GTC 2023 Keynote

Company Participants

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

Other Participants

- Aaron Rakers, Analyst, Wells Fargo
- Blayne Curtis, Analyst, Barclays Capital
- C.J. Muse, Analyst, Evercore ISI
- Joseph Moore, Analyst, Morgan Stanley
- Matthew Ramsay, Analyst, Cowen and Company
- Rajvindra Gill, Analyst, Needham & Company
- Stacy Rasgon, Analyst, Bernstein Research
- Timothy Arcuri, Analyst, UBS
- Toshiya Hari, Analyst, Goldman Sachs
- Vivek Arya, Analyst, Bank of America Merrill Lynch
- William Stein, Analyst, Truist Securities

Presentation

Simona Jankowski {BIO 7131672 <GO>}

Hi everyone, and welcome to GTC. It's Simona Jankowski, Head of Investor Relations at NVIDIA. I hope you all had a chance to view Jensen's news pack keynote this morning. We also published several press releases and vlogs detailing today's announcement. Over the next hour we will have an opportunity to unpack and discuss today's news with our CEO, Jensen Huang; and our CFO, Colette Kress in an open Q&A session with financial analyst.

Before we begin, let me quickly cover our safe harbor statement. During today's discussion, we may make forward-looking statements based on current expectations. These are subject to a number of significant risks and uncertainties and our actual results may differ materially. For a discussion of factors that could affect our future financial results and businesses, please refer to 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, based on information currently available to us, except as required by law, we assume no obligation to update any such statements.

We'll start today with a few brief comments by Jensen, followed by a Q&A session with Jensen and Colette. And with that let me turn it over to Jensen.

Jensen Huang {BIO 1782546 <GO>}

Hi, everybody, welcome to GTC. GTC is our conference for developers, to inspire the world on deposit -- or the possibility of accelerated computing and to celebrate the work researchers and scientists that use it. And so, please be sure to check-in on some of the conference sessions that we have. It covers some really amazing topics. The GTC Keynote highlighted several things. And let me -- before I go into the slides, what I'm going to do is, Colette, now will just cover basically the first slide, the rest of the slides we provided to you for reference.

And -- but let me make a couple of comments first. At the core of computing today, the fundamental dynamics at work is of course influenced by one of the most important technology drivers in the history of any industry, Moore's Law. And it's fundamentally come to a very significant slowdown, you could argue it's Moore's Law has ended. For the very first time in history, it is no longer possible using general-purpose computing, CPUs, to gain the necessary throughput without also the corresponding amount of increase in cost or power. And that lack of decreasing of power effectively or decreasing of cost is going to make it really hard for the world to continue to sustain the increased workloads, while maintaining sustainability of computing. So one of the most important factors dynamics and computing today is sustainability.

We have to accelerate all the workloads we can, so that we can reclaim the power and use whatever we reclaim to invest back in the growth. And so the first thing that we have to do is to not waste power, to not -- to accelerate everything we possibly can and was really focused on sustainability. I gave several examples of the workloads that we used, to highlight how in many cases we can accelerate an application by 50 times 40, 50, 60, 70 times or 100 times. While in the process, decreasing power by an order of magnitude, decrease in cost by a factor of 20. This approach is not easy. Accelerated computing, is a full stack challenge, NVIDIA accelerated computing, it's full stack. I've talked about that in many sessions in the past, it starts from the architecture to the system, to the system software, to acceleration libraries to the applications (inaudible).

We're a data center scale computing architecture and the reason for that is, because once you refactor in application to be accelerated, the algorithms are highly paralyzed. Once you do that, you can also scale out. So one of the benefits of accelerated computing from the work that we do, you can scale up, you can also scale out. The combination of it has allowed us to bring million X acceleration factors to many applications domain, of course, one of the very important ones is artificial intelligence.

NVIDIA's accelerated computing platform is also a multi-domain. This is really important because data centers, computers are not single-use devices. What makes computers such an incredible instrument is its ability to process multiple types of applications. NVIDIA's accelerated computing is multi-domain. Particle physics, fluid dynamics, all the way to robotics, artificial intelligence, so on and so forth, computer graphics, image processing, video processing, all of these types of domains consume an enormous amount of CPU cores today, enormous amounts of power. We have the opportunity to accelerate all of them and reduce power, reduce cost.

And then of course NVIDIA's accelerated computing platform is cloud to edge. This is the only architecture that is available in every cloud, it's available on-prem by just about every computer maker in the world and it's available at the edge for inferencing systems or autonomous systems, robotic self-driving cars, so on and so forth.

And then lastly, one of the most important characteristics about NVIDIA's accelerated computing platform is although we do at full stack, we design it and architect a data center scale, it's available from cloud to edge. It is completely open, meaning that you can access it from literally any computing platform, from any computing maker, anywhere in the world. And so this is one of the most important characteristics of a computing platform and it's because of its openness, because of our reach, because of our acceleration capability that the positive -- the virtuous cycle -- the positive virtuous cycle of accelerated computing has now been achieved. Accelerated computing and artificial intelligence have arrived.

We talked about three dynamics. One of them is sustainability, I just mentioned. The second is generative AI. All of the foundational work that has been done over the last 10 years, in the beginning a really big breakthroughs in computer vision and perception led to industrial revolutions in autonomous vehicles, robotics and such. That was just the tipping -- that was just the tip of the iceberg. And now with generative AI, we have gone beyond perception to now the generation of information. No longer just the understanding of the world, but to also make recommendations or generic content that is of great value. Generative AI has triggered an inflection point in artificial intelligence and it's driven a step-function increase in the adoption of AI all over the world and very importantly a step-function increase the amount of inference, that will be deployed in all the world's clouds and data centers.

And the third thing that I mentioned -- discussed in the keynote was digitalization. This is really about taking artificial intelligence to the next phase, the next wave of AI, where AI is not only operating on digital information, generating text and generating images but AI is operating factories and physical plants and autonomous systems and robotics. In this particular case, digitalization has the real opportunity to automate some of the world's largest industries. And I spoke about the digitalization of one particular industry, now gave examples of how Omniverse is the digital, physical operating system of industrial digitalization. And I demonstrated how Omniverse was used from the very beginning of product conception, the architecture, the styling of product designs, all the way to collaboration of the design, the simulation of the product, the engineering of the electronics to the setting up the virtual plants, all the way to digital marketing and retail.

In every aspect of a physical products company, digitalization has the opportunity to automate, to help them collaborate to bring the world of physical into the world of digital and we know exactly what happens to that. Once you get into the world of digital, our ability to accelerate workflows, our ability to discover new product ideas or ability to invent new business models tremendously increase. And so, I spoke about digitalization.

