Transcript for [NVIDIA CEO Jensen Huang Keynote at CES 2025](https://www.youtube.com/watch?v=k82RwXqZHY8) by [Merlin AI](https://merlin.foyer.work/)

0:06 - this is how intelligence is
0:10 - made a new kind of
0:14 - factory generator of
0:17 - tokens the building blocks of
0:21 - AI tokens have opened a new frontier the
0:24 - first step into an extraordinary world
0:27 - where endless possibilities are born
0:30 - [Music]
0:34 - tokens transform words into knowledge
0:37 - and breathe life into
0:41 - images they turn ideas into
0:46 - videos and help us safely navigate any
0:51 - environment tokens teach robots to move
0:54 - like the Masters
0:56 - [Music]
1:00 - Inspire new ways to celebrate our
1:02 - victories a martini pleas call light
1:06 - up thank you
1:09 - Adam and give us peace of mind when we
1:12 - need it most hi moroka hi Anna it's good
1:16 - to see you again hi Emma we're going to
1:19 - take your blood sample today okay don't
1:21 - worry I'm going to be here the whole
1:26 - time they bring meaning to numbers
1:30 - to help us better understand the world
1:32 - around
1:34 - [Music]
1:40 - us predict the dangers that surround
1:43 - [Music]
1:51 - us and find cures for the threats within
1:54 - us
1:56 - [Music]
2:01 - tokens can bring our Visions to
2:05 - [Music]
2:10 - life and restore what we've
2:12 - [Music]
2:12 - [Applause]
2:15 - lost
2:17 - Zachary I got my voice back
2:22 - buddy they help us move

2:26 - forward one small step at a time
2:29 - [Music]
2:35 - and one giant
2:38 - leap
2:39 - [Music]
2:53 - together and
2:56 - here is where it all begins
3:07 - welcome to the stage Nvidia founder and
3:09 - CEO Jensen
3:12 - [Music]
3:12 - [Applause]
3:13 - [Music]
3:14 - [Applause]
3:20 - Wong welcome to
3:24 - CES are you excited to be in Las
3:27 - Vegas do you like my Jack
3:32 - it I thought I'd go the other way from
3:34 - Gary
3:37 - Shapiro I'm in Las Vegas after all if
3:40 - does if this doesn't work out if all of
3:42 - you
3:44 - object well just get used to it I think
3:47 - I really think you have to let this sink
3:50 - in in another hour or so you're going to
3:53 - feel good about
3:56 - it well uh welcome to
4:01 - Nvidia in fact you're inside nvidia's
4:03 - digital
4:04 - twin and we're going to take you to
4:08 - Nvidia ladies and gentlemen welcome to
4:13 - Nvidia your
4:15 - inside our digital
4:20 - twin everything here is generated by
4:26 - AI it has been an extraordinary Journey
4:28 - extraordinary year here and uh it
4:32 - started in 1993 ready go with
4:38 - mv1 we wanted to build computers that
4:41 - can do things that normal computers
4:44 - couldn't and mv1 made it possible to
4:47 - have a game console in your
4:49 - PC our programming architecture was
4:52 - called
4:53 - UD missing the letter c until a little
4:56 - while later but UDA UniFi Unified device

5:00 - architecture and the first developer for
5:04 - UDA and the first application that ever
5:06 - worked on UDA was sega's Virtual
5:10 - Fighter six years later we invented in
5:15 - 1999 the programmable
5:18 - GPU and it
5:20 - started 20 years 20 plus years of
5:24 - incredible advance in this incredible
5:27 - processor called the GPU it made modern
5:31 - computer Graphics
5:33 - possible and now 30 years later sega's
5:37 - Virtual Fighter is completely
5:42 - cinematic this is the new Virtual
5:44 - Fighter project that's coming I just
5:46 - can't wait absolutely
5:49 - incredible six years after that six year
5:52 - six years after
5:53 - 1999 we invented Cuda so that we could
5:58 - explain or or expressed the
6:01 - programmability of our gpus to a rich
6:04 - set of algorithms that could benefit
6:05 - from it Cuda
6:08 - initially was difficult to explain and
6:11 - it took years in fact it took
6:13 - approximately six years somehow six
6:17 - years later six years later or
6:22 - so
6:25 - 2012 Alex kvki ilas sus and Jeff Hinton
6:30 - discovered Cuda used it to process
6:34 - alexnet and the rest of it is history AI
6:38 - has been advancing at an incredible Pace
6:41 - since started with perception AI we now
6:45 - can understand images and words and
6:47 - sounds to generative AI we can generate
6:51 - images and text and
6:52 - sounds and now agentic ai AIS that can
6:58 - perceive reason plan and act and then
7:02 - the next phase some of which we'll talk
7:04 - about tonight physical AI 2012 now
7:10 - magically
7:12 - 2018 something happened that was pretty
7:15 - incredible Google's Transformer was
7:19 - released as Bert and the world of AI
7:24 - really took off Transformers as you know

7:28 - completely changed the land landcape for
7:30 - artificial intelligence in fact it
7:32 - completely changed the landscape for
7:34 - computing
7:35 - altogether we recognized properly that
7:38 - AI was not just a new application with a
7:42 - new business opportunity but AI more
7:46 - importantly machine learning enabled by
7:49 - Transformers was going to fundamentally
7:51 - change how Computing works and
7:56 - today Computing is revolutionized in
8:00 - every single layer from hand coding
8:04 - instructions that run on CPUs to create
8:07 - software tools that humans use we now
8:09 - have machine learning that creates and
8:12 - optimizes new networks that processes on
8:16 - gpus and creates artificial
8:19 - intelligence every single layer of the
8:21 - technology stack has been completely
8:24 - changed an incredible transformation in
8:28 - just 12 years
8:30 - well we can Now understand information
8:33 - of just about any modality surely you've
8:37 - seen text and images and sounds and
8:39 - things like that but not only can we
8:42 - understand those we can understand amino
8:44 - acids we can understand physics we
8:47 - understand them we can translate them
8:49 - and generate them the applications are
8:52 - just completely endless in fact almost
8:55 - any AI application that you you see out
8:57 - there what modality is the input that it
9:00 - learned from what modality of
9:02 - information did it translate to and what
9:05 - modality of information is it generating
9:07 - if you ask these three fundamental
9:09 - questions just about every single
9:11 - application could be inferred and so
9:14 - when you see application after
9:16 - applications that are Aid driven AI
9:19 - native at the core of it this
9:22 - fundamental concept is there machine
9:24 - learning has changed how every
9:26 - application is going to be built how

9:28 - computing will be done and the
9:31 - possibilities Beyond
9:33 - well
9:35 - gpus gForce in a lot of
9:39 - ways all of this with AI is the house
9:42 - that GeForce built GeForce enabled AI to
9:47 - reach the masses and now ai is coming
9:51 - home to
9:52 - GeForce there are so many things that
9:54 - you can't do without AI let me show you
9:58 - some of it
9:59 - now
10:08 - [Music]
11:06 - [Applause]
11:08 - [Music]
11:14 - [Applause]
11:18 - [Music]
11:34 - that was realtime computer
11:44 - Graphics no computer Graphics researcher
11:47 - no computer scientist would have told
11:50 - you that it is possible for us to rate
11:52 - trce every single Pixel at this point we
11:56 - Ray tracing is a simulation of light the
11:59 - amount of geometry that you saw was
12:01 - absolutely insane it would have been
12:03 - impossible without artificial
12:05 - intelligence there are two fundamental
12:07 - things that we did we used of course
12:10 - programmable shading and Ray traced
12:13 - acceleration to produce incredibly
12:15 - beautiful pixels but then we have
12:18 - artificial
12:19 - intelligence be
12:21 - conditioned be controlled by that pixel
12:24 - to generate a whole bunch of other
12:26 - pixels not only is it able to generate
12:29 - pixels spatially because it's aware of
12:32 - what the colors should be it has been
12:35 - trained on a supercomputer back in
12:37 - Nvidia and so the neuron Network that's
12:39 - running on the GPU can infer and predict
12:43 - the pixels that we did not render not
12:46 - only can can we do that it's called
12:49 - dlss the latest generation of dlss also

12:53 - generates Beyond frames it can predict
12:55 - the future generating three additional
12:58 - frames for every frame that we calculate
13:01 - what you saw if we just said four frames
13:04 - of what you saw because we're going to
13:06 - render one frame and generate three if I
13:09 - said four frames at full HD 4K that's 33
13:13 - million pixels or so out of that 33
13:17 - million
13:18 - pixels we computed only
13:23 - two it is an absolute miracle that we
13:27 - can computationally comput tionally
13:29 - using programmable shaders and our R
13:31 - traced engine R tracing engine to
13:33 - compute 2 million pixels and have ai
13:36 - predict all of the other 33 and as a
13:40 - result we're able to render at
13:43 - incredibly high performance because AI
13:46 - does a lot less computation it takes of
13:49 - course an enormous amount of training to
13:51 - produce that but once you train it the
13:54 - generation is extremely efficient so
13:57 - this is one of the incredible cap
13:59 - abilities of artificial intelligence and
14:01 - that's why there's so many amazing
14:03 - things that are happening we used gForce
14:06 - to enable artificial intelligence and
14:08 - now artificial intelligence is
14:10 - revolutionizing
14:11 - GeForce everyone today we're announcing
14:15 - our next
14:16 - Generation the RTX Blackwell family
14:20 - let's take a look
14:29 - [Music]
14:45 - is
15:05 - [Music]
15:12 - [Music]
15:19 - here it
15:20 - is our brand new
15:23 - gForce
15:24 - RTX 50 Series Blackwell architect
15:30 - the GPU is just a beast 92 billion
15:34 - transistors
15:36 - 4,000 tops four pedop flops of AI three

