hello everyone welcome sigraph it is my first sigraph I'm so excited to be here I'm so excited to speak to all of you and I'm so excited to speak to Nvidia founder and CEO Jensen hang great to see you again thank you great to see you welcome to sigraph Lauren Welcome To My Hood you're regular here huh hey everybody great to see you guys Jensen it is like 99 degrees outside I know it's freezing in here isn't it I mean I don't know I'm shaking leather jacket yeah feels great looks good too all right so you have I'm wearing a brand new one oh a brand new one yeah how many of those do you have I don't know but Lori got me a new one for sigraph she said you're excited about cigar here's a new jacket go go do good job I look sharp thank you all right uh you have a long history of sigraph I mean when you think about the history of this conference which has been going on since 1974 and you think about the history of Nvidia from the 1990s onward where your DNA was really in you know know computer Graphics uh helping to make beautiful Graphics what is the significance of Nvidia being here today right now at a conference like sigraph well you know uh sigraph used to be about computer Graphics now it's about computer graphics and generative AI it's about simulation it's about generative Ai and and we all know that the Journey of Nvidia which started out in computer Graphics as you said um uh really brought us here and so I made a cartoon for you I made a cartoon for you of our journey did you make it or did J of AI make it I I I had it hang on a second I had it made I had it made um that's what CEOs do we don't do anything we just have it be done it kind of starts something like this hey guys would it be great if we had a cartoon and it Illustrated some of the most important milestones in the computer industry and how it led to Nvidia and where we are today and and so we Illustrated and also do it in three hours and do three hours right and and so this is this cartoon here is really terrific so these are some of the some of the most important moments in the computer industry uh the IBM system 360 of course the invention of modern Computing uh the teapot 1975 the Utah teapot 1979 Ray tracing um uh Turner wited uh one of the great researchers Nvidia researcher for a long time uh 1986 programmable shading uh of course uh most of the animated movies that we see today wouldn't be possible if not for programmable shading originally done on the CRA supercomputer uh led to uh what and then in 1993 Nvidia was founded Chris Curtis and I founded the company 19 1995 Windows PC revolutionized uh the personal computer industry uh put a personal computer in every home and every desk uh multimedia PC was uh was invented 2001 we invented uh the first programmable shading GPU and and that that really uh drove uh vast majority of nvidia's Journey up to that point But at the background of everything we were doing was accelerated computing and accelerate and and we believe that you could create a type of computing model that could augment the general purpose Computing so that you can solve problems that normal computers can't and the application we chose first was computer graphics and it was probably one of the best decisions we ever made because computer Graphics was insanely computationally in intensive and remained so uh for the entire 31 years that that Nvidia has been here and since the beginning of computer Graphics in fact required a crazy supercomputer to render some of the original scenes so it kind of tells you how computationally intensive it was and it was also incredibly high volume because we applied computer Graphics to an application at the time that uh wasn't mainstream 3D Graphics video games the combination of very large volume very complicated Computing problem led to a very large R&D budget for us which drove the flywheel of our company that observation we made in 1993 was spoton and it led us to be able to Pioneer the work that we're doing in accelerated Computing we tried it many times Cuda was of course the Revolutionary version but prior to that we had a Computing model we call CG C for graphics C on top of gpus and so we've been working on accelerated Computing for a long time uh promoting and evangelizing Cuda getting Cuda everywhere and putting on every single one of our gpus so that this computing model was compatible with the with any application that was written for it irrespective of which generation of our processors that was a great decision and one one day in 2012 we made our first Contact you know Star Trek first Contact with artificial intelligence that first Contact was alexnet it was in 2012 very big moment uh we made the observation that alexnet was an incredible breakthrough in computer vision but at the core of it deep learning was deeply profound that it was a new way of writing software instead of Engineers given input imagining what the output was going to be write algorithms we now have a computer that given an input and example outputs would figure out what the program is in the middle that observation and that we can use this technique to solve a whole bunch of problems that previously wasn't solvable was a great observation and we changed everything in our company to pursue it from the processor to the systems to the software stack all the algorithms Nvidia basic research pivoted towards working on deep learning uh by the way this is a great place for research as you know nvidia's uh passionate about about uh uh sigraph and this year we have 20 papers that are at the intersection of generative Ai and simulation and so in 20 2016 uh we introduced the first computer we built for deep learning and we called it djx1 and I delivered the first djx1 outside of our company I built it for NVIDIA to build models for self-driving cars and Robotics and such and and generative AI for graphics uh but we uh somebody saw an example of djx1 Elon reached out to me and said hey I would love to have one of those for a startup company we're starting and so I delivered the first one to a company at the time uh that knew nobody knew about called open a I and so that was 2016 uh 2017 was the Transformer uh that revolutionized modern uh machine learning modern deep learning in 2018 right here at sigraph we announced RTX the world's first realtime interactive R Tracer R tracing platform we call it RTX it was such a big deal that we changed the name of GTX which everybody refer to our graphics cards as to RTX uh another shout out for