There were five takeaways that we spoke about in the keynote. And we'll talk about today, and if you have questions in any of these areas, we love to entertain them. The first of course is that generative AI is driving accelerating demand for NVIDIA platforms. We came into the year full of enthusiasm with the Hopper launch, Hopper was designed with a transformer engine that was designed for large language models and what people now call foundation models. The transformer model -- transformer engine has proven to be incredibly successful. Hopper has been adopted by just about every cloud service providers that I know and available from OEMs.

And what is really signaling the increase in demand of Hopper versus previous generations and the accelerating demand for it, really signals a inflection for AI, because it used to be for AI research, which now with generative AI moving into the deployment of AI, into all of the world's industries, and very importantly a very significant step-function in the inference of these AI models. So generative AI is driving accelerating demand.

The second thing is, we talked about our new chips that are coming to the marketplace. We care deeply about accelerating every possible workload we can. And one of the most important workloads, of course, is artificial intelligence. Another important workload to accelerate as the operating system of the entire data center, you have to imagine that these giant data centers are not computers, but their fleets of computers that are orchestrated and operated as one giant system. So the operating system of the data center, which includes the containerization [ph], the virtualization, networking, storage, and very importantly, security, the isolation and in the future the confidential computing of all of these applications. It's a operating software defined layer, is a software layer that runs across the entire data center fabric. That software layer consumes a lot of CPU cores. And frankly, I wouldn't be surprised if for many, depending on the type of data centers that are being operated, I wouldn't be surprised of 20%, 30% of the data centers power is just dedicated to the networking and the fabric and all of the virtualization and the software-defined stacks, basically the operating system stack.

We want to offload, accelerate the operating system of modern software-defined data centers, and that processor is called BlueField. We announced a whole bunch of new partners and cloud data centers that have adopted BlueField, I'm very excited about this product. I really believe that this is going to be one of the most important contributions we make to modern data centers. Some companies design their own, most companies won't have the resources to design something of this complexity, and cloud data centers will be everywhere.

We announced Grace Hopper, which is going to be used for one of the major inference workloads, vector databases, data processing, recommender systems. Recommender systems as I've spoken about in the past, it's probably one of the most valuable and most important applications in the world today. And a lot of digital commerce and a lot of digital content is made possible because of sophisticated recommender systems are moving to deep learning. And this is a very important opportunity for us. Grace Hopper was designed specifically for that and give us an opportunity to get a 10x speed-up in recommender systems and large databases.

We spoke about Grace. Grace is now in production, Grace is also sampling. Grace is designed for the rest of the workload in a cloud data center, that is not possible to accelerate. Once you accelerate everything, what is left over is software, that really wants to have very strong single threaded performance. And those single threaded performance is what Grace was designed for. We also designed Grace not to just be the CPU of a fast computer but to be the CPU of a very, very energy-efficient cloud data center. When you think about the entire data center as one computer, when the data center is the computer, then the way you designed a CPU in the context of a accelerated data center AI first, cloud first data center, that CPU design is radically different.

We designed Grace CPU -- excuse me -- just slightly out of reach. The Grace CPU is designed, this is the entire computer module, this isn't just the CPU, but this is the entire computer module of a Grace Superchip. And this goes into a possibly called the system and you could rack up a whole bunch of Grace computers into a cloud data center, because it is so energy-efficient and yet so performance for single-threaded operation. We're really excited about Grace and its sampling now.

Let's see. We spoke a lot about generative AI and how it's a step-function increase in the amount of inference workload that we're going to see. And then one of the things that's really important about inference coming out of the world's data centers, is that really wants to be accelerated on the one hand. On the other hand, it is multi-modal, meaning that, there are so many different types of workloads that you want to inference. Sometimes you want to inference, you want to bring inference and AI to video and you're augmented with generative AI. Sometimes it's images, producing beautiful image and helping to be a co-creator. Sometimes you're generating text, very long text, so the prompts could be quite long, so that you can have a very long context or it could be generating very long text, writing very long programs. And so these applications, each one of the video, images, text, and of course also vector databases, they all have different characteristics.

Another challenge of course is in the cloud data center. On the one hand, you would like to have specialized accelerators for each one of those modalities or each one of those diverse generative AI workloads. On the other hand, you would like your data center to be fungible, because workloads are moving up and down, they're very dynamic, new services are coming on, new tenants are coming on. People use different services during different times of day and yet you would like your entire data centers to be utilized as much as possible.

The power of our architecture is that, it is one architecture, you have one architecture with four different configurations, they all run our software stack, which means that depending on the time of day, if one is under-provisioned or under-utilized, you could always provision that class of that configuration of accelerators to other workloads. And so this fungibility in the data center gives you the ability, or architecture -- one architecture, inference configurations, inference platform gives you the ability to accelerate various workloads to its best of your ability and then not have to perfectly precisely predict the amount of workload because the entire data center is flexible and fungible. So one architecture, four configurations.

One of our biggest areas of collaboration and collaboration partnership is Google Cloud GCP. We're working with GCP, across a very large area of accelerated workloads from data processing for data product, Spark RAPIDS to accelerate data product which represents data processing, probably represents some 10%, 20%, 25% of cloud data center workloads. It's a probably one of the most intensive CPU core workloads. We have an opportunity, we accelerating it, bring 20x speed-up, bring a lot of cost-reduction that customers can enjoy. And very importantly, a lot of power reduction, that's associated with that.

We're also accelerating inference with the Triton server. We're also accelerating their generative AI models. Google has world-class pioneering large language models that we're now accelerating and putting onto the inference platform L4. And then, of course, streaming graphics and streaming video, we have an opportunity to accelerate that. So our two teams are collaborating to take a large amount of workloads that could be accelerated in generative AI and other accelerated computing workloads and accelerating with the L4 platform, which has just gone public on GCP. So we're really excited about that collaboration, we have much more to tell you soon.

The third thing that we talked about was acceleration libraries. As I mentioned before, accelerated computing is a full stack challenge. Unlike a CPU, where a software is written and it's compiled using a compiler and its general purpose so all code runs, that's one of the wonderful advantages and breakthroughs of a CPU, it's general-purpose. The acceleration aspect of it, if you want to accelerate workloads, you have to redesign the application, you have to refactor the algorithm altogether and we codify the algorithms into acceleration libraries. Acceleration library all the way to linear algebra to FFT, to data processing, that we use to fluid dynamics and particle physics and computer graphics and so on and so forth. Quantum chemistry, inverse physics for image reconstruction, so on so forth. Each one of these domains require acceleration libraries. Every acceleration library requires us to understand the domain, work with the ecosystem, create an acceleration library connect them to applications in that ecosystem and power and accelerate the domain of use.

