Patrick Moorhead: Dave, welcome to the Six Five Summit 2021. And thank you for so much for speaking here on day four.

Dave Brown: Thanks, Pat. Thanks for having me. It's really great to be here.

Patrick Moorhead: Yeah, absolutely. So we are so excited that you're participating here and we're so excited you've chosen our summit to announce your new program, the Graviton challenge to help companies and developers adopt AWS design Graviton2 processors. So can you tell us more about Graviton2 processors and the Graviton challenge?

Dave Brown: Absolutely. Well, we built our AWS Graviton2 processors to really push the price performance envelope for our customers in EC two. And over the time we've seen thousands of customers getting up to 40% better price performance with AWS Graviton2 using our easy two based instances. And that's 40% better price performance over what they're able to achieve with other first generation x86 processors. And so we've been thinking about ways we can make it even easier for our customers to adopt Graviton2. And so we came up with this idea of the Graviton2 challenge, which is a four day plan that provides developers with a step-by-step process to move from the x86 workload to Graviton2. And you can sign up for this challenge. We'll even provide you with assistance in the migration. It is four days, but that's customizable. So you could do that over a number of weeks and as long as you do just four days and show us what you actually migrated at the end of that process, you can actually get some swag for the developers taking part, and we'll also call out to anybody that completed the contest and there are even prizes to be won as well.

And so that's the Graviton2 program or challenge really encouraging developers out there to see what they can do in just four days to move to Graviton2.

Patrick Moorhead: It looks super exciting, Dave, and we got the chance in the virtual green room to talk about this. And I think if you're a developer, the thought of being in a re-invent to present and get your prizes is pretty awesome. And pretty much you're going to be a rock star of your department.

Dave Brown: So that's the thing we're really excited about is getting developers to see what they can do in just four days. We've had customers do it in less time than four days to move to Graviton2. And then obviously the opportunity to win prizes and even being a re-invent announcement is also enticing. So we're pretty excited about the program.

Patrick Moorhead: Yeah. It's great to learn about the challenge and I can't wait to see about the winners. But this is not AWS's first rodeo when it comes to custom silicon. I've been tracking your custom silicon for a while, but can you talk to everybody who's watching about some of that history?

Dave Brown: Yeah, absolutely. So we actually started our custom silicon journey probably about 10 years ago and it actually came out of sort of us building EZ2, our computer engine, and then really starting to think about what do we have to do to really provide our customers with a level of performance that is at the level that they would expect. And at the time we were really struggling with jetta, really struggling to get the network performance and the IO performance that we needed. And so what we did is we essentially went and said, "What could we do? Is there a way that we could actually start to move some processes away from the central CPU and actually move those off to custom silicon?" And that's where we actually started to play with some of the sort of the advanced necks, smart necks back in the day, and starting to do some custom silicon work around what processes could we use to run some of these latency sensitive workloads that we needed for networking and storage and that sort of thing.

And so over the years, we started with some initial silicon, we got very involved with a company called Annapurna Labs. In 2016, we actually ended up acquiring Annapurna Labs. And internally we became really good at writing software for ARM-based chips or CPU's that sat on these offload cards that we were able to move all of our required processing to those cards. And so we learned internally how to code for ARM. We worked a lot with the ARM team themselves and what it means to actually write software for it, but ultimately be a server chip. And all of that led to a world where we ultimately in 2017 launched the nitra system and the nitra system is still something we use today on all of our instances and gives us a level of performance and security that's really just an incredible enhancement over what we had initially in the cloud.

And it's allowed us to learn a lot about building custom silicon and writing code for these ARM-based chips. And then in 2018, we actually launched the very first Graviton CPU, which we actually used our fourth generation networking chip for that first Graviton CPU. And the idea was really just to get it out there into the market, just to see a signal to the world that ARM had arrived in the cloud. And that Graviton was a reality and something we were working on and spark the ecosystem. And then in 2019, we launched our Graviton2 processor, which we'd actually been working on for several years. And that just brought incredible 40% price performance improvements over other x86 alternatives. And really just started to see a big migration of customer workloads to this new ARM-based Graviton2 processor.

