Nvidia 2024 AI Event: Everything Revealed in 16 Minutes - YouTube
https://www.youtube.com/watch?v=bMIRhOXAjYk

I hope you realize this is not a concert you have arrived at a developers conference there will be a lot of science described algorithms computer architecture mathematics Blackwell is not a chip Blackwell is the name of a platform uh people think we make gpus and and we do but gpus don't look the way they used to this is hopper Hopper changed the world this is Blackwell it's okay Hopper 28 billion transistors and so so you could see you I can see there there a small line between two dyes this is the first time two dieses have abutted like this together in such a way that the two CH the two dies think it's one chip there's 10 terabytes of data between it 10 terabytes per second so that these two these two sides of the Blackwell Chip have no clue which side they're on there's no memory locality issues no cach issues it's just one giant chip and it goes into two types of systems the first one is form fit function compatible to Hopper and so you slide a hopper and you push in Blackwell that's the reason why one of the challenges of ramping is going to be so efficient there are installations of Hoppers all over the world and they could be they could be you know the same infrastructure same design the power the electricity The Thermals the software identical push it right back and so this is a hopper version for the current hgx configuration and this is what the other the second Hopper looks like this now this is a prototype board this is a fully functioning board and I just be careful here this right here is I don't know10 billion the second one's five it gets cheaper after that so any customer in the audience it's okay the gray CPU has a super fast chipto chip link what's amazing is this computer is the first of its kind where this much computation first of all fits into this small of a place second it's memory coherent they feel like they're just one big happy family working on one application together we created a processor for the generative AI era and one of the most important important parts of it is content token generation we call it this format is fp4 the rate at which we're advancing Computing is insane and it's still not fast enough so we built another chip this chip is just an incredible chip we call it the mvy link switch it's 50 billion transistors it's almost the size of Hopper all by itself this switch ship has four MV links in it each 1.8 terabytes per second and and it has computation in it as I mentioned what is this chip for if we were to build such a chip we can have every single GPU talk to every other GPU at full speed at the same time you can build a system that looks like this now this system this system is kind of insane this is one dgx this is what a dgx looks like now just so you know there only a couple two three exop flops machines on the planet as we speak and so this is an exif flops AI system in one single rack I want to thank I want to thank some partners that that are joining us in this uh aw is gearing up for Blackwell they're uh they're going to build the first uh GPU with secure AI they're uh building out a 222 exif flops system we Cuda accelerating Sage maker AI we Cuda accelerating Bedrock AI uh Amazon robotics is working with us uh using Nvidia Omniverse and Isaac Sim AWS Health has Nvidia Health Integrated into it so AWS has has really leaned into accelerated Computing uh Google is gearing up for Blackwell gcp already has A1 100s h100s t4s l4s a whole Fleet of Nvidia Cuda gpus and they recently announced the Gemma model that runs across all of it uh we're work working to optimize uh and accelerate every aspect of gcp we're accelerating data proc which for data processing the data processing engine Jacks xlaa vertex Ai and mujo for robotics so we're working with uh Google and gcp across whole bunch of initiatives uh Oracle is gearing up for blackw Oracle is a great partner of ours for Nvidia dgx cloud and we're also working together to accelerate something that's really important to a lot of companies Oracle database Microsoft is accelerating and Microsoft is gearing up for Blackwell Microsoft Nvidia has a wide- ranging partnership we're accelerating could accelerating all kinds of services when you when you chat obviously and uh AI services that are in Microsoft Azure uh it's very very very likely nvidia's in the back uh doing the inference and the token generation uh we built they built the largest Nvidia infiniband super computer basically a digital twin of ours or a physical twin of ours we're bringing the Nvidia ecosystem to Azure Nvidia DJ's Cloud to Azure uh Nvidia Omniverse is now hosted in Azure Nvidia Healthcare is in Azure and all of it is deeply integrated and deeply connected with Microsoft fabric a NM it's a pre-trained model so it's pretty clever and it is packaged and optimized to run across nvidia's install base which is very very large what's inside it is incredible you have all these pre-trained stateof the open source models they could be open source they could be from one of our partners it could be created by us like Nvidia moment it is packaged up with all of its dependencies so Cuda the right version cdnn the right version tensor RT llm Distributing across the multiple gpus tried and inference server all completely packaged together it's optimized depending on whether you have a single GPU multi- GPU or multi- node of gpus it's optimized for that and it's connected up with apis that are simple to use these packages incredible bodies of software will be optimized and packaged and we'll put it on a website and you can download it you could take it with you you could run it in any Cloud you could