for uh uh a great researcher his name is Steph Parker uh many of you know he's been coming to C for a long time uh he passed this year and uh he was a a he was one of the one of the core uh Pioneer researchers behind realtime R raate tracing and we miss them dearly and so anyways uh and you mentioned last year during your sigraph keynote that RTX R tracing extreme was one of the big important moments when computer Graphics met AI That's right but that had been happening for a while actually so what was so what was so important about RTX in 2018 well RTX in 2018 so you know we uh we accelerated uh Ray traversal and bounding box uh detection and and um uh and we made it possible to uh use a parallel processor to accelerate rate tracing um but even then we were rate tracing at about you know frame every call it you know 10 frames maybe every second let's say maybe five frames every second depending on on how how many how many Rays we're talking about tracing and we were doing it at 1080 resolution uh obviously video games uh need a lot more than that obviously real-time Graphics need more than that this crowd definitely knows what that means but for the folks who are watching online who don't work in this field I mean this is basically a way of really manipulating light in computer Graphics simulating how a light interacts with versus through right happening in real time that's right the rendering processes used to take a really long time when you were making something it used to take a Cay supercomputer to render just a few pixels and now we have our RTX to accelerate that rate tracing but it was interactive it was real time but it wasn't fast enough to be uh a video game and so we realized that that we needed a big boost probably something along the lines of 20x or so maybe 50x or so boost and so uh the team uh invented dlss which basically renders one pixel while it uses AI to infer a whole bunch of other pixels and so we basically taught an AI that is conditioned on what it saw and then say fills in fills in the dots for everything else and now we're able to render fully Ray Trace fully path Trace simulation ations at 4K resolution at 300 frames per second made possible by by Ai and so 2018 came along uh 2022 as we all know chat GPT came out but the and what's that again chat GPT you know that yes open AI chat GPT a revolutionary revolutionary new uh new capability Ai and fastest growing service in history um but the two things that I wanted to highlight since since jgp the industry researchers um many of them in the room has figured out how to use AI to learn everything not just words but to learn the meaning of images and videos and 3D chemicals protein physics thermal dynamics fluid dynamics particle physics it's figured out the meaning of these uh all these different modalities and since then not only have we learned it we can now generate it and so you could that's the reason why you can go from text to images text to 3D images to text 3D to text text to video so on so forth text to proteins text to chemicals and so now generative AI has been uh made possible and this is really the Revolutionary time that we're in just about every industry is going to be affected by this just based on based on some of the examples I've already given you whether it's scientific Computing trying to do a better job uh predicting the weather with a lot less energy to uh augmenting and collaborating with with us uh creators to generate images or you know generating virtual scenes for industrial digitalization and very importantly robotics self-driving cars are all going to be transformed by generative Ai and so here we are in this brand new way of doing things and so let me just very quickly Lauren if you look at where we started in the upper left in 1964 uh the way that software was programmed human Engineers programming software now we have machines that are learning how to program the software de writing software that no humans can solving problems that we we could barely imagine before and now because we have generative AI a new way of developing software you know you you you I don't know if you know do you know Andre karpathy he's a really really terrific researcher I met him when he was at Stanford and and uh he coined the original way of doing software software 1.0 machine learning to be software 2.0 and now really we're moving toward software 3.0 because these generative AIS in the future instead of using machine learning to learn a new AI for every researcher you'll probably start with pre-trained models Foundation models that are already pre-trained and the way that we develop software could very very much be like assembling teams with Experts of various AI capabilities some that are using uh tools some that are able to generate you know special things and then a general purpose AI that's really good at reasoning that's connecting this this network of AIS together solving problems like team solve problems and so software 3.0 is here I've gotten the sense from talking to you recently that you are optimistic that this these generative AI tools will become more controllable more accurate we all know that there are issues with hallucinations uh low quality outputs that people are using these tools and they're maybe not getting exactly the output that they're hoping for right meanwhile they're using a lot of energy which we're going to talk about why are you so optimistic about this what is what do you think is pointing Us in the direction of this generative AI actually becoming that much more useful and controllable well uh the big breakthrough of Chad GPT uh was reinforcement learning human feedback which was uh the way of using humans to produce the right answers or the best answers to align the AI on our core values or align our AI on the skills that we would like it to perform that's probably the just the extraordinary breakthrough that made it possible for them to open chat GPT for everyone everyone to use other breakthroughs have have uh arrived since then Guard railing which uh which causes the AI to focus its energy or Focus its response in a particular domain so that uh it doesn't wander off and pontificate about all kinds of stuff that you ask it about it would only focus on the things that it's been trained to do align to perform and um uh that it it has deep knowledge in the third the third breakthrough is called uh retrieval augmented generation which basically is vectorized or data that has been uh embedded so that we