EPISODE 1177

[INTRODUCTION]

[00:00:00] JM: BGP or border gateway protocol is a protocol designed for routing and reachability between autonomous systems on the internet. BGPmon is a tool for assessing the routing health of your network, which allows for a network administrator to understand network stability and risk of data. Andree Toonk is the founder of BGPmon and joins the show to talk about BGP, how to monitor routing data and his work at Cisco.

If you want to become a supporter of Software Engineering Daily, you can become a paid subscriber by going to softwaredaily.com and subscribing. You can get ad free episodes with that paid subscription.

[INTERVIEW]

[00:00:41] JM: Andree, welcome to the show.

[00:00:44] AT: Thank you. Thanks for having me, Jeff.

[00:00:47] JM: Simple question, what is BGP?

[00:00:49] AT: Yeah. So BGP is for many probably one of the more important protocols on the internet that you've never heard of. It's the control plane of the internet. You can think of it like that. So essentially it's a routing protocol that all the big routers between all of the ISPs talk to each other and it provides reachability information for all the IP addresses on the internet. So you can think of it as Google Maps for IP addresses, and that's how when you send a packet to, say, facebook.com, you send it to your Comcast router. Comcast sends it to their internal network. But at some point they need to know should I go left or right? And BGP is the protocol that provides you with that information. So it's a really important protocol, but very few people have heard of it.

[00:01:36] JM: Why is it important for the average developer to know about BGP?

[00:01:41] AT: I don't know if it's important for the average developer. So most of what's happening in BGP is taken care of by network engineering teams, infrastructure teams on the internet. And so those teams do a great job for just making it work. And it's not easy, to be honest. It can be very error prone. It's a bit of a complicated system. It's hard to get in. And once you get going, you probably make a few mistakes and then you become a hardened engineer and become an expert.

But unfortunately, every now and then we do see challenges. And what the effect is on users and developers is maybe a temporary issue somewhere, or a routing challenge, or maybe you're running resources in amazon and temporarily that becomes unavailable. But for the most part, it works really well. It's been out there for a long, long time and obviously the internet's still holding up. But it is a bit of an outdated protocol, if you will. It's built on when the internet was still built on trust, and BGP relies heavily on trusting your neighbors. And those times are a little bit over on the internet. Maybe the other thing why it's interesting or why developers should know about it, I think you see BGP more and more coming out of the traditional infrastructure teams and network engineering teams into the cloud. So folks like amazon are exposing BGP more and more. Like if you're connecting with amazon to over a, let's, say side to side VPN or you're using new services such as their transit gateways and things like that. It's very likely that you get exposed to things like BGP because that's how they tie a lot of these things together.

[00:03:22] JM: And what's the developers interface for interacting with BGP?

[00:03:25] AT: Well, so developers don't interact with BGP typically directly. Again, most of this is done by network engineers that log into routers and execute CLI commands mostly. Luckily most of the network engineering folks are now transitioning to network automation. And so APIs are being built around essentially these big core routers. But the cloud providers obviously, they have APIs as well. So mostly there are now APIs being developed. But

traditionally, and certainly in some of the big CLI or big ISPs, it's all still CLI-driven, which is just interesting I think having grown up in the network engineering world. I think the network engineering world is lagging quite a bit behind sort of the SRE world, if you will, where we automated a lot of that stuff away a long time ago. And the network engineering world is now making that transition. And partially because a lot of the gear that the networks are built with, like switches and routers, are very challenging to automate, but that's happening right now as well. But I think for the average developer, you don't have to interact with it and you just have to trust that your network engineering teams are doing the right thing or that your cloud providers are doing the right thing. And maybe if you are setting up infrastructure in Amazon, then you might get exposed to that either directly or by one of your teams.

[00:04:46] JM: How does the global health of BGP get monitored?

[00:04:50] AT: Yeah, that's a great question. So maybe we should pull back a little bit and say, okay, so BGP, so it's this routing protocol that we use on the internet between all of the ISPs. And so in the world of BGP, we call ISPs autonomous systems. So an autonomous system is basically a unique player in the world of BGP. And so there are about 70,000 autonomous systems out there. So what's an autonomous system? Like Google is an autonomous system, Amazon, Comcast, Charter, Rogers, T-Mobile, all these big ISPs or cloud providers or data centers, they're all autonomous systems. They're unique organizations on the internet.