We -- every single we're constantly improving the acceleration libraries we have, so that the installed-base benefits from all of our increased optimizations for all of their investments of capital already, their infrastructure already. So you buy NVIDIA's systems and you benefit from acceleration for years to come. It's not unusual for us, on the same platform to increase the performance anywhere from four to 10 times after you've installed it over its life. And so we're delighted to continue to improve the libraries and bring new features and more optimization. This year, we optimized and released 100 libraries and 100 models -- 100 libraries and models, so that you can have better performance and better capability.

We also announced several very important new libraries. One new library that I'll highlight is cuLitho. Computational lithography is an inverse physics problem that calculates the max -- that processes calculates to maximize the equation as it goes through optics and interacts with the photoresist on the mask. This ability to do basically inverse physics and image processing makes it possible for us, to use wavelength of light, that are much, much larger than the final pattern that you want to create on the wafer. It's a miracle in fact, if you look at modern microchip manufacturing, you're in the latest generation, we're using 13.5 nanometer light which is near X-ray, it's extreme ultraviolet and yet using 13.5 nanometer light. You could pattern a few nanometers, three nanometers, five-nanometer patterns on a wafer. I mean, that's basically like using a fuzzy light, a fuzzy pen, to create a really fine pattern on a piece of paper. And that ability to do so requires magical instruments like ASML's magical instruments, computational libraries from Synopsys, the miracle of the work that TSMC does. And this field of imaging called computational lithography.

We've worked over the last several years to accelerate this entire pipeline. It is the single largest workload in all of EDA today, computationally intense. Millions and millions of CPU cores are running all-the-time, in order to make it possible for us to create all of these different masks. Now this step of the manufacturing process is going to get a lot more complicated in the coming years, because the magic that we're going to have to bring to future lithography is going to get in increasingly high and machine learning and artificial intelligence will surely be involved. And so the first step for us is to take this entire stack and accelerate it.

And over the course of last four years, we've now accelerated computational lithography by 50 times. Now, of course, that reduces the cycle time and the pipeline and the throughput time for all of the chips in the world that are being manufactured, which is really quite fantastic because these are $40 billion, $50 billion investments in the factory, if you could reduce the cycle time by even 10%, the value to the world is really quite extraordinary. But the thing that's really fantastic is, we also save an enormous amount of power. In the case of TSMC and the work that we've done so far, we have the opportunity to take megawatts, tens of megawatts and reduce it by factors of 5 to 10. And so that reduction in power, of course, makes manufacturing more sustainable and it's very important initiative for us. So cuLitho, I'm very excited about.

Lastly, I'll talk about the single largest expansion of our business model in our history. We know that the world is becoming heavily cloud-first. And cloud gives you the opportunity to engage your computing platform quickly, instantly through a web browser. And over the last -- over the last 10 years, the capabilities of clouds have continued to advance. To the point where it started with just CPU and running Hadoop or MapReduce or do inquiries in the very beginning, to know, they're high performance computing, scientific computing systems, AI supercomputers in the cloud.

And so we are going to partner with all of the world's cloud service providers. And starting with OCI, we've also announced cloud partnership with Azure and GCP. We're going to partner with the world's leading cloud service providers to implement to install and host NVIDIA AI, NVIDIA Omniverse and NVIDIA DGX Cloud, in the cloud. The incredible capability of doing so, is on the one hand you get the fully optimized multi-cloud stacks of NVIDIA AI and NVIDIA Omniverse. And you have the opportunity to enjoyed in all of the world's clouds in its most optimized configuration. And so you get all of the benefits of NVIDIA software stack in its most optimal form. You have the benefit of working directly with NVIDIA computer scientists and experts. So for companies who have very large workloads and who would like to have the benefit of acceleration, the benefits of the most advanced AI, we now have a direct service where we can engage the world's industries.

It's a wonderful way for us to combine the best of what NVIDIA brings and the best of all the CSPs. They have incredible services for security, for cloud -- for security, for storage, for all of the other API services that they offer. And they very well could be -- likely already the cloud you've selected. And so now for the very first time we have the ability to combine the best of both worlds and bring NVIDIA's best and combine it with the CSPs best and make that capability available to the world's industries.

One of the services that we just announced, that's in the -- that's platforms or service, NVIDIA AI, NVIDIA Omniverse and infrastructure as a service NVIDIA DGX Cloud. We also offered -- announced a new layer. We have so many customers that we work with, so many industry partners that we work with, to build foundational models. And the -- if a customer of an enterprise, if an industry would like to have access to foundation models, the most obvious and most accessible thing is to work with world leading service providers like Open AI or Microsoft and Google. These are all examples of AI models that are designed to be highly available, highly flexible and useful for many industries.

There are companies that want to build custom models, that are based specifically on their data and NVIDIA has all of the capabilities to do that. And so for customers who would like to build custom models based on their proprietary data, trained and developed and inference in their specific way, whether it's the guard rails that they would like to put, implement or the type of instruction tuning they would like to perform or the type of proprietary datasets that they would like to have retrieved. Whatever the very specific requirements that they have in language models, generative image models in 2D, 3D or video or in biology, we now have a service that allows us to directly work with you to help you create that model, fine-tune that model and deploy that model on NVIDIA DGX Cloud.

And as I mentioned, the DGX Cloud runs in all of the world's major CSPs. And so if you already have a CSP of your choice, I'm pretty certain that we'll will be able to hosted in it, okay. And so NVIDIA cloud services is going to expand our business model and we offer infrastructure as a service, DGX Cloud, platform as a service, NVIDIA AI and NVIDIA Omniverse. And we have new AI services that are designed to be custom, essentially the foundry of AI models that are available to the world's industries. And all of it in the world -- in partnership with the world's leading CSPs.

So that's it, those are those are the announcements that we made. We have a lot to go through. Thanks for joining GTC. And with that, Colette and I will answer your questions for you.

Questions And Answers

A - Simona Jankowski {BIO 7131672 <GO>}

Thank you, Jensen. Let me welcome our financial analysts to the Q&A session. We're going to be taking questions over Zoom, so please use the raise hand feature on Zoom if you would to like to ask a question and then unmute yourself when called upon. I'll pause for a moment here to review the queue before we take our first question. And our first question is from Toshiya Hari with Goldman Sachs.

Q - Toshiya Hari {BIO 6770302 <GO>}

Hi, Jensen and Colette, can you hear me okay?

