E25: NVIDIA's 7 Breakthrough AI Chips Change Everything

27 MIN · YOUTUBE · HTTPS://WWW.YOUTUBE.COM/WATCH?V=62IA7NBDNZM
https://www.youtube.com/watch?v=62IA7NbdNZM

SUMMARY

Nvidia's latest innovations, including the Vera Rubin systems and Groq 3 LPUs, promise significant advancements in AI performance, particularly in inference capabilities. The discussion highlights how these technologies enhance throughput, intelligence, and operational efficiency, paving the way for a new era of AI applications.

- *The Groq 3 LPUs provide a 35x improvement in throughput per megawatt and enhance the ability to handle trillion-parameter models.*
- *Vera Rubin systems are designed for extreme efficiency, featuring modular components and 100% liquid cooling, leading to faster maintenance and higher uptime.*
- *The integration of Groq chips with Nvidia's existing architecture allows for improved memory bandwidth and processing capabilities, crucial for AI workloads.*
- *The architecture supports both high throughput and low latency, making it suitable for diverse AI applications, including agentic systems that can perform tasks autonomously.*
- *The advancements in AI performance are expected to unlock new workloads and industries, moving towards more intelligent and capable AI agents.*
- *The focus on modular design and liquid cooling reflects a shift in data center infrastructure to support high-performance computing needs.*
- *Nvidia aims to empower developers and businesses to leverage AI more effectively, potentially transforming user expectations and experiences in various sectors.*

Today, you're joining me for something really special, a full breakdown of Nvidia's next-generation Vera Rubin systems, including the cutting-edge Groq 3 LPUs. I'm joined by Dion Harris, senior director of AI and HPC infrastructure at Nvidia and frequent guest of the channel. Also joining us is Stewart Pitts, who came from Groq and is now a senior manager of accelerated computing products at Nvidia. You're about to get an exclusive look at the mind-blowing hardware powering the biggest AI models on the planet for training, for fine-tuning, and for inference. I asked Dion and Stewart every question I could think of and they had some surprising things to say about where the AI market could be headed next. Your time is valuable, so let's get right into it. I'm super excited to talk to you both. Dion, it's great seeing you again.

>> Stewart, it's a pleasure to meet you. I'm going to immediately throw you into the fire and ask you all about the Groq chip that just got announced during the keynote. Can you walk me through at a high level even like what it is, what it does? Yeah, here you go. Absolutely. So, this week in the keynote with Jensen, we announced Nvidia Groq 3 LPX. This is a rack-scale inference accelerator for the Vera Rubin platform. It pairs with NVL72 for prefill Okay. and boosts fast decode together with Rubin GPUs. So, the combination is up to a 35x

improvement in throughput per megawatt while delivering significant intelligence gains, trillion parameter models up to a million million input tokens and also also improves speed. So, really with Nvidia now for inference, you're getting speed, intelligence, and extreme throughput all together. Yeah, so help me understand like where Groq fits into sort of the larger ecosystem because when I think of Nvidia's Rubin GPUs, we we know that they're great at training and they've already had high benchmarks for inference performance. So, is there a different kind of problem that the Groq chip uniquely solves or Yeah, certainly. I mean, you heard from Jensen in the keynote this week, you know, Nvidia is the inference king. The inference king. And Vera Rubin is the world's supercomputer. Like, let's let's not be confused. Rubin GPUs are critical. They deliver the most extreme throughput per watt, low cost per token in the industry. But, what we're seeing is really a new category, a new developer expectation for inference that is that is more intelligent, that is smarter, and also faster, right? So, this is a new class of token. We believe this is a new business opportunity up to a 10x revenue opportunity for token factories who are interested in providing premium, like ultra-premium inference as a subcategory of their overall inference offering.

Got it. And so, you have the base chip, and then you also have the LPX platform, right? That sits on top of that. So, can you walk me through like how one chip sort of scales up at the tray level? Yeah. Well, if only we had one right here, right? Let me show you. So, this is the Nvidia Groq 3 LPX compute tray. I mentioned that LPX is the rack-scale solution for the Groq LPU, the third-generation product. So, each tray has eight LPUs horizontally opposed, an FPGA, a host CPU, and a BlueField 4. Um, so when you deliver 32 trays at rack-scale, you get 256 LPUs. That's 40 TB per second of memory bandwidth, and up to 640 PB per second of scale-up bandwidth. So, this is an enormous memory boost to the Vera Rubin platform, and it pairs beautifully with Nvidia Dynamo, which is the software orchestration layer between Rubin GPUs and NBL-72 and the LPX rack. So, these are It's the ultimate peanut butter and jelly story. These two things go hand-in-hand, and it takes throughput, intelligence, and interactivity even further.