understand the meaning of that data and so it's a more authoritative data set it goes beyond just the trained data set and it actually pulls from other sources that's right it's not just pre-ra data source it's and something you know for example uh it might be uh all of the articles that you've ever written all of the papers that you've ever written and so now it becomes uh something an AI That's authoritative on your uh and it could be essentially a uh a chatbot of you uh so everything that I've ever written or ever said could be vectorized and then created into a semantic database and then before an AI responds it would uh figure it would look at look at your prompt and it would uh it would uh search uh the appropriate content from that Vector database and then augment it um in its gener generative process and you think that that is one of the most important factors these three combinations really made it possible for us to do that with text now the thing that's really cool is that we are now starting to figure out how to do that with visuals right and you know sigraph is really about a lot about images and and generation and so if you look at today's generative AI uh you could give it a a prompt and it goes into uh in this particular case this is a uh edify uh AI model that Nvidia created it's a 2d text to 2D Foundation model it's multimodal and um uh we used uh we partnered with GTI to use uh their library of data uh to train uh an AI model and so this is a a text to uh 2D image and you also created this slide personally right I well I had I personally had the slide created and so imagine I'm the prompt and then there's a team that that's kind of like a generative Ai and then magically uh this slide shows up and so so here here's a prompt uh and this could be a prompt for somebody who owns a brand it could be a brand of um for in this case Coca-Cola it could be a car it could be a luxury product it could be anything and so uh you you use the prompt and generate the image however as you know it's hard to control this prompt and it may hallucinate uh it may um uh create it in such a way that it's not exactly what you want and to fine-tune this using words is really really hard because as you know words is very low dimensional ity it's extremely compressed in its content but it's it's very imprecise and so the ability for us to now control that image uh is difficult to do and so we've created a way that allows us to uh control and align that with more conditioning and so the way you do that is we create another model and this model for example allows us to text to 3D on the bottom it's called and it's edify 3D one of our foundation models we've created this AI Foundry where Partners can come and work with us and we create the model for them with their data we invent the model and they bring their data and we uh create a model that they can take with them is it their data only uses their data so this only uses all of the data that's available on Shutterstock uh that they have have the rights to to use the train and so we now use prompt generator 3D we put that in Omniverse Omniverse uh as you know is a is a place where you could compose uh data and content from a lot of different modalities it could be 3D it could be AI it could be animation it could be materials and so we use Omniverse to compose all of these multimodality uh data and now you can control it you could you could change the pose you could change the placement you could change whatever you like and then you use that image out of Omniverse to condition The Prompt okay so you take what comes out of Omniverse you now augment it with the prompt it's a little bit like augment retrieval augment a mented generation this is now 3D augmented generation Getty uh the edified model is multimodal so it understand the image understands the prompt and it uses it in combination to create a new image so now this is a controlled image and so this way we can use uh generative AI as a as a collaborator as a you know as a partner to to work with us um and uh uh we can generate images exactly the way we like it how does this translate to the physical world how does it translate to something like robotics well we're going to talk about robotics but one of the things that I would love to show you and I had this made not by myself well I had it made myself okay this is an Incredible video and this is a uh this is a work that is done by wpp uh Shutterstock uh working with uh some of the brand uh world class world famous brands that you'll you'll know let's run the video Let's show the build me a table in an empty room surrounded by chairs in a busy restaurant build me a table with tacos and bowls of salsa in the Morning Light build me a car on an empty road surrounded by trees buy a modern house build me a house with hills in the distance and bales of hay in the Evening Sun build me a tree in an empty field build me hundred of them in all Direction with bushes and Vines hanging in between building a giant brain forest with exotic flowers and rays of sunlight isn't that incredible and so so let so this is so this is this is what happened we taught we taught an AI how to speak USD open USD and so the young the the the girl is speaking to Omniverse Omniverse generates the USD and use USD search to then find the catalog of uh 3D objects that it has it composes the scene using words and and then um generative AI uh uses that augmentation to generate uh to condition the generation process and and so therefore uh the work that you do could be much much better controlled you could even collaborate with people because you can collaborate in Omniverse and you can collaborate in 3D it's hard to collaborate uh in 2D and so we can collaborate in 3D augment the generation process I imagine a lot of people in this room who aren't just technical but they're also storytellers this is a very technical room storytellers see something like this there's like 90% phds in here and think I'm not even going to ask you to do a raising of your hand but I'm sure that would be fascinating so they see something like this I see something like this and I think okay that's pretty amazing you are speeding up rendering times you're creating images out of nothing there are probably just as many people thinking what does this mean for my job where do you draw the line between this is augmenting and helping people where do you see the line being drawn and this is replacing certain things that humans do well that's what tools