And then each unique organization on the internet claims or announces in the world of BGP one or more IP prefixes. And prefixes is basically a range of IP addresses like networks. And so there is about 850,000 of those IPV4 networks out there today and 100,000 IPV6 ones. So now you have these autonomous systems that are claiming ownership of these prefixes and say, “Hey, if you need to get to this IP address, you can send it to me because I know what to do with that.” And that's kind of how BGP works.

And so as I said earlier, that's been around for a long time and it's likely based on trust. It's a gossip based protocol. And because there're so many ASNs out there, like 70,000 today, you can't possibly connect directly to every ASN. So even big ISPs connect to other big ISPs,

autonomous systems, and then eventually you have this big graph where all the nodes are the autonomous systems and they distribute that IP reachability information.

And so this is all based on trust. If you are telling me, “Hey, I own IP address 8888. You can send it to me because I know what to do with it.” I have no real other option than to trust what you're saying. But I don't really know if you are saying the truth. You could be lying essentially. And that is one of the things that at BGPmon and other BGP monitoring services do. That has been the challenge. Like you might experience an outage and you don't really know what's going on. And that's what BGP monitoring provides. It provides you with a bunch of stuff. So one of them is what we call BGP hijack detection. It's like, “Hey, this IP address should always be announced by Google. Tell me if all of a sudden that ownership changes. If someone else is claiming ownership of the Google, say 8888 IP address.” Because if that happens, I expect this traffic for this IP address to go to Google. But all of a sudden it goes somewhere to China, or Russia, or Germany, or whatever, right? It's going to where it shouldn't go. And all of a sudden they've set up a man in the middle or a black hole or something like that. So if that happens, folks want to know about this and that's why it's important to check what is happening in the control plane of the internet, which is BGP.

But there're other things you can do as well. So that's sort of the more nefarious ones. The things that you may want to watch out for. But once you start studying what is happening in the world of BGP, you find a whole bunch of other interesting things. So for example, because you're literally looking at the control plane, you could see IP addresses being born essentially, new things coming alive. IP addresses that previously weren't announced in the world of BGP, which means you couldn't reach them all of a sudden becoming available or all of a sudden becoming unavailable. And once you start correlating and collecting all that data, you can start doing interesting things with it.

So for example you might detect that a set of IP addresses all of a sudden is withdrawn from the system. The BGP rather say, “Hey, I previously told you I could reach this IP address, but I no longer can. So I’m withdrawing that.” And maybe at the same time a whole bunch of other people are doing the same and maybe those IP addresses all belong to the same ISP, Google,

or to the same country. They were geo-located all to, say, Egypt, or whatever. And so once you start studying that data and run some algorithms on it, all of a sudden you can detect very large scale outages as well. So these are some of the things that you can do when you start monitoring for BGP events.

[00:09:09] JM: What is the implementation of BGP monitoring software look like in some detail?

[00:09:15] AT: Yeah, I mean that's obviously very implementation specific, but there's a few primitives that you need. So first of all you need a lot of data, because the BGP system is a very large system. There're a lot of players. And if you take a snapshot of the control plane, which is essentially what BGP monitoring software needs to do. It needs to get a copy of the control plane data, like a BGP feed we call that. So I can take a BGP feed. So I'm based out of Vancouver, Canada from an ISP here in Vancouver, Canada maybe, right? But that only gives me a regional view of what the ISPs maybe here in Vancouver see. But that might be a little bit different than in Europe or even in San Francisco or China or whatever, right? So the idea is that you need to get as many of these feeds as possible because it gives you as complete as possible view of the internet. Certain IP networks, for example, are only visible in certain regions or only visible in certain regions as originated by a certain autonomous system. So that's the first thing. You need to have lots of BGP feeds, copies of the internet routing table and what do they look like. And so that's your input.

The other thing a lot of folks need is sort of a source of truth. And so depending on how you do that, a lot of times when you sign up for a service like this, you will need to basically say, "Hey, this is me. I am just using Google as a random example. And these are all my prefixes. And I expect them to always be originated by this autonomous system, Google, or whoever else, right? So now you have a source of truth. So now you have these two things. You have a source of truth. This is my expected state. And you have sort of what's happening in the real world.