That's awesome. I'm going to put you a little bit on the spot here. I'd love You pointed to an FPGA. So, I'd be curious to know just like what that is, what it does, why it's in the tray. >> Yeah. Sure. So, the the FPGA basically sequences eight LPUs per tray and also interfaces with GPUs that live in an NVL72 rack. Okay. So, is that is it roughly an equivalent role to like the CPU in the compute tray or is it a pretty different role than what we see when we see the four GPUs and the two CPUs? Uh you know, each each serves a different job in the tray, but part of what makes LPUs unique is the ability for racks of LPUs, you know, fleets of LPUs, thousands of LPUs to operate simultaneously as a single superchip.

This is what in in tandem with the GRAC compiler allows for consistently low latency. So, this isn't like low latency as a party trick, right? Through the sequencing of hardware and software and together with Dynamo in tandem with NVL72, we're able to deliver very little latency variance. And this is critical, right? You have to have tokens that are smart and fast. If it's smart and unpredictably performant, that's not good for an agentyc system. It needs to be reliable reliable at scale. And I think just around the question you raised around the FPA, one of the core purposes it serves, similar to how we have our ML link switch chip, which allows those GPUs to talk to each other, the FPA's FPGA serves a similar function in allowing those those LPUs to talk to each other as well as go outside the rack. So, there's another connection point that gets it outside the rack as well. So, it's a very complementary architecture to some what we've done on the the Vera Rubin side as well. Yeah. And I I

actually think that's a perfect segue. So, I'd I'd love to talk about So, we saw this tray, right? I'd love to talk about the equivalent Vera Rubin tray. Sure. Yeah.

So, of course, we've talked before about Grace Hopper Blackwell and what that did to the entire rack scale to the entire data center, I should say. It took compute, networking, disaggregated, put it in a rack scale system, and it It just made you re-architect or rethink how you build a data center. Vera Rubin takes that a step further. So, we have the same sort of compute tray. If you recall, we did that for for Grace Blackwell. But, what you'll quickly see is a much more elegant simpler design.

So, there are no exposed cables or wires. Essentially, it's all connected through a PCB midplane within the trip within the tray itself. >> Okay. All of these components are all individually removable, which has a huge implication not just for manufacturing and building them, but also for field replaceable units for having less inventory on hand when you're servicing a lot of this equipment. So, as a point of reference to disassemble this tray in our Blackwell time frame, it took about 2 hours.

To do a single tray using Vera Rubin, it takes about 5 minutes. Woah. >> So, that type of efficiency shows up in a number of different facets, but it also helps deliver better reliability and resilience. So, as you're getting these systems stood up and as you have millions of endpoints in these gigabyte AI factories, it's going to deliver better performance, better resiliency, and better operational efficiency as well. And I mean, just at a high level, uh better uptime, right? Like, if it takes me 2 hours simply to disassemble it, then I still have to find the issue, troubleshoot the issue, reassemble it, things like that. You know, that costs servers money, right? Which costs the data centers money. So, what you're really saying is higher uptime just as a function of easy easier swap times, easier maintainability.

>> In fact, there's a metric in the data center. It's called goodput. Goodput. >> And what that means is what percentage of your time are you fully utilizing your computer? >> Yeah. And the reason why I looked at it that way, it's like you said, it's not just uptime and downtime, but it's how long did it take to repair, how long did it take to, you know, swap out components, and what percentage of your time is spent computing and delivering tokens. So, goodput is really a core focus, not just throughput or performance, but also goodput.

>> And I just want to make sure when we talk about good put, you know, we're talking about this tray, but we're also talking about this as well, right? Like everything we're all these metrics got it. Okay. I just want to make sure. Yeah. So, we used to see like two super chips. We used to see Can you walk me through like Is it still the same thing? What am I even It kind of looks like Darth Vader, right?

