Patrick Moorhead: Hi, this is Pat Moorhead and we are back for The Six Five Summit 2023, day one, talking about my two favorite topics. Cloud infrastructure and silicon. And back for another year is Dave Brown with Amazon EC2. Dave, how you doing?

Dave Brown: Yeah, I'm really well Pat. Thanks for having me. It's always good to be back.

Patrick Moorhead: It's always good. I mean, whether it's remotely, I think the last time we saw each other was in Spain.

Dave Brown: That's right.

Patrick Moorhead: And sometimes we even bounce into places we didn't even know we were going at the same time. So that's fun too.

Dave Brown: That's right. I was surprised to see you in Spain and you were probably surprised to see me there as well.

Patrick Moorhead: We bounced into each other in the studio and then breakfast, but it was good stuff. We all get around. That's the great thing about global business and commerce. And if there's one thing that just excites me and hopefully it comes across here, is how much the cloud has changed the way that customers are doing what they want to do with their businesses, and also how that has transferred over to the consumer. And one thing that I put a lot of time and a lot of research into is how you have innovated with your silicon. And in fact, you just keep making news, which is fun. I mean, I'm not a reporter, I'm an analyst, but I have a lot of fun and I like to write about it and do videos about it. But yeah, I mean you've made a ton of news. You want to talk a little bit about it?

Dave Brown: Yeah, absolutely. Well, I mean, you talk about the cloud and the impact that it's had on customers out there, right? It's just brought compute and storage and higher level services into the hands of any developer and able to get that really, really quickly. And that's the story that we all know about cloud, right? It's been around 16 years now. But the thing that we continue to do is to say what are the workloads that we need to be able to support? And so we learn a lot from our customers. We now have over 600 different instance types across EC2, which is crazy when you think about just 15 years ago, we started with one.

And the other thing is customers will never stop looking for more performance at a lower price. And as you know, when we've spoken in the past, that is the other thing that drives us is to really say, what can we bring to customers, whether it's third party silicon or from our partners like Intel and AMD and Nvidia, or whether it's building our own custom silicon in this quest to be able to bring more performance, lower price to customers. That's the thing that really makes a difference at the end of the day. And honestly, it allows them to do more. It allows them to get more for every dollar spent and ultimately deliver more value to their customer.

Patrick Moorhead: Yeah. And one of way the ways that you've done this is with AWS Graviton. I mean, and absolutes and black and whites are always a challenge for us analysts, but when it's absolutely true, it's easy to say. I mean, you are unique. You're the first and only major public cloud

provider to build its own custom Arm-based processors. And quite frankly, for me as an analyst, put Arm in the data center on the map. I mean it was on the Edge, it was in smartphones, tablets. When it came to doing hardcore processing, AWS really enabled it, quite frankly, for the entire industry. Can you talk a little bit, remind everybody why you built this, why you made this investment not only in hardware but also with all the software that had to go with it?

Dave Brown: Yeah. So our story, it's one of these things where you're innovating on behalf of the customer and it sort of takes you in a direction that you probably wouldn't have set out in going that direction. For us, the whole Arm journey really started with us trying to sort out network latencies back in the early days of the cloud. And if you remember those early days, maybe if you were doing this back then you would've been talking about virtualization and saying, well, can it actually ever support real world workloads or are these network latencies going to be a problem? And that's the world we were living in in like 2007 and 2008. And so, we knew we had to get into hardware and one thing led to another and we ended up buying Annapurna Labs in 2015. And they began building custom silicon for us. Very small chips at that stage, but chips that we were using to do network offloading, storage IO offloading, and ultimately we offloaded everything.

But in doing that, we learned to build and develop and write operating systems and write C code for Arm because that's the chip that we were actually using, the Annapurna Lab had done. And we started to realize that. And working very closely with the Arm team themselves, we started to realize that there was definitely some value in bringing Arm into the data center and building server chips. And so one thing led to another, in 2018 we launched our first Graviton processor, and that machine was actually not the first Graviton processor. It wasn't meant to be, it was actually just a bunch of network cards we put together.

Patrick Moorhead: It was based on the same ones in the Nitro system, right?