And basically as this data comes in, and these can be literally hundreds, if not more thousands of updates per second, you start correlating the source of truth which you expect with what is out there. And then if there's a delta, you can run some analysis on it and say, “Well, this delta is significant enough to inform whoever wants to be notified of any changes.” And so you can look for origin changes. Like, again, this prefix should always be announced by Google. All of a sudden it's announced by someone else that violates my expectations and then you trigger an alert and then you can build an alerting system. And there's multiple ways to do that. And so that is sort of for the alerting and informational purposes.

And similarly, if you want to look for outages, you can do similar things, withdrawal message. So the BGP monitoring system is all based on deltas. So it's like, “Hey, I previously told you this. That's no longer the case. I’m withdrawing that.” Or, “I previously told you this prefix was announced by this network. Now it's announced by this network.” So you need to somehow maintain a bunch of state to keep track of what's going on. And if you see, for example, all of a sudden a whole bunch of withdrawal messages, meaning prefixes are becoming unavailable, you can correlate that as well. So you have a bunch of processes running that figure out if this is significant or not but. But that's sort of the high-level thing. An expected state where a customer or a user goes in to define what is expected. And then so that's your source of truth. And then sort of the real life information, you compare the deltas. If it's significant enough, it goes into an alerting module and you can alert users however you want to. It could be email, phone calls, SMS messages, web hooks, however you would like to develop this.

[00:12:42] JM: Can we talk a little bit more about BGP from the perspective of a hyper scale or like an Amazon Web Services?

[00:12:49] AT: Yeah. So I’m not sure exactly what you're looking for, but I can give it a shot. So, Amazon, like any other cloud provider, Google, whoever, relies on BGP as much as a small ISP in wherever, the middle of Rural America or something like that. Like we all need BGP to participate in the internet routing tables. So there's really no big difference there, but the one thing that is different is interesting that you see more and more cloud providers providing a BGP interface, if you will, with their customers. So for example Amazon and Google, they're all

providing direct connect services now, where essentially you can connect with one gig port or 10 gig port and probably soon 100 gig port directly to Amazon. That allows you to really make this hybrid infrastructure and connecting your on-prem data centers to Amazon.

And so the protocol that is used over that direct connect link is also BGP. And so it's a little bit widening its scope now where it used to be really sort of the geeks of the network engineering world that were involved in the core of the internet, to now also more to, yeah, enterprises connecting your on-prem connection to amazon or whoever else.

And then I think what is very common in larger cloud providers, and literally any larger network, is the concept of traffic engineering. And so these folks like Amazon and others, obviously they have a ton of traffic, many terabits of traffic. And it's very expensive to get traffic from A to B if you have that much traffic. And so they might have all kinds of interesting traffic engineering technology. And so what's traffic engineering? In a simple way, it's like you have two ISPs, ISP 1 and 2, and for whatever reason you may prefer certain traffic. Let's say your Zoom traffic over ISP 1, because maybe it's cheaper or it has better performance or whatever.

And so what a lot of these larger organizations have is teams that build these traffic engineering systems. And so they look at certain inputs, type of traffic, or cost, or whatever and say, “Okay, this type of traffic should go over ISP 1 versus ISP 2.” Or maybe they look at latency. It's like, “This one – This ISP is very cheap. So that's where I’m going to send all my backup and inter-data center data transfer over. But this one is actually of higher quality, provides me with lower latency links. So that's where I’m going to send my real-time traffic over.” And then there's the concept of continuously monitoring these links and making decisions over how do you want to move traffic from ISP1 to ISP2. And a lot of times that is done by interacting with the BGP system and making sure that you move traffic to the right ISP depending on whatever your inputs are.

[00:15:50] JM: Are there any other challenges around maintaining high availability that are worth discussing around BGP and these hyperscale clouds?

[00:15:59] AT: Again, the hyperscalers, like anyone else, have similar challenges with BGP as anyone else like in terms of redundancy or high availability. They'll just have more and more ISPs. So if you start an ISP, you probably need at least one internet service provider. So ISPs need ISPs as well, right? And so you start with one and then all of a sudden there's an issue. Then you have two and three and you have some redundancy. And the same goes for the hyperscalers.

