EPISODE 1270

[INTRODUCTION]

[00:00:00] JM: AWS Outposts is a fully managed service that offers the AWS infrastructure, AWS services, APIs and tools to virtually any data center, co-location space or on-premise facility for a truly consistent hybrid experience. AWS Outposts is ideal for workloads that require low latency access to on-prem services, local data processing, data residency and migration of applications with local system interdependencies.

In this episode, we talk with Joshua Burgin, general manager of AWS Outposts at Amazon Web Services. Joshua own strategy, roadmaps, customer experience, pricing and demand generation for the AWS Edge/Hybrid Compute business, including full P&L responsibility. Joshua was previously a senior director of technology platform and services at Zynga and a senior manager of product building at RPI before that.

A few announcements before we get started. One, if you like Clubhouse, subscribe to the club for Software Daily on Clubhouse. It's just Software Daily. We'll be doing some interesting Clubhouse sessions within the next few weeks. Two, if you are looking for a job, we are hiring a variety of roles. We're looking for a social media manager. We're looking for a graphic designer, and we're looking for writers. If you are interested in contributing content to Software Engineering Daily, or even if you're a podcaster, and you're curious about how to get involved, we are looking for people with interesting backgrounds who can contribute to Software Engineering Daily. Again, mostly we're looking for social media help and design help. If you're a writer or a podcaster, we'd also love to hear from you.

You can send me an email with your resume, jeff@softwareengineeringdaily.com. That's jeff@softwareengineeringdaily.com.

[INTERVIEW]

[00:01:50] JM: Joshua, welcome to the show.

[00:01:51] JB: Hey! Thanks for having me, Jeffrey.

[00:01:53] JM: You work on AWS Outposts. We've done a show on Outposts before. Can you give a brief overview for Outposts and the typical use cases for them? Then we'll go deeper on them.

[00:02:05] JB: Sure. At a high-level, Outposts is a fully managed service from AWS, and it offers the same in AWS infrastructure, services, APIs tools that you can deploy in virtually any customer data center, or co-location space or kind of an on-premise facility and you get that truly consistent hybrid experience of using AWS but in locations that you couldn't use it before. The typical use cases we're seeing are honestly a pretty wide set, but fundamentally, we designed Outposts to serve workloads where ultralow latency, data residency or local data processing are really important. Anyone with these workloads, you're going to find yourself with the need to maintain compute and storage on-prem or on a co-lo.

[00:02:51] JM: Can you give a little bit more detail as to the use cases, like maybe give a few examples?

[00:02:55] JB: Yeah, sure. We're already seeing a lot of those kind of use cases that I mentioned in market segments like financial services, retail, telecom, gaming, manufacturing, healthcare, life sciences and also in — I mean, as you can imagine, the public sector including our government and educational customers around the world. To kind of dive into that a little. One of the things I found interesting is that, even customers that are already operating on AWS, some with quite large deployments, they often still have workloads that need to remain on-prem for the foreseeable future.

One example might be a customer like Riot, who make games like League of Legends, and they just released Valorant and Wild Rift. They want to deliver amazing gameplay experiences around the world. They're using Regions where we have them, but they also need to use Outposts or they found a use for Outposts, and they're actually exploring some of our local zone offerings as well. So that no matter where the player is, they get that kind of consistent under 25 millisecond latency. That's an example where if we had a Region in every country of the world or everywhere inside the United States, maybe they wouldn't need to use the Outposts.

On the other end of the spectrum, you have somebody like Morningstar, which is one of the world leading financial services companies. They're looking at Outposts as a bridge. Long-term, their vision is that most of their applications run on AWS in the Regions, with some of the applications that need to remain co-located with on-prem infrastructure, like a mainframe, those will stay on the outposts. But that bridge and that migration is going to be maybe three, five, ten-year journey and so they didn't have to choose between waiting until they modernize all their applications. They can get started with Outposts in a way that feels pretty similar to what they're doing on-prem. Then kind of migrate things, kind of a tronch at the time and pick them whether they stay on the outposts or move to the region. They get like that huge leg up by embracing AWS APIs and services through Outposts as part as this long-term migration.

[00:05:01] JM: If I set up an Outposts in my on-prem deployment, what is the division of labor between the outposts and the cloud? Because Outposts is obviously interfacing with remote cloud resources.

