Patrick Moorhead: Dave it's great to see you. And thank you for coming to The Six Five Summit, again. This is great. Last year you announced the AWS Graviton Challenge. And boy seems like you've been making a lot of progress since then.

Dave Brown: Well, Pat, thanks very much for having me. And it's really great to be back. It's hard to believe it's been a year already time flies. But last year was huge. As we announced the Graviton Challenge right here, and that was a huge success for us. After announcing it, we had thousands of developers actually download the four day plan to move to Graviton. And so we were very, very happy with what happened. But as you say it's been an incredibly busy year for us as well. And it was just a couple of weeks ago that we actually launched our 500 EC2 instance, which you know, was quite a milestone for us. That obviously is from 2006 to 2022 being able to launch 500 different types of instances. From computes, to memory, to networking, storage capabilities. All to be able to best meet what our customers are trying to do with their workloads.

And some of those things include recently launching both Intel Ice Lake and AMD Milan processors. We've also expanded our accelerators to include Habana accelerators as well as NVIDIAs 810g and the A100 GPUs. We've also improved some of our storage optimized instances and actually added a new feature to our Nitro services with the AWS Nitro SSD. Which is the first time that we've actually built our own SSDs as well now, as part of what we do at AWS. And that's allowed customers to get much better I/O performance. And then obviously as we spoke about last time, Graviton has taken a huge leap forward with the launch of the AWS Graviton3 processor. Which is on now our first seventh generation instance that's available. So it's been quite a year.

Patrick Moorhead: Yeah. It's fun. I'm an ex-compute guy and appreciate everything that you do. And the way that I like to describe it is you're you've de-commoditized compute. Gosh, I remember, I think 10 years ago people are like, "It's a commodity." And then it's like, "Nope. It's not." And I'm a believer that you commoditize yourself and what you're doing with a combination of internally designed silicon and you're partnering with companies like AMD, Intel, and NVIDIA is pretty awesome. So I mean, a few weeks back, you cranked out yet another generation of Graviton. Can you tell us a little bit more about that?

Dave Brown: Absolutely. As you say we want to make sure we support whatever our customers need. So whether it's Intel or AMD, obviously Graviton. But speaking about commoditizing compute, the thing that drives us is bringing customers more performance at a lower price. And honestly, that's been the thing that's driven us year over year, over year is how do we do that? And the Graviton processors is really coming out of that drive that we've had for price performance to really see what can we do for customers to improve performance significantly at a lower price, given better price performance for as wider variety of workloads as possible. And so with Graviton2 that came out in 2019, that actually provided the best price performance across 12 different instance families on EC2 in 23 different regions.

And that's actually the best price performance that we have available on EC2 today. Over the last year and a bit, we've also expanded Graviton2 to be able to support other AWS services. So whether it's Amazon Aurora, Elasticache EMR, we even have Lambda that's come out recently

and Fargate in the container space. And so it allows customers to take advantage of that price performance on the Arm, very, very simply. And we've seen tens of thousands of customers do that. Actually 48 of our top 50 EC2 customers. I actually use Graviton2 today in a meaningful way in their workloads. And so it's been very, very exciting to see that. But the thing that customers always do for us is they always have that next workload that needs more compute. It's more compute intensive. They're looking for more price-performance improvement as well.

And so that's where Graviton3 came in. And Graviton3 is actually a huge leap forward in the technology. And so we're not slowing down at all from what we had in Graviton2, we learned from that there were a couple of areas we wanted to improve. And that's what you see in Graviton3. And so Graviton3 actually offers up to 25% better performance over what you got with Graviton2. And that 25%... We're not a company that believes in benchmarks where we've tweaked the benchmark and got it just right, but it's not something that any customer would really see. That's probably on the lower end of what we expect most customers to actually see. It also brings you two times higher floating-point performance, which is really important for a lot of compute workloads. And then two times faster cryptographic workload performance as well.