Dave Brown: It was. It was the same as our Nitro Force service used. And you mentioned it earlier, because you mentioned the software side. And the reason we did it in 2018 was we realized we had to spark the ecosystem. And so we put it out there, it was our A1 instance, it's still around today, you can still play with it, it's still network cards. And its performance is okay, but it's certainly not what Graviton, what we know now is Graviton2, which would've been Graviton 1 had you not done the previous one. What Graviton2 brought to the table. And Graviton2 was just a massive step forward. I think law was the thing that really drove the silicon, at least the CPU industry for so many years. And in the early 2000s, we kind of got to this place where we weren't seeing that generation performance and price improvement that we had seen coming into the 2000s.

And what we'd seen on Arm was we could definitely spark more than a 10, 15% performance improvement generation over a generation. And so, Graviton2 gave you a 40% price performance improvement, what you could get from x86 at that time. And the big thing was also just using less power. And in the world we live in today, not only from a sustainability point of view, but also from just access to power, the world is consuming the power that we have at a frightening rate in many, many geographies. And so being able to run the same computer at lower power is not just green, but it's also just better use of a limited resource. So we've been very happy in Graviton3, we launched those instances earlier this year as well. And that's given

you 25% performance improvement over Graviton2, and you gain about 40% price performance versus what you get from x86.

So we are very happy. We're not slowing down in innovation at all on Graviton. Our customers are certainly in the economic times we find ourselves in where you are looking in many cases to cost optimize, you're trying to do more with less, right? After 10 years of get big fast, everybody's saying, hey, how do we cost optimize? And we really believe AWS is the best place for that with Graviton. And our customers have heard the message. And so we are literally trying to get as many Gravitons into our data centers to ensure that we can keep up with the demand from customers. So it's been a fun journey and there's a lot more work to do.

Patrick Moorhead: Yeah. And a lot of the work, quite frankly, the entire industry is going to benefit from. What are the things that being a former product marketeer, I always like to drill down into, or kind of the claims, and how long do those claims last? And what's been fun about watching your journey of Graviton and the other processors I hope we'll be talking about is I can see what you said and then what you delivered and then look maybe one or two years later, and it's still true. And I can tell you how, I mean, I know you own a lot of the parameters on both ends, but I'm impressed that what you have said maybe a year beforehand ended up being true on GA Day and then followed down the road. And having that consistency to me would be a very trustworthy thing for customers. But one of the things too, it's hard to have some competing parameters happen at the same time, right? Whether it's better price performance, lower cost, and lower energy. So is this something that customers can expect at the same time?

Dave Brown: Absolutely. When we build Graviton, obviously when you build these chips, you take them out and well, you simulate them first before you get the tape. So you have an idea of what this would be and what the performance should be, what the power consumption should be, what the price performance should be. And obviously, you only really get to see what it is once you've taped it out and you start to get the silicone back. It's always a nervous time. But with Graviton, we've been incredibly successful. We've been able to do that first time every time and see really, really good performance. And one of the things, when we give these 40% price performance improvement or 60% performance improvement or 60% power saving, we're not a benchmarking company, right? A lot of companies will try and benchmark and get the absolute best number that you can get, but most customers don't see that because they don't want to run the application in that way.

And it becomes a bit of a benchmarking competition. And so what we actually do is we just run a whole lot of customer workloads. We run things like obviously spec end, but we also run real world workloads like Memcached and we run Engine X and we run a whole lot of the database platforms and we kind of look at it and say, it kind of looks, it's about 40%. And to your point, customers have seen that. Some of them see a little bit less, some of them see a little bit more. It obviously depends on the application. Sometimes you might be IO bound and not CPU bound and then it might be a little different.

But on the whole, that 40% has kind of lived up to the claim. Obviously the power consumption is such an important part for us as well. So when we design the chip, we're really trying to make sure that it is as power efficient as possible, not only from, as I said earlier, from a sustainability point of view, but we know what it means to consume power in a data center and we know the impact of having a lower power consuming part and it's just better for all involved.

