EPISODE 516

[INTRODUCTION]

[0:00:00.3] JM: Employees often find themselves needing to do work outside of the office. Depending on the sensitivity of your task, accessing internal systems from a remote location may or may not be okay. If you're using a corporate application that shows the menu of your company's café on your smartphone, your workload is less sensitive. If you're accessing the proprietary code base of your company's search engine, your workload is more sensitive.

As Google grew in headcount, the different cases of employees logging in from different places grew as well. Google developed a fine-grained adaptive security model called BeyondCorp to allow for a wide variety of use cases. Whether you're an engineer logging in from a Starbucks or a human resources employee logging in from your desk, the BeyondCorp system uses the same access proxy to determine your permissions.

The BeyondCorp architecture is also built upon the assumption of a zero trust network. A zero trust network is a modern enterprise security architecture where internal servers do not trust each other. Zero trust networks assume that the network has already been breached. If you are writing an internal application, your default assumption should be to distrust any incoming request from someone else on the network unless they have properly authenticated.

The zero trust model is in contrast to an outdated security model of enterprises, which is that the hard outer defense of a firewall that purports to prevent attackers from ever making their way into the vulnerable inside of a network is impervious. The firewall model assumes that all of these servers within the firewall can trust each other.

Several papers have come out of Google discussing the BeyondCorp security model, and these papers describe the network architecture and the security philosophies of BeyondCorp. Since the releases of these papers, an ecosystem of security providers has sprung up to provide implementation services for companies that want BeyondCorp security in their enterprise. Google has also productized its BeyondCorp system with an identity aware proxy product, which is tied into their Google cloud product.

Max Saltonstall is the technical director of information technology in the office of the CTO at Google where he has helped to facilitate the widespread adoption of the BeyondCorp program.

In this episode we talk about enterprise security from remote employee access to zero trust networks. We also talk about implementing the BeyondCorp model. Why enterprises should consider it and how to do it yourself if you're interested in implementing the BeyondCorp model.

We've done lots of past shows about security from topics like car hacking to more sophisticated stuff, like smart contracting vulnerabilities. We've had discussions with luminaries, like Bruce Schneier and Peter Warren Singer, and to find all of our episodes about security, you can download the Software Engineering Daily app for iOS or android. These apps have all 650 of our episodes in a searchable format. We've got recommendations and categories in discussions around the episodes and it's all free and open-source. If you're interested in getting involved in our-open source community, we have lots of people working on the project and we do our best to be friendly and inviting to new people coming in looking for their first open-source project. You can find it at github.com/softwareengineeringdaily. You can join our Slack, and we'd love to see you there.

So with that, let's get on with this episode.

[SPONSOR MESSAGE]

[0:03:51.7] JM: If you are on call and you get paged at 2 a.m., are you sure you have all the data you need at your fingertips? Are you worried that you're going to be surprised by things that you missed, errors or even security vulnerabilities because you don't have the right visibility into your application? You shouldn't be worried. You have worked hard to build an amazing modern application for your customers. You've been worrying over the details and dotting every I and crossing every T. You deserve an analytics tool that was built to those same standards, an analytics tool that will be there for you when you needed the most.

[INTERVIEW]

[0:05:35.6] JM: Max Saltonstall, welcome to Software Engineering Daily.

[0:05:38.7] MS: Thanks very much for having me.

[0:05:39.8] JM: You're the technical director of information technology in the office of the CTO at Google. Could you explain what your role is? What do you do there?

[0:05:50.2] MS: Sure I help companies who are coming to Google, usually new to our cloud platform, and I help them figure out how to best use Google's technologies across all of clouds, that's Google cloud platform, G Suite and many other things that we make, and help them figure out where it's going to fit into their business goals. So how can they improve the way they do IT operations or user support. How can they improve their security? Reduce some of their operational overhead by moving to public cloud, or get just a more streamlined set of tools to speed up their developer's workflows.

[0:06:29.7] JM: Makes sense. And today we're talking about BeyondCorp. That's the focus of the conversation, but maybe we can explore some of those other topics that you just outlined.

BeyondCorp is an approach that Google developed around enterprise security. It was pioneered in 2014, and I want to get into discussing how Google reframed enterprise security, but I think in order to do that, we should first talk about some of the problems with traditional IT security. I think many of the people listening are software engineers and they don't know much about IT security. For example, I don't even really know much about what a firewall is, but I've used it for a long time. I think almost every company uses a firewall. I think that's a perfect example of something that is problematic with traditional IT security. What is a firewall and what are the problems with a firewall?

[0:07:21.9] MS: Sure. So the traditional model for enterprise security is I usually liken it to a castle. You've got the important stuff on the inside that you want to keep safe and the people who are inside the castle, you trust them to get access to all those things, and then there're people outside the castle and you want to keep them out. So what do you do? You build a big wall, you build a moat. Maybe you have a very closely guarded front gate. So if you've seen any Lord of the Rings type movie, you can have this mental picture, and that is the traditional model for IT security. I build a big wall around my stuff. This is a digital wall, and so it's restricting what kind of communication can come in and out, and then I make sure that all of the important stuff, whether it's finance or human resources or source code or whatever, is inside that wall, and I put as many kinds of perimeter protections around it so that I restrict who can come in, or who can get access, or who can get on the network. This wall metaphor really is translated to kind of a network space so that you keep people outside your network who don't belong there, but anyone who's inside, say, they're an employee, at the office, on a work machine, they can get to pretty much anything, and that's where the model starts to break down.