A - Jensen Huang {BIO 1782546 <GO>}

Perfectly. Nice to see you. Nice to (Multiple Speakers)

Q - Toshiya Hari {BIO 6770302 <GO>}

Yeah. Thank you so much for hosting this follow-up. Jensen, I guess, I had one question on the inference opportunity. Obviously, you dominate the training space and you've done so for many, many years now. I think on the inference side, the competitive landscape has been a little bit more mixed, given incumbency around CPUs, but obviously very encouraging to see you introduce this new inference platform. I guess, with the criticality of recommender systems that you spoke to the growth at LLMs and your work with Google, it seems like the market is moving in your direction. How should we think about your opportunity and inference, call it in the three to five year versus where you stand today and how should we think about Grace playing a role there over the next couple of years? Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Toshiya, thank you. First of all, I'll work backwards. And in three to five years, the AI supercomputers that we are building today, which is unquestionably the most advanced computers, the world makes today. It is of course of gigantic scale, it includes computing fabrics like NVLink, computing -- large computing -- large-scale computing fabrics like InfiniBand and very sophisticated networking that stitches that altogether.

The software stack, the operating system of it the distributed computing us, software, it's just computer science at the limits. And so there what's really going to be quite exciting is how AI supercomputer is going to go beyond research and extending into essentially AI factories because these AI models that people develop are going to be fine-tuned and improved basically forever. And I believe that every company will be in intelligence manufacturer. At the core of all of our companies, we produce intelligence. And the most valuable data we have are all proprietary, they're inside the walls of this Company. And so we now have the capability to create to build AI systems that helps you curate your data, package your data together that could then be used to help you train your proprietary model -- custom model, which can accelerate your business. That system, that AI training system is continuous.

The second inference, inference has largely been a CPU oriented workload. And the reason for that is because most of the inference in the world today are fairly lightweight. They might be recommending something related to shopping or (inaudible) or -- so on and so forth. And these kind of recommendations are largely done on CPUs. In the future there are several reasons why even video is processed on CPUs today. In the future what is likely to happen, our two fundamental dynamics that are inescapable at this point and it's an -- it was inevitable for quite a long-time, it is now inescapable. One of them is just sustainability. You can't continue to take these video workloads and process them on CPUs. You can't take these deep learning models, even if the quality of service was a little bit lesser good using CPUs to do it, it burns just -- it's just burns too much power. And so the first reason why we have to accelerate everything is for sustainability, we have to accelerate everything because Moore's Law was ended.

And that sensibility is now permeated, just about every single cloud service provider, because the amount of workload that they have that requires acceleration has increased so much. And so the -- their attention to acceleration, their alertness to acceleration has increased. And then secondarily, just about everybody has their power limits. And so in order to grow in the future, you really have to reclaim power through acceleration and then put it back to growth. And then the second reason is generative AI has arrived. We're going to see just about every single industry benefiting from augmenting from co-creators, co-pilots, that accelerates everything we do from text, the text we create, chatbots we interact with, spreadsheets we use, PowerPoint and Photoshop, and so on and so forth. They're all going to be you're going to be augmented by, you're going be accelerated by, inspired by a co-creator or co-pilot.

And so I think that the net of it all is that AI is for training, AI supercomputers will become AI factories, and every company will have it -- will have either on-prem or in the cloud. And secondarily, just about every interaction you have with computers in the future, we'll have some generative AI connected to it and therefore the amount of inference workload will be quite large in. My sense is that, inference will on balance be larger than inference, larger than training, but training is going to be right there with it.

A - Simona Jankowski {BIO 7131672 <GO>}

Thank you. Our next question comes from C.J. Muse with Evercore.

Q - C.J. Muse

Good morning. Good afternoon. Can you hear me?

A - Jensen Huang {BIO 1782546 <GO>}

Yes, C.J. Nice to talk to you.

Q - C.J. Muse

Perfect. Well, thank you for today. I just want to put my question on, I'd like to focus on Grace. In the past you've mostly discussed the benefit of Grace and Hopper combined. Today, you're also focusing a bit more on Grace on a standalone basis than what I was kind of expecting. Can you speak to whether you've changed your view on your expected service (Technical Difficulty)? And how should we think about potential revenue contributions over time? Particularly as you think about Grace standalone, Grace Superchip and obviously Grace Hopper combined?

A - Jensen Huang {BIO 1782546 <GO>}

I'll start from the punch line, work backwards. I think Grace will be a big business for us, but it will no -- it will be nowhere near the scale of accelerated computing. And the reason for that is because we genuinely feel that every workload that can be accelerated must be accelerated. And everything from data processing, of course, computer graphics to video processing to generative AI. Every workload that can be accelerated must be accelerated, which basically leaves workloads that can't be accelerated, meaning the converse of that, another way of saying that is, a single-threaded code. That's single-threaded code because Amdahl's law still prevails, everything that is left becomes the bottleneck. And because the single-threaded code is largely related at this point to data processing, fetching a lot of -- moving a lot of data. We have to design a CPU, that is really good at two things. Well, I mean, we just say -- let me, those are two things, plus a design point. The two characteristics that we really want for a CPU is one that has extremely good single-threaded performance. It's not about how many cores you have, but it's about how good the single threated cores you do have.

And number two, the amount of data that you move has to be extraordinary. This one module here moves one terabyte per second of data, that's just an extraordinary amount of data that we move. And you want to move it, you want to process that data with extremely low power, which is the reason why we innovated this new way of using cellphone DRAM, enhanced for data center resilience and used it for our servers. It's cost effective because obviously cellphone volume is very high, the power is one-eighth the power and moving data is going to be so much of the workload that is just vital to us that we reduce it.

And then lastly, we designed a whole system instead of building just a super-fast CPU core -- CPU. We designed a superfast CPU node. By doing so, we can enhance the ability for data centers that are powered limited to be able to use as many CPUs as possible. I think that the net of it all is that accelerated computing will be the dominant form of computing in the future because Moore's Law has come to an end.

But what is going to remain are going to be heavy data processing, heavy data movement and single-threaded code. And so CPUs will remain very, very important. It's just the design point would be different than the past.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question will come from Joe Moore with Morgan Stanley. Please go ahead.

Q - Joseph Moore {BIO 17644779 <GO>}

Great. Thank you so much. I'm told to follow-up on the inference's question. This cost per query is becoming a major focus for the generative AI customer and they talking about pretty significant reductions in the quarters and years ahead. Can you talk about what that means for NVIDIA? Is this going to be in H100 workload for the longer-term and how do you guys work with your customers to get that cost down?

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. There's a couple of dynamics that are moving at the same time. On the one-hand, models are going to get larger. The reason why they're going to get larger is because we wanted to perform tasks better and better and better. And there's every evidence that the capability, the quality and the versatility of the model is correlated to the size of model and the amount of data that you train the model with. And so on the one-hand, we want it to be larger and larger, more versatile. On the other hand, there are so many different types of work workloads. I remember you don't need the largest model to inference every single workload and that's the reason why we have we have 530 [ph] billion parameters models, we have 40 billion parameter models, we have 20 billion parameter models and even 8 billion parameter models. And these different models are created in such a way that some of them are the largely -- you always need a large model, and the reason why you need a large model is at the very minimum, the large model is used to help improve the quality of the smaller models. In case, it kind of like you need a professor to improve the quality of the student and improve the quality of other students and so on so forth.