run it in your own data Center you can run in workstations if it fit and all you have to do is come to ai. nvidia.com we call it Nvidia inference microservice but inside the company we all call it Nims we have a service called Nemo microservice that helps you curate the data preparing the data so that you could teach this on board this AI you fine-tune them and then you guardrail it you can even evaluate the answer evaluate its performance against um other other examples and so we are effectively an AI Foundry we will do for you and the industry on AI what tsmc does for us building chips and so we go to it with our go to tsmc with our big Ideas they manufacture and we take it with us and so exactly the same thing here AI Foundry and the three pillars are the NIMS Nemo microservice and dgx Cloud we're announcing that Nvidia AI Foundry is working with some of the world's great companies sap generates 87% of the world's global Commerce basically the world runs on sap we run on sap Nvidia and sap are building sap Jewel co-pilots uh using Nvidia Nemo and dgx Cloud uh service now they run 80 85% of the world's Fortune 500 companies run their people and customer service operations on service now and they're using Nvidia AI Foundry to build service now uh assist virtual assistance cohesity backs up the world's data their sitting on a gold mine of data hundreds of exobytes of data over 10,000 companies Nvidia AI Foundry is working with them helping them build their Gia generative AI agent snowflake is a company that stores the world's uh digital Warehouse in the cloud and serves over three billion queries a day for 10,000 Enterprise customers snowflake is working with Nvidia AI Foundry to build co-pilots with Nvidia Nemo and Nims net apppp nearly half of the files in the world are stored on Prem on net app Nvidia AI Foundry is helping them uh build chat Bots and co-pilots like those Vector databases and retrievers with enidan Nemo and Nims and we have a great partnership with Dell everybody who everybody who is building these chatbots and generative AI when you're ready to run it you're going to need an AI Factory and nobody is better at Building endtoend Systems of very large scale for the Enterprise than Dell and so anybody any company every company will need to build AI factories and it turns out that Michael is here he's happy to take your order we need a simulation engine that represents the world digitally for the robot so that the robot has a gym to go learn how to be a robot we call that virtual world Omniverse and the computer that runs Omniverse is called ovx and ovx the computer itself is hosted in the Azure Cloud the future of heavy Industries starts as a digital twin the AI agents helping robots workers and infrastructure navigate unpredictable events in complex industrial spaces will be built and evaluated first in sophisticated digital twins once you connect everything together it's insane how much productivity you can get and it's just really really wonderful all of a sudden everybody's operating on the same ground truth you don't have to exchange data and convert data make mistakes everybody is working on the same ground truth from the design Department to the art Department the architecture Department all the way to the engineering and even the marketing department today we're announcing that Omniverse Cloud streams to The Vision Pro and it is very very strange that you walk around virtual doors when I was getting out of that car and everybody does it it is really really quite amazing Vision Pro connected to Omniverse portals you into Omniverse and because all of these cat tools and all these different design tools are now integrated and connected to Omniverse you can have this type of workflow really incredible this is Nvidia Project Groot a general purpose Foundation model for humanoid robot learning the group model takes multimodal instructions and past interactions as input and produces the next action for the robot to execute we developed Isaac lab a robot learning application to train Gro on Omniverse Isaac Sim and we scale out with osmo a new compute orchestration service that coordinates workflows across djx systems for training and ovx systems for simulation the group model will enable a robot to learn from a handful of human demonstrations so it can help with everyday tasks and emulate human movement just by observing us all this incredible intelligence is powered by the new Jetson Thor robotics chips designed for Gro built for the future with Isaac lab osmo and Groot we're providing the building blocks for the next generation of AI powered [Applause] [Music] robotics about the same size the soul of Nvidia the intersection of computer Graphics physics artificial intelligence it all came to bear at this moment the name of that project general robotics 003 I know super good super good well I think we have some special guests do [Music] we hey guys so I understand you guys are powered by Jetson they're powered by Jetson little Jetson robotics computer inside they learn to walk in Isaac Sim ladies and gentlemen this this is orange and this is the famous green they are the bdx robots of Disney amazing Disney research come on you guys let's wrap up let's go five things where you going I sit right here Don't Be Afraid come here green hurry up what are you saying no it's not time to eat it's not time to eat I'll give I'll give you a snack in a moment let me finish up real quick come on green hurry up stop wasting time this is what we announce to you today this is Blackwell this is the plat amazing amazing processors MV link switches networking systems and the system design is a miracle this is Blackwell and this to me is what a GPU looks like in my mind