The one thing that's different with hyperscalers and like examples are Amazon, is if there's an issue with BGP, like someone is hijacking Amazon's prefixes, for example, then it doesn't just affect maybe a few hundred people. It affects thousands, and if not millions of people depending on the services that run on these IP addresses. So not only is amazon sort of use a compute provider. It's also becoming a network provider. And more and more you could see Amazon moving into really the network as a service type environments where they're now providing client VPN software, IPSEC, side-to-side VPNs, firewalls as a services and all that kind of stuff. So they're becoming more and more an ISP as well. And so it's very important for them to make sure that they monitor the state of the BGP. Someone else announcing their prefixes. Is traffic going over the right links? And continuously monitoring that information.

But in essence, the information or the requirements and the risks are the same, it's just that the impact is so much bigger if something happens.

[00:17:36] JM: Are there any attack vectors that we can discuss that come through BGP like DDoS attacks?

[00:17:44] AT: Yeah. So attack vectors for BGP are mostly around what I've previously called BGP hijacks. That's the biggest kind of challenge in BGP. So as I said, BGP has been around for a long time and it's the only protocol on the internet that that provides reachability information. So there is no real alternative to a routing protocol between ISPs. So that's kind of what we're stuck with. And like the old days on the internet, everybody trusted everyone. You could send everyone an email. There was no validation. And all the spamming came along. And

then we had to fix that, and the same with DNS. It's all untrusted then they built DNSSEC, same with SSL and TLS.

BGP is just going down that track right now. I mean I guess it started 10 years ago, but it hasn't gotten too much traction. Although right now about 25% of the routes are protected. And what does that mean? So like I said earlier, every BGP speaker claims ownership of certain prefixes and says, "Hey, if you have traffic for this IP address, you can send it to me." But there's really very little I can do in validating that what's coming in in terms of BGP messages is true. And there're plenty of cases where people are lying, and that's what a BGP hijack is.

And so one of the most common examples, for example, is an incident that happened, or most popular example about 10 years ago or so where YouTube was hijacked by Pakistan Telecom. And what happened there was I remember this pretty vividly. It was Sunday night. Everybody's watching their cat videos and all of a sudden the videos just stop. YouTube was broken. And the interesting part was it had nothing to do with Google or YouTube. Like their systems were running fine. But from their perspective, traffic had just stopped coming in.

And what happened was that the Pakistani government had asked their national Telco, Pakistan Telecom, to somehow block YouTube. There was some content on there that the government didn't like or didn't agree with and said, "You need to block this." So now how do you do you block YouTube or any other type of service in a whole country? How do you do that? So you have to have firewall rules everywhere or ACLs or whatever.

Well, one of the things you can do is play around with the control plane of the internet. And basically that's what they did. And so in BGP they made a new announcement for YouTube and they did it in such a way that it was better. And so they forced all this traffic towards them and then they just black holed it, dropped it on the floor. And this was supposed to only stay within Pakistan. So that all that traffic would come to them. They would drop it on the floor. And now they had achieved a country-wide outage.

But what in what happened in reality is that they exported the censorship. So this more specific route for YouTube, this more attractive route, accidentally was announced to the ISPs of Pakistan telecom as well. And they very rapidly within seconds made this available to the rest of the world. And all of a sudden everybody watching their cat videos saw these videos hanging. They stopped working. And so now you have a situation where arguably one of the most powerful and high knowledge companies, internet companies in the world, Google, was basically down, their YouTube services and it had nothing to do with them. And there was very little they could do to remediate this either. They were basically relying on Pakistan Telecom to undo this.

And so phone calls had to happen, yadi-yadi-ya. And so this lasted for I think two hours or so before it was undone. And to me that is such an interesting example where if someone like Google with like so many smart people can't fix, it that really shows some of the Achilles heels of the internet. So that's what a BGP hijack is.