Patrick Moorhead: Yeah. I remember when you first kicked off your custom silicon, people who are looking in figuring out what are they going to do with it. And then I know some people were thinking, "Oh, there's no way that AWS can do this." And then here we are multiple generations later. So what is the customer feedback and reaction been so far for the new Graviton processors?

Dave Brown: Well customers obviously love it. I mean the big enticing point for them as the 40% price performance improvement. That's something we've stated publicly that customers can see that. And in most of the use cases I've seen customers are actually getting that all better for most of their workloads. And so being able

to move from an x86 processor to Graviton2 and then get that sort of payback is just huge for customers. So there's been a lot of interest from that. Customers have also let us know that the migration has actually a lot simpler than they expected. And so normally, customers say to me, "Dave, moving to ARM, that's going to be a huge lift," but what we're actually finding is in most cases it's a lot simpler. We've also done a great job in working with the ecosystem to make sure that the various libraries and ISV software packages and that sort of thing are available.

And it's actually relatively quick to move. One of our customers SmugMug, they had actually used Graviton one back in the day, they were early to point to that and very excited about the possibility of ARM. And when Graviton2 came out, they actually port it in a single day. And we're seeing about 40% price performance improvement with that migration. We recently had another very large customer port an entire enterprise application in just four days. And so that might've been where the Graviton challenge came from and actually moved to production and also see an improvement. So a lot of excitement across the customer base from the smallest of startups, all the way to the largest of enterprises. And we've seen really strong adoption, which is really exciting.

Patrick Moorhead: When customers see or hear that 40% number, they have to go and investigate it. So let's do the double click on that. How are your customers able to get this from your silicon? Maybe you can talk about more about this and how you achieve this.

Dave Brown: Yeah, absolutely. So at the raw numbers level, so Graviton2 provides you with about a 40% raw performance improvement for most workloads over x86, the equivalent to x86 processor, and it's about 20% of the price. And so if you put those two numbers together and you look at most workloads, it's about a 40% price performance improvement is what we talk about, right? So if you had a set of 100 machines historically, and you move to Graviton2, you probably only need about 60 machines or so to run that workload. And so that's the improvement that customers are seeing now when we designed the Graviton CPU, we obviously... It's a lot of work on our side to really optimize that chip for what you need in the cloud. The other thing is obviously working closely with the ARM team to provide the ARM core that gives you the performance, but also sort of streamlining that process.

One of the benefits we had is we don't have to bring along years of backwards compatibility. It's a new process. And so we can really streamline it for just the workload that customers are looking to run in the cloud. The other thing is we get a lot of benefit from the nitra system. So it's not just about the processor. When we look at that 40% improvement, it's a full instance type. And one of the things we get there is that the improvement that comes from the nitra system. By offloading all of those workloads I spoke about earlier from the central CPU to these offload cards, we're actually able to give the customer more of the CPU than what you might normally get because of all these other processes running

on it. And we see that also gives customers about a 10% to 20% boost in performance, even with equivalent CPU's between cloud providers.

And so all of that together and also just the breadth of that improvement. So it's really not just specific workloads where customers are seeing this, but it's across a broad breadth of use cases and workloads that customers are tested on Graviton2, from web apps all the way through to open-source databases are now and there to some of our caching applications and customers are seeing about the 40% price performance, which is more than enough for customers to justify some of the work that goes into all the work that goes into actually doing the migration.

Patrick Moorhead: Yeah. One thing I really do appreciate about Graviton is that this is your processor with the features that are right for your environment based on what your customers are looking for. And there's not too many people that can talk about having that customization. Any other things around workloads that your customers are running? Is it a narrow set? Is it a broad set? And what kind of benefits are they getting?

Dave Brown: Yeah, as I said earlier, it really is the price performance they're seeing on a pretty broad set of workloads, right? And so it's coming from some of the sort of normal sort of what you call general purpose workloads for things like web applications where they might be running something like engine X or Apache and serving a website. And the day they get in about the 40% price performance improvement. A new area for us is actually source databases. We found that Graviton2's actually was incredibly good with some of the open source database workloads. And we've actually also made Graviton2 available with RDS, our relational database service, which is a fully managed database service. And for the open source databases managed by RDS, you can not just select Graviton2 to as one of those options without having to do any sort of migration yourself, which is a huge improvement.