15:41 - times higher than the last generation
15:43 - Ada and we need all of it to generate
15:46 - those pixels that I showed you 380 Ray
15:50 - tracing Tera flops so that we could for
15:53 - the pixels that we have to compute
15:54 - compute the most beautiful image you
15:56 - possibly can and of course 125 Shader
16:00 - teraflops there is actually a concurrent
16:03 - Shader teraflops as well as an Inger
16:05 - unit of equal performance so two dual
16:09 - shaders one is for floating point one is
16:11 - for integer G7 memory from Micron 1.8
16:16 - terabytes Per Second Twice the
16:18 - performance of our last generation and
16:20 - we now have the ability to intermix AI
16:24 - workloads with computer graphics
16:26 - workloads and one of the amazing things
16:28 - about this gener eration is the
16:30 - programmable Shader is also able to now
16:34 - process neuron networks so the Shader is
16:37 - able to carry these neuron networks and
16:39 - as a result we invented neurot texture
16:42 - compression and neurom material shading
16:45 - as a result of that you get these
16:47 - amazingly beautiful images that are only
16:50 - possible because we use AIS to learn the
16:53 - texture learn a compression algorithm
16:56 - and as a result get extraordinary
16:57 - results okay so this is this is uh the
17:01 - brand
17:02 - new
17:05 - RTX Blackwell
17:10 - 9
17:11 - now even even the even the mechanical
17:15 - design is a miracle look at this it's
17:17 - got two
17:18 - fans this whole graphics card is just
17:21 - one giant fan you know so the question
17:24 - is where's the graphics card is it
17:25 - literally this
17:27 - big the voltage regul to design is
17:30 - state-of-the-art incredible design the
17:33 - engineering team did a great job so here
17:35 - it is thank

17:42 - you okay so those are the speeds and
17:44 - fees so how does it
17:46 - compare
17:48 - well this is RTX
17:53 - 490 I know I know many of you have
17:57 - one I I know it look it's
18:01 - $1,599 it is one of the best investments
18:04 - you could possibly
18:06 - make you for
18:08 - $15.99 you bring it home to your
18:12 - $10,000 PC
18:15 - entertainment Command Center isn't that
18:18 - right don't tell me that's not true
18:21 - don't be
18:23 - ashamed it's liquid
18:25 - cooled fancy lights all over it
18:29 - you lock it when you
18:33 - leave it's it's the modern home theater
18:36 - it makes perfect sense and now for
18:38 - $1,500 and99
18:40 - $15.99 you get to upgrade that and
18:42 - turbocharged the living Daya lights out
18:44 - of it well now with the Blackwell family
18:46 - RTX 570 490 performance at 549
18:53 - [Applause]
19:01 - impossible without artificial
19:03 - intelligence impossible without the Four
19:07 - Tops four ter Ops of AI tensor cores
19:12 - impossible without the G7 memories okay
19:14 - so 5070 490 performance $549 and here's
19:19 - the whole family starting from 5070 all
19:22 - the way up to 5090 5090 twice the
19:25 - performance of a 4090
19:30 - starting of course we're producing at
19:33 - very large scale availability starting
19:35 - January well it is incredible but we
19:39 - managed to put these in in gigantic
19:43 - performance gpus into a laptop this is a
19:47 - 570 laptop for
19:51 - $12.99 this 570 laptop has a 4090
19:55 - performance I think there's one here
19:57 - somewhere
20:00 - let me show you
20:02 - this this is a look at this thing here

20:06 - let me
20:07 - here there's only so many
20:10 - pockets ladies and gentlemen Janine
20:14 - [Applause]
20:17 - Paul so can you imagine you get this
20:19 - incredible graphics card here Blackwell
20:21 - we're going to shrink it and put it in
20:23 - put it in there does that make any
20:26 - sense well you can't do that without
20:29 - artificial intelligence and the reason
20:30 - for that is because we're generating
20:32 - most of the pixels using pixels using
20:34 - our tensor cores so we retrace only the
20:37 - pixels we need and we generate using
20:40 - artificial intelligence all the other
20:41 - pixels we have as a result the amount of
20:44 - the Energy Efficiency is just off the
20:46 - charts the future of computer Graphics
20:49 - is neural rendering the fusion of
20:51 - artificial intelligence and computer
20:53 - graphics and what's really
20:57 - amazing is oh here we go thank
21:00 - you this is a surprisingly kinetic
21:04 - keynote and and uh what's really amazing
21:07 - is the family of gpus we're going to put
21:08 - in here and so the 1590 the 1590 will
21:13 - fit into a laptop a thin laptop that
21:15 - last laptop was 14 14.9 mm you got a
21:19 - 5080 5070 TI and
21:22 - 5070 okay so ladies and gentlemen the
21:26 - RTX Blackwell family
21:30 - [Applause]
21:37 - well GeForce uh brought AI to to the
21:41 - world democratized AI now ai has come
21:45 - back and revolutionized GeForce let's
21:48 - talk about artificial intelligence let's
21:51 - go to somewhere else at
21:57 - Nvidia this this is literally our office
21:59 - this is literally nvidia's
22:03 - headquarters okay so let's talk about
22:05 - let's talk about AI the
22:08 - industry is chasing and racing to scale
22:13 - artificial intelligence int artificial
22:16 - intelligence and the scaling law is a

22:20 - powerful model it's an empirical law
22:23 - that has been observed and demonstrated
22:25 - by researchers and Industry over several
22:28 - Generations ations and this the the
22:30 - scale the scaling law says that the more
22:34 - data you have the training data that you
22:37 - have the larger model that you have and
22:39 - the more compute that you apply to it
22:41 - therefore the more effective or the more
22:44 - capable your model will become and so
22:48 - the scaling law continues what's really
22:51 - amazing is that now we're moving towards
22:54 - of course and the internet is producing
22:56 - about twice twice the amount of data
22:59 - every single year as it did last year I
23:01 - think the in the next couple of years we
23:03 - produce uh Humanity will produce more
23:05 - data than all of humanity has ever
23:08 - produced uh since the beginning and so
23:10 - we're still producing a gigantic amount
23:13 - of data and it's becoming more
23:15 - multimodal video and images and sound
23:18 - all of that data could be used to train
23:21 - the fundamental knowledge the
23:23 - foundational knowledge of an AI but
23:26 - there are in fact two other scaling laws
23:30 - that has now emerged and it's somewhat
23:32 - intuitive the second scaling law is post
23:36 - trining scaling law posttraining scaling
23:39 - law uses Technologies techniques like
23:41 - reinforcement learning human feedback
23:44 - basically the AI produces and generates
23:47 - answers the hum based on a human query
23:51 - the human then of course gives a
23:53 - feedback um it's much more complicated
23:55 - than that but the reinforcement learning
23:56 - system uh with a fair number of very
23:59 - high quality prompts causes the AI to
24:03 - refine its skills it could find tune its
24:07 - skills for particular domains it could
24:09 - be better at solving math problems
24:11 - better at reasoning so on so forth and
24:13 - so it's essentially like having a mentor
24:17 - or having a coach give you feedback um

24:20 - after you're done going to school and so
24:22 - you you get test you get feedback you
24:24 - improve yourself we also have
24:26 - reinforcement learning AI feedback
24:29 - and we have synthetic data generation uh
24:32 - these techniques are rather uh uh Ain to
24:36 - if you will uh self-practice uh you know
24:40 - you know the answer to a particular
24:41 - problem and uh you continue to try it
24:44 - until you get it right and so an AI
24:46 - could be presented with a very
24:48 - complicated and difficult problem that
24:50 - has that is verifiable U functionally
24:53 - and has a has an answer that we
24:55 - understand maybe proving a theorem maybe
24:57 - solving a solving a uh geometry problem
25:00 - and so these problems uh would cause the
25:03 - AI to produce answers and using
25:05 - reinforcement learning uh it would learn
25:08 - how to improve itself that's called post
25:11 - training post training requires an
25:12 - enormous amount of computation but the
25:14 - end result produces incredible models we
25:18 - now have a third scaling law and this
25:21 - third scaling law has to do with uh
25:24 - what's called test time scaling test
25:26 - time scaling is basically when you're
25:28 - being used when you're using the AI uh
25:32 - the AI has the ability to now apply a
25:35 - different resource allocation instead of
25:37 - improving its parameters now it's
25:40 - focused on deciding how much computation
25:43 - to use to produce the answers uh it
25:46 - wants to
25:47 - produce reasoning is a way of thinking
25:50 - about this uh long thinking is a way to
25:52 - think about this instead of a direct
25:54 - inference or One-Shot answer you might
25:57 - reason about you might break down the
25:59 - problem into multiple steps you might uh
26:02 - generate multiple ideas and uh evaluate
26:05 - you know your AI system would evaluate
26:07 - which one of the ideas that you
26:08 - generated was the best one maybe it