>> Exactly. It's It's a little cloak and dagger right now, but it's essentially the same core components. You have a Vera Rubin super chip here, which has two Rubin GPUs connected to a Vera CPU. And then so you have four of those in the full compute tray, right? And so now Vera itself is a brand new CPU, which is connected to your your Rubin GPUs, and again delivers two x more performance per watt compared to the Grace CPU. So, what's really interesting and exciting is when you look at the CPU and the importance and the role it plays, especially in a lot of the agnetic workflows, CPUs are becoming much more important because you have all this

infrastructure that's leveraging CPUs for tool calling, it's leveraging CPUs for reinforcement learning. And so to the extent that you can make this performant as possible, you remove bottlenecks. And so that's why when we redesigned Vera, we wanted to make sure that it could power the performance required because if you speed up the GPU by two x, you need to make sure the CPU doesn't become a further bottleneck. So, that was really a design principle that went into this system.

>> That's a really awesome insight actually. I would think that as inference and AI took over, CPUs would become less important. But hearing that makes me kind of rethink the whole pieces. So, sorry, keep going. That's I just didn't expect you to say that. >> No, that that's astute because like I said, as we look at designing these systems, we're thinking about the full system. We're thinking about the full architecture. And to the extent that you have a CPU in the workflow, it's going to it needs to be performant. So, that was really a focus of re-architecting Vera. But then you look at the other parts of the trays, you'll notice there's no air fans here on the front side.

In In we have our CX We actually have um eight CX-9s here compared to the previous gen. Um so you can basically see where the CX-9s were actually integrated into the super chip with Grace Blackwell. Now they're completely modular. So if that happens to fail, you don't have to replace this entire compute chip. You can basically hot swap these out, right? And so that's really the modular design that went into this entire compute tray. And if I remember right, this is the first compute tray that's 100% liquid cooled as opposed to 80% for the previous generation, right?

>> Correct. So in the previous generation, we had fans on the front end. And so if you were to look at this system, you would see air vents because you still need to allow air cooling to go in to cool off the the fans on on the front end, which you had like I said, you had your NICs, you had your BlueField's on the front. But now this is completely liquid cooled. You can see the hot the cold plates going on on top of these.

There's cold plates um as well here. And this is actually where the liquid cooling apparatus goes in and out. You have two layers of of liquid cooling for moving in the cold the cool 45° C uh cooling and evacuating out the warm temperature. >> water almost, right? Exactly. Exactly. Exactly. But that ability to use 45° coolant saves tremendous in terms of overall power and cooling and efficiency because you don't need to run your chillers nearly as much. And so that allows you to save more power, which you can then do more compute. So in that same power envelope, you're getting much more performance per watt.

>> I figured you just move all the data centers to the North Pole, right? That's the answer. >> Or space. Or space, yeah. So same same question real quick. Is this 100% liquid cooled? Yeah, it is. And so all of the innovation that Deion just spoke about from the NBL 72 rack and tray carries forward to the LPX rack and tray. And so this is this is extreme co-design in action. People have asked me today, "How did you go from signing the agreement with Groq in less than 90 days?" I haven't counted exactly, but a very short amount of time to delivering a rack scale LPX solution like this. All this innovation carries forward. So, this is a fully liquid cooled tray, standard MGX racks architecture. And this isn't only beneficial for NVIDIA, it's beneficial for token factories.

Makes it easier to operate, faster time to value, easier to maintain. That's awesome. The The reason I'm asking about liquid cooling is obviously if everything here is 100% liquid cooled, it pushes data centers to be liquid to support that liquid cooling, right? So, do you find that NVIDIA's really leading this transition into liquid cooling for the data centers? Can you like help me understand are they ready for this >> As we went from Hopper to Blackwell, it was a huge transformation for the data center. Because we were essentially saying to deliver the best performance and token per watt, liquid cooling is the way to go. And so, that shift happened largely in the last year or two.