[00:05:16] JB: Yeah. I mean, that is an important detail. It's actually part of the core value proposition. We designed it to remain connected to the AWS Regions so you get the benefit of all of the services we have there. Fundamentally, your hardware is obviously local and once your application is up and running, that application is running locally on the EC2 instance and EBS volumes. Your S3 bucket data is all local. None of that leaves the outposts, unless of course you tell it to leave the outposts. You can send data back to the region if you want. When you make mutating calls, that's like if you have to make an API call, we have to launch new instance of starting and stopping an EMR or an EKS cluster. That's the container services, the EKS from Kubernetes. Those reach back to the regional control planes for those services.

I mentioned the core value proposition of that consistent hybrid experience. So having that regional control plane is actually one of the ways we offer more services to customers and reduce the overhead that you need on-prem. You don't need dozens and dozens of racks just to run the services. You get all of the compute and all the storage for what you're actually doing locally. For most customers, even people in banking like FAB, First Abu Dhabi Bank and they're using Outposts in the UAE region to kind of do both, disaster recovery and meet local business continuity requirements for the — they're in banking obviously. They're still pretty happy with the fact that only the metadata like instance IDs and so forth leave the outposts. Having a control plane nearby, but not on the outposts as long as they have kind of good connectivity and redundancy, that still meets all their data residency requirements.

It's a little different, if I could dive in if that's okay. We also offer series of services through our Outposts Service Ready program. Those are about 55 or 60 different ISVs who offer services. A lot of those are available in the AWS marketplace in our Regions as well. Of course, a lot of these services run on-prem. A lot of these services and I'll kind of give you a list, they're on locally on the EC2 instance, there's nothing running in the region. That can be like networking, or backup services from people like Commvault, Veritas. Pure Storage offers their FlashArray and FlashBlade services. NetApp, those are storage services. Jenkins, Datadog, Terraform, Dynatrace, PagerDuty. Some of these services run entirely on your Outposts or SaaS, so they might be running in a Region. Trend Micro for security, Sisense for analytics, **[inaudible 00:07:52]** DB and Mongo, our database services, those are running directly on your Outposts.

Depends how the service kind of actually is designed to run and what makes sense, but it is a really good question. It's different than your traditional purchase a server from a vendor, and then put a hypervisor on top, and then deploy services yourself. We kind of remove all that undifferentiated heavy lifting. One of the ways we do it is by connecting back to a Region. You can connect to actually to any of our 23 Regions around the world outside China. As long as the latency works out for you, you can actually connect to regions that are not immediately next to you.

[00:08:36] JM: Tell me more about the networking infrastructure that goes into an Outposts and what you're actually looking to optimize for. Are you looking to optimize for connectivity to a variety of Regions or are you looking to optimize for latency? How do those different trade-offs in network infrastructure play out?

[00:08:59] JB: Networking is one of the most interesting and challenging spaces as you can imagine. There's a wide variety of configurations that we're dealing with someone's legacy on-prem data center, to a high-end co-location facility. We actually have a program where we work with these co-los to certify the work for Outposts, which makes the installation easier. We kind of look at their security and operational posture, as well as their networking configuration. There's kind of two different pieces to the Outposts in terms of hardware when you're thinking about networking. The first is we have redundant, top of the rack switching gear, which is reusing much of what we have developed over the last 15 years to operate at scale on our regions. We can support a really high degree of interconnect performance between the EC2 instances and the storage that's running inside the rack, so that's sort of one thing.

We all support 110, 25, and 100 giga bit per second networking uplinks, so you can configure this a bunch of different ways. You can use our direct connect service, which is kind of dedicated, don't have to leave the VPC connectivity back to one of our Edge locations. You can send outpost traffic directly out over the Internet. You can connect to pretty much anything inside of your own network. Again, we support that construct I mentioned earlier, the VPC, which is what a lot of our customers tell us is pretty important. They want the security if they're connecting back to the AWS Region, that all of their resources are in that secure private network and can connect to services that use private link, which is a way of extending other services such that they appear to be inside the customer's VPC.

Again, if you're familiar with AWS, the Outposts should feel very familiar to you as well. Your resources are inside the VPC. If you have multiple Outposts, you can configure them so they're in different VPCs connected to different availability zones for reliability and redundancy. You can use direct connect. But if you're not using AWS, that doesn't make sense for you, you don't have to do that. You could configure your network security and the layer two connectivity inside of your own switching gear in your data center as well.

