Daniel Newman: Igor, welcome to The 2022 Six Five Summit. So excited to have you join us.

Igor Arsovski: Glad to be with you, Daniel.

Daniel Newman: So it's been a great show, big week, topics that are going to be certainly near and dear to your heart. AI, data, semiconductors, all in focus here at the summit. But I've wanted to start off with something a little bit fun with you. So, most of the people that speak at our summit are a little long in the tooth at their current companies. Meaning, they've got some years under their belts. You're pretty brand new over at Groq coming in as its newest fellow. Just kind of give me the quick background, your story. How did you land here?

Igor Arsovski: Yeah, yeah. So let me give you a quick one. So I started back at IBM in 2003. I started as a memory designer. I quickly moved through the ranks and ended up being the CTO of the ASIC business unit that ended up moving from IBM to global foundries first, and then it ended up as the ASIC business unit at Marvell. Now, during this time in the ASIC business unit, I actually first met Jonathan, the CEO of Groq, in 2016 when he was looking for an ASIC vendor to build his first chip, basically. This was the V1 or what he called the Allen.

So since then, we basically got that chip I think, manufactured and taped out in a very short time. I think that from final net list to tape buddy was about 13 months and we had a first time, right? Silicon, which is currently shipping with the Groq customers. Since then, I kind of transitioned through a system company. So I worked for Google, kind of running their Google Cloud technology development. So this was the group that was basically feeding technology into the TPU, the VPU, and many different VUs we used that are yet to be announced. And then about two months ago or three months ago, Jonathan reached out to me and asked me to come join Groq. So I couldn't resist and here I am, two months at Groq right now.

Daniel Newman: Well from fairly significant experience working for some of the largest names in tech. It's probably a little bit of a different pace at Groq, joining a startup that's really trying to challenge in a field that really is open for some new innovators to come and be disruptive. There are some companies that are certainly seen as the leaders today. And I think Groq wants to and sees itself as a very capable challenger. I remember talking to Jonathan last at our summit in 21, he's also going to join us here, so hopefully all of you are going to tune into that session as well. But I kind of like the lens that you're going to provide here, being fresh, making a move from one of the world's largest cloud companies right now over to Groq.

But AI's a big challenge. I mean, we are in a stage where it's becoming pervasive. Everybody starting to understand that it's going to be part of our lives. There are very practical applications for AI, and there are some that are often touted that are much further away in terms of truly being AI than the way they're often positioned in Hollywood or in movies, or. But I'm curious through your lens, as an engineer, as a designer, that's building Silicon, what do you see the big challenges facing the AI industry?

Igor Arsovski: Yeah. Yeah. I mean, some of the biggest challenges are just the sheer growth in model sizes. So these are the AI models that are growing at an incredible pace. And if you try to match them to the growth

of just transistors per chip, just to try to match those two, you'll find that there's a big divergence. I think the doubling of models was happening, I don't know, there's speculations that are anywhere from three months doubling to maybe a year or so that the models are doubling. If you look at Moore's law, we're not getting that type of scaling, we used to get it every 16 months, it's not longer the case. So I think that just the discrepancy between the model growth and the Silicon kind of improvements is a big challenge. So what that drives is better architectures.

So how do you differentiate? Do you need to come up with a better architecture, something that's more friendly to scaling into the next technology node or the next more than more scaling with chips and 3d stacking and things like that. And there's other challenges that go beyond the chip level, right? If you look at, to handle some of the biggest models right now, data centers are deploying hundreds of thousands of SOCs working together to kind of train on a specific model or implement a specific inference. This is no longer just a chip challenge, it's really coming up with a primitive, like a chip that could be scaled in an efficient way, in an energy proportional way to actually implement something truly massive from the aspect of like kind of system development.

Daniel Newman: Yeah. There seems to be an overwhelming, consistent answer from designers to implementers about the volume of data and the complexity that it brings, right? Oftentimes models are created using a limited subset because there's not enough processing power to process all the data. And certainly as we add more data, that's going to enter in real time, that adds more complexity. So, continuous learning is a goal, but the movement of data is both expensive. It can be complicated and it's not just a chip level problem, it's a fabric problem, a network problem, a compute problem. And it's also what's driving, I think a lot of the disaggregation in semis.

You of course have a lot of background in ASIC. You mentioned that you were basically been building them from the onset. You did it both on the cloud side, but also on the pure chip and cloud optimization side over at Marvell. But I'm curious, I want to kind of do a double back there. Because you said Jonathan called you and you just felt you couldn't resist. I think it was the term you said, someone's got to go back and play that and see if that was right. But point is you basically said, "Couldn't resist came over to Groq."

What convinced you? Because you were working at a company that's breaking things, moving fast, innovating also somewhere Jonathan was in his past. But with all your experience, working somewhere that I'm sure in Google is going to be investing in building its own Silicon more and more. I mean, as we've seen the cloud scale companies all building their own, why make this leap and do it at Groq? What really drove you? What was the most convincing thing he said? Or can you share that?

Igor Arsovski: Yeah, yeah. Yeah. I mean, leaving Google was not an easy task, just to be fair. I was leading the technology development group there, and it was really exciting to work on some of the biggest problems in the world. What really pulled me towards Groq is really the unique approach to solving the problem that Jonathan has taken with Groq. If I look at the architecture, it's a different approach than most AI companies out there. Everybody's kind of pushing down the path of a GPU like architecture, multiple cores, all contending for the same memory resource, all contending for the same IO resource. And by doing that, you're creating these conflicts where these cores are needing

to basically fight for these resources. What Groq has done is something unique. They've kind of started from scratch and they've gone down this path of this one core approach.

