Daniel Newman: Jonathan Ross CEO from Groq, welcome back to the 2022 Six Five Summit. So, excited to have you here once again. How are you?

Jonathan Ross: Good. Good, love being here. Thanks for having me again.

Daniel Newman: Yeah, it was a lot of fun having you on last year. Hoping this conversation will be once again, very entertaining in a field AND an area of a lot of interest. Semiconductors obviously has had a ton of attention over the pandemic and AI really can't go a day now, even as the average person, not in technology and not hear about how AI is going to change our lives. So, it is companies like yours that are building technologies that are going to really bring a lot of that possibility to reality. And so, that's what we're going to talk about here today. Now it has been a year, you are a startup and a year is an eternity in the world of a startup. So, any quick highlights, anything over the last year, since we last talked that you really want the world to know?

Jonathan Ross: Sure. And I think a year in startup terms is like seven in big company terms or something, right?

Daniel Newman: Like dog years.

Jonathan Ross: Exactly, exactly. Yeah. So, lots of things to highlight our compilers making tons of progress. We started with our compiler. We've actually had some experiences recently where we talked to a customer, they've spent six months working with Nvidia to tune their models, to get them to work performantly. We just had one where it was six months for Nvidia and it was 10 days for us. And then we beat Nvidia by 6X. So, it's all about software, right? That's super important. Hiring's been great. I know people have been saying it's one of the hardest times to hire ever in history. I don't think that really matters when you're going for talent density, really great people just want to work with other really great people. Although of course, hiring might get a little bit easier soon and we've started shipping some product to a whole bunch of different verticals and getting feedback and learning and that's what it's about. The speed of iteration is the speed of innovation. So, we just want to get stuff out there, get in people's hands, find out and iterate.

Daniel Newman: Well, you've got customers signing up, which means you're not just an idea, you've got revenue.

Jonathan Ross: We hit our revenue target for the first time last quarter.

Daniel Newman: I mean, that's pretty massive and by the way, you kind of subtly talked about talent getting more accessible. And that's because the market is likely going to see some contraction, which is going to separate the winners and losers in many ways, especially in startups because it's going to get harder to come by money, that free flowing capital is not going to be there. Features that were becoming companies will probably not get funded and they're going to have to get back to the drawing board. And the companies that are really robust are going to see a wave of investors that are going to want to get involved because they're going to all be kind of chasing the same few, right? The ones that are building good technology that have customers that are winning business and that have a lot of long term promise.

And that's what we're seeing here and of course you also introduced a cloud solution. So, for the customers that want to be able to access Groq in a more consumable way, that's been launched as well. So, that was pretty cool. I think I heard about that at the recent Groq Day. So, I want to start off because we don't have a ton of time, Jonathan and I do want to cover a lot of ground. So, I want to talk a little bit about supply chain. The GPU has been one of the things that has really struggled under the resiliency issues in the supply chain. Of course, you're just starting to ship as you mentioned, but you are. How are you kind of navigating that and how are customers of Groq navigating what's going on in the supply chain?

Jonathan Ross: Well, first off every company is suffering, big companies, small companies don't matter. If you're shipping product, then you're having supply chain issues period. The real difference is how the supply chain issue manifests. And what I mean by that is if we're going to ship a certain amount of silicon, if we can get more out of that silicon for customer's problems, then you don't need as much silicon and therefore you can more easily solve it. So, as an example, we have one customer that was able to get about a 300X performance boost by using our chip, that was Argonne National Labs. We did some press around that and that means that you can ship one wafer worth silicon rather than 300, that's a massive difference and that's a way to reduce pressure on the supply chain. We have another customer that got a 2000X, right? And so, the idea here is it's not just about how much silicon is shipped, it's about what's on the silicon.

Daniel Newman: Yeah, it's volume, right? I mean, we hear a lot about this. It's like I had an interesting conversation about memory recently and it's like you can have all the cores in the world, but if you don't have enough memory to support it, you're never going to get the performance if you can't get access to the memory to get the data. So, it's these things that people don't always think about. We love volume, we love size, we love big numbers, but in the end, that's not always the result. I'm going to come back to that, by the way, those performant numbers that you just shared. I'm going to come back to that.