But we've seen it across the board. I often talk to customers all the time who have consumed Graviton and starting to use Graviton. And the amazing thing is it's free across the spectrum. So it's from the smallest startups to really the largest enterprises. Some of them that honestly I didn't expect to be consuming Graviton at this point in its life cycle. They are making incredible progress. And to give you a few customers, like Sprinkler is a customer experience management company, I think we all know them. They moved to Graviton and they saw a 36% performance improvement in query, as well as a 40% price performance improvement, versus what they were seeing previously with x86. So it was already, that's a good one. Wealthfront as well is one of these robo investing companies on the financial side. They actually were able to save eight hours of data processing every day by moving to Graviton.

And then we just see a lot of customers who are looking to scale but don't want to consume the cost. And Graviton's a really great way of being able to do that. And so Instructure, it's an education technology company, and they were actually able to scale up and they wanted to make sure they could scale without compromising performance while keeping costs under control. And so they used our C7G, our compute optimized Graviton instances, and they actually saw 30% performance improvement as well as reducing their cost by 20%. So it's across the spectrum. And we are always interested if a customer comes to us and says, I didn't see it, right? Graviton is not what you claimed it to be. We're always happy to get involved. And normally it's something in the programming language, they're running an older version of Java or they haven't optimized something and we can normally help them get close to that 40%. So it's been really broad and really, really good.

Patrick Moorhead: Well, the other thing you did is you didn't oversell the initial versions to be good for all use cases. Like you really said, okay, this is what this chip does well. Hey, let's extend this with version two. And oh, by the way, with version three, and also you also put the corresponding IO and storage around it to be able to handle those workloads as well. One thing I'm explaining to you, a lot of your customers, potential customers, and I think other analysts too is, hey, don't be confused. Graviton is not merchant silicon, it's not built around, it's optimized for the AWS architecture and their data centers. And you put that in a full stack on top of it, essentially it is a custom processor for you that speaks native Arm language, so it's compatible with the rest of the world. And I think that that's something that some people miss is they're like, how do they do this type of thing?

I think it's fascinating. So Amazon uses a lot of AI and has done it for a long time. And whether it's Alexa, whether it's recommender engines on amazon.com, but also just by the data, AWS runs more machine workloads than any other cloud out there. And it has been fun to see the resurgence in this conversation about machine learning, particularly AI. And this year I saw a lot of really interesting announcements that I covered on generative AI. I talked about Bedrock, I talked about Code Whisper, talking about the GA of Inferentia2 instances, and then Trainium, TRN1 instances. And as we know, that has this gigantic capability to connect the GPUs together based on some special technology that you've brought to the table. Can you talk a little bit more about these special purposes, special purpose silicon for training, for inference, and instances that customers can utilize?

Dave Brown: Yeah, absolutely. And you're right. I mean, we've been doing machine learning for many, many, many years now, both within the more traditional Amazon business, but also in AWS. And we've

had a lot of customers doing machine learning on us as well. And as I said at the start, when we look at any workload, we're trying to understand how do we improve performance while lowering cost for the customer? And there's no innovation or no amount of work or no complexity in that innovation that'll stop us from doing that. And so when we looked at machine learning a few years ago, we started to see that it was definitely going to be playing a meaningful part, compute as we know it, and more and more customers were starting to use some of those early models. But we also knew that customers were not going to be able to deploy these models if we didn't get the cost under control.

And so, if we couldn't sort of have this leap of price performance improvement, customers were not going to do as much image recognition, they weren't going to do as much text translation. All the stuff they were doing now obviously with generative AI and what it can do and the way that's been made visible to so many people through some of these chat clients. There's enormous potential. But if you cannot get the cost under control, there's actually very few companies in the world that could afford to train a model or to do anything at the scale of what you've seen out there today. And so we've launched a number of different custom silicon products. Our first one was Inferentia, which is really focused at inference. When inference is sort of the second part of machine learning, you train your model, and once you have your model, you deploy it and then you look at real world data and you actually infer from that data some sort of outcome.

And that process is called inference. And you go back three years ago and the largest cost of machine learning was inference. And so that's why, it's about 90% actually of the cost companies were seeing with the models at the time. And so, Inferentia allowed us at launch to actually reduce the cost of inference for things like burden and resnet five to 38% versus what you get from a GPU. And interestingly today, that same chip actually gives you 72% price performance improvement, and that's only because of software optimization. So it's this incredible journey of you build the silicon and then you actually optimize the SDK and the software to get this performance at the same price so you can get more price performance over time. And that really, really helped a lot of customers. Alexa moved to that and saw not only improved latencies, but also about a 30% reduction in cost.