And that was one of the areas in Graviton2 that we really wanted to improve. And then actually three times better machine learning, workload performance. And so you can see it's just a huge step forward in all of those metrics. It's also the first instance in the cloud that's generally available to actually support DDR5 memory. As you know, that's the next version of memory, RAM out there that's going to be available. And so we're excited to do that, which because of that, it brings 50% more memory bandwidth as well. And then super importantly we are all looking at our carbon footprint today, and many of our customers are working with us and saying, "How do we optimize? How to become carbon neutral? And Graviton3 actually offers 60% less energy consumption to actually perform the same workload on Graviton3 versus the equivalent x86 based instance uses 60% less energy.

And so it's really becoming a meaningful part of our energy efficience internally within AWS, as we look at a lot of workloads running on Graviton. But also for many of our customers as they want to do that. And so, as you mentioned just a few weeks ago, we did launch our C7g instance. So I was very happy to get that out there. It's been very exciting over the last few weeks to watch Twitter and see what customers are doing. And so we're excited to see where that goes.

Patrick Moorhead: Just one of the things that I really appreciate is your conservatism with your claims. Dave, I've spent 10 years in systems and 11 years at a processor company. I know how to pick apart numbers. And maybe I've been involved in a little benchmarketing before in the day. But what you did is you really made conservative commitments. You under committed and over delivered. And as an industry analyst, I appreciate that. But I also know that customers appreciate, particularly when you bring something out new, it's important to engender trust. The one interesting thing that's unique about when you make your performance claims is this is inside your environment. And there are some really cool things that you can do to get even bigger levels of performance. But let's talk about Graviton3 let's drill down. You said 25% increased performance, two to three X floating-point and crypto workloads. Can you talk a little bit more in depth maybe from what the customers are seeing from a C7g instances?

Dave Brown: Absolutely. And we're happy with the way that we talk about these numbers as well. We want to make sure that as broader set of customers actually see the performance claim. As I said earlier, it's not something we specific benchmark. But we really actually run real world workloads. And then we look at those workloads and say what do we think is a number that most customers would see for their workload? And so it's been really good to see that both on Graviton2 and then on Graviton3 during the private preview, which we started last year at reinvent in December, we've had literally hundreds of customers actually use Graviton2 and benchmark and then run the application. What I mean by benchmark. So customers like Epic Games, Formula 1, Next Role, Snap, Sprinklr, and even Twitter have actually evaluated their workloads on C7g instances over the last couple of months.

And so some of the numbers have been really interesting. Twitter, they were testing on Graviton2. And so we gave them a Graviton2 C7g instance, and they actually ran a variety of workloads. And what they saw is they saw a performance improvement from 20% to 80% for some of their workloads. So some of their workloads were actually 80% faster on Graviton3 than they were on Graviton2. And then they actually also saw a reduction in tail latencies. So up to 35%. So not only was it faster, it was just a lot more consistent as well. And Formula 1 is one of my favorite forms of motor racing. And I enjoy working with that team. And they've been doing a lot of computational fluid dynamics work on AWS over the last few years. And I think you can see it in this year's racing, how it's changed some of the wheel to wheel racing that we're now seeing, but they actually tested Graviton3 as well and they saw a 40% performance improvement over C6g.

And Sprinklr observed 27% better performance. And then one of the companies that's really gone all in on, on, on Graviton is actually Honeycomb IO. They actually just announced a hundred percent of their fleet is now running on Graviton instances. But when they tested Graviton3, they actually saw 35% better performance. And then also a 30% reduction in latencies. It's again, similar to what Twitter had seen. So really it's just across the board and it's one thing for us to benchmarket. But I think what customers really love is actually seeing some of the real world workloads and results from other customers. Whether it's on Twitter or in blog posts or things we put out there, that's the thing that really makes Graviton so popular.

Patrick Moorhead: You had me at Formula 1, Dave. Maybe after the show, we can compare notes on that. But I'm a big Formula 1 fan and appreciate how much technology goes into the design. Literally changing the design of the cars every year. Some companies were able to react to it better than others, as we've seen this season. It's it's incredible. So in the spirit of conservatism, one thing that I noticed was in first generation Graviton, you were very clear, you bracketed the types of applications that the customer should use. And then in Graviton2, you widened it a little further. So what more can I do in C7g with Graviton3 Or maybe in other words, what workloads are you recommending for Graviton3?