I haven't run into very many situations we can't support. There are always more features to build, whether it's network segmentation or self-service configuration for customers in terms of their networking set up, BGP support and so on. If you look at customers in the telco space, which are some of the most demanding in terms of network, I think that's why you end up with people like Dish who just announced the joint partnership with making pretty big bets on outposts and infrastructure that's enabled by Outposts, like our local zone offering to deliver nationwide cloud native 5G network built on open RAM.

[00:11:58] JM: I remember asking this question last time I did an interview about Outposts. But as far as delivering a holistic AWS experience using outposts, it's kind of a tall order, because there's so much going on in the cloud and replicating all that. Through Outposts infrastructure, it would be tough to do. How far along is that Venn diagram of AWS infrastructure that's available on Outposts versus that which must be accessed through the cloud?

[00:12:28] JB: I mean, that's a good question. We're fond of saying here that it's still day one. The way I would look at it is, it's very early for the Outposts business and it's honestly still pretty early for AWS if you look at total technology spending, a significant percentage of it. Somewhere between 75% and 96% might still be considered to be on-prem. As AWS continues to evolve Outposts will along with it. I think to start with the — the overlap between the hardware is actually pretty far along. We built this from the ground up, so that it reuses and it makes use of our nitro system, which is the custom silicon offload cards, the virtualization stack is on that. We have the security chip and then the lightweight hypervisor. That's exactly the same thing you're running in Regions. The EC2 instances and over the next year I think what you'll see is that, pretty much all of EC2 instances are going to be available on Outposts in the rack footprint as well.

We can support it, I don't know if it was clear in the last interview, but each Outposts rack doesn't have to be a separate logical Outposts. It can be if that's what customers want. But you can bundle together multiple racks up to about 96 racks right now to be a larger logical Outposts. In many ways, that rivals some of our largest installations as well. The EC2 instances, the EBS hardware, the networking, the Nitro system, the power supplies. That's all really similar to the region. We were in complete lockstep with our engineering and infrastructure teams. I think that's a real benefit to customers, as they don't really have to think about that. It's not a universal overlap. There are some instance types like our P4, instance types for machine learning training. We don't see a lot of people doing that on-prem right now. You need a really large installation of that for it to make sense. Most of the customers that we're talking to are like, "I'd rather do that in Region and kind of make it your problem."

Obviously, there's some differences with the hardware, because we're in a customer's location. Physical security is little bit different. We have a fully enclosed rack with tamper detection, inside the region of course and our availability zones and data centers, We're handling that for people. Security is a little bit different. In terms of the service basket, though, what we've been focused on in kind of the short run is that, what are the core services that people want. Over the last year, I think we've shipped a pretty decent amount of those, EC2, EBS, S3 was a top asked for service. EMR, EKS, ECS, ALB, ElastiCache, Cloud **[inaudible 00:15:09]**. These are the things that we hear from almost every customer. RDS as well. We just announced SQL Server. Before that we, had Postgres and MySQL support. I think we're going to continue in the next 12 to 18 months to really focus on the core services that everybody tells us they want.

A lot of folks use these third-party services on-prem, and so they're looking to continue using those, so that's why we had the Outposts Service Ready program. Then of course, you can already use many of the services, especially the management and governance ones that people are already using in Region. On Outposts, you can put an instance in autoscaling group. You can use CloudTrail or CloudWatch. These things kind of just work. I wouldn't say we're anywhere near done. I think there's — this has been my experience and I've been at AWS seven years and Amazon, a total of 10. When you talk to customers, nobody is ever fully satisfied with what you've done and they're always giving you new ideas. In the Outposts case, for new form factors, new services. We're going to continue to do things over the next 12 months to up the game there.

I don't think we'll ever offer all 185 or 200 plus AWS services on Outposts. Customers aren't really asking us for that. In that case, we also have the local zones offering, which are built on Outposts and Wavelength as well, which is our 5G-enable edge compute offering. Those are already built on Outposts. They have a wider range of service offerings and they're installed in metro areas. We have four of the local zones right now, and we've announced that we'll have support for 15 of those across the US this year.

We hear more from customers. If they really want a big basket of services, they're either pretty happy running in the region or they would like us to run a more elastic local zone, somewhere near them. They don't want to necessarily have an entire data center replicating what they could get from AWS.

[00:17:08] JM: You are the general manager for Outposts, and that involves a lot of moving pieces. Can you tell me about how the team is structured and how you manage it?