And so are -- because there are so many different use cases, you're going to have different sizes of models. And so we optimize across all of those, you use the rightsized model for the rightsized application. Our inference platform extends all the way from L4 to L40. And one of the ones that I announced this week, is this incredible thing, this is the Hopper H100 NVLink, and we call it H100 NVL. This is basically two Hoppers connected with NVLink. And as a result, it has 180 gigabytes -- 190 gigabytes, almost 190 gigabytes of HBM3 memory.

And so this 190-gigabyte memory gives you the ability to inference modern large-size inference language models, all the way down to, if you would like to use it in very small configurations, this dual H100 system solution lets you partition down to 18, is it 18 -- 16 different -- correct me if I'm wrong later, 16 or 18 what we called multiple instance GPUs mix. And those miniature GPUs --fractions of a GPUs could be inferencing different language models or the whole thing could be connected or four of these could be put into a PCI Express server -- commodity server, that can then be used to distribute a large model across it. This has already reduced, because the performance is so incredible. This has already reduced the cost of language inferencing by a factor of 10, just from A100. And so we're going to continue to improve in every single dimension, making the language models better, making the small models more effective, as well as making each inference more cost effective and with new inference platforms like NVL. And then very importantly, the software stack, we're constantly improving the software stack over the course of the last couple of two, three years, we've improved it so much. I mean, orders of magnitude in just a couple of years and so we're expecting to continue to do them.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question will come from Tim Arcuri with UBS. Please go ahead.

Q - Timothy Arcuri {BIO 3824613 <GO>}

Thanks a lot. Jensen, I think, I thought I heard you say that, Google's inferencing large language models on your systems. I wanted to confirm that that's what you were saying, and I guess, does that mean that they're using the new L4 platform? And if that -- is that brand new? So in other words, they were using TPU [ph] but they are now using your new L4 platform? Just curious more details there. Thanks.

A - Jensen Huang {BIO 1782546 <GO>}

Our partnership with GCP is a very, very big event and it is a inflection point for AI, but it's also an inflection point for our partnership. We have a lot of engineers working together to bring the state-of-the art models that Google has to the cloud. And L4 is a versatile inference platform, you could use it for video inferencing, image generation for generative models, text generation for large language models. And I mentioned in the keynote, some of the models that we're working on together with Google to bring to the L4 platform.

And so, L4 is going to be -- it's just a phenomenal inference platform, it is very energy-efficient, only 75 watts. The performance is off the charts and it's so incredibly easy to deploy. And so this -- between the L4 on the one end, I'll show it here, between L4, this is an L4 -- this here is an L4 and this is the H100. Okay? So this is the L4, this is the L4. And this is between these two processors, it is about 700 watt and this is 75 watt. And so this is the power of our architecture, one, software stack, can run on this as well as this. And so depending on the model size, depending on the quality of service, you would like to deploy, you could have this in your infrastructure, and they are fungible. And so I'm really excited about our partnership with GCP and the models are going to bring to the inference platforms on GCP is basically across the Board.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question will come from Vivek Arya with Bank of America. Please go ahead.

Q - Vivek Arya {BIO 6781604 <GO>}

Thanks for taking my question and thank you Jensen and Colette, for a very informative event. So I had a near-term and longer-term question. On near-term, just curious about availability of Hopper, how you doing in terms of supply? And then long-term, Jensen, we heard about the range of software and service innovations. How should we track their progress, right? Of the last number I think we heard in terms of software sales was about a few 100 million, so about 1% of your sales. What would you consider success over the next few years? What percentage of your sales do you think could come from software and subscriptions overtime? Thank you.

A - Colette Kress {BIO 18297352 <GO>}

So let me first start, Vivek with your statement regarding supply on our H100. Yes, we do continue building out our H100, for our demand that what we've seen this quarter. But keep in mind, we're also seen stronger demand from our hyperscale customers for all of our data center platforms as they focus on generative AI. So even in this last month since we've talked about earnings, we're seeing more-and-more demand. So we feel confident that we will be able to serve this market, as we continue to build the supply, but we feel we're in a good space at this time.

A - Jensen Huang {BIO 1782546 <GO>}

I think that software and services will be a very substantial part of our business. However, as you know, we serve the market at every layer, we're full stack company, but we're in open platform, meaning that if a company would like -- if a customer would like to work with us at the infrastructure level, at the hardware level, we're delighted by that. If they would like to work with us at the hardware plus library level, we're delighted by that. At the platform level we're delighted by that. And if a customer would like to work with us all the way at the services level or at any of the level, all inclusive, we're delighted by that.

And so we have the opportunity to grow all three layers, the hardware layer is of course already a very large business, and as Colette mentioned, that part of our business, generative AI is driving acceleration of that business. And at the platform layer -- these two layers are just being stood up as cloud services. For companies that we'd like to have an on-prem, they're going to be based on subscription. However, as we all know that today with the world being multi-cloud, you really need the software to be on cloud, as well as on-prem. And so the ability for us to be multi-cloud, hybrid cloud is a real advantage and real benefit for our two software platforms and that is just beginning. And then lastly our AI foundation services are just analyze and just beginning.

I would say, that the model that we presented last time includes our sensibility that we're talking about today. We've been talking about, laying the foundations and the path towards today. This is a very big day for us and the launch of a --probably the biggest business model expansion initiative in the history of our company. And so I think the $300 million of platform and platform software and AI software services, that today has just been pulled in. And -- but I still think that it's -- the size of it is consistent with what we've described before.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question will come from Raji Gill with Needham.

Q - Rajvindra Gill {BIO 16383656 <GO>}

Thank you, Jensen, and thank you for this presentation. Just a question from a technological perspective, regarding the relationship between memory and compute. As you mentioned these generative AI models are creating huge amounts of compute, but how do you think about the memory models? And do you view memory as a potential bottlenecks, so how do you solve the memory disaggregation problem? That'll be help to understand. Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. So, well, it turns out in computing everything is a bottleneck. And if you push to the limits of computing, which is what we do for a living, we don't build normal computers, as you know, we build extreme computers. And when you build the type of computer as we build processing is a bottleneck. So the actual computations are bottleneck. Memory bandwidth is a bottleneck. Memory capacity is a bottleneck. Networking -- or the computing fabric is a bottleneck. The networking is a bottleneck. Utilization is a bottleneck, everything's a bottleneck. We live in a world of bottlenecks, we're surrounded by bottles.

And so, the thing that, that is true, as you were mentioning, is the amount of memory that we use, the memory capacity that we use is increasing tremendously. And the reason for that is, of course, most of the generative AI work that we do in training the models require a lot of memory. But in inferencing requires a lot of memory. The native -- the actual inferencing of the language model itself doesn't necessarily require a lot of memory.