So now to – And there's plenty of other examples, but that's a very popular example or well-known example. Now that was about 10 years ago. There is now an initiative that is called RPKI, resource public key infrastructure. And essentially very high-level, it's like DNSSEC sac for BGP announcements or TLS for BGP announcements. And really what it solves is for BGP participants, routers on the internet, to validate that what you are saying to me is actually true. So BGP is this gossip-based protocols. We all relay each other's information. And traditionally you couldn't really validate what you were saying was true. Now I can go out of band and check, “Hey, actually Jeff's telling me he owns these prefixes. Is that actually true?” And there's this RPKI system, this PKI system that says, “Oh yeah, this prefix or this network really belongs to Jeff. He can make certain claims about that.” So that is something that's been in the works for a long time and now we're just starting to see that being kind of rolled out over the internet starting a few years ago. And I think we're at about 200,000 signed prefixes as we call them, or route origination at the stations that you can cryptographically validate that that's true. And so about 25% of the IP addresses on the internet now have this in place making BGP hijacks a lot harder to pull off. So everybody should sign their prefixes.

[00:23:07] JM: What is your current level of interest or involvement with BGP today?

[00:23:13] AT: Oh. Well, my interest is always high. So I think I've been in this world for about 20 years. I started at a large internet exchange point in the Netherlands, Amsterdam Internet Exchange. And then I've always worked for ISPs and infrastructure providers. So I've always been poking with BGP and at some point I realized, "Hey, actually, if you start collecting this control plane information, you can study it." And so that's what I started to do. And then I was kind of amazed by what you could find in it.

And so in 2008 or so I started this project called BGPmon, which back then was just a free service where people could sign up and basically say, "Hey, if these are my prefixes and this is what my expected state is and if something changes I would like to get an alert." And so we built this little monitoring service. And that turned out to be very popular and many thousands of people, our network engineers started to use that. Eventually that became a company. And eventually that's now part of Cisco. But that's kind of my level of interest and sort of involvement.

And so as the other thing I did as we did this when we had all this data, we worked with many service providers and cloud providers out there to study sort of incidents and out of bands like, "Hey, did you know this happened?" And it was not uncommon for people to hear. I would reach out to them and say, "Hey, I noticed that between 2 and 3 p.m. some of your networks were rerouted somewhere else." And it wasn't uncommon for people to say, "Oh, thank you for reaching out. We did actually see a drop in traffic, but we didn't really know what happened. We were looking at all of our logs and firewall rules and monitoring services and it all looked fine, but yet we still got customer complaints."

And to me it was kind of eye-opening how this was sort of an aspect of monitoring that was – It wasn't really taken into account, because it was almost too low in the stack and there's not much you can do. If traffic's just not coming in, if your ISP is just not giving you the traffic that you expect, you're kind of like, "Yeah, everything looked fine. We just saw the number of requests drop." And so that was a really interesting kind of eye-opening moment for me. And

then we kind of helped a lot of these organizations get to the bottom of these things sort of either by providing these monitoring services or just working with them directly and studying some of these really interesting events.

[00:25:39] JM: And what elements of BGP are you working on today? Sorry, BGPmon?

[00:25:45] AT: So BGPmon is now part of Cisco. I'm less and less involved in it. So actually just recently left Cisco. So I'm not actively involved in BGPmon anymore, but I'm still actively involved in studying the data and just going out and figuring out what's going on. Working with trusted partners and friends in the industry to continuously keep an eye on what's going on, because to me it's just such a fascinating world. I know there's not a lot of people that either are interested in it or are exposed to it or understand it really in depth. So it's a small group of people that really keep an eye on this. And I really enjoy doing that, just digging into the data. It's such a massive source of valuable information that surprisingly very few people are looking at.

And so a lot of people want to know simple information you can find out of this. It's like I want to know all the IP addresses that amazon owns. Well, the simplest way is to just look into the BGP routing tables to see what prefixes Amazon is claiming ownership of. So those are some basic stuff, but all that stuff basically starts at the BGP level.

Other fascinating things you can see with that is I remember clearly a large scale outage in Egypt. So in Egypt we had the Arab Spring. This was probably in 2009 or so. I'm not sure. It's about 10 years ago. And so this was one of the big protests that was organized using social media, right? And now we see this all over the place. But that was one of the first ones. And at that point the Egyptian government had ordered the government or the ISPs in Egypt to shut down the internet. To essentially try and squash this by you know preventing communication.

And so by looking at these BGP routing tables and events that we were happening, we could see exactly – And I think there's like four or five major ISPs in Egypt when the shutdown happened. So at 8:00 we saw ISP1. And at 8:20 we saw ISP2. And at 8:30 we saw ISP3. And

within an hour, all of them were down and we could exactly say 99% of the country's internet is down and these are the time stamps and this is how it happened.