Dave Brown: Absolutely. Graviton1, or Graviton before we added the number, that was really just a a processor that we put out there. It was our A1 instance that was a single instance that was never a plan to make it available more broadly. And it was ready to spark the ecosystem. It was an experiment, on our part really, to tell customers and to tell the ecosystem ISVs open source

software operating systems that Arms coming to the cloud. And it had an enormous impact just in a year we saw massive explosion of growth. Graviton2, we said we really want to make this available on any instance type that we think having Arm based or Graviton based processor actually makes sense. And so you saw us do that on 12 different instance types, all the way from compute to memory, to storage and memory optimized, all sorts of things.

And Graviton3 will absolutely go the same way. So you'll see our launch Graviton3 across a number of different instances types as well. Each one carefully designed for what the customer needs to do. C7g though, is the first of these Graviton3 instances. And that's a compute optimized instance. So what we mean by that is it actually has one to two vCPU to memory ratio. And so for most of our instances, customers are actually looking at how much memory do I get in terms of gigabytes of memory. Do I get to every virtual CPU, virtual core, essentially, that I have on the processor? And so this is a one to two ratio being compute optimized a little bit less me than you'd have on some of our other instance types. But it's really focused on those applications that are compute intensive.

And so anything that you have that is compute intensive or maybe needs higher compute power CPU, power, higher floating-point performance and obviously that cryptographic performance. If you're doing any sort of SSL or anything along those lines, you'll see a big improvement. And so applications that can take better advantage of the faster memory bandwidth also DDR5 is something that would work a lot work very well. But anything along those lines. So it gets to the specifics. what we see customers run on these types of instances are typically computing intensive application service and microservices. So the parts of your distributed system and parts of your control plan, you're typically run. Obviously in a world where analytics is a big thing, distributed analytics applications, ad serving on the internet. We have a lot of customers that run ad serving applications, high performance computing talking about Formula 1.

This has been a huge area for Formula 1 to do computational fluid dynamics is the HPC space. And then even machine learning whether it's inference or some lower levels of training and machine learning, media and coding. And then obviously we have a lot of customers that run gaming as well. So it's anything that needs a little less memory, but an enormous amount of compute. And so that's well suited. Over time, you'll absolutely see us do more with Graviton3. But that's where we started right now.

Patrick Moorhead: If somebody would've told me 10 years ago that a processor that was internally developed by a cloud company would be doing the highest level of performance. I just would've said, "You're nuts. That's just not going to happen. Good luck. Good luck with that." So the fact that it can do high performance computing, HPC, media coding, gaming is super impressive to me. And certainly it's a 10 year journey of in-house silicon. I mean, don't just start this thing up and call it a day. I think some people forget the decade of investment that you put into it. So that's exciting stuff. How do developers who are watching right now, how do they get started?

Dave Brown: It's really simple. I mean, as with using anything on EC2, it's literally an API call or a consult click away. And we do have that portfolio of 13 different instance families. And so the most important thing is find the instance that you want to start with, that you think is best suited for your

workload. Whether that's burstable. We have our T instance family, T4g on Graviton, which is much, much cheaper on the lower end and burstable performance. And then compute optimized, memory optimized, storage optimized, even accelerated computing workloads. Last year we launched our first Graviton and a GPU based instance for the gaming space, which is really interesting. So you just pick it up and work out what it is and available across the 23 regions. So there's one close to you. And with C7g specifically, right now, it's available in two of our largest us regions.

And we'll be expanding that over the coming months. The other thing we also have available is we actually have a free trial available as well with Graviton. We've actually just improved that we actually increased the size of the instance that you get as part of that free trial. And so you actually get a T4g small instance that you can use for a month, several or 750 hours per month. And use that to really test your workload, test the operating system, recompile. And so it's a way to actually do the port to Graviton completely free from Amazon. If you want to look at managed services as well, that's the weather place that customers really just find it super easy. Specifically like the database space say, "Hey, I can get that performance improvement.

And on my SQL Postgre database, let just go use Amazon Aurora. Or in the container space let me use Fargate or even Lambda serverless. Let me go use Lambda for that." So it was just a lot of simple ways to actually get going. We also have a very growing Graviton partner network. Now where many, many companies out there looking to partner with us. And so if you go to the marketplace, you can find a lot of applications ready to run on Graviton. Whether it's security, monitoring, CICD, even container services all there ready, packaged, and ready to go. So it's very simple. It's gone from the A1 instance that was very specific to really just a broad set of ways that you can actually make use of Graviton very simply in your environment and get that price performance benefit.