[00:17:18] JB: I can give you a little insight there. I mean, the general manager role at Amazon emerged from the concept of many years ago, what we used to call two pizza teams and single-threaded leadership. Where we really make sure that each team can operate relatively independently and be loosely coupled with other teams and communicate through hardened APIs rather than stuffing everything under single larger groups, which tends to slow things down especially on the innovation front. The GM role can be different from team to team, but fundamentally, you're responsible for the entire business, which includes product and engineering, and business results and so forth. The rest of the structure, we do what make sense for employees and their careers. It can fluctuate from team to team.

[00:18:08] JM: But I'm curious at a deeper level if you can talk about it, like how do you orchestrate all the cross-functional work? Because there's so much hardware, and software integration and moving pieces. I'm just curious about the details a little more.

[00:18:25] JB: Yeah. I mean, what I can share, that's pretty common at any larger enterprise. We have good relationships with our sister teams. At Amazon, there is a process that — and people have talked about this in public before, that we call OP1, is our annual planning process, where we all get together, and stand back, and take a look at the next 18 months and three years and so forth. That's where a lot of the coordination happens. We make sure that initiatives that rely on teams working together get the proper attention and get funded. I wouldn't say this is a solved problem at AWS or Amazon. No different than anywhere else. There's always more that we want to do for our customers than we can get to in any time period.\

But I think what it comes down to is that the teams look at our customers, and the overall business and what are the exciting new areas. I think Outposts is one of those, but it's not the only one. We try and rationalize our investments across all those opportunities. In the case of Outposts, I think one of these we'd benefit from, I've mentioned this earlier, is the fact that it's — I don't want to trivialize the work here, but it's just EC2. It's the same EC2. It's the same EBS. It's the S3. The hardware engineering work, we didn't invent hardware engineering suddenly for Outposts. We've been in the hardware engineering and the custom Silicon business for quite some time. Back in 2015, we purchased a company called Annapurna and there are some of the geniuses and hard-working individuals behind the Nitro system, and Graviton, and Inferentia and Trainium, which are custom chips that give people just amazing price performance.

We get to leverage all of that work. We don't have to go spin up a hardware engineering team or an infrastructure operations team. The fact that we're already operating around the world at scale and have logistics experience and supply chain management experience, Outposts is drafting off of that. I'm not saying it's identical. There's a little bit different work. But that's what I think makes it more possible for us to move quickly. We didn't just start yesterday.

[00:20:39] JM: Describe the different models of Outposts and how they differ from one another?

[00:20:44] JB: We call those — I think what you're talking about is the form factors.

[00:20:48] JM: Yeah. Well, the 1U and the 2U.

[00:20:51] JB: Got you. Yeah. That's a good —back in Reinvent in 2019, what we launched, the first sort of form factor for outposts is a 42U kind of full rack form factor. This is an industry-standard 80-inch tall, 24-inch wide, 48-inch-deep rack. That one takes up one rack position, it can take up more rack positions. If you order multiple racks, it's designed with security in mind, fully enclosed with locks, tamper detection, redundant power supplies, requires between five KVA and 15 KVA of power. There's also a redundant networking built in. I mentioned this earlier. Depending on what SKU you order, there's between 2 and 11 or so. EC2 instances, and EBS and S3 storage per rack. That's obviously people who are in a controlled environment who need a lot of compute and storage, pretty traditional data center co-lo type folks.

What we actually heard after we announced that was — I mean, there are a lot of people who thought this was great. I talked about banking, and gaming, healthcare, life sciences, telco. People said, "This is great. We love it. can you make it even smaller, because we have locations like restaurants, and telco like that edge sites, the cell towers, manufacturing? Could you shrink AWS and allow us to put it in place where we really only need a couple of servers? It might be virtual machines, but a couple of physical servers." 1U and 2U, again, these are just rack units, which is an industry standard way of measuring how much space server takes up.

1U and 2U are much smaller than 42U as the number would indicate, so they fit in the same standard EIA 310, 19-inch racks. Usually, alongside other equipment like other servers or network switches. The way to dimensionalize them is, they're about one and two pizza boxes in height. The 1U is about an inch and three quarters tall, and 24 inches deep, which is kind of shallow. The 2U is three and a half inches tall and designed to fit into 36-inch-deep racks. Both of those have lower power and network bandwidth requirements than a full rack outpost. You're not going to be able to build a full rack or a multirack outposts out of these 1U and 2U devices. It's not really what they're designed for.