And what's really interesting is like, “Well, what is the one percent that is remaining? That's weird. Why is that not down? Well, let's take a look.” And there were maybe a few dozen prefixes or IP ranges that were still available. And turns out those were of very high value. Those are prefixes owned by the military, major banks and the Egyptian Stock Exchange. And so there's just some real valuable and interesting intelligence information you can find in things like that. And so even though I’m less and less involved with that, I’m still super fascinated by that and keep an eye on events like that. But mostly out of personal interest.

[00:28:47] JM: Do you have any predictions for how the world of BGP will change in the next five to ten years?

[00:28:53] AT: Well, let's see. So the big thing that's happening in the world of BGP is a little bit what we just talked about, RPKI. So the problem of route hijacks is fairly well understood now. I mean it's been something that's been around for a long time. But as we are building this multi-trillion dollar economy on the internet, it's kind of amazing that it still relies on an old protocol like this that you can so easily lie in. And so it's a bit of a shaky foundation, right?

So RPKI, where we sign these BGP announcements, is very important. This is something that can help prevent the lying essentially or it gives us a way to validate that you're not lying. So that is a trend that will continue to go. And so I think we are at 25%. So hopefully over the next few years we'll get to close to 100% of prefixes that are signed, and that should really help with these BGP events.

And then, yeah, for any of the tech nerds or network nerds that are listening, this only prevents sort of what we call route origination validation. But it's fairly easy to spoof what's called AS paths as well, so sequences of AS numbers. And so that's sort of the next part that the industry will need to take a look at, which is called path validation. But RPKI should take care of the far majority of the incidents that we have. And then there is other things such as path

validation that are on the horizon. If the industry gets there or not, we'll see. The next few years will show that.

And the other interesting thing is, yeah, just to see in the world of BGP IPV4. Addressing is obviously very important. So as you know, we're running out of IPV4 addresses. So normally when you needed IPV4 addresses, you would go to your regional internet registry, something like ARIN in North America or RIPE in Europe. But today if you're starting a new business and you need IP addresses, you can't really get them anymore. They basically used to give them to you for free if you were a member of that organization. And you had to pay a small membership fee and maybe a small fee, but it wasn't very expensive.

Now if you need IPV4 addresses, you kind of have to go out to the secondary market and buy IPV4 addresses from someone else. And so that market is super active right now. And prices are going up. And so a lot of folks don't know this, but yeah, a lot of the big cloud providers like Amazon, Alibaba, Google, but also Zoom, they're buying IPV4 addresses on the market basically from people that want to sell them, which is a very interesting kind of market to follow as well. We'll see how long that goes for. Obviously everybody's been waiting uh for IPV6 to really take off. And when IPV4 ran out, folks are saying this is the time. But really what's happening now is there's this market where people are still buying IPV4 addresses off for significant amounts of money.

I recently did a blog post about it and studied some of the IP addresses that Amazon owned. Yeah, I think we valued that at almost two billion dollars. Like just the value of the IPV4 addresses that they own. So that's very, very significant. So hopefully IPV6 will continue to take off. Just this week I think in the world of BGP and IPV6, we reached a bit of a milestone where we saw a hundred thousand IPV6 prefixes in the internet routing tables. So that's pretty good. And so we'll see how that goes. And that will continue to grow. But those are some of the things that I think we'll see happening. RPKI, maybe at some point, path validation to really make it more secure, and then, yeah, the move towards sort of the IPV4 market and IPV6 taking off and continuing to grow.

[00:32:42] JM: One area I'd like to explore a little bit more is the interaction between DNS and BGP. Can you dive into that in more detail?

[00:32:52] AT: Sure. So DNS and BGP are both sort of the most critical protocols on the internet to make the internet work, right? And so they're also sort of the Achilles heel of the internet, because even though they're – majority of the traffic goes over HTTP, HTTPS, that kind of stuff. And so DNS, most people will know that like in order to go to google.com, I first have to do a DNS query. And so in the past what we've seen is certainly a lot of interest to sort of DNS infrastructure is considered as critical infrastructure. What does it take to take down the DNS infrastructure? Because if you take that down, then you don't have to worry about taking Google down or Facebook that have massive infrastructure. If I can take down the root DNS service, for example, then you've basically taken down the whole network. So that's critical infrastructure and something that a lot of people are paying attention to and similarly with BGP.