They've removed all the arbiters, they've removed all the caches, all the reactive type elements, which in the CPU world have become like a really a gut send for performance improvement. But for the ML world, this has become like a really simple, clean architecture. It's very uniform in nature of the chip. The data movement is very short. As you mentioned earlier, data movement is a big problem where the power is coming from. So the chip looks like a group of effectively conveyor belts that are basically moving data and from east to west, and then from north to south, there is instructions that are being issued, executing these operations. So that architecture looks really, really interesting from just data movement and regularity, which is naturally scales well into next generation nodes. But the biggest Delta, the improvement that I see is this a hundred percent determinism that is there behind every Groq architecture.

And what that does is, allows the software team to actually predict exactly what operation will be executed in what operational functional unit at every nanosecond of the operation. So basically they have a full predictive view of what's happening on the chip. Now you can say, this is maybe not a big deal, but it is really understated as we move to more and more advanced nodes, and we're pushing higher and higher frequencies on chip, this determinism is actually key to predict how much current needs to be supplied to the chip to avoid brownouts, and then how much cooling and how much operations can be issued to maintain under a specific thermal power, basically envelope.

That's not found in any of the other solutions that I've seen so far. And that to me is kind of key, especially as we move to this more than Moore's law, it would allow significant improvement in performance and really predictable deterministic type outcomes out of the chip. So that was really one of the big pulls for me through Groq. I really believe that the architecture has a lot of scaling potential and a lot of value moving forward.

Daniel Newman: So, I got to ask a follow on that, it sounds to me like this would be materially, a shift in data center architecture, right? So this deterministic architecture would change a lot of things, could have impacts on power envelope, it could have impacts on, even I know ESG is a popular topic, but if we're creating more efficiency or being very deterministic, very specific, I could see some benefits there. Of course, giving the adequate resources for what's required to compute a certain workload and to manage. I mean, what are some of the ways that you see this level of determinism really impacting the data center?

Igor Arsovski: Yeah, yeah. No, so I agree with you, like at the chip level, you get these benefits where you can really manage the current delivered to the chip to be matching the requirements that you ask for the chip to execute. So you're kind of flattening out the current demand really tightening down the power demand, and there's not a lot of brownouts and overshoots and voltage, so you can actually really optimize power. At the data center level, this determinism allows you to really synchronize multiple chips to act like one giant core. So by all of these chips communicating at the same time and executing in a very deterministic fashion, you actually can get a lot more communication, effective communication between chips, between CTCs and IO, and kind of created this really super chip in a

data center. Something that would've required a lot of synchronization between chips in non-deterministic nature.

So what that does, it allows almost linear scaling with the added number of chips into the network. And I think we have an ISCA paper that's coming up in about a month. I think Dennis Abts, who is our chief architect at Groq is going to be presenting the details of that and providing some hard data behind some of the statements I'm making right now.

Daniel Newman: So I'm curious as someone that is a developer, designer going from bigger company to small, you mentioned some of the specific things. What are the things kind of in the broader development of next generation Silicon solving future challenges for AI? What are the things that really interest you? What are the types of projects, concepts, things that you're maybe working on or want to work on that interest you and that you think are going to have the biggest impact on the world?

Igor Arsovski: Yeah, that's a good question. So I think for me, the short term kind of optimization has always been to try to kind of tie up as much across the vertical stack. So when you operate at the chip company, you kind of optimize the chip and you kind of stop at the module level, you deliver to your customer, they optimize the board and somebody else optimizes the system and software and so on. I think the exciting thing about being in a startup basically like Groq, is that not only can you optimize at the chip level, but really affect the entire vertical stack. So some of the biggest efforts right now are really co-optimization efforts all the way from the chip level to the software level.

So I have regular meetings with our software compiler group, where we're actually figuring out how to basically optimize across that whole vertical stack. So that is really exciting. Beyond that, I think that co-optimization will become almost essential as we move down to these more than more architectures, which have a lot of hazards like thermal budgets, power supply noise budgets, that co-optimization is going to be critical to kind of enable the next generation of systems, which have reduced more less and less margins, so we can get more and more efficiency and deliver green data centers that everybody's kind of pushing for. And I really am excited about that piece, especially. Yeah.

Daniel Newman: Well, one of the things we love is competition here. We believe that it definitely is what drives great innovation. So it's been great to spend some time talking to Jonathan, speaking with you, learning about what Groq is doing. There's so many complexities with the future development of AI and it is going to solve so many of the world's biggest challenges. And I think as we continue to sort of educate the market and the world, it is going to be a technological challenge, but it's going to solve so many problems. And it just really excites me Igor, and it excites me to see folks like you taking the risk, going from the big and the safe and exciting to the fast, and sometimes less safe environment, but that's where so much of our innovation in the world comes from.

So congratulations on your move. Congratulations on being the newest fellow at Groq. And really appreciate you joining me here for this year's Six Five Summit Igor. We're going to have to have you back on, in about a year or maybe even sooner, because I want to hear how things are going.

Igor Arsovski: Sounds good, Daniel. Thank you so much. So this has been a pleasure. Absolutely.