>> Okay. And especially as you're building out new AI factories, it's now the presumed the presumed standard. That you're a liquid cooling That you're you're building liquid cooling based solutions. But, in the the event that they don't have a liquid cooling based infrastructure, we've worked with partners like Laidon, like Schneider Electric, like Vertiv, like Delta Electronics to build reference architectures that allow them to retrofit existing data centers. Or if you have a greenfield, you can build a reference design that help helps you understand from beginning end what that spec looks like. So, we really looked at this from end to end figuring what the customer needs to do in order to deploy these rack scale solutions. Got it.

There's a huge transition to liquid cooling. There's also a huge transition in networking, right? So, I'd love to talk about this third tray if you don't mind. >> Certainly. And And tell me all of the updates to the networking. Yeah, so like you like we talked about, um we looked at what's going to drive the best performance per watt. And so, to the extent that we can drive up compute density, that's what went into the design principle of this rack scale architecture. So, we have all the compute consolidated in one tray. We have all the scale up networking in its own set of trays. And so this is the NVLink switch tray, which essentially has four NVLink switch chips. They're all completely liquid cooled. And again, they're delivering roughly about 3.6 teraflops per second or terabytes, I should say, per second of all to all bandwidth. Each one has that. When you scale this up into an entire rack, 260 terabytes per second. And again, we talked about the speed of the internet.

It's actually about The speed of the internet has actually increased. And now we're still be able to move the entire speed of the internet in about a second on on this in full rack scale system. So incredible bandwidth. Yeah. Blows my mind. So okay. So this tray is a lot emptier than the previous ones I've seen, right? So what why leave any free space at all? What What's the The thing is, so in in the previous generation, you had this space here as well. So the switch tray itself is really about moving data between the GPUs. That that's the core function. And so the point of this, because you still have the same rack footprint, right? So we have the same MGX rack footprint, you could have made it less shallow or you could have made it more shallow or less deep, but But it still needs the You still need the space. So this was just as a result of the MGX form factor, but the performance that you that you get out of these liquid cool systems is incredible. And the density that you require here is exactly what's needed to power those 72 GPUs.

>> And is that why like in the Vera Rubin Ultra, a lot of this space goes away, right? Like it's a much denser like style of tray? In the Vera Rubin Ultra, when we go to the Kyber style solution, it's slightly different, right? So that's where you have I believe we may have one over there. So that's where you have essentially um your switches, but they plug into a backplane or midplane, I should say. So Jensen had them on stage. One of those

big They they look like a huge compute tray and it literally plugs in vertically into the system and that's what connects all of those GPUs in that fully dense 144 GPU connected system. Got it. Okay, let me see if I get my bearings here because now you guys are co-designing seven chips per generation, right? Okay, we've talked about the GPUs and the CPUs, right? The Rubin GPUs and the Vera CPUs.

We've talked about the Grock LPUs and a little bit about this FPJ which is not one of those seven, right? Um we've talked about the NVLink switch chips and the CX 9s. I'm missing one I believe. >> Bluefield 4. So we actually have Bluefield 4 in both this one and this guy as well. What does that do? Yeah, so Bluefield 4 is a data processing unit. And so just like we talked about, you need to get data to and from these processors. So it actually commands a lot of the north-south communication within the network. So when north-south meaning either data coming in from an external user or going out and and being retrieved from the data data plane in terms of storage. And so the reason why you want a separate processor is it creates not just performance benefits, but also security isolation because you can basically isolate anything coming in out of the trays or going to and from storage. And so that's one of the key elements of the DPU, the data protection is for performance isolation and security isolation. Got it. So it's really like the common input output from each of these trays. Got it. Perfect.

That makes a lot of sense. So obviously each chip is a big performance jump from the previous generation, right? Um I'd love to know how as we scale up to a full rack, what that means in terms of performance benefits at the rack level. Can you help walk me through Absolutely. Absolutely. So when we went from Blackwell, it was roughly about 1.4 exaflops of AI performance. And again, we're talking that's super supercomputer level, you know, scale of performance.

For Vera Rubin, it's about 3.6 exaflops of AI performance in a single rack. 3.6 okay. >> Yes. So so again, almost a 4x improvement, three to 4x improvement, but the power went up by roughly about 50%. So what's really incredible with that demonstrates the value of this co-design, we're delivering more performance, a disproportion amount of performance for the additional power. So the power efficiency actually went up. >> Exactly. >> Yeah. And so in like practical terms, right? What does AI that performance leap and B the addition of the Groq chips, what does that let us do that we couldn't do before? Like what kind of workloads, please?