And so we were very happy with that. And so, we said we should go after training as well as inference. And so our latest chips are Inferentia2, which is the next big leap forward in inference. And then Trainium1, which is our first training chip, so really going after training. Now, what actually happened is about three years ago, that 90% on inference started to decrease. And it wasn't the big cost we were seeing for a few customers. And what was happening is LLMs were coming in sort of behind the scenes, and these training models were just getting larger and larger and larger. And three years ago, customers are training on single GPUs and single servers with eight GPUs. And now you're running these training models in tens of thousands of GPUs and many thousands of servers with low latency networking. So we knew we had to go after solving the training cost as well, and that's what Trainium really does.

And so, Trainium1 actually is a 50% price performance improvement over what you'd get from a comparable GPU that you typically use for training. And then Inferentia gives you gain about a 50% price performance over GPU as well for inference. And both of those are very well suited for large language models, which is what regenerative AI is based on. The networking side, you mentioned that briefly. That's been interesting because many folks we use InfiniBand for these large language models. And that's really come from the HBC space, where InfiniBand made its

name. It's a very, very low latency network. And we knew when we did this, we wanted to try and see whether we could do it with ethernet, and it's really paid off. So it's our elastic fabric adapter. It's a brand new networking protocol based on ethernet, but think of it as sort of like UDP meets TCP in an interesting way.

And what we're able to do is we're able to get very, very large flows across the network, so make much more efficient use of the available network, but also reduce latency significantly. So get down to the very low 10 to 15 microseconds of latency. And what that means at the application level is you can actually scale these clusters and get into many with our latest. Even on our Nvidia, and we obviously support Nvidia as well, even with the latest Nvidia A100 and we'll have the H100 coming out very soon, we can actually scale up to about 20,000 GPUs in a massive cluster for one of these massive training runs. And so yeah, it's been an enormous amount of work and we are very, very excited about the custom silicon and obviously the price performance improvement that customers will see from that, whether they're training models or taking a model out there, fine tuning and then running inference. I think it'll really pay off.

Patrick Moorhead: What I love is how you got ahead of large language models, and I know some people kind of pretend like it was something new, but you've been working on this for years. Obviously you show up with a piece of silicon. That's not something that just happens and you can crank this out like a piece of software. It took a lot of thinking and I did kind of read the cookie crumbs that y'all had left on talking about the future and what the future looks like. And then, here we have a chip that can support this, but even more importantly, a network that can support that plus bedrock and tools like Code Whisper for the developer. So, what's powering all this performance and cost savings? How are you doing this?

Dave Brown: Well, obviously, with our custom silicon, with our team from Annapurna, we're able to build, instead of having a chip that is, and it's very similar to Graviton as well. Instead of having these chips that have many, many different functions and a lot of complexity and consume a large amount of power, we are able to really focus in on just the parts that are actually needed. And so that's the thing with Graviton. I don't have to take what's 30 years of x86 features and functionality and still have that available in the chip today. And as you mentioned earlier, I'm able to design this with exactly what I need within my data center as a cloud provider, which means that there's a lot of stuff we don't have to build in. And obviously there's a lot of innovation that happens as well. And obviously a cost point of view, we're able to do it ourselves and so we're able to bring it to customers at lower margins just because we're building these directly from the suppliers.

And so all of that goes into making sure that we're able to give customers much, much, much better performance at a lower price. The other thing is building our servers. I wish I could show you some of these servers, but we always think of a server as being like a one or two blade machine. When you get into the generative AI space, very often, a rack only has two or four servers in it. And so you can think of a server taking many, many parts of the rack. And it's just, whether it's the GPUs that are in there, in the case of our own custom silicon, we actually have 16 Trainium accelerators and Inferentia accelerators in the machine. And so it's about driving that performance. And so we're able to give you 3.4 petaflops of compute with Trainium1 and 2.3 petaflops in Inferentia2. And then being able to do that in a power efficient way and a cost efficient way as possible.