26:11 - solves the problem step by step so on so
26:13 - forth and so now test time scaling has
26:16 - proven to be incredibly effective you're
26:19 - watching this sequence of technology and
26:22 - this all of these scaling laws emerge as
26:24 - we see incredible achievements from chat
26:28 - GPT to 01 to 03 and now Gemini Pro all
26:33 - of these systems are going through this
26:36 - journey step by step by step of
26:38 - pre-training to posttraining to test
26:41 - time scaling well the amount of
26:43 - computation that we need of course is
26:45 - incredible and we would like in fact we
26:48 - would like in fact that Society has the
26:51 - ability to scale the amount of
26:52 - computation to produce more and more
26:55 - novel and better intelligence
26:57 - intelligence of course is the most
26:59 - valuable asset that we have and it can
27:01 - be applied to solve a lot of very
27:02 - challenging problems and so scaling law
27:06 - it's driving enormous demand for NVIDIA
27:08 - Computing it's driving an enormous
27:10 - demand for this incredible chip we call
27:14 - Blackwell let's take a look at Blackwell
27:17 - well Blackwell is in full
27:21 - production it is incredible what it
27:24 - looks like so first of all there's some
27:27 - uh every every single cloud service
27:29 - provider now have systems up and running
27:31 - uh we have systems here from about 15 uh
27:35 - 15 15 U uh excuse me 15 computer makers
27:39 - it's being made uh about 200 different
27:42 - SKS 200 different configurations they're
27:45 - liquid cooled air cooled x86 Nvidia gray
27:48 - CPU versions mvlink 36 by 2 MV links 72
27:53 - by1 whole bunch of different types of
27:55 - systems so that we can accommodate just
27:58 - about every single data center in the
27:59 - world well this these systems are being
28:03 - currently manufactured in some 45
28:06 - factories it tells you how pervasive
28:08 - artificial intelligence is and how much
28:11 - the industry is jumping onto artificial

28:13 - intelligence in this new Computing
28:16 - model well the reason why we're driving
28:19 - it so hard is because we need a lot more
28:22 - computation and it's very clear it's
28:25 - very clear that that um
28:37 - Janine you know
28:40 - I it's hard to tell you don't ever want
28:43 - to reach your hands into a dark
28:47 - place hang a second is this a good
28:50 - idea all right
28:56 - [Applause]
28:58 - [Music]
29:08 - wait for
29:11 - it wait for
29:16 - it I thought I was
29:23 - worthy apparently yor didn't think I was
29:26 - worthy all right
29:29 - this is my show and tell this is a show
29:31 - and tell so uh this mvlink system this
29:36 - right here this mvlink system this is
29:39 - gb200 MV link 72 it is 1 and 12
29:44 - tons 600,000
29:47 - Parts approximately equal to 20
29:51 - cars 12 12 120 kilow
29:59 - it has um a spine behind it that
30:02 - connects all of these GPU
30:04 - together two miles of copper
30:08 - cable 5,000
30:11 - cables this is being manufactured in 45
30:14 - factories around the world we build them
30:18 - we liquid cool them we test them we
30:20 - disassemble them shiping parts to the
30:24 - data centers because it's 1 and A2 tons
30:27 - we reassemble it outside the data
30:29 - centers and install them the
30:30 - manufacturing is insane but the goal of
30:33 - all of this is because the scaling laws
30:35 - are driving Computing so hard that this
30:38 - level of computation Blackwell over our
30:41 - last generation improves the performance
30:44 - per watt by a factor of four performance
30:47 - per watt by a factor of four perform
30:50 - performance per dollar by a factor of
30:52 - three that's basically says that in one

30:55 - generation we reduce the
30:58 - cost of training these models by a
31:00 - factor of three or if you want to
31:02 - increase um the size of your model by a
31:04 - factor of three it's about the same cost
31:06 - but the important thing is this these
31:09 - are generating tokens that are being
31:11 - used by all of us when we use Chad GPT
31:14 - or when we use Gemini use our phones in
31:16 - the future just about all of these
31:18 - applications are going to be consuming
31:19 - these AI tokens and these AI tokens are
31:22 - being generated by these
31:24 - systems and every single data center is
31:26 - limited by power
31:28 - and so if the perf per watt of Blackwell
31:31 - is four
31:33 - times our last
31:36 - generation then the revenue that could
31:38 - be generated the amount of business that
31:40 - can be generated in the data center is
31:41 - increased by a factor of four and so
31:43 - these AI Factory systems really are
31:46 - factories today now the goal of all of
31:48 - this is to so that we can create one
31:51 - giant chip the amount of computation we
31:54 - need is really quite incredible and this
31:56 - is basically one giant chip if we would
31:58 - have had to build a chip one here we go
32:02 - sorry
32:03 - guys you see that that's
32:06 - cool look at that disco lights in
32:11 - here right if we had to build this as
32:14 - one chip obviously this would be the
32:15 - size of the wafer but this doesn't
32:17 - include the impact of yield it would
32:19 - have to be probably three or four times
32:21 - the size but what we basically have here
32:23 - is 72 Blackwell gpus or 144 dieses this
32:28 - one chip here is 1.4 exop flops the
32:32 - world's largest supercomputer fastest
32:34 - supercomputer only recently this entire
32:37 - room supercomputer only recently
32:38 - achieved an exf flop plus this is 1.4

32:42 - exf flops of AI floating Point
32:44 - performance it has 14 terabytes of
32:47 - memory but here's the amazing thing the
32:49 - memory bandwidth is 1.2 petabytes per
32:52 - second that's basically basically the
32:56 - entire internet traffic that's happening
32:59 - right
33:01 - now the entire world's internet traffic
33:04 - is being processed across these chips
33:08 - okay and we have um 103 130 trillion
33:12 - transistors in total
33:15 - 2592 CPU
33:17 - cores whole bunch of networking and so
33:20 - these I wish I could do this I don't
33:22 - think I will so these are the black
33:25 - Wells these are our
33:29 - connectx networking chips these are the
33:32 - mvy link and we're trying to pretend
33:34 - about the Envy the the Envy Ling spine
33:37 - but that's not possible okay and these
33:40 - are all of the hbm memories 12 ter 14
33:44 - terabytes of hbm memory this is what
33:46 - we're trying to do and this is the
33:47 - miracle this is the miracle of the
33:49 - Blackwell system the blackwall dies
33:52 - right here it is the largest single chip
33:54 - the world's ever made but yet the
33:57 - miracle is really in addition to that
34:01 - this is uh the grace black wall system
34:03 - well the goal of all of this of course
34:05 - is so that we can thank you
34:10 - thanks boy is there a chair I could sit
34:12 - down for a
34:25 - second can I have a m AO
34:39 - Ultra how is it possible that we're in
34:42 - the mobe ultra
34:46 - Stadium it's like coming to Nvidia and
34:49 - we don't have a GPU for
34:54 - you so so we need an enormous the
34:57 - computation because we want to train
34:59 - larger and larger models and these
35:02 - inferences these inferences used to be
35:04 - one inference but in the future the AI
35:06 - is going to be talking to itself it's

35:08 - going to be thinking it's going to be
35:10 - internally reflecting processing so
35:12 - today when the tokens are being
35:14 - generated at you so long as it's coming
35:17 - out at 20 or 30 tokens per second it's
35:20 - basically as fast as anybody can read
35:22 - however in the future and right now with
35:25 - uh gp1 you know with the new the pre
35:29 - Gemini Pro and the new GP the the 0103
35:32 - models they're talking to themselves we
35:34 - reflecting they thinking and so as you
35:37 - can imagine the rate at which the tokens
35:40 - could be ingested is incredibly high and
35:43 - so we need the token rates the token
35:44 - generation rates to go way up and we
35:47 - also have to drive the cost way down
35:49 - simultaneously so that the C the quality
35:52 - of service can be extraordinary the cost
35:54 - to customers can continue to be low and
35:57 - uh will continue to scale and so that's
35:59 - the fundamental purpose the reason why
36:01 - we created MV link well one of the most
36:04 - important things that's happening in the
36:05 - world of Enterprise is a Genentech AI a
36:08 - Genentech AI basically is a perfect
36:10 - example of test time scaling it's a AI
36:13 - is a system of models some of it is
36:16 - understanding interacting with the
36:18 - customer interacting with the user some
36:20 - of it is maybe retrieving information
36:22 - retrieving information from Storage a
36:24 - semantic AI system like a rag uh maybe
36:28 - it's going on to to the internet uh
36:30 - maybe it's uh studying a PDF file and so
36:33 - it might be using tools it might be
36:34 - using a calculator and it might be using
36:36 - a generative AI to uh generate uh charts
36:39 - and such and it's iter it's taking the
36:42 - the problem you gave it breaking it down
36:44 - step by step and it's iterating through
36:45 - all these different models well in order
36:48 - to respond to a customer in the future
36:50 - in order for AI to respond it used to be
36:52 - ask a question answer start spewing out