Then specifically, what we announced at Reinvent this year or back at the end of 2020 was that the 1U instance is using that Graviton 2 processor. That's like really great price-performance up to 40% better than kind of comparable Intel offerings. These are a good match for what we're hearing from customers where both spaces constrained and they only need kind of precise amount of compute and storage that will fit in 1U. The first versions of these will support up to 64 vCPUs, 128 gigabytes of memory and 4 terabytes of local MVME storage. That's about it. The 2U offering is actually based on an Intel processor, and will support up to 128 vCPUs, 512 gigs of RAM and about 8 terabytes of local MVME storage.

The benefit of the 2U server as well is that we can fit in some optional accelerators like GPUs or our AWS Inferentia inference chip for ML workloads. Can't really fit those in 1U, just to sort of power and space constraint. But that's actually a lot of flexibility if you think about it. You've got everything from as little as a single 1U outposts or you can kind of bundle the 1U or 2U outpost together to get about six of them. Or you can install a single rack, or up to 96 racks. I think that covers a lot of ground with what we're going to be releasing later this year with the small form factors.

[00:24:42] JM: Give a sense for what the overall market demand for Outposts-like systems is.

[00:24:50] JB: That's an interesting question. I think what we've said earlier and what we've said before and I think this is still true. The vast majority of technology spending is still on-prem. somewhere between 75% and 96%. Depending on how you measure the industry, it's between $1 and $4 trillion a year in spending. These are just kind of standard industry terminology or kind of measurements. What I would say is, I think in the fullness of time, most people are not going to want to remain, retain their own data centers and they're going to find themselves migrating the bulk of their workloads to one of our regions. The cost benefits, the operational benefits, the scale, the elasticity, all those services you get, it's just going to make sense.

We've only got 25 regions right now, we'll have more over time. We've announced five more. That's 80 availability zones right now and there's I think 16 more that have been announced, but they're not in every country. So if you kind of go back and say, "All right. Well, what workloads could run on an outpost?" I think you got to give two different answers to this. There are workloads that will remain on the outposts for the foreseeable future. Maybe forever, right? Until we can — as I like to joke, until we can solve the speed of light problem. If you need really low latency, where there's a regulatory reason why something can't leave a facility, or a country, or in the cases of iGaming or mobile betting, you'll have all these municipalities actually saying to people like Tipico, which is one of our Outposts customers, "You can run your application and make it available in our state for mobile devices or online betting, but it has to be in a specific facility, or specific cage or that sort of thing." We can't control regulations. We can only make sure that we comply with them.

I think those workloads or a manufacturing facility where you can't really tolerate downtime based on a connectivity going away, those workloads are going to remain a good target for Outposts for until such point where connectivity or regulations are no longer a requirement. Then what's interesting to me is, the rest of that spending that's on-prem, if you've got a workload where the latency isn't a factor, there's not a data res residency requirements, you don't need to do the data processing locally, like near a hospital or lab equipment, or for an autonomous vehicle garage. Those people might still find themselves with the opportunity to embrace AWS by running the workload on an Outposts. I thing long-term, most of those people will be comfortable moving to a local zone or a region, but is not something we need to be prescriptive about.

The benefit of being AWS and offering people choice, and having Outposts be the foundation of local zones and Wavelength, which we deploy with our telecom partners and the 5G partners around the world like Vodaphone and Verizon, and SKT, and KDDI is — we can serve all of these needs. If you want to run something in your own data center, we've got your back. If you want to run it in a metro area, we'll have local zone there, or if you need a 5G application to support roaming, you can put it in a Wavelength zone. If you get comfortable running it in a

Region like the data warehouse or your HPC cluster might run in the Region or your machine learning training.

I kind of look at it not as an either or and some of it is going to depend on when you are planning to do the migration and what the rules are locally in your country, and what the comfort level is inside your organization.

[00:28:33] JM: What are the hardest engineering problems you've worked on so far with Outposts?

[00:28:38] JB: Oh boy! We love to do hard things at AWS overall. In my last seven years, every time I feel like we've solved whatever the hardest problem was right then and new one emerges because we're never satisfied, because our customers are never satisfied. If I had to answer that question for Outposts, I'd actually say that security is a hard problem. We often say securities job zero here, and although it's a shared responsibility with our customers and we provide services and tooling, when you're deploying AWS infrastructure in somebody's facility, we take the security of that infrastructure and your data as seriously as we do operating anything in our region.