do uh we invent tools here this you know this conference is about inventing technology that ultimately ends up being a tool and that Tool uh either accelerates our work um uh collaborates with us so that we could do uh better work or even bigger work uh do work that's uh impossible before and so I think what you're going to what you you'll likely see is that generative AI uh is now going to be more controllable than before we've been able to do that with using uh Rags retrieval augmented generation to control uh text generation better reducing hallucination now we're using Omniverse with generative AI to control generative uh images better and reduce hallucination both of those tools uh help us be more productive and do things that we otherwise can't do and so I think I think um for all of the artists in the world I what I would say is is uh jump on this tool give it a try um imagine the stories that you're going to be able to tell uh with these tools and um uh and with respect to jobs uh I would say that it is very likely all of our jobs are going to be changed in what way well my job is going to change um the way in the future uh I'm going to be prompting a whole bunch of AIS uh everybody will have an AI that is an assistant and so every single company every single company every single job within the company will have AIS that are assistant to them our software programmers as you so you know now have AIS that help them program uh all all of our software Engineers have AIS that help them debug software uh we have AIS that help our chip designers design chips uh without without AI uh Hopper wouldn't have been possible without AI Blackwell wouldn't be possible you know today we're this week we're sampling we're sending out engineering samples of Blackwell all over the world they're under people's chairs right now I think if you just look if you get a GPU and you and I yeah you get a GPU you get a GPU yeah that's right supply chain we all we all we all wish yeah yeah and so so um I none none of the work that we do would be possible anymore without without generative Ai and uh that's increasingly the case with uh uh uh Our IT department helping our employees be more productive uh it's increasingly the case with our supply chain team optimizing Supply uh to be as efficient as possible um or our data center team using AI to manage the data center to save as much energy as possible you mentioned Omniverse before yeah uh that's not new but the idea that more generative AI would be within the Omniverse helping people create these simulations or digital twins yeah that's what we're announcing this week by the way Omniverse now Omniverse now uh understands uh text to USD um it could understand text to uh and has a semantic database so that it could do search of all the 3D objects um and uh that's how that that young lady was able to to say fill fill uh the scene with a whole bunch of trees uh describing how she would like the trees to be organized and somehow populates it with all these 3D trees then when that when that's done that 3D scene then goes into a generative AI uh uh model which turns turns it into a photorealistic model and if you want the the Ford truck to not be augmented but to use the the the actual brand um brand ground Truth uh then it would it would honor that and keep that uh keep that in the or in the final scene and so so I think if you if you if you do that uh so one of the things that that we talked about is how every single every single group in the company uh will have will have ai assistance and and um there's a lot of question lately about about um uh whether all this infrastructure that we're building is leading to productive work in companies I just gave you an example of how generative AI is impossible without without uh Nvidia nvidia's designs would be impossible without generative AI so we use it to transform the way we work but we also use it uh in many examples that I've just shown you in creating new products and new technology that either makes possible uh rate tracing in real time uh or Omniverse that we can now uh uh imagine and help us uh create much larger scenes um or uh our self-driving car work or our robotics work uh none of that none of that new capability would be possible without it and so one of the things that that we're announcing here uh this week is uh the concept of uh digital uh agents uh digital AIS uh that will augment every single job in the company and so uh one of the one of the the most important use cases that people are discovering is customer service and every single group every single company has customer service every single industry has customer service uh and in the future uh it's going today it's it's humans uh doing customer service but in the future my guess is that it's going to be human still but AI in the loop and the benefit of that is that you'll be able to uh uh retain uh the the um uh the experiences of all the customer service agents that you have and capture that institutional knowledge that you can then run into analytics that you can then uh use to create uh better services for your customers uh the the just now I showed you a a Omniverse augmented generation for images this is a rag this is a uh retrieval augmented generative Ai and the thing that we're doing is is uh We've created this customer service basically uh microservice that sits in the cloud and it's going to be available I think today or tomorrow and you can come and try it and we connected to it a digital human front end basically an IO uh the io of an AI that has the ability to uh speak make eye contact with you um an animate in an empathetic way um and uh uh you could decide to connect your chat chat gbt or your AI I to the digital human or you can connect um uh your digital human to our uh retrieval augmented generation uh customer service AI uh so however you like to do it we're a platform company so irrespective of which piece you would like to use uh they're completely open source and you can come and use the pieces that you like uh if you would like the incredible digital human rendering technology that we've created for uh rendering beautiful faces uh which requires subsurface scattering with path tracing this breakthrough uh is really quite incredible and it makes it possible for us amazing Graphics researchers welcome to sigraph 2024 so it makes it possible to animate uh using an AI so uh you you chat with the AI it generates text that text then is translated uh to sound text to