Jillian Forde (00:00):

Hey, everyone. re:Invent is coming right around the corner, so we're going to give you a recap coverage of the keynotes that are happening every single day, starting Monday, November 27th, with Peter DeSantis keynote, and all the way through Thursday, November 30th, with Werner Vogels keynote. So starting in December, we're going to be doing deep dives into all of these releases from the re:Invent keynotes. So, you'll definitely want to make sure you've hit subscribe on the AWS Podcast, so you can keep up with all of the exciting announcements that happen at re:Invent.

(00:41):

Welcome everyone to the AWS Podcast. I'm your host for today, Jillian Forde, and this one's for all the GPU lovers out there, if you've want to use some P5 instances and be able to reserve them for a short duration of time, then this episode you want to stick around for. Because we're going to be talking about a really exciting launch that happened this year, which is Amazon EC2 Capacity Blocks for machine learning. And I've got the one and only expert on this topic, Jake. So Jake, why don't you introduce yourself to the AWS Podcast listeners and what you do at AWS?

Jake Siddall (01:21):

Hey, sure. Really excited to be here. So yeah, my name's Jake Siddall. I'm a senior product manager within the EC2 part of AWS. Been at Amazon for a little bit over two years now and I work on our capacity products. So my team works on products like On-Demand Capacity Reservations. And I was the product manager for this new product we recently launched that we're going to talk about now, EC2 Capacity Blocks for machine learning.

Jillian Forde (01:45):

This is super exciting. So, let's get right into it. So what exactly are EC2 Capacity Blocks?

Jake Siddall (01:51):

Capacity Blocks are a new Amazon EC2 usage model. They are meant to further democratize machine learning by making it easy for any customer to access GPU instances to train and deploy machine learning and generative AI models. So with EC2 Capacity Blocks, you can reserve GPU instances for the amount of time that you need to train and run your machine learning workloads, meaning that you have capacity when you need it, without having to hang on to GPU instances all the time when you're not actually using them. Capacity Blocks are currently available for P5 instances, which are powered by the latest NVIDIA H100 Tensor Core GPUs and Capacity Blocks are also currently available in the U.S., East Ohio region. Capacity Blocks are the first and only reservation model available in the industry, today, where you on your own can go and schedule GPU instances to be available on a specific feature date for just the amount of time that you need those instances for.

Jillian Forde (02:45):

Yeah, I mean, I really can't wait for customers to start being able to use this. Tell us more about why it is that the team decided to actually create this product.

Jake Siddall (02:54):

There's a bunch of excitement right now in the machine learning space, due to some of the new ways that foundation models are being applied in both consumer facing and enterprise applications. This has driven a bunch of growth and demand for GPU capacity to train and fine tune machine learning models,

to run experiments, and to handle surges and demand for models that customers have deployed in their applications. So all this demand has outpaced industry-wide supply, making GPUs a pretty scarce resource and hard to come by to support these workloads. So access to GPU capacity is, the biggest obstacles that many customers are facing when they want to train and deploy these models for their applications. The problem is especially challenging for customers whose capacity needs for GPUs fluctuate depending on the research and development phase that they're in, because the demand for GPU capacity has resulted in a lack of really flexible options to just spin up a GPU cluster for a short period of time to run a job.

(03:47):

We built this product, so that customers have a more flexible option to predictably scale a cluster of GPUs for just a short period of time to run a small set of machine learning jobs. Before you had this product, customers who needed assured access to GPU capacity would've to reserve capacity just up to their peak needs and hold onto a lot of underutilized capacity, or underutilized GPUs in between the periods when they would have peak demand. So, this led to a lot of waste for a lot of customers who had these fluctuating needs for GPUs. And this waste was, it was expensive, customers would've to be paying for capacity that they weren't actually using and this was cost prohibitive for a lot of customers. So, we think that this new product is going to be more cost-efficient for customers who have these fluctuating needs for GPU instances.

Jillian Forde (04:30):

Yeah, there's so much really to unpack here. So let's really dive into the actual use cases, because that these Capacity Blocks, you reserve and it's for a short period of time. So what are some of the use cases that you're seeing, where it makes sense to use this for a short period of time? And then if customers actually need this for a longer period of time than that, what do you recommend?