>> Yeah. Yeah. Well, to build on that, you know, what Jensen talked about yesterday in keynote was when you pair an LPX rack scale accelerator with an NBL 72 Vera Rubin system, you achieve up to a 35x improvement in token throughput. So to bring that back to your questions, that's up to a 35x tokens per megawatt, right? So it means for the same unit of energy, you're getting 35 It's hard to even conceptualize. 35 times the token volume per system, right? But it's not just throughput, it's also intelligence, right? And speed. So with NVIDIA you get all three. That's the way to think about it. Um firstly, the second thing I'd build to what you were speaking about um is when you when you asked about the sort of incremental gains that an LPX rack brings to Vera Rubin system, the memory capacity increase when you pair a an LPX rack with an NBL 72 rack is nearly 40x improvement in memory bandwidth per rack. And you can think about that in terms of like a professional catering company, if you will. It's like adding a fleet of line chefs, you know, who are there to execute and deliver on the desired result consistently and very quickly. That's the type of value add that from a memory

bandwidth perspective LPX adds to NBL 72. Yeah, thanks. That's a really great example. That like really clarifies. So what I'm what I'm really hearing is like it's some combination of being able to serve people 30 to sorry, 35 to 40 times faster, serve 35 to 40 times more people with the same workload, or any combination of the two.

Okay. And I mean the way I'd also characterize it, when you think about an AI factory or a data center, you have a mix of workloads, right? And so there's a segmentation that every operator has to think about. In other words, you want to optimize for throughput Yeah. and driving down the cost per token, or do you want to optimize for latency or intelligence? And so that type of calculus makes inference extremely hard. And so what this solution offers is through the Vera Rubin platform, you can deliver high throughput, you can deliver low latency, you can actually bridge the gap and have low latency high throughput solutions, all of which are delivering state-of-the-art performance and and intelligence. And so it's really having this 1 + 1 = 3 story because essentially they have you different characteristics.

Like this guy has 288 GB of HBM4. Yeah. The SRAM has 500 MB. So for anything that's going to be dependent on the memory size of footprint like having lots of parameters, lots of MB KB cache is going to be great for this. Anything that's going to be very dependent when I say anything, we're looking at at model by model, layer by layer. Where can you leverage these different architectures to fully exploit the benefits while also leveraging the benefits of the other architectures. So really figuring out how can you tell a complimentary story.

Yeah. No, that that makes a lot of sense. I have one question that like I'm trying to untangle in my head here, right? So you talk about high throughput, you talk about like high speed, right? And then there's this third dimension that you guys keep mentioning which is in more intelligent tokens. Can you help me understand like what what does that axis even mean? Like what is a more intelligent token? So what's really important to think about is the way that we deliver intelligence is a couple of different ways.

One way is a larger model. Right? So, in other words, as you look at the scaling laws, you'll see that as as models have gotten more intelligent, they've grown in parameters sizes. And the other one is your context window. So, the ability to handling and process more contextual information to give you a better output. And so, you can scale on a couple of those dimensions in terms of model size and in terms of context length to produce a better output or intelligent level. Now, the one thing that you have to think about is as you're looking at And then the other thing is actually just speed. Being able to do things quicker allows you to iterate faster to return more intelligence within the same time window. So, you think speed, model size, and context.

When you look at these two architectures, they have complementary sort of values that this is going to be great at model size, right? So, this allows you to fit a lot of the KB cache that you have as a result of computing that input token. But, this is going to be incredibly great at generating responses fast, especially anything that's going to be bandwidth limited. And so, you're able to get the best of both worlds, so to speak, in terms of generating intelligent responses. Yeah, please.

>> And I would just build to say, you know, with everything Deion said, we're solving a critical problem. You talk to any developer, they'll tell you this, which is today if you're building an agentic system, agents are either fast and expensive, like unstainably unsustainably expensive at with the intelligence required, or slow and intelligent. So, agents need tokens that are smart and fast, you know, not not slow and intelligent and slow, right? And so, so that's what Deion is saying is with Nvidia, you get all three. I think I mentioned this, you're getting you're getting quantity of tokens. You know, agent multi-agent systems require up to 15 times more tokens than a singular purpose-built AI chatbot. So, they require large token volume, but they also need the intelligence, the context, and the speed to prove economic value, right? For agentic systems to make a dent on society, that that these are the critical ingredients we're delivering.