speech that speech the sound then animates the face and and then RTX path tracing um does the does the rendering of the digital human and so all of this is available for developers to use and you could you could decide which parts you would like to use how are you thinking about the ethics of something like this you're unleashing this to developers to Graphics artists but these are being Unleashed into the world y do you think a chatbot like that a a very human like visual chatbot should say that it's a chatbot what is it so human that people start mistaking it for humans they're emotionally vulnerable it's it's still it's still pretty it's still pretty robotic and I think that that's not a terrible thing you know we're going to stay we're going to be robotic for for some time I think it we've made we've made uh this digital human technology uh quite realistic but you and I know it's still a it's still a robot and so I I think um uh that's not a that's not a a horrible way um it is the case that there are many many different applications where the human engagement is much uh much more engaging uh having a human representation or near human representation than a text box uh maybe somebody needs companion or Healthcare needs to advise um somebody who is an outpatient uh who had just gone home uh you know helping elderlies um you know there's a whole bunch of applications uh a tutor which to educate a child um all these different applications are better off having somebody who is much more human and being able to connect with uh with the audience that's interesting yeah what I hear you talking a lot about today these are software developments right they're relying on your gpus But ultimately this is software this is NVIDIA going further up the stack meanwhile there are some companies some folks in the generative AI space who are in software and Cloud services but they're looking to go further down the stack right they might be developing their own chips or tpus that are competitive with what you are doing how crucial is this software strategy to Nvidia maintaining its lead and actually fulfilling some of these promises of growth that people are looking at for NVIDIA right now well um we've always been a software company um and even first and the reason for that is because accelerated Computing is not general purpose Computing general purpose Computing can take any c program or C Supply program Python and just run it and almost everybody's uh program can be compiled to run effectively unfortunately when you want to accelerate fluid dynamics you have to understand the the algorithms of fluid dynamics so that you could uh refactor it in such way that it could be accelerated and you have to design an accelerator you have to design the cou at GPU so that it understands the algorithms so that it could do a good job accelerating it and the benefit of course is that we can by doing so by redesigning the whole stack we can accelerate applications 20 40 50 times 100 times for example uh we just put um Nvidia gpus in in uh gcp uh running pandas uh which is the world's leading uh data science platform and we accelerated from 50 to 100x over general purpose Computing uh in the case of deep learning over the course of last 10 to 12 years or so uh We've accelerated deep learning by a million times which is the reason why it's now possible for us to create these large language models a million times speed up a million times reduction in cost and energy is what made it possible for us to make General generative AI possible but that's not that's that's by designing a new processor a new system tensor core gpus the mvlink switch fabric uh is completely groundbreaking for AI of course the systems itself the algorithms the distributed computing uh uh uh Library we call Megatron that everybody uses um tensor rtlm those are algorithms and if you don't understand the algorithms the the applications above it it's really hard to figure out how to design that whole stack what is the most important part of Nvidia software ecosystem for nvidia's future well every single one of it takes a new library we call it dsl's domain specific Library um in in uh in generative AI that DSL is called qnn uh for SQL processing uh data frames is called qdf and so if you go SQL pandas uh CDF is what makes it possible for us to accelerate that for Quantum emulation it's called K Quantum uh uh C fft we got a whole bunch of coups uh computational lithography which makes it makes it possible for us to uh help the industry Advance the next generation of processed technology called K litho um you know the the number of coups it goes on and on every time we introduce a domain specific Library it exposes accelerated Computing to a new market and so as you see it takes that collaboration the the the the the um uh the full stack of the library and the architecture and the go to market and developers and the ecosystems around it to open up a new field and so it's not just about building the accelerator you have to build a whole stack nvidia's dependent on a lot of things going right your your foray into the future your Innovation depends on a lot of things going right you have to continue pushing the laws of physics you do have competitors who are nipping at your heels at all times we've talked about this what keeps you up at night uh you're also somewhat relling on the geopolitical stability last night just so you know elevation drink some water that's what they told me but it was it was too late by the time I learned about it this morning I I woke up with a a terrible headache elevation that's what was last night okay so elevation uh but truly you have to keep pushing the laws of physics you have competitors who are nipping at your heels both on the software and Hardware side you are somewhere Aliant on the geopolitical stability of the South China Sea uh geopolitics uh so much going on right now you're reminding me it's it's it's super hard building a company it is you but you've had a lot of you're making me nervous I was fine before there's so many things that he wants to go back to showing you slides uh but truly you've and you've had you've had a lot of Tailwinds and I'm just you know how optimistic are you that these the things are going to keep trending in your direction things have never trended in our Direction uh you have to will the future into existence accelerated Computing you know the world wants general purpose computing and the reason for that is because it's easy you just you just have the software it just runs twice as fast every single year don't even think about it and you know every five years is 10 times faster every 10 years is 100 times faster what's not to love but of course you can shrink a transistor but you can't shrink an atom and eventually uh the CPU architecture uh ran its course and so we it's not sensible anymore uh as the technology doesn't give us those leaps that a general purpose instrument could be good at everything you know could be good at these incredible things from Deep learning to Quantum simulation to molecular Dynamics the fluid dynamics right so on the computer graphics and so we created this accelerated Computing architecture to do that um but that