However, if you have -- if you want to connect to a retrieval model, that augments the language model, augments the chatbot, with proprietary very well curated data, that is custom to you, proprietary to you, very important to you. Maybe it's healthcare records, maybe it's about a particular type of domain of biology, maybe has something to do with chip design. Maybe it's an AI, it's a database that has all of the domain knowledge of NVIDIA and what makes NVIDIA click and where all of our proprietary data is embedded inside the walls of our Company. Can now be using a large language model, we could create these datasets that can then augment our language model. And so, increasingly we need not just large amounts of data, but we need large fast data. Large amounts of data, there are many ideas for that, of course, all of the work that's done with SSDs, all of the work that's people are doing with CXL and basically affordable attached disaggregate memory. All of that is fantastic, but none of that is fast memory. That's affordable memory, that's large amounts of accessible hot memory, but none of it's fast memory. What we need is something like what Grace Hopper does.

We need 1 terabyte per second of access to 0.5 terabyte of data. And if we had 1 terabyte per second to 0.5 terabyte of data, if you wanted to have a petabyte of data in the distributed computing system, just imagine how much bandwidth we're bringing to bear. And so this approach of very high speed, very high capacity data processing is exactly what Grace Hopper was designed to do.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question will come from Stacy Rasgon with Bernstein Research. Please go ahead.

Q - Stacy Rasgon {BIO 16423886 <GO>}

Hey guys, thanks for taking my question. I appreciate it. I was wondering if you could go a little bit into the economics of the DGX Cloud business. So like who actually pays for the infrastructure and they have just the cloud vendor pay debt (inaudible) and then you lease it back, so you're running it? Or I guess just how does that work and then how does the customers paid who gets the upside in the economics from the customers, how you pricing? Anything you can give us on how that actually works and impacts the model would be super helpful.

A - Jensen Huang {BIO 1782546 <GO>}

Yes. Stacy, thank you. First of all, the process goes like this, we presented the NVIDIA DGX Cloud partnership to our CSP partners, and they are all super excited about it. And the reason for that is because it's -- we are the onboarding, if you well of very important customers and very large partners that would then consume their storage, security, a whole bunch of other application APIs. And so when we presented this idea that we would like to rent Nvidia DGX Cloud from them and we would take the instances -- the reserved instances, it was called reserved instances to the market ourselves and engage customers, they were super, super happy about that.

Obviously, NVIDIA has very deep relationships with many large vertical ecosystems in the world, I highlighted too in the slide deck that I sense you guys in healthcare and drug discovery, we have very deep relationships with many companies there. We have very deep relationships with just about every car company on the planet. And these two industries, particularly have a great deal of urgency to take advantage of the latest generation of AI -- generative AI or Omniverse digitalization. So the first thing is, we present the idea, the proposal that proposal partnership to them. And then if they're interested and so far they've been incredibly enthusiastic, they would then purchase systems that are -- that include other people's gear but also includes our gear to standup the DGX Clouds. And so the cloud service providers, procure whatever infrastructure, power, networking, storage, so on and so forth, in order to standup the infrastructure and hosted and manage it. Okay, so that's step two.

And then step three is we take the DGX Cloud services to market. And in combination of all the value that we would deliver we would set the price and engage the customers and directly engage the customers business.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question comes from Aaron Rakers with Wells Fargo. Please go ahead.

Q - Aaron Rakers {BIO 6649630 <GO>}

Yeah. Thanks for taking the question. I want to go back, I think maybe with C.J.'s question earlier, just kind of the breadth of the Grace, maybe not the Grace Superchip, but just the Grace CPU strategy. As we think about kind of the evolution, maybe you can help us appreciate, how much of data center cloud workloads are single threaded performance? And in that context, do you foresee a situation where actually your -- see server partners deploying, the Grace CPUs without necessarily deploying your H100 or subsequent versions of GPUs. Do you see actual single CPU deployment as a market opportunity for you?

A - Jensen Huang {BIO 1782546 <GO>}

And I'll look backward again, I appreciate the question, the answer is yes. However, Grace is really, really targeted at a niche market and let me just be clear about that. The x86 -- we use x86 as all of our company. And we use x86, obviously our PCs, our workstations, we have exciting go-to-markets with Intel Sapphire Rapids for all the new workstation lines. We're using Sapphire Rapids in our DGX, we're using Sapphire Rapids in our OVX servers. And so because they're single-threaded performance of Sapphire Rapids is really quite good, it's excellent in fact.

And as I mentioned, when you take an application and you accelerate all of the workload that is -- all the work -- the parts of the curve that you can accelerate with (inaudible). Then really all that's left, a single-threaded code. And that single-threaded code is either doing control or oftentimes it's moving memory around -- giant amounts of memory is managing memory. And the amount of data that is managing is growing quite tremendously. And so as I mentioned, Grace is really designed for the type of applications where the data center is largely accelerated as well as moving a lot of data. Now having said that, and for customers who need x86, obviously, which represents a large part of the world and remains a large part of the world. And I expect to continue to be so, x86 is the predominant platform and it doesn't make sense to move enterprise computing to ARM, necessary to Grace necessarily.

And so I think that where we're focused are the applications that I mentioned. However, in some of the CSPs where they are already going to move to ARM [ph], because they would like to, they would like to build CPUs that are bespoke to their needs and their requirements. And Grace is really a great companion and the reason for that is, because the design point that we designed Grace for is very different than the design point that almost any other CPU that I've known about has been designed for. And so I think that for cloud data centers that are moving in the direction of ARM, this was really a wonderful way to either accelerate that, they benefit from the entire software expertise and the systems ecosystem and the peripherals ecosystem that we've brought to Grace and the design point is so special. And it's really designed for a energy efficient, extreme energy efficient cloud data center. And so these, anybody who is interested in these particular areas which is not everyone in the world, but it's also a very important segment of the world. I think Grace is going to be very successful for them even as an independent and standalone CPU.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question is from Matt Ramsay with Cowen.

Q - Matthew Ramsay {BIO 17978411 <GO>}

Thank you, Simona. Thanks, Jensen and Colette for doing this. I guess, I have two questions Jensen and one of which is kind of a follow-up to something that I'd asked you about, in some prior calls. Which is, there this transition happening, I think from -- in your data center business, from selling accelerate our cards to selling systems. And I'm really interested in what that means for the economics of your data center business, in terms of margins long-term?

And I guess the second question is a little bit related and extends a bit on this DGX Cloud opportunity. I just wanted to --one of the questions I've been getting a lot over the last month, month and a half since you announced is maybe take Microsoft as an acute example, like how are you guys partnering with them? Is there -- is it really a partnership? Is there some friction point into who might want to own the customer relationship, so how did that evolve over time? If you could just kind of walk us through, it seems like they would want to own the AI customer, you guys were -- are going to now go to them directly with rented space from the cloud from them. And just this, how does that evolved overtime and the relationship between yourselves and your largest CSP customers as you bring that business to market? Thanks.