It also in-memory caches as well. So elastic cache and those sorts of things available on Graviton2, and it basically, the other ones are super interesting is in the compute intensive workloads, even things like HPC. We're actually starting to see some customers starting to experiment with HPC, with Graviton2 and our high-performance networking giving you up to 100 gigabytes per second on our Graviton2 compute optimized instances. In terms of suitability it's like Linux-based workloads with programming languages, such things as such as Java Python, PHP. Those are relatively simple to move because you're dealing with... It's virtualized, you've got a hypervisor there that you're running on, but it's pretty much across the board. The other thing is we've worked a lot with the ecosystem to make sure that libraries and that are available.

And often when customers do find there's a library that's not ported, we'll actually work together with them and the library developer to actually get it ported to ARM relatively. And obviously examples of customers, I mentioned SmugMug earlier, but also another customer next role in the ad tech space,

they've seen a significant price performance benefits as much to at Next Role is up to 50% price performance improvement in moving. And they let us know that the adoption process was relatively seamless for them as well as they look to move over. And so it is really a broad set of workloads. The migration in most cases is relatively easy and obviously customers love the price and performance gains.

Patrick Moorhead: Yeah. And let's talk about how much time and effort it takes for customers to adopt Graviton2. And by the way as a side note, I did appreciate that you didn't over commit with Graviton1. You suggested a more narrow set of workloads. I was a little surprised that you opened it up with Graviton2, but I was also really happy because customers have to think a little bit less about what types of applications they can bring into the environment. So let's talk about time and effort it takes to get in. And you mentioned it a little bit of some of your customers and this may have been the genesis for the challenge, I hear?

Dave Brown: Yeah, exactly. And so we've heard most customers taking anywhere from a couple of days, as I said, there was a customer just a couple of weeks back that took four days for a pretty large application. And that was very fast.

Patrick Moorhead: I think I knew who that might be.

Dave Brown: We had SmugMug do it in a day as well, which I think that was the record we've seen. Some of them takes a few weeks, a month, a month or two, in some cases. It depends on the workload and how much work there is and how much of the ecosystem are those ISV tools applications are available for you. I think the one thing to say is in most cases it is actually a lot simpler than customers expect. And so what I always do when I'm talking to customers, as I say, take an engineer or two and see what you can do in two weeks. And I think customers always come back and they were very surprised at the progress they could make and how rich the ecosystem is for ARM. And they were able to make significantly more progress than they expected.

And then from that point onwards, it's a few more weeks of testing and then some qualification and eventually they're up and running on Graviton2, which is normally the process that folks have been taking. Some of the things that impact the time that it takes for the application to move, obviously the architecture of the application itself, its dependencies, support for infrastructure services they might be using in the production environment. And in many of those cases, what happens is if they do get stuck we have a team of folks that are available to work with them and expand the ecosystem, which is already very, very robust. It's pretty broad. But if there's anything that's missing, we're happy to work with customers to make sure it's available.

Patrick Moorhead: Well, I've been watching the advent of the, I'll call it the general purpose ARM ecosystem. Some people think that just this just started happening over a few years. Now, I use the Graviton instance as the trigger point that says, "Okay, we are here, but this has been going on for 10 years. I've been watching this a

rollout." Can you talk a little bit about the current ecosystem support for Graviton2?

Dave Brown: Yeah, absolutely. And that was one of the big reasons why we actually launched Graviton1, or just Graviton as was called at the time, was really despite the ecosystem. And so while ARM had done incredibly well in the mobile space and there was a growing ecosystem as you call out for many years, we needed to let the world know that we were going to be bringing ARM to the cloud and to the service space because it was another set of applications that really needed to migrate and really needed to move. And that was one of our biggest surprises, honestly, with the Graviton1 processor was firstly how popular it was with customers. We had debates internally at the time, is this going to be a university project? Are our customers really going to adopt this? And customers really did adopt it in ways that even surprised us.