that that fights that fight that's headwind do you see what I'm saying because general purpose Computing is the easy way to do it we've been doing it for 60 years why not keep doing it and so so accelerated Computing was only possible because we we deliver such extreme extraordinary speedups at a time when energy is becoming more scarce um at a time when when um uh we no longer could just ride the CPU curve any longer dard scaling is has has really uh ended and um I and so we need another approach and and that's why that's why we're here but notice every single time we want to open up a new market like qdf in order to do data processing data processing is probably what a third of the world data a third of the world's Computing every company does data processing and most companies data is in data frames you know in tabular format and so in order to create an acceleration library for tabular formats was insanely hard because what's inside those tables could be floating Point numbers 64-bit integers it could be you know numbers and letters and all kinds of stuff and so we have to figure out a way to go compute all that and so so you see that almost every single time we want to grow into something you have to go and learn it that's the reason why we're working on robotics that's the reason why we're working on autonomous vehicles to understand the algorithms that's necessary to open up that market and to understand the Computing layer underneath it so that we can deliver extraordinary results and so as you could see each each time we open up a new market um healthare digital biology the work with the amazing work we're doing there with biion Nemo and uh uh parab bricks for Gene sequencing every single time we open up a new market it just requires us to reinvent everything of that Computing and so there's nothing easy about it generative AI takes up a lot of energy I'm just saying my job's super hard but your assistance you're your AI assistants are going to make it easier right what's that somebody's got to Pat my back hey little Applause you guys Cher on yeah go ahead uh let's talk about energy yeah generative AI incredibly energy intensive y uh I am going to read from my note cards here uh According to some research chat gbt a single query takes up nearly 10 times the electricity to process a single Google search uh data centers consume 1 to 2% of overall worldwide energy but some say that it could be as much as 3 to 4% some say as much as 6% by the end of the decade uh data center workloads tripled between 25 uh 2015 and 2019 that was only 2019 um AI generative AI is taking up a large portion of all of that is there going to be enough energy to fulfill the demand of what you want to build and do yeah um yes and um a couple of observations so first there there are two or three um or or three or four uh model makers that are pushing to Frontier a couple of years ago uh they're they're probably three times that many this year but it's still it's still single digit you know it's very high single digit but call a 10 that are pushing the frontiers of um of models and the size of the models are uh call it uh twice as large every year maybe maybe uh faster than that and in order to train a model that's twice as large you need you know more than twice as much data and so the computational load is growing um probably uh by you know call it a factor of four each year just for simple simple simple thinking now that's one of the reasons why Blackwell uh is is uh so highly anticipated because we accelerated the application uh so much using the same amount of energy and so this is an example of accelerating applications uh uh at constant energy constant cost you're making it cheaper and cheaper and cheaper now the important thing though is I've only highlighted 10 companies the world has tons of companies and their data centers everywhere and Nvidia is selling gpus to a whole lot of companies and a whole lot of different data centers and so question is what's happening at the core the first thing that's actually happening is the end of CPU scaling and the beginning of accelerated Computing data processing um uh just text completion speech recognition all of those kind of basic AI things that are that are used um recommender systems uh that are used in in data centers all over the world they are moving everyone is moving from CPUs to accelerated Computing because uh they want to save energy accelerated Computing helps you save so much energy 20 times 50 times in doing the same processing so the first thing that we have to do you know as a society is accelerate every application we can can if you're doing spark data processing ex run it with accelerated spark so that you could reduce the amount of energy necessary by 20 times if you're doing SQL processing do SQL um accelerated SQL so that you could reduce the power by 20 times and so uh uh if you're doing weather simulation accelerate it when you're doing uh whatever scientific simulation you're doing accelerated uh image processing accelerated a lot of those applications used to be uh running on CPUs and general purpose Computing all of that should be accelerated going forward that's the first thing that's happening now it is that reducing the amount of energy being used all over the world absolutely the density of our gpus and density of accelerated Computing is higher energy density is higher but the amount of energy used is dramatically lower so that's the first thing that's happening of course then generative AI generative AI is probably consuming let's pick a very large number probably you know 1% or so of the world's energy but remember even if the data centers uh consume 4% of the world the goal of generative AI is not training the goal of generative AI is inference