But in terms of that criticality, like which one's more critical? Well, in order for you to go to google.com, you do a DNS query and that query may go to your DNS server and maybe you're using something like OpenDNS or Google's 8888 or something like that. But in order for that DNS query to leave your computer into Comcast or whatever, it needs to then go to 8888. It relies on BGP. So you need to have BGP first to distribute that routing information to figure out where 8888 lives.

And so you cannot do DNS queries if BGP isn't working correctly. So that's sort of really the first kind of layer. So without BGP, nothing works. DNS won't work either. So that's sort of the ordering in terms of criticality and bootstrapping the internet. You need to have a working BGP system. So you have BGP or IP network layer information. I know I can build this topology database and then I know how to get to all the IP addresses, and some of them are DNS. But I think they're both super critical for the internet. They're both considered as critical infrastructure and should be well-protected.

One of the interesting things though like – So this is a little bit of a side track, is a lot of DNS operators and even CDNs use BGP quite heavily internally to provide high availability. So for example, think of the root name servers, right? They got a lot of traffic. And so the root name servers is where you find where .com lives .net and then you have this whole tree. So it's very important that those are alive and highly available. And so there're many of them. We call those Anycasted services. So 8888, for example, is one IP address. But it actually exists in many, many different places in the world. So normally we know an IP address has to be unique, can only exist once. But with any cost, you kind of turn that up onto his head and you say, "Actually, if you know what you're doing, you can replicate this IP address in many different places around the world." And the way you do that is simply by BGP announcing this IP address out of all of your data centers, like all of Google's data centers or all of the OpenDNS data centers or all the locations that the root name servers have.

So all of a sudden if you want to go to that particular IP address that is Anycasted, you end up to the one that is closest to you, which is really great. So this is what we call Anycast, because if there's now a DDoS trying to take down a popular DNS server, for example, that traffic gets sort of load balanced, right? So if you have attacks originating from all around the world, it doesn't go to one server. It's kind of distributed to all the servers around the world.

So if this is then – I don't know, let's say a hundred gig attack, right? So it's a fair number. If you only had that server in one data center and you had to have a hundred gig in that data center, which is a fair amount of capacity to bring in. But if you have 10 data centers around the world, then all of a sudden you only – And then assuming it distributes evenly, you only need 10 gig in all of your 10 gig data centers or 10 data centers and then you have 100 gig. So this distribution using BGP Anycast really helps with managing and absorbing these large DDoS attacks.

And what you're seeing more and more, and maybe this is interesting for your audience, is within data centers, assuming you own your own hardware and people are still building their own, you can have your servers talk BGP as well. And so maybe you have five servers all with the same IP address and they talk BGP with the data center router. And so if traffic arrives on

the data center router for a particular Anycasted IP address, it will load balance it automatically overall 5, 6, 10, 20, 100 service, however many you have. And this is what we call ECMP or Anycast within a data center. And a lot of CDNs use this. A lot of DNS operators use this. So if you ever wanted to take down a particular server or something, all you do is basically shut down this BGP session and the router now thinks, “Oh, I used to have five options. Now I only have four.” So you get this automatic load balancing, high availability both within the data center, but also globally. And so this is what you see operators of services that need to be highly available particular DNSs critical server use. So they have all these data centers around the world. We use Anycast BGP routing to make sure it distributes nicely and goes to the closest data center. And then even within the data center, they have multiple servers. They all talk BGP to a router. And this is how we provide load balancing within the data center over multiple services. So back to your question, DNS and BGP, how do they play together? I think they're both super critical to the internet, but they also rely on each other to provide high availability.

[00:38:49] JM: Well, Andree, we've covered a lot of ground. Is there anything else around DNS or BGP or any other subjects that you'd like to discuss?

[00:38:57] AT: No. I think we covered most of it. Yeah, I hope this was interesting to your listeners. I know it's a bit of a niche subject. Hopefully we demystified this a little bit. And yeah, if you guys are interested, just reach out to me. I’m always happy to chat with folks more about this topic.

[00:39:15] JM: Okay, thanks for coming to the show. It's been great talking.

[00:39:17] AT: All right. Thank you, man.

[END]