And that ultimately means that customers get that price performance. But the other place that's interesting is the software side. And so, it's not just about building hardware, as I mentioned earlier. It's about optimizing the software. And with Inferentia1, we went from 38% price performance to 72% price performance and that's just software optimization. And so we have teams of folks that are working with our neuron SDK, which is the SDK that works directly with Trainium and Inferentia and then plugs into PyTorch and TensorFlow and MXNet, whatever the framework you're using. So the porting is relatively simple, but we're able to optimize the neuron SDK to really get the most out of the hardware. And that's something that we'll continue to do. We work closely with customers, we work on every single model that comes out and we are really trying to get out every single last bit of performance that we can to really get those models optimized.

So obviously, there's partners as well. We recently announced a very deep partnership with Hugging Face which has probably the world's largest array of open source models that folks are allowed to use. And it's continuously working across the ecosystem. And that is one of the things that I think is so important with Bedrock as well, is there's not going to be one model through rule or models. We've seen this in the past with these models where they'll come out early, they're novel, they're new, but ultimately, folks will take that model and train it in different ways. And the model, you rarely want to be able to get access to many different types of models.

And then also know that you can train those models or fine tune those models without leaking your data. And those are two of a couple of really key tenets that we have for Bedrock, which is being able to say, can we bring many models to customers? Because we're going to have our own models, which we'll run shortly. We have Anthropic in there, we have AI 21, Hugging Face is in there as well. And so we want to make sure that there's many models available and they're going to innovate a different pace and customers are going to want to use different models and build their own models and all that's going to be possible with our custom silicon.

Patrick Moorhead: Yeah, it's clear to me that open models and then areas where companies are just really good at creating their own models. They might have a certain core competency in a certain area for various reasons and applying those. I actually like where this is headed and I believe that further democratization from Trainium and Inferentia and just making this available to more people. I think we've seen this movie before. It just propagates more innovation, better innovation and more cost-effective innovation. And by the way, let's not forget the amount of power that these new models, particularly on the training side, can take, where only lighting up the features that are absolutely required becomes more important and being able to then build the entire rack and then the entire fleet and the entire data center to optimize it for this usage model. It really puts in front of you the importance of architecting the entire solution of the data center for something like this to get most efficient.

Dave Brown: Yeah. And it's not just efficiency. The other one is availability and operational performance. It's another thing we don't talk a lot about. But making sure that the accelerators and the servers are always highly available and don't crash in the middle of a training run.

Patrick Moorhead: Yes, yes, something I glossed over, didn't I?

Dave Brown: Yeah. And we spent an enormous amount of time just getting what we call our annualized failure rate, which for us, we define, not as hardware failure, but any time a machine misbehaves. Getting that down and really helping customers. And so, that's another place where our custom silicon has allowed us to get levels of availability that we haven't been able to achieve before.

Patrick Moorhead: Yeah. So you've done a really good job with your announcements. I mean, all the way from, hey, here's the beta customers and here here's how they're doing and here's our production customers. And oh, hey, by the way, we're using this ourselves as well. We're not just trying to sell this to people. Can you talk about some use cases and some customers that you're seeing out there for Inferentia and Trainium that maybe we haven't heard or ones that are just, you want to keep talking about or are excited about?

Dave Brown: Yeah, it's early days. And just as with Graviton2, folks are taking the chip and starting to play with it and work. And we have many customers testing. There are a number that have achieved some things and I'm happy to talk about it publicly. One of them is Runway, where they do text summarization and image generation. They're an AI company and they do question answering and things like that. And so, they've been working with Inferentia2. Now, they did the work to go from what they were using previously in GPUs with PyTorch to move over and actually run with our neuron SDK, which is a relatively simple process. And the outcome of that for them was actually two times higher throughput on Inferentia2 than they were able to get from other EC2 instances at that time. And so, just higher performance, lower cost of inference. That all feeds into allowing them to not only reduce latencies but ultimately spend less, introduce more features, look at more complicated models.