and the inference ideally we create new models for predicting weather uh predicting new materials allow us to uh um uh optimize uh our supply chain reduce the amount of energy consumed and wasted gasoline uh as we deliver products and so the goal is actually to reduce the energy consumed of the 96% and so uh very importantly you have to think about gener about AI from a longitudinal perspective not just going to school but what happens after going to school you and I both went to Stanford Stanford's not inexpensive um I think you studied something slightly different though yeah yeah sure it's a big school it's worked out well for it's worked out well for both of us and so and so so the goal of course is is is going to school is important uh but of course the important thing is is really after school and all of the contributions that we're made we're able to make the society and so generative AI is going to increase productivity it's going to Ena enable us to discover new science make things more energy efficient um don't let me don't let me finish without without showing you uh the next and so so that the lights just came on because why we were talking about energy and all of a sudden it's like the Earth was like okay Tamp down the energy usage folks I thought I thought they were G am I getting Chang we still have a few minutes so I mean I hope so I mean I'm not I'm not going to get off the stage until Mark comes on here and kicks me off how about that he's not going to do that he's a great guy so anyways uh think think think about generative AI um longitudinally and all the impact of of generative AI the second thing the next thing I'll say about generative AI is remember in the the traditional way doing Computing is called retrieval based Computing everything is pre-recorded all the stories are written pre-recorded all the images are pre-recorded all the videos are pre-recorded and so and so everything is stored off in a data center somewhere pre-recorded generative AI reduces the amount of energy necessary to go run to a data center over the network re retrieve something and bring it over the network don't forget the data center is not the only place that consumes energy the world's Data Center only rep that is only 40% of the total Computing done 60% of the energy is consumed on the internet moving the electrons around moving the bits and bites around and so generative AI is going to reduce the amount of energy on the internet because instead of having to go retrieve the information we can generate it right there on the spot because we understand the context we probably have some uh content already on the device and we can generate the response so that you don't have to go retrieve it uh somewhere else well part of that is also moving atoms around right one last thing I got to tell you laen one last thing I remember AI doesn't care where it goes to school today's data centers are are built near the power grid where Society is of course because that's where we need it in the future you're going to see data centers being built in different parts of the world where there's excess energy it's just that it costs a lot of money to bring that energy to society maybe it's in a desert maybe it's uh in places that has a lot of sustainable energy but it's not a lot of water well uh there's plenty of water as well it just happens to be undrinkable water and so uh we can use we can use water that are uh we we can we can put data centers where there's less population and um more energy okay just don't don't forget that there's there's plenty there's a lot of energy coming from the Sun there's a lot of energy in the world and what we need to do is move data centers out closer to where there's excess energy and not put everything near population well part AI doesn't care where it's trained part I'd never heard that phrase before AI doesn't care where it goes to school and that's interesting yeah it's true I'm going to think on that yeah uh part of part of calculating the carbon emissions though is also considering the supply chain it's also considering it's going all down the line uh and it requires transparency don't don't move don't move the energy to the data center use the energy where the data center is and then when you're done uh you have a highly compressed model that is essentially the compression of all the energy that was used and we can take that model and bring it back hey can we talk about the next wave so the first wave of course the first wave is accelerated Computing I think I know that she's the interviewer and we're on we're we're we're doing this on her terms but I'm the CEO so so and and and so so no Lauren we we need to come and tell tell this this group about the work that we're doing that that that uh is is really really Court to I I have so many good questions for you I know I know I want to ask you about open source which I think you're going to be talking to Mark about I I want to ask you about I have it import by the way open source is really important it's incredible yeah if not if if not for open source if not for open source uh how would all of these industries and all these companies be able to engage Ai and so look at look at all the companies and all the different Industries they're all using llama llama 2 today today llama 3.1 just came out uh people are so excited about it uh We've we've made it possible to democratize Ai and engage every single industry in AI but the thing that I want to say is this the first wave is accelerated Computing reduces energy consumed allows us to deliver continued computational demand without all of the Power continued to grow with it so number one accelerate everything it made it possible for us to have generative AI the generative AI the first wave of it of of course is all the pioneers and we you know we know many of the Pioneers open AI anthropic uh Google uh Microsoft a whole bunch of amazing companies doing this uh uh X is doing this x. is doing this amazing companies doing this the next the next wave of AI we we didn't talk about which is uh Enterprise of course one of its applications is customer service and we hope that we can uh give every single organization the ability to create their own AIS and so everybody would be augmented and and have a collaborative AI that could um uh Empower them help them do better work the next wave of AI after that is called physical Ai and this is this is really really quite extraordinary this is where we're going to need three