The fact that we already had Nitro, and the Nitro controller and the security was pretty good, but we have people on our security team who think about this all day, every day and we work with them to do things like improve the physical security of the rack itself. I mentioned that it's fully enclosed, there's a locking back and front. There is tamper detection on the rack, there's tamper detection in every service. Anthony Liguori talked about this with you I think last time, but there is a key that you can remove that essentially destroys all the data on the server, because everything is encrypted at rest all the way down to the DRAM. I don't know if I want to make this kind of bold of a claim, but I would say that it's actually at least a secure, if not more secure from an infrastructure level than anything else you can deploy in your own data center.

We didn't just slough that responsibility off on people and say, "It's up to you to secure your own infrastructure." That's actually really hard problem, because security is not solved. There are always new attack vectors, and side channels and things to think about. Again, the good news there is we have lots of people who think about this probably even when they're sleeping. Some of the smartest people I've worked with.

I think the other one beyond security is that — you mentioned this earlier, how do we get the Venn diagram of overlap to be high. Delivering that consistent set of APIs and features while shrinking it down sometimes to a single rack or maybe a single service, that's interesting. It's not easy to think about how you deliver EBS, or ALB or RDS and give all of the features that people want, the ones at least that make sense, including high-availability features with outposts-to-outposts connectivity in an environment where sometimes the network is not managed between the devices. That's not true in our regions, where we obviously fully managed these things and build in redundancy. That's hard because is not the same for every service.

A lot of you mentioned earlier, the GM construct. A lot of my job is to make sure that we build primitives down at the EC2 layer, the networking layer, the EBS layer, communication between outpost. We build primitives that other service teams can use. I don't know if we've gotten this perfect just yet, but we're very conscious of the fact that we have to solve that problem, so it's not a problem for our service teams and then it's not something that the customer really needs to think about. That's what they expect if they're going to get something from AWS.

Maybe one more that just occurred to me is, the last year as we know has been one of the more unusual years in recent memory, due to of course the pandemic. If you think about Outposts, we have to install them. We actually designed this so that they're installed by our own technicians. We operate in — used to be 15 countries, then 20 countries. As of last week, it's about 57. So as you can imagine, some of these installations that we've done in the past year have been tricky because of COVID restrictions. I mean, we have to take the safety, the physical kind of health safety of our customers and our technician super seriously. Of course, we have to meet all the regulations that keep changing and we've got to meet customer deadlines.

Figuring that out, which maybe that's a temporary problem, that's been interesting. Installs in the outback in Australia, in New Zealand, countries like Taiwan that were under quarantine restrictions, the UK. We've met all of these requirements, but it's been a little touch and go there, and we've had to be creative.

[00:32:57] JM: Are there any customer use cases for outposts that have surprised you?

[00:33:04] JB: Surprised us? Again, one of the last things I've learned in the seven years is that we are fond of saying this. 90% of the time, customers, they know best what you should be doing and you should listen to them. Even on a newer service like outpost when you're launching. You've kind of learned to expect, to hear something new from customers. It's not like a huge shock. Even if you hear like you're on the wrong track, which happens from time to time. we can adjust course.

That said, I think pleasantly surprised, maybe not shocked by traction and a couple market segments. I mentioned iGaming and mobile betting, that's one that I've been — I think what I've surprised with is how quickly cities and states around at least the US and in some cases around the world, they're moving to legalize this and build regulatory frameworks. We've seen customers like Tipico that are based in Europe. They're finding it pretty easy to make a few changes to the application that are already running on AWS in Europe, and then deploy them on Outpost. They talked about how it's kind of a pivot from their plan, which was to use third-party hardware until we release outpost. I think I've been pleasantly surprised there that they've embraced it so quickly and we been able to meet the regulatory requirements, kind of working with different co-lo partners and different public policy people.

In terms of other surprises, I think also the early traction with telco has been a pleasant surprise. It's not like a shock, because we kind of knew that telco is a thing where you're going to have on-prem installation next to other equipment whether it's at a cell site, or a centralized facility. What I've been surprised with though, is the pace at which they're willing to kind of dive in with us, give us feedback and guide the roadmap. That's where we ended up, with that huge announcement with dish networks, where they're building this first of its kind, industry first, cloud native 5G network. They're the once that are going to be using the small form factor outpost, the 1U, 2U. At the edge, they're going to be using outpost in the nationwide network of local zones, and then of course, running some of their workloads back in the region. It's pretty validating to hear from customers like Dish that it's not just the outposts that attracts them to us. Maybe the outpost is the kicker. They want a standardize on a single vendor.