And that cost optimization really drives innovation. And the other one is Finch Computing as well. They are using Inferentia2 instances for headline generation, text summarization, all on these LLM models and obviously do more things over time. And then you did mention internally, Alexa's been using Inferentia1 now for many, many years. And as I said earlier, that was latency. And when you ask Alexa a question, the response that she generates back to you, both the actual text and the actual audible, the way that she speaks that sounds more human-like is all done with Inferentia. And then obviously Amazon search, Amazon ads as well, all internal teams using Inferentia and getting that improved price performance. But over time, we are very excited about some of the customers we can't speak about yet as well. And I think we'll see a lot more customer use cases coming out as customers continue to use it.

Patrick Moorhead: Yeah. Dave, always a pleasure to talk about how you are using your custom silicon to lower costs, maximize performance. Really it sounds a little marketing, but really unleash the developer capabilities out there and doing it different ways. And I have to give credit for consistency. I mean, I think this is the seventh chip you and I have talked about, not even including the three versions of Nitro before that that we didn't have a chance to talk about. Any thoughts or anything you can share with the audience about the future?

Dave Brown: Yeah. I think it is that consistency, right? And as I said earlier, there's really two things. When I think about my organization, EC2. In the broader AWS, the first thing is learning about new workloads from customers. And so, what is out there? What do customers want to achieve?

What are the things that they would like to do in the cloud today that they can't do? What are the things that are too expensive? Those are the things we really want to be able to go after. And then finding ways to give customers more performance at a lower price. Whether it's our own custom silicon was doing that with one of our partners like Intel or AMD where we've seen really, really good performance from them and great integration with the Nitro system, or Nvidia where we spoke about elastic fabric adapter networking and how well that's working with the Nvidia GPUs.

And so at the end of the day, whether it's our custom silicon or it's a partner's chip, we want to just be able to give customers, it's built to support their workload and bring them the best price performance. And the future's incredibly hard to predict. I think I've learned that in my time in the cloud. And generative AI is the latest thing and it could be one of the most disruptive things we've seen in many, many years. And I think many of our customers and your viewers are going to want to do something with it. And I think working out how to practically apply it is something that I think a lot of people are grappling with right now. And then once they get through that and say, hey, we have some use cases, we have some models we want to deploy. The very next thing is going to be performance and cost.

And we really want to make sure as AWS that we are able to bring as many models to customers for them to use through our Bedrock program, and then also through our custom silicon innovation, really get the price down. Because that's the thing that's going to unlock being able to deploy and to ultimately lead to better experience for our customer's customers in whatever they end up doing with generative AI and their applications. And so I think it's an incredibly exciting time and we're certainly not going to slow down in the work that we're doing around custom silicon.

Patrick Moorhead: Just love you're one of the few people that's not afraid to talk about lowering costs for its customers. I just admire that. I don't have many conversations like that, Dave, so I really appreciate that. But that is just the reality. When you live in a real world, not a virtual world, costs matter. And doing more with less, doing a lot more with the same, and tuning into a lot of your end of the quarter calls, talking about how you're shepherding your customers and helping them, which quite frankly will pay off for years into the future.

Enterprises and people looking for long term partners. And by helping them in times of need, they're going to remember who helped them, who helped them move them forward. And one of the ways that you're moving it forward is with your custom silicon in corresponding fashion with merchant silicon. And you're giving your customers a choice. But as an ex chip guy, now turned analyst, I used to be a systems guy too. I did that for a decade. There's a lot here that I see that I like. And Dave, it's always fun chatting with you, comparing notes. We should do it more often. But really want to thank you for coming on this Six Five Summit again to share all this great stuff.

Dave Brown: Yeah, Pat, it's always good to be here. Thanks for the invite back and we should definitely do it more often.

Patrick Moorhead: Would love to do that.

Dave Brown: Great.

Patrick Moorhead: This is Pat Moorhead for The Six Five Summit 2023, day one, talking about some of my favorite things, cloud infrastructure and cloud optimized silicon. Take care. Tune in for the rest of the series. Clear out day two. You've got an awesome day two and a day three coming. If you can't sit and just perpetually watch your videos all day long, you can do purview as well. Take care, have a nice morning, afternoon, evening, wherever you are on the planet. Take care.