Jake Siddall (04:53):

The use cases where Capacity Blocks make sense varies from, depending on where customers are in their research and development cycle. So Capacity Blocks make sense for the initial experimentation and prototype portion of development of a machine learning model. We have a range of Capacity Blocks, size and duration options that customers can choose from. So, they can start with a smaller size Capacity Block for a shorter duration when they're just running their initial experiments and building their early prototypes. And then as they progress and they move towards training a larger AI model, they can reserve a larger block of time, like a larger cluster for more instances, for a longer period of time to train an AI model. Capacity Blocks works really well if you are looking to train like a task specific model. And then after you have your model, Capacity Blocks also are a good fit if you want to fine tune your model with new data that you've collected over time.

(05:44):

And then lastly, Capacity Blocks can also work well if you have expected surges in demand, if you have a model that is deployed. Say, you're releasing a new feature or launching a new product and you expect a surge in demand for the days or weeks that follow this new product release, you can use a Capacity Block to scale up temporarily to meet that surge in demand, without having to make long-term commitments and reserve capacity for a long period of time.

(06:09):

I do just want to call out that we have another way to reserve capacity as well, On-Demand Capacity Reservations. So, Capacity Blocks aren't meant to be a substitute for On-Demand Capacity Reservations.

Capacity Reservations are great if you have long-term steady state usage needs for GPUs. You can cover your baseline usage with an On-Demand Capacity Reservation, offset that cost with a savings plan or a reserved instance discount, and then you can supplement that On-Demand Capacity Reservations with Capacity Blocks just when you need to burst above your baseline usage needs. So, you can basically reserve Capacity Blocks to tax for your year, to provide capacity assurance for your peak needs, just when you need to scale up to that peak capacity.

Jillian Forde (06:49):

That is such a good call out of being able to use an On-Demand capacity Reservation, and then being able to stack on top of that the EC2 Capacity Blocks. Because at least I know from my experience of working with customers, I'm also seeing a lot of companies are starting to figure out, okay, if my product's launching here and maybe we're an E-commerce company, I need to be able to scale out with our launch. How is it that we can do that when we're actually being able to use these GPU based instances? I really love all the different options here that you're really presenting for customers to be able to actually scale their GPU based needs.

Jake Siddall (07:25):

Yeah, definitely. We think that Capacity Blocks are going to allow customers to operate more flexibly when it comes to GPU capacity, in a way that they really haven't been able to anywhere in the industry for the past year or more.

Jillian Forde (07:37):

It definitely makes sense to me. I mean, hearing about a lot of these benefits, now that you can really start to get more creative and how it is that you're planning your capacity. And especially as our listeners here are thinking about, okay, 2024, what are their AI and GPU compute needs going to be? But I'd love for you to also call out, maybe there might be some other things that really are other benefits that you are seeing. So, is anything else that comes to mind?

Jake Siddall (08:03):

I do think it's just important to call out that the main benefits of using Capacity Blocks would be, you can ensure that you're going to have the GPU instance capacity that you need for your machine learning developments on a specific feature date, ahead of time. We know that a lot of customers today, if they're asking for GPU capacity anywhere they go, a lot of times, it can be kind of unclear when they're actually going to receive that capacity, which makes it hard to plan their machine learning development cycles.

(08:29):

So with this product, you get a lot of transparency in terms of when you're actually going to have that GPU capacity that you need, so you can plan around that time and avoid wasted time. You can allocate your engineering resources, your engineering effort very efficiently that way. Secondly, this product provides predictable access to P5 instances, which are, I think I mentioned before, they're powered by the NVIDIA H100 GPUs. And, these are the highest performing instances offered by EC2 for deep learning. So, you can access these instances now. Any customer can come in and access these instances through Capacity Blocks, without making any long-term commitments. And they can do this easily on their own through the EC2 console, through the click of a few buttons. And then lastly, Capacity Blocks are always delivered inside of an EC2 UltraCluster, which provide low latency, high throughput network

connectivity between all the instances that are part of that Capacity Block. And, that's just ensures that customers have the best distributed training performance available on EC2.