36:55 - in the future you ask a question a whole
36:57 - bunch bu of models are going to be
36:58 - working in the background and so test
37:01 - time scaling the amount of computation
37:03 - used for inferencing is going to go
37:06 - through the roof it's going to go
37:07 - through the roof because we want better
37:09 - and better answers well to help the the
37:12 - industry build agentic AI our our go to
37:15 - market is not direct to Enterprise
37:16 - customers our go to market is is we work
37:19 - with software developers in the it
37:21 - ecosystem to integrate our technology to
37:24 - make possible new capabilities just like
37:27 - we did did with Cuda libraries we now
37:29 - want to do that with AI libraries and
37:33 - just as the Computing model of the past
37:36 - has apis that are uh doing computer
37:38 - Graphics or doing linear algebra or
37:41 - doing fluid dynamics in the future on
37:43 - top of those acceleration libraries C
37:46 - acceleration libraries will have ai
37:49 - libraries we've created three things for
37:52 - helping the ecosystem build agentic AI
37:54 - Nvidia Nims which are essentially AI
37:58 - microservices all packaged up it takes
38:00 - all of this really complicated Cuda
38:02 - software Cuda
38:04 - DNN cutless or tensor rtlm or Triton or
38:09 - all of these different really
38:11 - complicated software and the model
38:13 - itself we package it up we optimize it
38:15 - we put it into a container and you could
38:17 - take it wherever you like and so we have
38:20 - models for vision for understanding
38:21 - languages for speech for animation for
38:24 - digital biology and we have some new new
38:28 - exciting models coming for physical Ai
38:30 - and these AI models run in every single
38:33 - Cloud because nvidia's gpus are now
38:35 - available in every single Cloud it's
38:36 - available in every single OEM so you
38:38 - could literally take these models
38:40 - integrate it into your software packages

38:42 - create AI agents that run on Cadence or
38:46 - they might be S uh service now agents or
38:49 - they might be sap agents and they could
38:52 - deploy it to their customers and run it
38:54 - wherever the customers want to run the
38:55 - software the next layer is what we call
38:57 - Nvidia Nemo Nemo is
39:02 - essentially a digital employee
39:06 - onboarding and training evaluation
39:09 - system in the future these AI agents are
39:13 - essentially digital Workforce that are
39:16 - working alongside your employees um
39:18 - working AI doing things for you on your
39:20 - behalf and so the way that you would
39:23 - bring these specialized agents into your
39:26 - these special agents into your company
39:28 - is to onboard them just like you onboard
39:31 - an employee and so we have different
39:33 - libraries that helps uh these AI agents
39:36 - be uh trained for the type of you know
39:39 - language in your company maybe the
39:41 - vocabulary is unique to your company the
39:43 - business process is different the way
39:45 - you work is different so you would give
39:46 - them examples of what the work product
39:49 - should look like and they would try to
39:50 - generate and you would give a feedback
39:52 - and then you would evaluate them so on
39:54 - so forth and so that uh and you would
39:57 - guardrail them you say these are the
39:58 - things that you're not allowed to do
39:59 - these are things you're not allowed to
40:01 - say this and and we even give them
40:03 - access to certain information okay so
40:06 - that entire pipeline a digital employee
40:09 - pipeline is called Nemo in a lot of ways
40:13 - the IT department of every company is
40:16 - going to be the HR department of AI
40:18 - agents in the
40:19 - future today they manage and maintain a
40:23 - bunch of software from uh from the IT
40:25 - industry in the future they will Main
40:27 - maintain you know nurture onboard and
40:31 - improve a whole bunch of digital agents

40:33 - and provision them to the companies to
40:34 - use okay and so your H your it
40:37 - department is going to become kind of
40:39 - like AI agent HR and on top of that we
40:42 - provide a whole bunch of blueprints that
40:45 - our ecosystem could could uh take
40:47 - advantage of all of this is completely
40:49 - open source and so you could take take
40:51 - it and uh modify the blueprints we have
40:53 - blueprints for all kinds of different
40:55 - different types of Agents well today
40:57 - we're also announcing that we're doing
40:58 - something that's really cool and I think
41:00 - really clever we're announcing a whole
41:03 - family of models that are based off of
41:06 - llama the Nvidia llama neotron language
41:10 - Foundation models llama 3.1 is a
41:14 - complete
41:16 - phenomenon the download of llama 3.1
41:19 - from meta 350 650,000 times something
41:23 - like that it has
41:25 - been der red and turned into other
41:29 - models uh about 60,000 other different
41:32 - models it it is singularly the reason
41:35 - why just about every single Enterprise
41:36 - and every single industry has been
41:38 - activated to start working on AI well
41:40 - the thing that we did was we realized
41:42 - that the Llama models really could be
41:45 - better fine-tuned for Enterprise use and
41:48 - so we fine-tune them using our expertise
41:50 - and our capabilities and we turn them
41:52 - into the Llama neotron Suite of open
41:56 - models there are small ones that
41:59 - interact in uh very very fast response
42:02 - time extremely small uh they're uh sup
42:05 - what we call Super llama neotron supers
42:08 - they're basically your mainstream
42:10 - versions of your models or your Ultra
42:13 - model the ultra model could be used uh
42:15 - to be a teacher model for a whole bunch
42:17 - of other models it could be a reward
42:20 - model evaluator uh a judge for other
42:23 - models to create answers and decide

42:25 - whether it's a good answer or not
42:27 - give basically give feedback to other
42:29 - models it could be distilled in a lot of
42:31 - different ways basically a teacher model
42:33 - a knowledge distillation uh uh model
42:36 - very large very capable and so all of
42:39 - this is now available online well these
42:43 - models are incredible it's a a number
42:46 - one in leaderboards for chat leaderboard
42:49 - for instruction uh lead leaderboard for
42:53 - retrieval um so the different types of
42:55 - functionalities necessary that are used
42:57 - in AI agents around the world uh these
43:00 - are going to be incredible models for
43:02 - you we're also working with uh the
43:04 - ecosystem these Tech all of our Nvidia
43:07 - AI Technologies are integrated into uh
43:10 - uh the it in Industry uh we have great
43:13 - partners and really great work being
43:14 - done at service now at sap at Seaman uh
43:18 - for industrial AI uh Cadence is during
43:21 - great work synopsis doing great work I'm
43:23 - really proud of the work that we do with
43:25 - perplexity as you know they
43:26 - revolutionize search yeah really
43:28 - fantastic stuff uh codium uh every every
43:32 - software engineer in the world this is
43:33 - going to be the next giant AI
43:36 - application next giant AI service period
43:41 - is software coding 30 million software
43:43 - Engineers around the world everybody is
43:46 - going to have a software assistant uh
43:48 - helping them code uh if if um if not
43:51 - obviously you're just you're going to be
43:53 - way less productive and create lesser
43:55 - good code and so this is 30 million
43:58 - there's a billion knowledge workers in
44:00 - the world it is very very clear AI
44:03 - agents is probably the next robotics
44:06 - industry and likely to be a
44:07 - multi-trillion dollar opportunity well
44:10 - let me show you some of the uh
44:12 - blueprints that we've created and some
44:14 - of the work that we've done with our

44:15 - partners uh with these AI
44:21 - agents AI agents are the new digital
44:25 - Workforce working for and with
44:28 - us AI agents are a system of models that
44:32 - reason about a mission break it down
44:34 - into tasks and retrieve data or use
44:37 - tools to generate a quality
44:40 - response nvidia's agentic AI building
44:43 - blocks Nim pre-trained models and Nemo
44:46 - framework let organizations easily
44:48 - develop AI agents and deploy them
44:51 - anywhere we will onboard and train our
44:54 - agentic workforces on our company's
44:56 - methods like we do for
44:58 - employees AI agents are domain specific
45:02 - task experts let me show you four
45:04 - examples for the billions of knowledge
45:07 - workers and students AI research
45:09 - assistant agents ingest complex
45:12 - documents like lectures journals
45:14 - Financial results and generate
45:16 - interactive podcasts for easy learning
45:19 - by combining a unet regression model
45:21 - with a diffusion model cordi can
45:24 - downscale global weather forecasts down
45:26 - from 25 km to 2
45:28 - km developers like at Nvidia manage
45:32 - software security AI agents that
45:34 - continuously scan software for
45:37 - vulnerabilities alerting developers to
45:39 - what action is
45:41 - needed Virtual Lab AI agents help
45:45 - researchers design and Screen billions
45:47 - of compounds to find promising drug
45:49 - candidates faster than
45:52 - ever Nvidia analytics AI agents built on
45:56 - an Nvidia metr blueprint including
45:58 - Nvidia Cosmos nimron Vision language
46:01 - models llama neaton llms and Nemo
46:05 - retriever Metropolis agents analyze
46:08 - content from the billions of cameras
46:11 - generating 100,000 pedes of video per
46:14 - day they enable interactive search
46:17 - summarization and automated