computers one computer to uh create the AI another computer toes simulate the AI both using synthetic gen for synthetic data generation as well as a place where the AI robot the humano robot or the the manipulation robot could go learn how to uh refine its Ai and then of course the third AI is the computer that actually runs the AI so it's a threea three computer problem you know it's a three body problem and so it's incredibly comp it's incredibly complicated and we created three computers to do that and we made a video uh for you for you to enjoy uh understanding this the thing that that we've done here is this uh in every each one of these computers depending on whether you want to use the software stack the algorithms on top or just the Computing infrastructure just a processor for the robot or the functional safety operating system that runs on top of it or the AI and computer vision models that run on top of that or uh just the computer itself any piece is any layer of that stack is open uh for robotics uh developers we created a quick video let's take a look at that is that okay that sounds great the era of physical AI is here physical AI models that can understand and interact with the physical world will embody robots many will be humanoid robots developing these Advanced robots is complex requiring vast amounts of data and workload orchestration across diverse Computing infrastructures Nvidia is working to simplify and accelerate developer workflows with three Computing platforms Nvidia AI Omniverse and Jetson Thor plus generative AI enabled developer tools to accelerate project Groot a general humanoid robot Foundation model Nvidia researchers capture human demonstrations seeing the robot's hands in spatial overlay over the physical world they then use robokos a generative simulation framework integrated into Nvidia Isaac lab to produce a massive number of environments and layouts they increase their data size using mimic gen M which helps them generate large scale synthetic motion data sets based on the small number of original captures they train the Groot model on Nvidia dgx cloud with the combined real and synthetic data sets next they perform software in the loop testing in Isaac Sim and the cloud and Hardware in the loop validation on Jetson Thor before deploying the improved model to the real robots Nvidia osmo robotics Cloud compute orchestration service manages job assignment and scaling across distributed resources throughout the workflow together these Computing platforms are empowering developers worldwide to bring us into the age of physical AI powered humanoid robot you know you know what's amazing Lauren at this conference sigraph is where all of this technology comes together isn't that right everybody researchers of sigraph isn't that right so whether it's computer Graphics or simulation artificial intelligence robotics all of it comes right comes together right here at sigra and that's the reason why I think you should come to sigra from now on me yes I'm happy to I'm thrilled to you I I I took I took uh uh I want 100% of the world's Tech press should come to Cigar we can get behind that just drink a lot of water um uh I went and saw some of the art exhibits last night upstairs in the exhibition effort fantastic just really really cool loved the literal spam Bots whoever created that one go check it out um I was actually listening to the sigraph spotlight podcast before this if folks haven't listened I I really recommend it uh the special projects chair was interviewing a couple of Graphics Legends including David yam yeah and one of the things that David M talked about was archives and this is kind of an existential question for this crowd right but people are creating this really amazing digital media all these these computer Graphics you are accelerating it with your technology it changes so fast now how do you ensure that everything folks are building lives into the future file formats archives accessing all of this work in the future the robots will live on y I have no concern they're going to take over y right Y what about what about the art that people are creating this is this is the existential question but all this well one of the one of the one of the uh that's an excellent question and one of the one of the um the formats that we deeply believe in is open USD open USD is the first format that brings together multimodality from almost almost every single tool and allows it uh to interact to be composed together uh to uh go in and out of these Virtual Worlds and so you could bring in just about any format ideally over time into it uh at this conference we announced that URF the universal robotics data format is now compatible with or can be ingested into into um uh open USD and so one one format after another format uh we're going to bring uh everything into this one common one common language using standards is one of the best ways uh to uh allow content and data to uh be shared be allow everybody to collaborate on it and to lift forever for example HTML without HTML uh it would have been hard for all of these different content from around the world to be accessible to everybody and so in a lot of ways open USD is the HTML of Virtual Worlds and um uh we we've we've been a early promoter of it um uh there's uh amazing companies that have joined and many other companies joining and um my expectation is every single design tool in the world will be able to connect to open USD and once you connect to that virtual world uh you can collaborate with uh anybody with any other Tool uh anywhere and so just like we with HTML you said this content can live forever are you going to build a gensen AI that lives forever absolutely there's a gensen AI in fact just about everything that I've ever said everything that I've ever written and ever done uh will likely be ingested into one of these uh uh generative AI models and I'm I'm I'm hoping that that happens and then uh in the future you'll be able to prompt it and and hopefully something smart gets said so jent and the might to be uh running your earnings calls in the future I hope so that's that's the first thing that has to go that's the first thing it has to go to a bot Jensen thank you so much I think we're probably going to get kicked off stage soon but you'll be back shortly with Mark Zuckerberg and welcome to your first cig ladies and gentlemen Lauren good thank you thank you it's really great shatting again thank you everybody we'll be right back