So, give me a second because I'm going to ask you about that. But last year and Jonathan, you're a little bit quippy. So, I want to use a Groq analogy, but last year you made a kind of a quip about or sorry not a quip, it's really your strategy, right? Bringing the cost of compute to zero. This year, I think it was a Groq Day you made another comment, you said "accelerating the accelerators" marketecture or what's going on there? What are you saying when you say "accelerating the accelerators?"

Jonathan Ross: Well, I think a great example of this is the whole supply chain discussion as well, right? So, if you think about it, CPUs can run at a certain rate and they're very general, they're very easy to program. And so, if you could run everything efficiently enough, you would use CPUs. But you can't, you can't get enough performance, right? And so, there are people who use GPS in order to get a little more performance on their workloads. Now those GPS won't actually accelerate every workload, but they'll accelerate a lot of the most expensive ones. And so, at Groq we've developed a chip, it's not a GPU, but it's really an accelerator for accelerators. As an example, so Argonne National Labs, which I mentioned, they just deployed 2,200 A100 GPUs, right? That's the newest thing available. If you wanted to order that today, you'd be waiting a year, right?

So, good luck. With 16 of our chips, they were able to take one of their very important workloads and accelerate it 50% faster than the 2,200 A100 are able to achieve. So, 16 chips versus 2,200, but 50% faster. And so, when you're able to do that, what you can do is you can actually save those GPUs, very expensive GPS, very precious GPS for the things you're actually good at, right? Just like you use CPUs for the things that they're actually good at. And so, we want to accelerate the accelerators, the GPUs.

Daniel Newman: It's interesting because the trend in a lot of tech right now, but especially in semiconductors is disaggregation. I mean, we're seeing right the monolithic chip sets are being segregated, networking chips, security chips. Of course the accelerators, you're you look at what cloud providers like AWS is building with Nitro. You're looking at these DPUs and data and infrastructure processing units, that's the thing. And it sounds like you're sort of following suit right now. You're following that trend line and saying let's focus on very specific chips for very specific things and outperform by a lot in these particular areas.

Now I have to ask, I think everyone that's watching this is probably asking this question, okay, you threw numbers out there by the way about Groq versus Nvidia for this specific purpose. Are you performance wise saying you blew them away in terms. So to me, I guess my question is in a time when this is so critical and companies are trying to streamline costs, they're trying to create efficiencies, they're trying to of course get more out of their data. I mean, why would this not just land immediately and everybody be clamoring for it? And I'm sure you're in demand. I mean, you're already saying such and such percent, but I mean, this seems like a no brainer. With this level of performance, why wouldn't everybody be signing up?

Jonathan Ross: So, I think the way to think about it is why don't people use GPUs for everything, right? The majority of silicon chip today is still CPUs. And the reason is there are a lot of things that CPUs are better at, right? Then GPUs are, the CPUs actually faster. There are things that GPUs are faster at and there are things that our chip is faster at. And we've been discovering that along with the customers. But we actually had a call just this week, where there was a customer who put in an order for GPUs. They couldn't get it fulfilled in the time that they needed it. And it was an enormous number of GPUs that they needed. And the expectation is that a much smaller number of our chips would be able to solve their problem.

And so, that connection had to be made, but the connection wasn't made by that customer, wasn't made by us, we didn't know how to find each other. What happened was a mutual vendor, a vendor that was blocked by the fact they couldn't get their GPUs introduced us because by introducing us, we could solve that customer's problem and therefore we could get that customer deployed much more quickly with that vendor.

Daniel Newman: So, it sounds to me like some of the challenge may merely just be that Groq isn't necessarily a first name on a list of knowns by the traditional buyers right now that are buying GPUs to do workload acceleration or AI projects. But obviously as you start to win deals, you mentioned Argonne earlier. I imagine that's not the only deal, you're starting to get into more deals and you're going to be able to say, Hey, first of all, it sounds like you can coexist very successfully with the big CPUs and even other GPU makers that have other specific benefits and be working

almost side by side as an accelerated accelerator. Are there other examples, are you seeing other instances where you're getting this kind of performance and you're starting to doors are opening up and Groq's being brought in?