But the other bigger surprise was really the way that the ecosystem adopted it. And it really was a signal to the world that ARM is going to being in the service space and in the cloud. And all of the open source repository started to move very, very quickly. We started to see ISV's moving and we started to see things like agents for logging and all that sort of thing, starting to move to support Graviton1 as well. And what that really did was by the time we came around to launching Graviton2, or announcing Graviton2, about a year later and then launching the first few instances about three months after that early 2020, the ecosystem was really there. So when customers started to think seriously about porting their applications, they weren't slowed down by having to wait for various applications or open source products. And so today most of the major Linux-based, even the BSD based, operating systems support Graviton2.

We have a growing list of ISV's from security to monitoring, to CICD containers, orchestration software, just to name a few. They've all added support for Graviton2.

So internally key AWS services like Amazon EKS, ECS code bold, code commit, Amazon inspector also support Graviton2 based instances today and that's a list of services that you can expect to grow over time as we look to bring Graviton2 to more AWS services.

Patrick Moorhead: The amount of services that you're supporting, it was an indication to me as an industry analyst that Amazon is all in, on Graviton. It made a big impression on me. So let's get down to brass tacks here. For a customer, what advice would you give to them looking to adopt Graviton2?

Dave Brown: Okay. Well, the first one, obviously the most simplest way, is if you are using a fully managed AWS service with a Graviton2 instances, that's always the easiest way and the easiest place to start. And so that would be something like move one of your open source databases to RDS using Graviton2. And there, you can see price performance improvements of 30 to 35% on some of those databases, which is really, really great. Secondly, is find some simple service that you have

on your side where you can take one to two developers and kind of put them on the problem for a week or two and see how much progress that it'll take. Customers often, as I said earlier, try and second guess how it'll take, they don't get started.

They think it's super complicated, but just putting an engineer on the problem for a week or two and seeing how much progress you can make really is eye opening in many cases. And many of the customers I worked with, that engineer has actually gotten the project or the program or the application migrated in that week or two's time. And then obviously with the Graviton challenge, the four day plan lays out a blueprint for how to be successful and how to successfully adopt Graviton2. We also have a free trial. And so we've actually made a free trial available. This is outside of the normal free trial that we have for AWS, where you actually get our T4G instance, which is one of our Graviton2 based instances, where you can actually take that absolutely free and run that instance and actually port your applications. And we're going to keep that free for some time so that customers can continue to easily play around with Graviton2, without having to incur any cost at all.

Patrick Moorhead: Yeah. I can't wait to see the outcome of the challenge here. So Dave, we've talked a lot about general purpose processors. We started off with the nitro layer, but you're also building chips for machine learning, training and inference. Can you talk a little bit about that?

Dave Brown: It's just grown incredibly over the last couple of years, and we've seen almost all of our customers using machine learning in some way, but one thing has also happened in that time is the models have become much more capable, but also more complex. And they often require a lot more processing power. And with that processing power often has increased cost as well. And so we saw a number of customers struggling with the custom machine learning because they were saying, "I'd love to be doing more, but really it's just it's cost prohibitive." And the thing that's really expensive in most of the use cases and makes up about 90% of the cost is actually inference. And so once you've built your model, you deploy the model as called inference and actually doing that, as I say, it makes up about 90% of the costs.

So we launched a custom silicon chip designed by AWS as well, called Inferentia. It's really targeted at reducing the cost of inference for customers. And so with Inferentia, which launched reinvent 2019, customers can see about a 40% price performance improvement as well, so it's the 40% number, when they move from a GPU instance to the Inferentia instance and actually move their workload there as well. And so that's been very, very good for us. The other thing we've recently announced at reinvent last year is a new chip that'll be coming out later this year called Trainium, which will be targeting the other half of machine learning. And so that is all focused on training and reducing the cost of training and we hope in about as much as what we saw with Inferentia. So we're working to get that out at the moment. All of these chips are all deeply integrated with the existing frameworks that are out there, right?