Patrick Moorhead: I'll say it again. And sometimes this sounds like the mutual a admiration society, but I know how hard silicon is to get right. And one thing about your strategy is it's not just about lower cost, but it's also about highest performance. So you're really trying to deliver both of them and do this not only from first party silicon, but also working with your partners. So last question, any parting thoughts for the audience?

Dave Brown: Yeah. I love the fact that you mentioned some of our partners as well. Whether it's Intel, who we've worked with for 15, 16 years on CT now. AMD, who we really brought to EC2 in a meaningful way in 2018. And the Milan processor from AMD is really performing very well. Or NVIDIA who we still use an enormous number of accelerators from them and customers love that. The important thing is our customers want those processes and those accelerators in the cloud. And they have workloads that run very, very well on those. And we run workloads on those ourselves and we'll continue to do for the foreseeable future. So that's selection that's available is very, very important.

And as you say, price performance is really, really good. What I often see Pat is a lot of companies will focus on the performance aspect actually. And then at a later point, they go and decide either they want to price it. And often more performance means higher price. What

we've really been pushing is well, more performance shouldn't mean higher price is this price-performance equation, that the thing is so important. So hopefully we'll see more of that across the broader industry over time. But you mentioned it earlier. This has been really a 10 year journey. I remember back in 2008, when we were looking at the cloud in the early days. And a lot of the analysts out there were saying, "Well, I'm not sure this cloud is ever going to work with 300 millisecond of networking latency.

And is it as good as bare metal?" And back in the day, the virtualization stacks really just weren't what they needed to be to support the workloads we had today. And that's what drove us from way back then to say how do we think about this differently? How do we offload things like networking and storage to these accelerator cards? And these accelerator cards actually happened to be Arm based because we ended up buying Emprenure Labs as a chip manufacturer. My engineers started writing code for Arm. We started building operating systems for Arm. And that was the journey that really got us to where we are today. And so the Nitro system now runs all of our instance types with these Nitro offload cards. We just launched that Nitro SSD. And then obviously as you continue that journey, you get to Graviton and what we've done in the processor space.

And then obviously what we've done in the accelerated space with Inferentia and Trainium as well. And so it's all been driven by this quest to give customers better price performance, and that'll continue to drive us. We continue to work backwards from the customer, understand what they need and really building reliable, secure, and then scalable cloud infrastructure is what we want to be at the end of the day. And so we can accelerate our customer's business innovation and allow them to bring pretty much any workload to the cloud and then run it in a very cost effective way. And as I said earlier on one of the things I know is customers are going to continue to come back and say, "Well, Graviton3 was great, but could you give us a bit more?" And so we're not going to slow down at all in terms of cost effectiveness and also that high performance. And so I expect a lot more innovation from us over the years to come.

Patrick Moorhead: So Dave, one of the byproducts of what you're doing for your customer also impacts the overall industry. I mean, I look at what you did with Nitro and now intelligent offload is the industry thing. Right after you created that and not just for the public cloud, but also for on-prem. That's the direction that they're going. So I feel like your investments are not only impacting your customers, but also raising the water level for the entire industry. And I think challenging the industry to do things better. And that sometimes you don't get a lot of credit for that. But I think those in the industry who know see what's happening and the response that the industry has had.

Dave Brown: Absolutely agreed. And if what we do does drive the industry to ultimately give better price performance in other areas. Or better solutions in other areas. And I have better processes or better accelerators. That's something that benefits my end customer as well. So that's a good thing.

Patrick Moorhead: Absolutely. So Dave, I'm looking forward to reinvent coming up and you always have more announcements that I can ever consume. But I'm a product guy. I love it. But looking forward to,

and thank you so much for coming on the show again. And I think it's been very valuable. And next year let's do this again. Appreciate that.

Dave Brown: Yeah. Appreciate that. Well, thanks Pat. It's always great to be here. And so thank you for the time. Thank you.