Jonathan Ross: Yeah. So in one case, for example, there's a CPU manufacturer who actually brought us into a deal because they didn't have any other way to solve the customer's problem. There is a large data center company that has done the same and we seem to be able to compliment what they're offering in a way that no other options can.

Daniel Newman: Yeah, that to me is going to be one of the telltale signs is when basically you're able to be kind of first in line to be brought in. We often talk about the fact that the race to be number two in the AI space is a wide open race right now. And of course some of your performances may be number one in some areas and the great thing about this business is it's pretty binary, Jonathan. It's not one of those things where I always say in some sports, I'll just use sports now like soccer or football, who's the best player on the field? Well, that's typically pretty subjective, right?

But in a track meet, it's pretty binary. You run the race, right? And whoever got to the line first is the fastest. And I think when it comes to performance in AI and accelerators and GPUs, performance will speak itself. Of course revenue, they might be a number one by revenue and it may not always be number one by performance, but it sounds to me like that's what you're setting out to do, be number one by performance when it comes to the particular workloads that you guys are trying to really gain market traction in.

Jonathan Ross: Well, and to be fair on that, you made a comment about is it just that we're not a household name? Is it hard to match us, right? And there are workloads where we get a 100X, there are workloads where we get 10X, there are workloads where we're at par with a GPU. So, why wouldn't you go with the household name? Which is why one of the things that we've committed to because we have limited resources is that if we know of a better solution for a customer than our own product, we will tell them, right? We just want to be that trusted voice. And right now there is so much noise in this industry. It's incredibly hard to find out what's actually going on and what can solve your problem, right?

And there's only one company that benefits from that's NVIDIA because if you don't understand what is available and what you can use, you're just going to go with the thing you've heard about. So, we're trying to reduce the level of noise. And if that means that another company gets that business, that's fine. There's more than enough to go around, right? But the supply constraint issue is a little bit artificial. I think it's just hard for people to make the connections to the better products.

Daniel Newman: Yeah and I think that's a high integrity way to operate. And of course a lot of credit to Nvidia for the successful franchise the company's been able to build, but it's also innovation and competition is critical to our ecosystem. I always say, it's an imperative that the leaders are being competed with at scale because if not, then we won't get the innovation we need. We need the challenge in the market to constantly be there. Companies like yours, challenging the incumbents to push them to develop better products and of course, open doors.

Jonathan Ross: And a great example of that was back when developing the TPU at Google, Nvidia wasn't anywhere close to a 100 Tera Ops and they didn't think it was possible. And we couldn't get anything from them that was that fast. We built that chip and then they followed and same with Groq, we built a PetaOp Chip and now they're trying to build PetaOp chips. I'm sure that what we release next, they will try and copy. And so, we're really helping to pull this entire industry forward.

Daniel Newman: Well, that's a pretty big deal. So to me, like I kind of alluded to before Jonathan, if you had the opportunity and the cost is right and the efficiency and the performance are all there, it seems kind of like no brainer. Why would you not, right? If it's available, if it works, if it's not exponential-

Jonathan Ross: No one ever got fired from buying from IBM, right?

Daniel Newman: Yeah. But-

Jonathan Ross: So, what it takes is some sort of structural issue like supply chain crisis. I can't do what I've always done, I have to solve this problem. And therefore, I'm now going to look at other options. And so, we're looking at a lot of comp... Actually what's interesting, many of the companies that we're finding are very interested in our product are the ones that are actually suffering the most in this financial situation that everyone's finding themselves in. The ones that are rolling in dough, they're fat and happy, they can do whatever they've done. The ones that are struggling to make their margins are the ones that are talking to us because they don't have an option. They have to do something that's better.

Daniel Newman: Well, and at this particular inflection point in the market, like you said, talent may become more accessible. Companies are going to be looking for new options. The Groq, the cloud offering for instance, to give customers another way to consume in a more Op X fashion, as opposed to the CapEx investment. To your point, if it takes 200 of one option and 10 of another, the economics have to play a part, especially if performance wise, all things are equal. So, kind of wrapping this up though, it sounds obvious to me but is it a limitation of how much you can build? Is it a limitation of how much awareness is out there? And as you're meeting new customers, are you meeting resistance or is it like once they see how this works, it's game on? I mean, what is sort of the constraints to seeing Groq be where it's at to being because like you said, the gap in the revenue size between you and the whole market is there's a lot of gap for you to catch up and a lot of opportunities. So, what's it going to take?