A - Jensen Huang {BIO 1782546 <GO>}

I really appreciate the question. First question, so you can't build software, you can't develop software. We can't generally develop software, if you're not a systems company. And the reason for that, you can't build software for a chip, the chip doesn't sit there and be a computer. And so you have to be a systems company. And especially, the type of software we develop, we're not trying to replicate somebody's software, we're building bespoke brand new software.

None of the software that we created existed before we created them. Even computer graphics with RTX and full path tracing and all of the AI generation that we used in modern computer graphics wasn't possible until we created the software. And so in order to create a software, you have to have a system. And as a systems company, what makes you NVIDIA unique as this, we build the entire system from the data center down. We literally start from the data center, not from a chip. We start from the data center, we build the entire computer. In the future the data center is the computer, the entire data center is the computer.

And that's -- it's something I've spoken about for coming up on a decade now, it's one of the reasons why our combination with Mellanox was so strategic, so important, I think people realize it today. The work that we did together to architect into our data centers is really quite foundational. And the way we think about the world, where I see the world is -- the entire data center, frankly, even on a planetary scale is the computer. And so you have to think about the world starting there. And that includes the computing elements, that includes the systems, that includes networking and storage and compute fabric and CPUs and so on and so forth, all the way up to systems software stack and very importantly the algorithms, the libraries.

We design it as a data center and the way we design it, we design it with discipline, so that we can then disaggregate it, fractionalize it.

So if a customer would like to buy, just the aged DGX GPU, which is right here, this is what a GPU looks like today. A lot of people think that a GPU looks like this, of course, this is NVIDIA's GPU and they both run the same software stack. That's kind of a miracle. This one runs the same software as that, it just went a slower and it takes a long-time to do it. And so the ability for us to design the entire data center, and then disaggregate it and let customers decide what's the best form factor for them, best configuration for them, best deployment methodology for them, some people use MPIs, some people use Kubernetes, some people use VMware and some people use containers with bare metal and the list goes on. And yet the distributed computing stack is affected by all of that.

And so we work we work with the industry across all of the layers of the software and then we disaggregate the system, the components, the system software. We just aggregate the libraries. You can run it anywhere you like from a workstation, a PC, all the way up to the cloud or supercomputer. We disaggregate the networking, we disaggregate the switches, we disaggregate literally everything. Or we're delighted to put it all together for you, if you would like us to stand-up a supercomputer for you and you like it in 30 days, it's possible, because we've productized the entire thing, we disaggregated and integrated into the world's industry standards, wherever we could. And as a result this computing platform is literally everywhere and it's binary compatible, that's the magic.

And I think that's one of the reasons why we were able to -- on the one-hand be a systems company and develop software and on the other hand, be a computing platform company that's available everywhere. With respect to the go-to-market, if we lose the customer, if the CSP would like to have direct customer relationship, we're delighted by that. And the reason for that is because they have a whole bunch of NVIDIA GPUs in their cloud, they've NVIDIA computing in their cloud and we have our software platforms in their cloud anyways. If a customer would like to use it in that way, they can download NVIDIA AI enterprise, they can run their stack, so on and so forth. Everything just works exactly as it does today.

However, there are many customers that we'd like to or need to work with us, because we refactor their entire stack. And we have the expertise because we understand the entire stack, how to take a problem, that otherwise is barely possible. Or it's barely possible in a multi-cloud configuration, meaning they would like to run it in the cloud as well they would like to run it in Azure or OCI or GCP, as well as on-prem.

We have the expertise to help them do that. And so in those cases they need to have direct access to our engineers and our computer scientists and is also the reason why we're so busy. We're working with industry leaders, who would like to build something very quite special or quite proprietary, based on their own platform and they just need our computing expertise to make it possible for them to either deploy it at the scale they want, at the reach of multi-cloud that they want or at the level of cost and power that they would like to reduce. And in which case they contact us.

Now notice, if we become the direct customer interface, we would still invite our CSP partners because we don't offer storage, we don't offer the rest of the APIs, we don't offer security. There are many industrial safety and privacy and data management regulations and standards that has to be complied with and the world's leading CSPs have those expertise. And so there is a lot of collaboration that's going to happen. If they come through us, terrific. If they go through the CSPs, fabulous. We're happy about it either way.

A - Simona Jankowski {BIO 7131672 <GO>}

Our next question will come from Blayne Curtis with Barclays. Please go ahead.

Q - Blayne Curtis {BIO 15302785 <GO>}

Hey, thanks for let me ask the question. I wanted to ask on inference and it's kind of two parts, because there's two (inaudible) here, small models and large. So, I guess, what I'm curious on, you had a T4 card back a while ago, I don't think you did anything with Ampere. So the L4 is kind of the new version of that. So I'm just trying to understand back when you did T4, I think we're talking about inference being similar to what you said now kind of a large market, maybe even as big as training. And I think it became CPUs. What's changed now, I guess, that you're feeling like those smaller models need to move to an accelerator?

And then just trying to understand a large side, the NVL is like 700 watts, so that seems like a lot of power to add to every server. So it's how our customers thinking about deploying this? It's huge model, need lot of horsepower, but it's not a one-for-one into every CPU. So kind of two parts of the equation there on inference and how do you guys monetize it?

A - Jensen Huang {BIO 1782546 <GO>}

Yeah. Thanks. T4 is one of the most successful products in our history. Millions of T4s are in the cloud. And however there are tens of millions of CPUs in cloud. And so there's still a lot of workload in the cloud, that's done on CPUs. Now there are two reasons why it really needs to be accelerated now. One, of course, is sustainability. You just have to accelerate every workload you can. And the world can't continue to consume more power for more CPU throughput. And so that's number one.

Number two, generative AI is an inflection point. There's just no question about that now. The capability of AI, the usefulness and all of these different industries, what generative AI has now been connected into. Just think about what has happened in the last couple of months. Generative AI has been connected into the most popular applications on the planet, office, teams, Google docs. Those are the most popular productivity applications in the history of humanity. And that's just been in terms of AI is just been connected into it. And all of that has to be inference somewhere.