But innovations like Graviton, multi foreign factors of Outposts, multiple deployment options like local zones as well. That's been surprisingly validating to hear the we're on the right track viscerally. I try not to believe my own hype, which is, I'm very proud of what the team has delivered in a relatively short amount of time. But I'm always pleasantly surprised when we hear from customers across all these different market segments and around the world that we're onto something and they want more of it. Usually, the unpleasant surprise is when you get silence. Like when a few customers telling you that you've delivered what they want. But most of what you're worried about is not hearing feedback from people. I'd rather hear feedback that people want more. That's definitely true in manufacturing, than people who don't really embrace at all the vision that we have.

[00:36:21] JM: What's the process of maintenance for Outposts? Like if I install this thing on my on-premise deployment, then is there enough transparency in the infrastructure that I can figure it out for myself, how this hardware works and I can do all the necessary maintenance. Or is it expected that it's just going to run so operationally smoothly that I won't really need to fix it or if I do need to fix it, I'll call a third-party or I call AWS and they come fix it?

[00:36:50] JB: Yeah. There's a mix except — the core value proposition is that it's fully managed by AWS. What we mean there is that, we're monitoring the environmental conditions, the temperature, whether things are working, hardware components, whether they look like they're failing. Obviously, individual pieces of hardware fail from time to time and we look for things like that. We monitor that. We'll kind of proactively notify you, and then figure out when we can replace most of the things. A few of the components are designed to be serviceable by the customer, like replaceable power supplies. But in terms of the adding additional servers or replacing a server, that's something — it's a part of the service that we actually provide for you. We kind of have a two-business-day objective for replacing things.

Of course, people who are doing high availability have redundancy built in to the Outposts to begin with. They have more than the bare minimum number of servers that they need. It's not something customers need to think a ton about. Ideally, if we're doing our job, which I think we do pretty good at, the customer can just sit there and be confident that we are looking at the hardware, that when they need to add additional servers, they can reach out to us through their normal channels and we will add things over a couple of weeks, which people tell us is actually pretty fast. They can be up and running much faster than with traditional hardware.

Then as far as maintenance, including things like — I don't know if people appreciate this, but the next time there's some kind of chipset level, heart bleed, firmware issue, it will be patched on your outpost just like it is in the region before you even hear about it. That, we're taking care of behind-the-scenes for people. It doesn't require any downtime.

[00:38:33] JB: I'd like to come back to an example that you mentioned, which is Riot Games. It's just interesting to me that you could deploy Outposts and use them to reduce latency of a multiplayer game, which has serious latency requirements. Can you tell me a little bit more about that specific use case and why was it that the deployment of on-prem infrastructure was able to reduce latency?

[00:39:02] JM: Yeah. I mean, with Riot and a lot of other customers and kind of the AAA gaming space, this might be a little bit less true, not entirely, but for mobile games but in the AAA gaming space. You have a speed of light problem. So when you think about multiplayer games, there's traditionally — you not necessarily have literally everybody playing on the same server at exactly the same time. The game is usually designed to scale out across multiple servers. In a traditional world, what these people would do, work with a co-lo provider or build up their own data center. They build a few of them around the world, and they would just kind of try to get like roughly speaking the best coverage for where they thought the most players would be.

Obviously, it's not cost-effective. If you're even for a pretty big company like Riot, it's not necessarily cost-effective to have as much infrastructure deployed as AWS house around the world, because you're the only one using it. In the case of Riot, what they were looking for was — they're going to release their new game, which is now live called Valorant, highly successful from their perspective. They're using a few of the AWS Regions, but there are other places like the center of the United States, the US and places around the world where our regions are just far enough away that people were close to the region might get 40 milliseconds of latency and people who are further away from that region, like if you can imagine using a region in Frankfurt and connecting from the Middle East, you'd be 40 milliseconds away from the Middle East and

25 milliseconds away, which actually gives you a little bit of an unfair advantage in some of these games that you're playing real time.

Riot already had the game server technology to distribute the players to a region that is closest to them from the latency perspective. What they were looking for was to give everybody that 25 millisecond experience. Outposts are essentially strategically placed around the world, not individual outposts. Three or four of these installs with multiple racks. They kind of give those players that equal 25 millisecond or give or take experience. If we had local zones in all of these locations, I think they would embrace those relatively agnostic long-term about what people embrace there as long as it's AWS.