46:20 - reporting and help monitor traffic flows
46:23 - flagging congestion or danger
46:28 - in industrial facilities they monitor
46:31 - processes and generate recommendations
46:33 - or
46:34 - Improvement Metropolis agents centralize
46:38 - data from hundreds of cameras and can
46:40 - reroute workers or robots when incidents
46:43 - occur the age of agentic AI is here for
46:48 - every
46:52 - organization okay
46:57 - that was the first pitch at a baseball
47:00 - that was not generated I just felt that
47:03 - none of you were
47:05 - impressed okay so ai ai was was created
47:09 - in the cloud and for the cloud AI is
47:12 - creating the cloud for the cloud and for
47:15 - uh enjoying AI on on phones of course
47:18 - it's perfect um very very soon we're
47:21 - going to have a continuous AI that's
47:23 - going to be with you and when you use
47:25 - those metag glasses you could of course
47:27 - uh point at something look at something
47:29 - and and ask it you know whatever
47:31 - information you want and so AI is is
47:34 - perfect in the CL was creating the cloud
47:35 - is perfect in the cloud however we would
47:38 - love to be able to take that AI
47:40 - everywhere I've mentioned already that
47:41 - you could take Nvidia AI to any Cloud
47:44 - but you could also put it inside your
47:45 - company but the thing that we want to do
47:47 - more than anything is put it on our PC
47:49 - as well and so as you know Windows 95
47:53 - revolutionized the computer industry it
47:55 - made possible this new Suite of
47:57 - multimedia services and it change the
47:59 - way that applications was created
48:01 - forever um Windows 95 this this model of
48:05 - computing of course is not perfect for
48:08 - AI and so the thing that we would like
48:10 - to do is we would like to have in the
48:13 - future your AI basically become your AI
48:15 - assistant and instead of instead of just

48:18 - the the 3D apis and the sound apis and
48:21 - the video API you would have generative
48:23 - apis generative apis for 3D and
48:25 - generative apis for language and
48:27 - generative AI for sound and so on so
48:29 - forth and we need a system that makes
48:32 - that possible while leveraging the
48:35 - massive investment that's in the cloud
48:38 - there's no way that we could the world
48:40 - can create yet another way of
48:41 - programming AI models it's just not
48:44 - going to happen and so if we could
48:46 - figure out a way to make Windows
48:50 - PC a worldclass
48:52 - aipc um it would be completely awesome
48:55 - and it turns out the answer is Windows
48:58 - it's Windows wsl2 Windows wsl2 Windows
49:03 - wsl2 basically it's two operating
49:06 - systems within one it works perfectly
49:09 - it's developed for developers and it's
49:11 - developed uh uh so that you can have
49:13 - access to Bare Metal it's been wsl2 has
49:16 - been
49:17 - optimized optimized for cloud native
49:20 - applications it is optimized for and
49:23 - very importantly it's been optimized for
49:25 - Cuda and so wsl2 supports Cuda perfectly
49:29 - out of the box as a
49:31 - result everything that I showed you with
49:36 - Nvidia Nims Nvidia Nemo the blueprints
49:41 - that we develop that are going to be up
49:43 - in ai. nvidia.com so long as the
49:47 - computer fits it so long as you can fit
49:50 - that model and we're going to have many
49:51 - models that that fit whether it's Vision
49:54 - models or language models or speech
49:55 - models or these animation human digital
49:58 - human models all kinds of different
50:01 - different types of models are going to
50:02 - be perfect for your PC and it would you
50:06 - download it and it should just run and
50:08 - so our focus is to turn Windows wsl2
50:12 - Windows PC into a Target first class
50:16 - platform that we will support and

50:19 - maintain for as long as we shall live
50:21 - and so this is an incredible thing for
50:23 - engineers and developers everywhere let
50:25 - let me show you something that we can do
50:27 - with that this is one of the examples of
50:28 - a blueprint we just made for
50:31 - you generative AI synthesizes amazing
50:35 - images from Simple Text prompts yet
50:38 - image composition can be challenging to
50:40 - control using only words with Nvidia Nim
50:43 - microservices creators can use Simple 3D
50:46 - objects to guide AI image generation
50:49 - let's see how a concept artist can use
50:52 - this technology to develop the look of a
50:54 - scene they start by laying out 3D assets
50:58 - created by hand or generated with AI
51:01 - then use an image generation Nim such as
51:04 - flux to create a visual that adheres to
51:06 - the 3D
51:07 - scene add or move objects to refine the
51:13 - composition change camera angles to
51:15 - frame the perfect
51:17 - shot or reimagine the whole scene with a
51:20 - new
51:24 - prompt assisted by generative AI and
51:26 - Nvidia Nim and artists can quickly
51:29 - realize their
51:30 - [Music]
51:33 - Vision Nvidia AI for your
51:37 - PCS hundreds of millions of PCS in the
51:40 - world with Windows and so we could get
51:42 - them ready for AI uh oems all the PC
51:45 - oems we work with just basically all of
51:47 - the world's leading PC oems are going to
51:49 - get their PCS ready for this stack and
51:52 - so aips are coming to a home near you
52:02 - Linux is
52:08 - good okay let's talk about physical
52:12 - AI speaking of Linux let's talk about
52:14 - physical
52:16 - AI So Physical AI imagine
52:22 - imagine whereas your large language
52:25 - model you give it your your context your
52:30 - prompt on the left and it generates

52:34 - tokens one at a time to produce the
52:37 - output that's basically how it works the
52:40 - amazing thing is this model in the
52:42 - middle is quite large has billions of
52:45 - parameters the context length is
52:47 - incredibly large because you might
52:49 - decide to load in a PDF in my case I
52:51 - might load in several PDFs before I ask
52:54 - it a question those PDFs are turned into
52:57 - tokens the attention the basic attention
53:00 - characteristic of a transformer has
53:02 - every single token find its relationship
53:05 - and relevance against every other token
53:08 - so you could have hundreds of thousands
53:10 - of tokens and the computational load
53:14 - increases quadratically and it does this
53:17 - that all of the parameters all of the
53:19 - input sequence process it through every
53:21 - single layer of the Transformer and it
53:23 - produces one token that's the reason why
53:25 - we needed blackw
53:27 - and then the next token is produced when
53:30 - the current token is done it puts the
53:32 - current token into the input sequence
53:34 - and takes that whole thing and generates
53:36 - the next token it does it one at a time
53:39 - this is the Transformer model it's the
53:41 - reason why it is so so incredibly
53:44 - effective computationally demanding What
53:47 - If instead of PDFs it's your surrounding
53:51 - and what if instead of the prompt a
53:53 - question it's a request go over there
53:55 - and pick up that that you know that box
53:58 - and bring it back and instead of what is
54:00 - produced in tokens its text it produces
54:04 - action
54:05 - tokens well that I just described is a
54:09 - very sensible thing for the future of
54:11 - Robotics and the technology is right
54:13 - around the corner but what we need to do
54:16 - is we need to create the effective
54:18 - effectively the world
54:21 - model of you know as opposed to GPT
54:24 - which is a language model and this World

54:26 - model has to understand the language of
54:28 - the world it has to understand physical
54:31 - Dynamics things like gravity and
54:34 - friction and inertia it has to
54:36 - understand geometric and spatial
54:38 - relationships it has to understand cause
54:40 - and effect if you drop something a fall
54:42 - to the ground if you you know poke at it
54:44 - it tips over it has to understand object
54:48 - permanence if you roll a ball over the
54:50 - kitchen counter when it goes off the
54:52 - other side the ball didn't leave into
54:54 - another quantum universe that that's
54:56 - still there and so all of these types of
54:59 - understanding is intuitive understanding
55:01 - that we know that most models today have
55:04 - a very hard time with and so we would
55:06 - like to create a world we need a world
55:09 - Foundation model today we're announcing
55:11 - a very big thing we're announcing Nvidia
55:14 - Cosmos a world Foundation model that is
55:18 - designed that was created to understand
55:21 - the physical world and the only way for
55:23 - you to really understand this is to see
55:25 - it let's
55:29 - [Music]
55:32 - flip the next Frontier of AI is physical
55:36 - AI model performance is directly related
55:39 - to data availability but physical world
55:42 - data is costly to capture curate and
55:46 - label Nvidia Cosmos is a world
55:49 - Foundation model development platform to
55:51 - Advance Physical AI it includes Auto
55:55 - regressive world found Foundation models
55:57 - diffusion-based World Foundation models
56:00 - Advanced
56:01 - tokenizers and an Nvidia Cuda an AI
56:04 - accelerated data
56:07 - pipeline Cosmos models ingest text image
56:11 - or video prompts and generate virtual
56:13 - world States as
56:14 - videos Cosmos Generations prioritize the
56:17 - unique requirements of Av and Robotics
56:20 - use cases like real world environments