So where the customers are using pie charge or Amexnet, or TensorFlow, it's relatively easy to actually go ahead and integrate with Inferentia. And obviously the same thing with Trainium. We've had customers like Autodesk and also our own Alexa actually use Inferentia for a wide range of ML use cases. In Alexa's case, they were actually doing all of the sort of responses that Alexa makes to help her voice sound more humanlike. They actually moved that from an GPU based workload or use case to Inferentia. And they actually saved about 35% in cost. And they also improve latencies by 20%. And so Alexa responds just a little bit faster because Inferentia is helping her talk in the background.

Patrick Moorhead: I've actually talked to your Alexa customers, your internal customers, twice. And one thing that I called out was that the performance gets has gotten better over time and not every chip is like fine wine, but can you talk a little bit about how you get better performance out of this same chip over time?

Dave Brown: Well, absolutely. So there's obviously the chip we built the chip for ML inference that delivers the lowest cost for machine learning inference in the cloud today. And obviously we achieved a certain level of throughput and performance with that chip when we started. And then obviously a large part of continuous innovation on that is working on the various SDKs. And Inferentia actually comes with an SDK called neurons. We call it the neuron SDK, and we actually do new releases of that neuron SDK every quarter. And in those releases, we're adding new features, support for new models, better integration with existing frameworks, but also more importantly optimizations for Inferentia chips. So we have a team of engineers and hardware engineers. They're not changing the chip.

We've obviously built the chip and got it out there, but they're able to do things through the SDK and the way that it uses the Inferentia silicon to actually achieve better performance, even with the same chip over time. And so there's been a significant improvement since the launch and that's something that we would to continue. So it's really about driving innovation in hardware, not just in the hardware, but also in the software that uses that as well.

Patrick Moorhead: Yeah. It's been fun watching the evolution of Inferentia and I cannot wait to see more details on Trainium. We got some little details, but I can't wait to see more as you start to roll this out. So one really interesting thing at the same time as you're doing your own first party silicon, you are announcing instances that are first to market with many of your other partners out there on the open market. Can you talk a little bit about the future of how do you envision the EC2 offerings based on these different processors overall inside of AWS?

Dave Brown: Absolutely. So our mission in AWS is to offer customers choice, to get the best instance for the workload that they want to do. And if you look across our instances, we actually just recently crossed 400 instances across our portfolio, right? With every combination of compute memory, networking, storage capabilities, and then also obviously choice of processes and accelerators in the case of Nvidia and some of the GPU's. And so we have a great history of

collaboration with our silicon partners, including Intel, AMD, and Nvidia. And we continue to add more whatever they have available, whatever's coming out. And we really strive to bring the best of that to our customers as quickly as we possibly can. And we see that continuing we'll continue to innovate internally on behalf of our customers. And we'll also continue to work with our partners to bring the very best silicon that they have available to market. As Amazon EC2 instances. Our customers rely on us to innovate at a very fast clip in the cloud, right?

That's one thing that's really set us apart is just how fast we can innovate and how fast we allow our customers to innovate on top of us. And so we're going to stay focused on that and we expect the innovation to continue at the fast pace that you've seen to date.

Patrick Moorhead: Yeah, it's funny. I sometimes get questions from people that say, "Do we really need all this choice?" And I explained to them, "Well, that's actually what enterprises want is, is they do want a lot more choice and really just the education to get them there." One thing that I found that enables you to offer all these choices is scale because scale matters in this game. I spent a bunch of years in the chip industry and you have to have scale to have this many choices and do this effectively. So hats off to you. Started off with nitro, moved into general purpose, now we're moving into machine learning, inference and training. Super, super exciting. So, Dave, I want to thank you for what I know is going to make day four of the Six Five Summit a lot better. And so appreciate breaking the news about the Graviton challenge.

I cannot wait to see some of the customer stories that come out of that, the people who are going to fight to be there in two days or the people that want to get those prizes and become the rock star and get up on the stage at AWS reinvent. So thank you very much.

Dave Brown: Well, Pat, thanks to you very much. We're also just as excited for the Graviton challenge and to see what happens. It's going to be really exciting and it's been a pleasure to be with you today. Thank you for having me.

Patrick Moorhead: And Dave, thanks again. So this is Pat Moorhead with Moor insights and Strategy, hoping that you a very great day four of the Six Five Summit 2021.