A lot of other gaming companies that we're talking too, it's exact same consideration. Out of a get cost-effectively, the right number of installations. Too many installations are actually bad. I mean if you devolve this all the way to the end and you had five players per outpost spread across a thousand locations, that's super low latency, but it's actually not better because people need to play together. You need to have kind of some concentration. It's kind of an interesting map that's fundamentally just based on overall internet conductivity, the AWS backbone, how latency sensitive the game is to begin with, how the matchmaking algorithms work for those games. Some games require higher density of players per location to make sense and other once, you can have very small deployments, five or six person per game server. So you could actually have more locations and even lower latency. That's going to be game dependent.

[00:42:28] JM: As we're winding down, I guess I'd like to get your perspective on the future, what you're focused on with Outposts today and where you see the product going?

[00:42:38] JB: I's an interesting question. I mean, in many ways, I'd like to joke that we already — my job as GM is to live in the future, which is, we're always looking ahead. I think Outposts in some ways came from an idea about where things are headed, which is we think the vast majority of workloads are still moving to the cloud. But extending the concept of the cloud to include deployments at people's on-prem locations or cities. I'm also focused on execution in the here and now. We still need to release our small form factors later this year. There's ongoing work at all times to keep a super high operational and availability bar. We've got a pretty exciting roadmap.

We've done some releases earlier this year for S3, and EBS local snapshots. I think the future looks a lot like the past, right? We listen to customers and all these market segments. We're going to deliver more form factors, support for more countries. We're up from 20 to 57. We're going from one form factor to three. A few instance types to many instance types, a few services to 15 to 20 or so. The ability to have accelerator cards and do inference at the edge. Which I think, if I look ahead two or three years, the biggest thing I'm hearing from customers is, they want to accelerate their migration to the cloud. They look at Outposts in many ways as filling a gap that was holding them back from moving workloads that won't run on the Outposts to the AWS Region.

That's been an interesting discovery for me, where — what they tell us is, "Look, I don't want to train my developers on three different stacks, or maybe I've tried some kind of hybrid service before that hasn't really panned out, because I'm still managing the hardware, and I'm still managing the cloud deployment, getting my developers train up on those APIs. They already know a lot of the workload will run in the Region just fine, but what they want is, all my developers are going to be using AWS and they'll be using and whether the workload is some kind of huge machine learning workload running in the cloud or an edge deployment on outposts in a restaurant, or retail environment or manufacturing firm.

The challenge for us if you say, what does the future look like is, the challenge is to stay on top of what customers are asking and deliver these new form factors, new instances with better performance, lower prices. Machine learning is just — I hear this kind of from probably 70% to 80% of the customers that they're looking to deliver machine learning inference workloads in factories, and restaurants and retail stores at the edge or on-prem and to kind of modernize their applications. It's up to us to move fast enough and we've got a pretty good track record there over the last 15 years, but you can't really ever rest.

[00:45:38] JM: That seems like a good place to close off. I guess, I'll close with one final question. What is the hardest problem that you are encountering in your job today, whether it's management or technical problems or whatever you want to mention?

[00:45:55] JB: I think the hardest problem in AWS is — and probably for my job, but I don't think it's just about Outposts. We really do spend a lot of time, I spend a ton of time listening to customers, perspective customers, people who are not customers and who tell us that we're on the wrong track. I've kind of made that my mission over the last seven years here. The hardest thing is that, they tell you all this exciting stuff, the customers that you could build and you never have enough resources to get it all done at the same time. Frankly, if you try to do all of it at the same time, you wouldn't deliver anything in the short-term. Balancing kind of fierce pace that we've been able to keep up with an expanded set of offerings, there's no silver bullet for that problem. I mean, you can't get infinite resources even if you can. You'd run into kind of classic development effort problems where you can't just throw thousand people at something and have it done in a day.

I think that's the hardest thing, is there's so many exciting ideas and you can't get them all done at the same time and then move on to the next set of ideas. It's a good problem to have, but I'm pretty aware that we need to keep serving our customers and delivering quickly to them.

[00:47:16] JM: Awesome! Well, thank you so much for coming to chat. It's been a real pleasure talking to you.

[00:47:20] JB Absolutely. Thanks for having me, Jeff.

[END]