56:23 - lighting and object permanence
56:27 - developers use Nvidia Omniverse to build
56:29 - physics-based
56:31 - geospatially accurate scenarios then
56:34 - output Omniverse renders into Cosmos
56:36 - which generates photoreal physically
56:39 - based synthetic
56:40 - [Music]
56:51 - data whether diverse
56:54 - objects or environments
56:58 - conditions like weather or time of day
57:01 - or Edge case
57:04 - scenarios developers use Cosmos to
57:07 - generate worlds for reinforcement
57:09 - learning AI feedback to improve policy
57:13 - models or to test and validate model
57:17 - performance even across multisensor
57:21 - views Cosmos can generate tokens in real
57:24 - time bringing the power of foresight and
57:27 - Multiverse simulation to AI models
57:30 - generating every possible future to help
57:33 - the model select the right
57:36 - path working with the world's developer
57:38 - ecosystem Nvidia is helping Advance the
57:41 - next wave of physical
57:45 - [Music]
57:48 - AI Nvidia
57:51 - Cosmos Nvidia
57:54 - Cosmos Nvidia Cosmos the world's first
57:58 - world Foundation model it is trained on
58:02 - 20 million hours of video the 20 million
58:06 - hours of video focuses on physical
58:09 - Dynamic things so n n Dynamic nature
58:12 - nature themes themes uh humans uh
58:15 - walking uh hands moving uh manipulating
58:19 - things uh you know things that are uh
58:22 - fast camera movements it's really about
58:24 - teaching the AI not about generating
58:27 - creative content but teaching the AI to
58:30 - understand the physical world and from
58:32 - this with this physical AI there are
58:35 - many Downstream things that we could uh
58:38 - do as a result we could do synthetic
58:40 - data generation to train uh models we

58:43 - could distill it and turn it into
58:45 - effectively the seed the beginnings of a
58:47 - robotics model you could have it
58:49 - generate multiple physically based
58:53 - physically plausible uh scenarios that
58:56 - the future basically do a doctor strange
58:58 - um you could uh because because this
59:01 - model understands the physical world of
59:02 - course you saw a whole bunch of images
59:03 - generated this model understanding the
59:05 - physical world it also uh could do of
59:08 - course captioning and so it could take
59:11 - videos caption it incredibly well and
59:14 - that captioning and the video could be
59:17 - used to train large language models
59:21 - multimodality large language models and
59:24 - uh so you could use this technology to
59:26 - use this Foundation model to train
59:28 - robotics robots as well as larger
59:30 - language models and so this is the
59:32 - Nvidia Cosmos the platform has an auto
59:35 - regressive model for real-time
59:37 - applications has diffusion model for a
59:39 - very high quality image generation it's
59:42 - incredible tokenizer basically learning
59:44 - the vocabulary of uh real world and a
59:48 - data pipeline so that if you would like
59:49 - to take all of this and then train it on
59:52 - your own data this data pipeline because
59:54 - there's so much data involved we've
59:56 - accelerated everything end to endend for
59:58 - you and so this is the world's first
60:00 - data processing pipeline that's Cuda
60:02 - accelerated as well as AI accelerated
60:04 - all of this is part of the cosmos
60:06 - platform and today we're announcing that
60:09 - Cosmos is open licensed it's open
60:12 - available on
60:19 - GitHub we hope we hope that this moment
60:23 - and there's a there's a small medium
60:24 - large for uh uh very fast models um you
60:28 - know mainstream models and also teacher
60:30 - models basically not knowledge transfer
60:33 - models Cosmo Cosmos World Foundation

60:36 - model being open we really hope will do
60:39 - for the world of Robotics and Industrial
60:41 - AI what llama 3 has done for Enterprise
60:45 - AI the magic happens when you connect
60:49 - Cosmos to Omniverse and the reason
60:51 - fundamentally is this Omniverse is a
60:56 - physics grounded not physically grounded
60:59 - but physics grounded it's algorithmic
61:02 - physics principled physics simulation
61:05 - grounded system it's a simulator when
61:08 - you connect that to
61:10 - Cosmos it provides the grounding the
61:13 - ground truth that can control and to
61:16 - condition the Osmos generation as a
61:19 - result what comes out of Osmos is
61:21 - grounded on Truth this is exactly the
61:23 - same idea as connecting a large language
61:25 - model model to a rag to a retrieval
61:28 - augmented generation system you want to
61:30 - ground the AI generation on ground truth
61:34 - and so the combination of the two gives
61:36 - you a
61:38 - physically simulated a physically
61:41 - grounded Multiverse generator and the
61:45 - application the use cases are really
61:47 - quite exciting and of course uh for
61:50 - robotics uh for industrial applications
61:52 - uh it is very very clear this Cosmos
61:56 - plus
61:57 - o Omniverse plus Cosmos represents the
62:00 - Third computer that's necessary for
62:02 - building robotic systems every robotics
62:05 - company will ultimately have to build
62:07 - three computers a robotics the robotics
62:10 - system could be a factory the robotics
62:11 - system could be a car it could be a
62:13 - robot you need three fundamental
62:15 - computers one computer of course to
62:17 - train the AI we call the dgx computer to
62:21 - train the AI another of course when
62:24 - you're done to deploy the AI we call
62:26 - that agx that's inside the car in the
62:28 - robot or in an AMR or you know at the uh
62:32 - in a in a stadium or whatever it is

62:34 - these computers are at the edge and
62:37 - they're autonomous but to connect the
62:39 - two you need a digital twin and this is
62:42 - all the simulations that you were seeing
62:43 - the digital twin is where the AI that
62:46 - has been trained goes to practice to be
62:50 - refined to do its synthetic data
62:52 - generation reinforcement learning AI
62:54 - feedback such and such and so it's the
62:57 - digital twin of the AI these three
62:59 - computers are going to be working
63:01 - interactively nvidia's strategy for uh
63:04 - the industrial world and we've been
63:05 - talking about this for some time is this
63:07 - three computer
63:09 - system you know instead of a three three
63:12 - body problem we have a three Computer
63:14 - Solution and so it's the Nvidia
63:22 - robotics so let me give you three
63:25 - examples
63:26 - all right so the first example is uh uh
63:29 - how we apply apply all of this to
63:32 - Industrial digitalization there millions
63:36 - of factories hundreds of thousands of
63:38 - warehouses that's basically it's the
63:41 - backbone of A50 trillion doll
63:43 - manufacturing industry all of that has
63:46 - to become software defined all of that
63:48 - has has to have Automation in the future
63:51 - and all of it will be infused with
63:53 - robotics well we're partnering with Keon
63:56 - the world's leading Warehouse automation
64:00 - Solutions provider and Accenture the
64:03 - world's largest professional services
64:05 - provider and they have a big focus in
64:08 - digital manufacturing and we're working
64:10 - together to create something that's
64:12 - really special and I'll show you that in
64:14 - the second but our go to market is
64:16 - essentially the same as all of the other
64:18 - software uh platforms and all the
64:20 - technology platforms that we have
64:22 - through the uh developers and ecosystem
64:26 - Partners uh and we have just just a

64:29 - growing number of ecosystem Partners
64:31 - connecting to Omniverse and the reason
64:34 - for that is very clear everybody wants
64:36 - to digitalize the future of Industries
64:38 - there's so much waste so much
64:40 - opportunity for Automation in that $50
64:43 - trillion doar of the world's GDP so
64:45 - let's take a look at that this one one p
64:47 - one example that we're doing with Keon
64:49 - and
64:52 - Accenture Keon the supply chain solution
64:55 - company Accenture a global leader in
64:58 - Professional Services and Nvidia are
65:01 - bringing physical AI to the $1 trillion
65:05 - warehouse and Distribution Center Market
65:08 - managing high- Performance Warehouse
65:10 - Logistics involves navigating a complex
65:13 - web of decisions influenced by
65:15 - constantly shifting variables these
65:18 - include daily and seasonal demand
65:20 - changes space constraints Workforce
65:23 - availability and the integration of of
65:25 - diverse robotic and automated systems
65:28 - and predicting operational kpis of a
65:31 - physical Warehouse is nearly impossible
65:34 - today to tackle these challenges Keon is
65:38 - adopting Mega an Nvidia Omniverse
65:40 - blueprint for building industrial
65:42 - digital twins to test and optimize
65:45 - robotic fleets first Keon's warehouse
65:48 - management solution assigns tasks to the
65:51 - industrial AI brains in the digital twin
65:54 - such as moving a load from from a buffer
65:56 - location to a shuttle storage
65:58 - solution the robot's brains are in a
66:01 - simulation of a physical Warehouse
66:03 - digitalized into Omniverse using open
66:06 - USD connectors to aggregate CAD video
66:09 - and image to 3D Light Art to point cloud
66:13 - and AI generated data the fleet of
66:16 - robots execute tasks by perceiving and
66:20 - reasoning about their Omniverse digital
66:22 - twin environment planning their next
66:24 - motion and acting