Yeah, that that really clarifies it. Go ahead, please. >> And I was going to say one thing, if you look at there's a number of other solutions out there that may be SRAM based or can deliver fast tokens, but without being able to put the full model in memory, without being able to process the full context, the output isn't nearly as good. In fact, they measured purely, you know, sort of latency-driven SRAM based solutions, they're about 50% in terms of accuracy, whereas traditional other models delivered, you know, with full capabilities can deliver up to 80% accuracy. So, what you start to realize is you don't want an agent that's fast and dumb. Yeah, right. You want an agent that's fast and can leverage the state-of-the-art intelligence to deliver that that response.

>> That that makes a lot of sense. Yeah, that that's very clear, and I think that's a perfect segue into my last question, which is um as you look ahead, right? As these systems aren't just getting faster and serving more people, but they're also getting more intelligent. Is there something you guys personally are specifically looking forward to seeing solved, like a new class of problem being solved, a new kind of workload, a new industry that's leveraging these solutions? Well, I'll take a stab and I say, you know, one of the things that we've been talking about here at at the show is sort of this new class of agentic workloads that's happening, and I think we have to be very clear of what we mean by agentic.

Okay. We've been integrating we've been sort of integrating AI and applications. We've been working with AI, you go to it, you ask it a question, it gives you a response. But, we're at the stage now where AI can go and do stuff for you, right? It can actually do real work. It can actually do real You give it a task, and it'll go and say, "How do I optimize that task? These are the different steps I need to take. These are the skills I want to build to go and deliver that task." And I can produce an output that is an actual value to the person that has it. So, I think as we move down this path of a Gentic, like you said, it's going to have a different um sort of requirement in terms of hardware, software, security, but the value and the opportunity to really bring about what people have talked about as AGI, but I'll call it as really AI that's delivering value and doing something.

And that's the whole idea around Clause. It's actually going out and getting information. It's going out and using tools. It's running scripts. It's actually doing stuff on the behalf of the user. So, that is really exciting, and so a lot of the work that we've talked about here at the show is going to unlock a lot of those capabilities. >> Yeah. Tell me, what what's exciting you the most? >> Yeah. So, what I'm personally excited about for myself is how much more productive I can be overnight while I'm sleeping.

Um that's my post-GTC project. It's how do I get How do I get Clause to work for me? Um but beyond myself, you know, the thing I get most excited about is the power of an individual developer. And sure, you know, I can be more productive with Clause working for me overnight, but the scale, smarts, and access that an individual developer what they can create with an Agentic system on this type of supercomputer, I can't even quite conceptualize it. I can't wait to see what people build with this stuff. I can't wait to see customers get this Vera Rubin and LPX in their hands cuz I think it's just going to really change the personal expectations that people have of the types of experiences they can build. Yeah, I I feel like, you know, people keep calling it the chat GPT moment, but I really like what Jensen said or it was more of a big bang moment for like the next era of AI with AI agents for everyone, not just developers. So, I'm really excited to see that, and I'm really honored to be able to talk about the system that helps power that. So, thank you very much for your time.

>> Thanks, Alex, as always. Thanks so much for your time. A huge thank you to Dion and to Stewart for walking us through NVIDIA's Vera Rubin ecosystem, how they integrated the Grok 3 LPU so fast, and the huge impacts this next generation of hardware will have on throughput, power, and intelligence. This is much bigger than just language models. It affects everything from image and video generation to robots, AI agents, and even open claw, which is already taking many of the markets we invest in by storm. Thank you to the Nvidia team for flying us out to California, for supplying us with press passes for GTC, and for making this interview possible.

And of course, thank you for watching and supporting the channel. Without you, I wouldn't get opportunities like this in the first place. And if you want to see what else I learned at Nvidia GTC and what I'm investing in, check out this video next. Or if you want more science behind the stocks, then this video is for you. Either way, thanks for watching and until next time, this is Ticker Symbol You. My name is Alex. This is Ticker Symbol You. My name is Alex, reminding you that the best investment you can make is in you.