And I think the NVIDIA platform is really the ideal platform to inference all that, because we can handle video, we can handle text, we can handle images, we can handle 3D, we can handle video, we can handle well (inaudible), we can handle just by everything you throw at us. And so I think that is really an inflection point. With respect to this, the reason why (inaudible) is nothing in today's cloud data centers. And the thing that is really spectacular about this is with us, you get to replace hundreds of CPU servers. That's the reason why you accelerate. The reason why you accelerate as you spend 700 watts, so that you can save 10 times of that, so 700 watts or 7 kilowatts. And that's the amount, you want to accelerate everything you can, because all of a sudden you reclaim 6.9 kilowatts and you can then reinvest that into future workloads. Okay? So that's -- this is the motion if you will, the conservation of energy motion that the world CSP is going through. We just accelerate workloads, reclaim power, invested into the new growth. The one, two, three step, and the way to do that is to put GPUs into the world CSPs which is easy PC [ph] to do today.

A - Simona Jankowski {BIO 7131672 <GO>}

Thank you. We just have time for one last question and that will come from Will Stein with Truist. Please go ahead.

Q - William Stein {BIO 15106707 <GO>}

Great. Thanks so much for squeezing me in. Jensen, several years ago you really introduced to the world to accelerate it, and were also in offload parallel processing compute, maybe reintroduced, it's something that used to exist long-time ago. But certainly in modern times, this is like big computing, revolution in a way. But now these other products that you're talking about in particular the Grace CPU and the BlueField DPU. Can you talk about, what your vision is for a modern data center? And what do you envision a typical architecture looking like maybe three years, five years from now? Is DPU more relevant with your Grace CPU or more relevant with the more traditional x86? Do you see those X86 servers, continuing to perpetuate and enterprises for traditional enterprise software or do you see it going away? I'd love any sort of long-term view on that. Thank you.

A - Jensen Huang {BIO 1782546 <GO>}

Really appreciate that. I believe the data centers in the next five to 10 years, if we start from 10 years and work our way back. Or even five years and work our way back, will basically look like this. There will be an AI Factory insight and that AI Factory is working 24x7. That AI Factory will take data input, it will refine the data and it will transform the data into intelligence. That AI Factory is not a data center, it's a factory. And the reason why it's a factory is doing one job, that one job is either refining and improving and enhancing our large language model or a foundation model or a recommender system. And so that factory is doing the same job every single day, engineers are constantly improving it enhancing it, giving a new models, new data to create new intelligence. And so every data center will have number one in AI Factory. It will have an inference fleet, that inference fleet will have to support a diverse set of workloads. And the reason for that is because we know that video represents some 80% of the world's Internet today. And so video has to be processed, it has to generate text, it has to generate images, it has to generate 3D graphics.

The images and 3D graphics will populate virtual world. And these virtual worlds will be -- will run on Omniverse type of computers. In these Omniverse computers, we'll of course simulate all of the physics insight, it will simulate all of the autonomous agents inside. It will enable and connect different applications and different tools and it would be able to do essentially virtual integration of plants, digital twins of fleets of computers, self-driving cars, so on and so forth. And so there'll be types of virtual world simulation computers. All of these types of inferencing systems, whether it's 3D inferencing, in the case of Omniverse or physics inferencing. In case of Omniverse to all of different domains of generative AI that we do. Each one of the configurations will be optimal for the domain. But most of them will be fungible, meaning that, each one of the architecture should be able to receive and offload the work from something that's over-provisioned -- oversubscribed -- excuse me, oversubscribed and take -- pickup some of the workloads.

Okay, so the second part is the inference workloads. Every single one of the nodes will have Smart NICs on it, like a DPU, a data center operating system processing unit. And that is going to offload all of -- and offload and isolate, it's really important to isolate it, because you don't want the tenants of the computer, which all are basically inside. You have to think about the world in the future as zero trust. And so all of the applications and all of the communications has to be isolated from each other. They're either isolated by encoding, they're isolated by virtualization. And the operating system is separated from the -- control plane [ph] separator from the compute plane. The control plane, the operating system of the data center will run, be offloaded, accelerated on the DPU, on the BlueField. So that's another characteristic.

And then lastly, whatever is left, that's not possible to accelerate because you're -- the code is just ultimately single threaded. Whatever is left, you need to run it on a CPU, that is the most energy-efficient that you can possibly do, not at the CPU level only, but at the entire compute node really. And so and the reason for that is because people don't operate CPUs, they operate computers. And so it's nice that the CPU is energy-efficient at the core, but if the rest of the data processing and the I/O [ph] and the memory it consumes a lot of power then what's the point. And so the entire compute node has to be energy-efficient.

Many of those CPUs, will be -- a lot of them will be x86 and a lot of them will be ARM. And I think these two CPU architectures will continue to grow in the world's data center because ideally we've reclaimed power through acceleration, which gives the world a lot more power to grow into. And so that acceleration reclaim then grow three-step process, is really vital to the future of data centers. I think this represents a canonical data center, of course, different sizes and scales. You now know, you can now see, as this question kind of reveals our mental image of what a data center does and which also explains why it's so vital that we -- the one thing I forgot to say, and is really vital, is all of this is being connected to two types of networks. There is one type of network, that's the computing fabric, NVLink and InfiniBand our computing fabrics. And they're really intended for distributed computing, moving a lot of data around, orchestrating the computation of all these different computers. And then another layer of networking, Ethernet, for example, for the control for the multi-tenancy for the orchestration, workload management, so on and so forth, the deployment of the service to the users and that's done on Ethernet.

The switches, the NICs, super sophisticated some of it, copper, some of it direct drive, some of it is long-reach fiber. And all of that layer that fabric is vitally important. Now you see, why it is that we invest and what we do. When we think about the data center scale and we start from the computation, the acceleration of it. As we continue to advance it at some point, everything becomes a bottleneck. And whenever something becomes a bottleneck and we have a very specific viewpoint about the future and nobody else is building it in that way or nobody else could build in that way. We would take -- tackle the endeavor and go remove the bottleneck for the computing industry. One of those important bottlenecks of course is NVLink, another one is InfiniBand, another the DPU, the BlueField. I just talked to you about Grace and how it removes bottlenecks for single threaded code and very large data processing code.

And so this entire momento [ph] model of computing, I think in some degree will be implemented very, very quickly in the world CSPs. And the reason for that, it's very, very clear. The two fundamental drivers of computing in the near future. One of them is sustainability, acceleration is vital to that. And the second is generative AI, AI computing is vital to them.

I want to thank all of you for joining GTC. We had a lot of news for you to consume and I appreciate all the excellent questions. And very importantly I want to thank all of researchers and scientists who took the risk and who had the faith in the platform that we were building, that over the last 2.5 decades as we continue to advance accelerated computing. Have used use this technology and used this computing platform to do groundbreaking work. And it's because of you and all of your amazing work that has really inspired the rest of the world to jump on through accelerated computing.

I also want to thank all of the amazing employees of NVIDIA for just the incredible company that you felt build-in and the ecosystem that you've built. Thank you, everybody. Have a great night.

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

2024, BLOOMBERG LP. All rights reserved. Any reproduction, redistribution or retransmission is expressly prohibited.