now doing and able to invest and are chief sure of that now every two years or 18 months kind of you look at.

## Q - Hans Mosesmann {BIO 1522582 <GO>}

That's good to know. But we got a bunch of questions just come in. We don't have a lot of time. This is a tactical one. I'm not sure if you can answer. Maybe Simona can come in. Can you please talk about efforts to source supply for the second half of the year? And how does NVIDIA define significant as mentioned in the latest conference call?

## A - Ian Buck {BIO 18454865 <GO>}

Yeah. Simona can comment a little bit more on the conference call details and I can follow.

## A - Simona Jankowski {BIO 7131672 <GO>}

Sure, happy to do so. And I hope you guys can hear me okay. So we commented on the earnings call that we are going to have substantially higher supply in the second-half relative to the first half of the year, and that essentially backs up the extended demand visibility that we see stretching out a few quarters into year-end as well.

As we commented, we have seen a pretty steep increase in demand through the quarter, all the way leading up to the current time. And so we're working closely with customers to ensure that we have supply for them. That also helped underpin the strong guidance we gave for the second quarter, and then even with that higher baseline in the second quarter, we commented a substantially higher level of supply second-half versus first-half.

We haven't been more granular on the exact linearity between Q3 and Q4, so just give us a bit of time as we get closer to the back half of the year, we'll be able to provide guidance quarter-by-quarter.

## Q - Hans Mosesmann {BIO 1522582 <GO>}

Okay.

## A - Ian Buck {BIO 18454865 <GO>}

No, that's pretty much. I don't have much more to add to that. We are certainly -- our biggest customers of course playing with us and everyone is swarming generative AI, but as part of that. We are able -- we can -- we are working, of course, on plan -- doing that planning with them and continue to do that with them as we're doing all the things that Simona mentioned at the same time.

## Q - Hans Mosesmann {BIO 1522582 <GO>}

Okay. Here's another question along the same lines, maybe you can answer this. What is the biggest bottleneck for NVIDIA or GPUs more broadly? How much time do you think it'll take for the industry to build sufficient inventory or supply levels?

## A - Ian Buck {BIO 18454865 <GO>}

You know the -- well, I'm not going to comment on specific bottlenecks from a supply standpoint. I think the challenge or bottleneck perhaps and further adoption, it's not really a bottleneck. It's just where is it going is the broadening. We're seeing now the enterprises pick up AI, and for that to happen, the people of the providers of AI including NVIDIA need to meet the enterprises where they are and some of them, they are -- they have their on AI expertise. They have either through acquisition or hired or brought in-house are working with -- closely with the startup or others to adopt AI to their -- to influence or improve their business. And you see that in some of the large language model startups, for example, providing that.

I really like the work that AI is doing, for example, you know, making it easier to use the older software. With AI, that can click all the buttons and check all the boxes instead of having on existing software, and that's the degree we are doing those things. [ph] But after meeting the needs of those enterprises, where they are in terms of perhaps service or taking appropriate model and fine-tuning it to a something useful, where really the only thing the enterprise needs to do is provide the right kind of data and take and convert appropriate model into their own virtual assistant or chat capability.

So a lot of the activity right now is about helping them adopt AI into their workflows, into their products, and some of the work we're doing on our own email product, for example, is exactly that, where it's checking the easy-to-use. You can provide a few -- hundred up to maybe a thousand or two examples, text and text out and you can fine-tune the GPT model all way up to 175 or larger to answer questions in that format and in that context, instead of just asking a generic ChatGPT question which continue generic answer from a generic human or a generic -- how the unit you would answer it, you can have a -- you can answer it like a financial expert or a support call expert or other such things, and connecting AI with information retrieval systems.

So when you ask a question, you don't just get an answer, which may or may not be right, and certainly, it can -- we can't make ChatGPT lie or write code that actually looks right but it's something made up, but actually get to the actual sourcing information. And we see that a little bit with the work that Bing is doing, but more broadly, generative AI is useful if you can only generate an answer, but tell you with sources so you can further explore the results.

So that democratization in the -- is the -- and connecting all those GPUs with industry is a big push right now, and you're starting to see the early movers in that space. That's where the next -- that next wave of GPU usage and also revenue from services and things like DGX great efforts as well as our partners is going to -- is moving the needle.

## Q - Hans Mosesmann {BIO 1522582 <GO>}

Last question and we got a minute. Let's see if we can keep it 10 minute. How he conversations with biggest clients, hyperscalers or enterprise changed after your last earnings call because it seems from what we hear that this was a real wake-up call for many decision-makers on how to make serious investments on AI. So question is, how quickly has this changed since the conference call which is basically two-three weeks ago?

## A - Ian Buck {BIO 18454865 <GO>}

I don't know that's changed from the conference call. It's certainly changed in the ChatGPTmoment so -- and the generative AI moment. And the converse probably only continues to be amplified with activities we're seeing in the Street. But the opportunity for generative AI in every one of their services, in every one of their capabilities. And the -- seeing NVIDIA not as a supplier of GPUs, which we are or supplier of infrastructure, but a partner at many levels.

We always were a partner in their hyperscale efforts. The development of their servers, the design of their data centers, and how to build something even more capable and optimized and power-efficient and at-scale. Every one of them has their unique challenges and capabilities and their own technology that they can contribute to and working with NVIDIA to make it work well.

Amazon's EFA, in their Elastic Adapter, and you can tell a lot of work to make sure that that can work at scale. Other hyperscalers that have their own, also they use InfiniBand. They're working with them to scale out InfiniBand they're going through their platforms. [ph] And we've always been a partner on the data center partner. That's only amplified our side. [ph] Now, it's about Grace and Grace Hopper and CPU land and what we can do in that space, which is very exciting. And step up was the software side.

So all of the capabilities in the software and the infrastructure and integrating into all the different frameworks, and the core capabilities broadening across other services, so other developers and researchers can get access to that infrastructure and meet them, that there is only amplified. And we always we're partnering with (inaudible) and JAX and others, and that certainly has continued and grown.

What we're seeing now is more and more of their service groups seeing NVIDIA as also a partner to optimize their further latest platform or what we have to offer in generative AI. Seeing the opportunity to do work less, moving workloads that were even still on CPU for easing a little bit of AI or simplified AI, to using much more intelligent larger models to improve the quality service to have a better interaction

with the device, even if it's talking to a hockey puck on your kitchen counter, or the cloud, and then -- but we're also a partner with them to optimize those models.

That 2.5X, 3X that we did with A100 wasn't just us working in the back office. It was the optimizations we were doing because our customers -- our biggest customers were giving us challenges and side by side working with them to optimize those workloads, which get obviously reflected in things like benchmarks and elsewhere. So those are amplified and certainly, we have our engagements with their service teams of deploying AI and figuring out how to run and use Hopper, use Hopper at scale to inference better and more efficiently, and move more workloads to the GPU structure they have and plan for growth moving forward is a big part of that. That definitely has gone up quite a few checks.

## Q - Hans Mosesmann {BIO 1522582 <GO>}

Well, I could Imagine. Well, Ian, thank you so much. Very enlightening. Looks like you're going to be hiring another thousand software engineers. Hopefully, you don't have to do all the interviews yourself, but exciting times. Simona, thank you as well. And we look forward to the group session later this afternoon. Have a great day, and thanks.

## A - Ian Buck {BIO 18454865 <GO>}

Thank you, and hope to see you again in person.

## Q - Hans Mosesmann {BIO 1522582 <GO>}

You got it.

## A - Simona Jankowski {BIO 7131672 <GO>}

Thank you. Bye-bye.