Jillian Forde (09:26):

This is super cool. All right, so lets kind of get into how it actually works for a customer that wants to make a reservation. So, can you walk through what that's like?

Jake Siddall (09:35):

I think a good analogy here is that Capacity Blocks work kind of similar to booking a hotel room, or creating a reservation in a hotel room. So with a hotel reservation, you specify the date and duration that you want that room for and the size of the room that you want, so the number of beds that you want. For example, you might want a queen bed or a king bed, a couple of twin beds. Likewise, with Capacity Blocks, you can select the date range and duration that you require, the date range that you need a reservation within, as well as the duration of time that you need that reservation for. You can also specify the size of the reservation that you need, so that's the number of GPU based instances that you require. And that's kind of how you search for an available capacity block that you want to reserve, that works for you.

(10:18):

And then once you actually reserve, once you find a Capacity Block and you reserve it, on your reservation start date, you'll be able to access the reserve EC2 Capacity Block. You'll be able to launch your P5 instances into that Capacity Block. At the end of the reservation, EC2 will terminate any instances that are still running inside of the Capacity Block. And that way, we can make sure that, that capacity is available for the next customer who has reserved the Capacity Block. And that way, every customer who uses Capacity Blocks will get really reliable on-time delivery of their capacity at the time that they've reserved it.

Jillian Forde (10:50):

Wow, that is really super cool and I love the analogy of being able to reserve it like reserving a hotel room. Tell me more about some other customers who are using this and share some of those use cases.

Jake Siddall (11:06):

We've been working with a company called OctoML. They are an AI platform that help customers easily run, and tune and scale generative AI applications. They help customers optimize model execution using automation to scale their services and reduce energy burden. We've been working with them on Capacity Blocks, because they think that Capacity Blocks are going to be super useful when their customers have a new product launch that requires temporary bursts of GPU based capacity, or GPU based instance capacity to support large scale inference workloads.

(11:36):

So, they're super excited about Capacity Blocks. We've been working with them to help them get started. The feedback that we've gotten from OctoML is that this is going to help them operate more efficiently, which is kind of in line with the mission of their company. We've been working with another company called Snorkel. Snorkel provides a data development platform for AI workloads. It helps enterprises quickly create and use AI. Capacity Blocks helps Snorkel scale up for bursts of GPU capacity to train models over the course of several weeks, and it helps them make sure that they're going to have these bursts of GPU capacity to meet the timelines that they've promised to their end customers.

Jillian Forde (12:09):

It's super interesting. And I think a theme that I think we're going to really uncover from the rest of this conversation that we're having today is really, as people are planning out, okay, what is 2024? What is your AI strategy? I know a lot of people are thinking about, "Okay, how does generative AI fit into their business?" And now with Capacity Blocks, how far ahead are you going to start scheduling it? And, I think that can lead into some other questions that I've got too, related to how people can think about pricing and that can help. And of course, the farther out they are from being able to plan it out, the more, I guess, strategic people can be in really optimizing just their overall cloud costs.

(12:50):

I did have some other questions that I want to make sure that I get in. So this is probably one I'm sure that people might be wondering, as they're hearing about Capacity Blocks and reservations. I'm sure there's a scenario maybe where there's someone who's doing a machine learning training job. Let's say they've scheduled it for 10 days and now on the 5th day, they realize, "You know what?" "It's just not going the way it was," that they thought it was going to be. So if they cancel the reservation, do they have to restart the job? Do they have to create a new reservation? Can they modify the existing reservation?

Jake Siddall (13:27):

Yeah. Something that's important for customers who are using Capacity Blocks to understand is that once a Capacity Block is reserved, it cannot be canceled or modified. So, would recommend that customers who are using Capacity Blocks reserve the minimum amount of time that they think they'll need to start. And then if they realize that they need more time during their reservation, while they're running their workload, while they're training their model, then they can easily go back into the console and reserve another block of time to finish the workload later on. You can look at different Capacity Blocks before you actually reserve anything. You can see the different start dates and the different prices of Capacity Blocks before you reserve, so that allows customers to get really comfortable and make sure that what they reserve is going to meet their needs, before they actually make that reservation commitment.