66:26 - the robot brains can see the resulting
66:28 - State through sensor simulations and
66:30 - decide their next action the loop
66:33 - continues while Mega precisely tracks
66:36 - the state of everything in the digital
66:38 - twin now Keon can simulate infinite
66:42 - scenarios at scale while measuring
66:44 - operational kpis such as throughput
66:48 - efficiency and utilization all before
66:50 - deploying changes to the physical
66:53 - Warehouse together with Nvidia
66:56 - Keon and Accenture are Reinventing
66:58 - industrial
67:00 - autonomy in the future is that that's
67:03 - incredible everything is in
67:05 - simulation in the future in the future
67:09 - every Factory will have a digital twin
67:12 - and that digital twin operates exactly
67:14 - like the real factory and in fact you
67:17 - could use Omniverse with Cosmos to
67:20 - generate a whole bunch of future
67:22 - scenarios and you pick then an AI
67:24 - decides which which one of the scenarios
67:26 - are the most optimal for whatever kpis
67:28 - and that becomes the programming
67:30 - constraints the program if you will the
67:33 - AI that will be uh deployed into the
67:35 - real factories the next example
67:37 - autonomous vehicles the AV revolution
67:39 - has arrived after so many years with weo
67:43 - success and Tesla's success it is very
67:46 - very clear autonomous vehicles has
67:48 - finally arrived well our offering to
67:51 - this industry is the three computers the
67:54 - training systems the training the AIS
67:56 - the simulation systemss and and the and
67:58 - the synthetic data generation systems
68:00 - Omniverse and now Cosmos and also the
68:03 - computer that's inside the car each car
68:06 - company might might work with us in a
68:08 - different way use one or two or three of
68:10 - the computers we're working with just
68:12 - about every major car company around the
68:14 - world whmo and zuk and Tesla of course

68:17 - in their data center byd the largest uh
68:20 - EV company in the world jlr has got a
68:22 - really cool car coming Mercedes because
68:24 - a fleet of cars coming with Nvidia
68:26 - starting with this starting this year
68:27 - going to production and I'm super super
68:30 - pleased to announce that today Toyota
68:33 - and Nvidia are going to partner together
68:35 - to create their next Generation
68:43 - AVS just so many so many cool companies
68:47 - uh lucid and rivan and Shi and of course
68:50 - uh Volvo just so many different
68:52 - companies Wabi is uh building uh
68:54 - self-driving trucks Aurora we announced
68:57 - this week also that Aurora is going to
68:59 - use Nvidia to build self-driving trucks
69:02 - autonomous 100 million cars build each
69:05 - year a billion cars vehicles on a road
69:08 - all over the world a trillion miles that
69:10 - are driven around the world each year
69:13 - that's all going to be either highly
69:15 - autonomous or you know fully autonomous
69:18 - coming up and so this is going to be a
69:20 - very L very large industry I predict
69:22 - that this will likely be the first
69:24 - multi-trillion dollar
69:26 - robotics industry this IND this business
69:28 - for us um notice in just just a few of
69:33 - these cars that are starting to ramp
69:34 - into the world uh our business is
69:36 - already $4 billion and this year
69:39 - probably on a run rate of about $5
69:40 - billion so really significant business
69:42 - already this is going to be very large
69:44 - well today we're announcing that our
69:46 - next generation processor for the car
69:49 - our next generation computer for the car
69:51 - is called Thor I have one right here
69:53 - hang on a second
69:57 - okay this is
69:58 - Thor this is
70:01 - Thor this is this is a robotics
70:05 - computer this is a robotics computer
70:08 - takes sensors and just a Madness amount

70:11 - of sensor information process it you
70:15 - know een teed cameras high resolution
70:20 - Radars Liars they're all coming into
70:22 - this chip and this chip has to process
70:24 - all that sensor turn them into tokens
70:27 - put them into a Transformer and predict
70:30 - the next PATH and this AV computer is
70:34 - now in full production Thor is 20 times
70:38 - the processing capability of our last
70:40 - generation Orin which is really the
70:42 - standard of autonomous vehicles today
70:44 - and so this is just really quite quite
70:47 - incredible Thor is in full production
70:49 - this robotics processor by the way also
70:51 - goes into a full robot and so it could
70:53 - be an AMR it could be a human or robot
70:56 - could be the brain it could be the
70:58 - manipulator this Rob this processor
71:00 - basically is a universal robotics
71:04 - computer the second part of our drive
71:08 - system that I'm incredibly proud of is
71:10 - the dedication to safety Drive OS I'm
71:14 - pleased to announce is now the first
71:17 - softwar defined programmable AI computer
71:21 - that has been certified up to asold D
71:24 - which is the highest standard of
71:27 - functional safety for automobiles the
71:30 - only and the highest and so I'm really
71:33 - really proud of this asold ISO
71:36 - 26262 it is um the work of some 15,000
71:40 - engineering years this is just
71:43 - extraordinary work and as a result of
71:45 - that Cuda is now a functional safe
71:49 - computer and so if you're building a
71:51 - robot Nvidia Cuda y
71:58 - okay so so now I wanted to I told you I
72:00 - was going to show you what would we use
72:03 - Omniverse and Cosmos to do in the
72:06 - context of self-driving cars and you
72:09 - know today instead of showing you a
72:11 - whole bunch of uh uh videos of of cars
72:14 - driving on the road I'll show you some
72:16 - of that too um but I want to show you
72:19 - how we use the car to reconstruct

72:22 - digital twins automatically using Ai and
72:25 - use that capability to train future am
72:29 - models okay let's play
72:34 - it the autonomous vehicle Revolution is
72:37 - here building autonomous vehicles like
72:40 - all robots requires three computers
72:44 - Nvidia dgx to train AI models Omniverse
72:48 - to test drive and generate synthetic
72:50 - data and drive agx a supercomputer in
72:54 - the car
72:55 - building safe autonomous vehicles means
72:58 - addressing Edge scenarios but real world
73:01 - data is limited so synthetic data is
73:04 - essential for
73:06 - training the autonomous vehicle data
73:09 - Factory powered by Nvidia Omniverse AI
73:12 - models and Cosmos generates synthetic
73:15 - driving scenarios that enhance training
73:18 - data by orders of
73:20 - magnitude first omnimap fuses map and
73:24 - geospatial data to construct drivable 3D
73:31 - environments driving scenario variations
73:34 - can be generated from replay Drive logs
73:36 - or AI traffic
73:39 - generators next a neural reconstruction
73:42 - engine uses autonomous vehicle sensor
73:45 - logs to create High Fidelity 4D
73:48 - simulation
73:49 - environments it replays previous drives
73:52 - in 3D and generates scenario Vari ations
73:55 - to amplify training
73:57 - data finally edify 3DS automatically
74:01 - searches through existing asset
74:04 - libraries or generates new assets to
74:07 - create Sim ready
74:12 - scenes the Omniverse scenarios are used
74:15 - to condition Cosmos to generate massive
74:18 - amounts of photo realistic data reducing
74:20 - the Sim toore
74:22 - Gap and with text prompts generate near
74:26 - infinite variations of the driving
74:30 - scenario with Cosmos neotron video
74:33 - search the massively scaled synthetic
74:36 - data set combined with recorded drives

74:39 - can be curated to train
74:43 - models nvidia's AI data Factory scales
74:47 - hundreds of drives into billions of
74:49 - effective miles setting the standard for
74:52 - safe and advanced autonomous driving
74:55 - [Music]
74:59 - is that incredible
75:03 - we take take thousands of drives and
75:08 - turn them into billions of miles we are
75:11 - going to have mountains of training data
75:14 - for autonomous vehicles of course we
75:16 - still need actual cars on the road of
75:18 - course we will continuously collect data
75:21 - for as long as we shall live however
75:23 - synthetic data generation using this
75:26 - Multiverse physically based physically
75:29 - grounded capability so that we generate
75:32 - data for training AIS that are
75:34 - physically grounded and accurate and or
75:36 - plausible so that we could have an
75:38 - enormous amount of data to train with
75:40 - the AV industry is here uh this is an
75:43 - incredibly exciting time super super
75:45 - super uh uh excited about the next
75:47 - several years I think you're going to
75:48 - see just as computer Graphics was
75:51 - revolutionized such incredible pace
75:53 - you're going to see the pace of Av
75:55 - development increasing tremendously over
75:57 - the next several
76:08 - years I I think I think
76:13 - um I I think the next part is is
76:17 - robotics so um
76:26 - human
76:31 - robots my
76:36 - [Applause]
76:38 - friends the chat GPT moment for General
76:42 - robotics is just around the corner and
76:44 - in fact all of the enabling technologies
76:46 - that I've been talking about is going to
76:50 - make it possible for us in the next
76:52 - several years to see very rapid break
76:54 - breakthroughs surprising breakthroughs
76:56 - in in general robotics now the reason