Jonathan Ross: So, if we were to approach every single customer and try and sell them and they all said, yes, we wouldn't be able to deliver that supply. We're just not a large enough company to do that. And so, right now our focus has been let's take the customers where we've got a 1000X, a 100X or some sort of very large multiple because architecture is pretty fundamentally different and then focus on them because that allows us to have the largest impact on the market. And it allows us to have the largest impact given the amount of silicon supply that we can get. And so, that means that we're focused, right?

But we're going to be building out from that over time and that's just going to take iteration, time. But if you are stuck, if you have a problem like where we're seeing a lot of success, any sort of RNNs, any sort of your audience may not know what these are. So recursive, neural nets, graph neural nets, any sort of NLP models, these are areas where we're doing a lot better than GPUs. And if you're limited, you can come to us. Cybersecurity is another big one where we've been doing particularly well and we can solve these problems better than the existing GPUs.

Daniel Newman: Yeah. So, I mean that's a pretty exciting time. It sounds to me like to some extent, it's going to be your own little economy of supply and demand that you can't fully release the power of Groq to the market yet because if a real wave of interest came in, you would have to disappoint some people. But if I'm catching what you're saying right, is you've effectively identified some specific workloads that your performance is exponentially better than what's available in the market and you are able to deliver that now.

The data from the test and the benchmarks will prove it that it's not marketing, its science. And in the end I guess you'll start to what? Deliver that creates scale and then over time you'll work your way from the 1000X to the 100X to eventually the 50X, which by the way, are all big numbers. You can say like, oh, it's only 10 or only 50. Well in a training exercise, 10 times or 50 times faster with a big set of data for something like NLP or a neural network, that's pretty darn big. So, a thousand is massive, but let's not belittle all the work that you've already done to even hit 10 or 50X.

Jonathan Ross: Yeah, and I think you said that this is about performance and really it's not. So, what matters is being able to take the workload that you're working on, whatever program you're developing or application you're developing, and to be able to put it into production. And what you want is product that allows you to do that. If you can't get the silicon, you can't do that. If you can't write the software, you can't do that. If it's too hard to maintain in production or it's not reliable enough, or any of these other things, then you can't take your workloads and put them into production. So, people aren't buying chips, they're buying a path to deploying their applications into production, right? It's developer velocity and that's all that matters because right now the one thing that's more supply constrained than silicon is actually engineers, especially ML developers. And so, if you can get more out of the ML developers that you hire, then you're going to be able to achieve a lot more than you otherwise could.

Daniel Newman: Well, what do they say? People processes systems and technology or people, you know what I'm saying? It kind of sounds like the amalgamation of all these things are going to be the limiting factor or the catapult of scale for the future in this particular space. Jonathan, I want to thank you so much for joining me at this year's Six Five Summit.

I'm really excited to kind of give a year out and see where things are at because one thing is for sure, I talked to so many companies and the demand for AI, the demand for data in this economic potential downturn, things like automation, things like artificial intelligence, the utilization of data at scale is going to be the key to helping companies weather the storm. And then beyond that, just whether it's cyber security, whether its national security, whether it's supply chains, whether it's education our children get, the drugs that we discover to solve the

next pandemic, all of this is going to be backed by the types of technologies and the problems that you're trying to solve. So, keep going, thank you so much, Jonathan. And let's have a conversation sooner than a year from now.

Jonathan Ross: Thanks. And as a little bit of a call to action here, we're here to help. We're open for business. If you need any help, it doesn't even have to be our silicon, we just want the relationship. We want to talk to you. We want to help you and that's it. That's what we're here for.

Daniel Newman: Absolutely. So, check it out. Definitely going to be links in the YouTube show notes when they're out there, but not hard to find the Groq, right? The Groq, G-R-O-Q, you search it. It'll find you for sure. So, thanks again for tuning in everybody. And Jonathan, thanks for joining me. We'll see you soon.

Jonathan Ross: Thanks for having me.