Jillian Forde (14:12):

Got it. Okay. And I know we've got a few minutes left, so let's get into really explaining how the pricing works. I think that can definitely help people to understand how they should think strategically about being able to use Capacity Blocks.

Jake Siddall (14:24):

Yeah. The price of the Capacity Block is dynamic, and it depends on supply and demand at the time that you book it. So you pay for a Capacity Block all upfront and you reserve it. And then during the reservation time, the only thing that you would pay for is if you're using a premium operating system, while you're running your instance, you pay a small operating system uplift fee based on your usage. And then to determine the price of the Capacity Block itself, we look at the demands on each day of our eight-week reservation horizon, and we look at the demands on each of those days. So on days when we see more demand, relative to other days, the price is going to be a little bit higher for the unbooked Capacity Blocks on those days.

(15:01):

And then on days when we see less demand, the price will be a little bit lower. This is the price for just unbooked Capacity Blocks. Once you actually book or reserve the Capacity Block, that price that you saw, that you were shown at the time that you reserved it, that's locked in, that will not change after you book it. The reason that we have this dynamic pricing is to try to make more capacity available to customers. Its, we want to try to kind of match the supply that we have with the demand the customers have, so that there's more available capacity when customers need it. To get the lowest price that you can with Capacity Blocks, I recommend that you search across the widest date range possible when you're actually looking for an available Capacity Block, because we always show you the lowest price block in the time range that you've provided, that matches your reservation specifications.

Jillian Forde (15:44):

And even better for customers to really start thinking way in advance, so they can really start to be more strategic of when it is that they can start to actually do their planning and making sure that they have the compute they need at the lowest price possible. Love it. Okay. So, re:Invent is coming up. Any sessions that you want to share that you're doing?

Jake Siddall (16:06):

Yeah, I have a session. If you look up the code, CMP105 in the re:Invent app. It's going to be on Monday at 11:30 AM in The Venetian. So, we're going to deep dive into Capacity Blocks a little bit more, just how customers can kind of optimize, ensuring that they have GPU capacity available for their machine learning workload needs.

Jillian Forde (16:24):

Super cool. And for all the listeners here, we're going to be doing deep dives of all the launch announcements that happen at re:Invent, so that entire week, you'll definitely want to listen to the AWS Podcast, make sure you've got it downloaded. Hopefully, we'll see that there's exciting announcements from Jake's team. We'll certainly find out. But Jake, anything else that you can just share with the listeners about EC2 Capacity Blocks?

Jake Siddall (16:49):

It's pretty early on with Capacity Blocks. We just launched roughly a week or two ago. I mean, we launched this product at the end of October, so we are still working on some exciting new things based on the feedback that we've been collecting from customers. We know that a lot of customers have been asking us, "When Capacity Blocks will support some other instance types and when we're going to roll it out into some different regions?" We're taking this feedback in and we're planning on introducing some feature enhancements that we think are going to be valuable to some more of our customers.

Jillian Forde (17:15):

Super exciting. I'm really excited for all the listeners here. Any place where people can go to go and get started?

Jake Siddall (17:22):

Can get started in the AWS Management console. If you go into the Ohio region, you can navigate to the EC2 console, click on Capacity Reservations in the navigation panel. That will take you to a screen where you can get started with the Capacity Blocks. We also have a Command Line Interface, or APIs that you can use through the Command Line Interface, or the AWS SDKs. I'd recommend if customers want to get

started, you can go to our user guide documentation for Capacity Blocks, or you can check out a news blog that we published on October 31st when we watch the product. It gives a nice walkthrough of how to get started with the Capacity Blocks feature.

Jillian Forde (17:54):

This is awesome. Jake, thank you so much for being here on the AWS Podcast.

Jake Siddall (17:59):

Yeah, thanks a lot of having me, had a good time here.