76:58 - why General robotics is so important is
77:01 - whereas robots with tracks and wheels
77:03 - require special environments to
77:05 - accommodate them there are three
77:09 - robots three robots in the world that we
77:11 - can make that require no green
77:15 - fields Brown field adaptation is perfect
77:19 - if we if we could possibly build these
77:20 - amazing robots we could deploy them in
77:23 - exactly the world that we've built for
77:25 - ourselves these three robots are one
77:29 - agentic robots agentic AI because you
77:33 - know they're information workers so long
77:34 - as they could accommodate uh the
77:36 - computers that we have in our offices is
77:37 - going to be great number two
77:40 - self-driving cars and the reason for
77:42 - that is we spent 100 plus years building
77:44 - roads and cities and then number three
77:47 - human or robots if we have the
77:50 - technology to solve these three this
77:53 - will be the largest technology industry
77:54 - IND the world's ever seen and so we
77:58 - think that robotics era is just around
78:01 - the corner the critical capability is
78:04 - how to train these robots in the case of
78:07 - human or
78:08 - robots the imitation information is
78:12 - rather hard to collect and the reason
78:14 - for that is uh in the case of car you
78:16 - just drive it we're driving cars all the
78:17 - time in the case of these human robots
78:20 - the imitation information the the human
78:23 - demonstration is rather laborious is to
78:25 - do and so we need to come up with a
78:27 - clever way to take hundreds of
78:30 - demonstrations thousands of human
78:32 - demonstrations and somehow use
78:35 - artificial intelligence and
78:37 - Omniverse to synthetically
78:40 - generate
78:42 - millions
78:44 - of
78:46 - synthetically generated motions and from

78:49 - those motions the AI can learn uh how to
78:52 - perform a task let me show you how
78:54 - that's
79:05 - done developers around the world are
79:08 - building the next wave of physical AI
79:10 - embodied robots
79:13 - humanoids developing general purpose
79:15 - robot models requires massive amounts of
79:18 - real world data which is costly to
79:20 - capture and
79:21 - curate Nvidia Isaac Groot helps tackle
79:25 - these challenges providing humanoid
79:27 - robot developers with four things robot
79:30 - Foundation
79:31 - models data
79:33 - pipelines simulation
79:36 - Frameworks and a Thor robotics
79:40 - computer the Nvidia Isaac Groot
79:43 - blueprint for synthetic motion
79:44 - generation is a simulation workflow for
79:47 - imitation learning enabling developers
79:50 - to generate exponentially large data
79:52 - sets from a small number of
79:55 - demonstrations first Groot teleop
79:58 - enables skilled human workers to portal
80:01 - into a digital twin of their robot using
80:04 - the Apple Vision
80:05 - Pro this means operators can capture
80:08 - data even without a physical robot and
80:10 - they can operate the robot in a
80:12 - risk-free environment eliminating the
80:14 - chance of physical damage or wear and
80:18 - tear to teach a robot a single task
80:21 - operators capture motion trajectories
80:23 - through a handful of teleoperated
80:26 - demonstrations then use Groot mimic to
80:28 - multiply these trajectories into a much
80:31 - larger data
80:33 - set next they use Gro gen built on
80:37 - Omniverse and Cosmos for domain
80:39 - randomization and 3D to real
80:42 - upscaling generating an exponentially
80:45 - larger data
80:48 - set the Omniverse and Cosmos Multiverse

80:51 - simulation engine provides a massively
80:53 - scaled data set to train the robot
80:57 - policy once the policy is trained
81:00 - developers can perform software in the
81:02 - loop testing and validation in Isaac Sim
81:05 - before deploying to the real
81:08 - robot the age of General robotics is
81:11 - arriving powered by Nvidia Isaac
81:18 - Groot we're going to have mountains of
81:20 - data to train robots with
81:24 - Nvidia Isaac group Nvidia Isaac group
81:28 - this is our platform to provide
81:30 - technology platform technology elements
81:32 - to the robotics industry to accelerate
81:34 - the development of General
81:36 - Robotics and um well I have one more
81:39 - thing that I want to show you none of
81:41 - none of this none of this would be
81:43 - possible if not for uh this incredible
81:47 - project that we started uh about a
81:49 - decade ago inside the company what
81:51 - called project project digits deep
81:54 - learning GPU intelligence training
81:58 - system
82:00 - digits well before we launched it uh I
82:05 - shrunk it to
82:06 - dgx and to harmonize it with
82:09 - RTX agx ovx and all of the other X's
82:13 - that we have in the company and and um I
82:18 - and and it really revolutionized uh djx1
82:21 - really
82:22 - revolutionized where where's djx1
82:25 - dgx-1 revolutionized artificial
82:28 - intelligence the reason why we built it
82:30 - was because we wanted to uh make it
82:33 - possible for researchers and startups to
82:35 - have an out-of-the-box AI supercomputer
82:38 - imagine the way supercomputers were
82:39 - built in the past you really have to uh
82:42 - build your own facility and you have to
82:43 - go build your own infrastructure and
82:45 - really engineer it into existence and so
82:48 - we created a supercomputer for AI for AI
82:51 - development for researchers and and

82:52 - startups that comes literally one out of
82:54 - the box I delivered the first one to a
82:56 - startup company in 2016 called open Ai
83:00 - and Elon was there and and Ilia sus was
83:02 - there and many of Nvidia Engineers were
83:05 - there and and um uh we we celebrated the
83:07 - arrival of djx1 and obviously uh it
83:12 - revolutionized uh artificial
83:14 - intelligence and Computing um but now
83:16 - artificial intelligence is everywhere
83:18 - it's not just in researchers and and and
83:20 - startup Labs you know we want artificial
83:22 - intelligence as I mentioned in the
83:23 - beginning of our
83:25 - this is now the new way of doing
83:27 - Computing this is the new way of doing
83:28 - software every software engineer every
83:30 - engineer every creative artist everybody
83:33 - who uses computers today as a tool will
83:37 - need a AI
83:39 - supercomputer and so I just wished I
83:42 - just wish that djx1 was smaller and
83:49 - um you know so so um you know imagine
83:55 - ladies and gentlemen
84:04 - our this is nvidia's latest AI
84:12 - supercomputer and and it's finally
84:15 - called project digits right now and if
84:19 - you have a good name for it uh reach out
84:20 - to us um uh this here's the amazing
84:25 - thing this is an AI supercomputer it
84:27 - runs the entire Nvidia AI
84:30 - stack all of nvidia's software runs on
84:33 - this dgx Cloud runs on
84:36 - this this
84:38 - sits well somewhere and it's wireless or
84:41 - you know connect it to your computer
84:43 - it's even a workstation if you like it
84:44 - to be and you could access it you could
84:47 - you could reach it like a like a cloud
84:50 - supercomputer and nvidia's AI works on
84:53 - it and um it's based on a a super secret
84:56 - chip that we've been working on called
84:58 - GB 110 the smallest Grace Blackwell that
85:02 - we make and I have well you know what

85:05 - let's show let's show everybody insight
85:34 - isn't it just isn't just it's just so
85:37 - cute and this is the chip that's
85:40 - inside it is in it is in
85:43 - production this top secret chip uh we
85:46 - did in collaboration the CPU the gray
85:48 - CPU was a uh is built for NVIDIA in
85:52 - collaboration with mediatech
85:55 - uh they're the world's leading s so
85:56 - company and they worked with us to build
85:58 - this CPU this CPU s so and connect it
86:02 - with chipto chip mvy link to the
86:04 - Blackwell GPU and uh this little this
86:08 - little thing here is in full production
86:11 - uh we're expecting this computer to uh
86:14 - be available uh around May time frame
86:17 - and so it's coming at you uh it's just
86:19 - incredible what we could do and it's
86:22 - just I think it's you
86:26 - really I was trying to figure out do I
86:28 - need more hands or more
86:30 - pockets all right so so uh imagine this
86:33 - is what it looks
86:35 - like you know who doesn't want one of
86:38 - those and if you if you use
86:41 - PC Mac you know anything because because
86:46 - uh you know it's it's a cloud platform
86:48 - it's a cloud computing platform that
86:49 - sits on your desk you could also use it
86:51 - as a l Linux workstation if you like uh
86:54 - if you would like to have double
86:56 - digits this is what it looks like you
86:59 - know and you you connect it you connect
87:01 - it together uh uh with connectx and it
87:05 - has
87:06 - nickel GPU direct all of that out of the
87:10 - box it's like a supercomputer our entire
87:12 - supercomputing stack uh is available and
87:15 - so Nvidia Project digits
87:20 - [Applause]
87:28 - okay well let me let me let me tell you
87:31 - what I told you I told you that we are
87:33 - in production with three new Blackwells
87:38 - not only is the grace Blackwell

87:40 - supercomputers mvlink 72s in production
87:43 - all over the world we now have three new
87:46 - Blackwell systems in production one
87:49 - amazing AI foundational M World
87:53 - Foundation model the world's first
87:55 - physical AI Foundation model is open
87:58 - available to activate the world's
88:00 - industries of Robotics and such and
88:03 - three and three robotics three robots
88:07 - working on uh agentic AI uh human or
88:10 - robots and self-driving
88:12 - cars uh it's been an incredible year I
88:15 - want to thank all of you for your
88:16 - partnership uh thank all of you for
88:18 - coming I made you a short video to
88:20 - reflect on last year and look forward to
88:22 - the next year play please w
88:36 - [Music]
89:26 - [Applause]
89:30 - [Music]
89:51 - [Music]
89:58 - [Music]
90:10 - [Music]
90:15 - [Applause]
90:15 - [Music]
90:56 - [Music]
91:14 - have a great C us
91:17 - everybody happy New
91:19 - Year thank you