[0:08:43.3] JM: Okay. So you've described this model of the castle around which we are deploying defenses of a firewall or some other kind of moat. What's wrong with that model of having this castle where you have the internals being the juicy important details of an organization, and that's okay, because it's surrounded by a moat and it's surrounded by these high walls of the castle and everything is secure because it's surrounded by that castle. What's wrong with that model?

[0:09:14.4] MS: Sure. There's a bunch of flaws that we started to see that made it hard to use. One main one is it's got that moist chewy to center that you mentioned, so anyone who slips by

— Maybe to carry this metaphor perhaps too far, anyone who sneaks in on a caravan of what looks like food and is actually a spy or an assassin or something, once they're inside the castle, there is no more defense. There is no way to stop someone who gets past that big perimeter from doing bad things. So there's essentially one barrier to some sort of malicious actor.

In the office environment that could be as simple as putting some malware on someone's machine that they accidentally downloaded from a malicious website, and once they bring that machine to work and they sign in to the work network and they're at the desk, now your attackers inside, and they penetrated the wall. They can get access to all kinds of stuff, all kinds of confidential corporate resources.

So the assumption that anything inside your castle is safe is incorrect, and especially as people are using phones, tablets, laptops, devices are moving around, they're moving in and outside those walls. So in addition to be able to pick up malware and then bring it inside, you also have people who want to go and work from outside. I might need to travel to a conference and get some work done or go to a client or just be on an airplane, and if I'm outside the castle in that traditional model, there is no way for me to get to the stuff I need to get my work done, whether it's finance records or Salesforce or whatever it is.

So you actually decrease usability in that model, because people have to get inside, usually through a VPN, which pretends that my computer is inside that privileged corporate network, even if it's actually outside, or they're stuck outside because it doesn't work and they can't get any work done. The castle model is dangerous in terms of letting attackers past just one perimeter and it also inhibits people from getting things done when they're outside the office.

[0:11:28.5] JM: This world where we have a more porous network, and the reality is that devices are passing in and out of the corporate network. How does that reality change the policies that an enterprise needs to take around how people should be able to access the internal network?

[0:11:50.0] MS: Well, what we did with BeyondCorp — The BeyondCorpe name stands for beyond the corporate network. So what Google decided to do was instead of changing the way you could get in or not get in to that special internal network, we said, "Let's actually flip the

model so that anybody can get access to the important corporate tools from any network." It no longer matters where you're sitting or where you're connected from. What matters is who you are and what you're using.

We take the network out of the equation, and the mandate for BeyondCorp is that you can get access to any tools you need to do your job from anywhere without having to use a VPN or other special software. The way we do that is by looking at the context of your device, your user, your session, rather than the network, and that's where the trust comes from.

[0:12:47.2] JM: Let's start to explore this model. So I log into — Can I log into the internal network with any device or can I only use a device that is issued to me by the company that I'm working for?

[0:13:02.0] MS: So we require a managed device, because we need to know about that device in order to give it some trust. So at the center of this is an accurate inventory of devices that we trust or at lease that we know about. We don't necessarily trust them. So when I opened up my laptop and want to connect, and it doesn't matter if I'm at work, at home, at a coffee shop, on an airplane, same exact process. My computer is going to check via various mechanisms and agents with some of that management suite and is going to be feeding data to our device inventory service. So it's going to say, "Hey! Here's who I am. This is Max's laptop, I'm running this version of the OS. I last checked in that time. Here's my patch level." etc., etc.

That inventory system is going to be feeding information to the access proxy to say, "Here's the kind of state that Max's laptop is in," and so I can authenticate in a really strong way, so I use my username, password and a universal two-factor with strong cryptographic handshake second factor and my device is in a good state. It's not missing some key security patch, it still got a password lock, it's encrypted, all of the sort of standard things we would expect, then I'm able to earn a high level of trust and every time access a resource, whether it's our HR system or codebase or anything, the proxy is comparing how much trust does Max need to look at this thing? How much can he earn right now? Is that enough to grant access or am I just going to give him a forbidden error? But the internal-external network distinction doesn't matter.

[0:14:48.5] JM: Okay. What kinds of data that is going through the user's device, what kinds of data contribute to the level of trust? How is the level of trust associate with that user and that device evaluated?

[0:15:05.7] MS: So there's a lot of data points from a lot of different systems, and we had to do a fair amount of work actually to corroborate, because you might have one system that's just reporting on a machine based on its MAC address and another system uses its serial number, and these two systems might not actually have much overlap in what they report back, and so you can't always tell, "Have I gotten around two records from two different laptops or is this one laptop that I just know disparate things about?"

So we ingest data from a lot of different systems and then have to do that corroboration to de-dupe and merge records about hosts so that we can paint a full picture of as specific device. So some of that is about software, right? Am I up-to-date of my patches? Was to reduce zero day vulnerability? So now all of a sudden everybody is unsafe until we get them up to speed on the patch that just came out 10 minutes ago.

[0:16:06.2] JM: Fair enough. So let's take a top-down example. Let's say I have a company laptop, I sit down in Starbucks and I use a single sign-on system to log into my BeyondCorp network identity. Explain the authentication process from the enterprise's point of view, from Google's point of view.

[0:16:29.0] MS: So when you are connecting, you're going to hit an external, accessible to the internet, single sign-on. We're using Google identities, because we're Google, so you're going to login as if you are logging in to any kind Google service and establish that identity, and all of the services you might need to get to are behind this access proxy, which is also exposed to the internet. So your attempt to say — I don't know, take a vacation day, because you want to go into our HR system. It's going to resolve to the proxy, which is serving as a frontend for all of these corporate services. You're going to authenticate. It's either going to already know who you are or it's going say, "I don't know who you are." Send you to the single sign-on. You login, you establish identity. Your machine has a certificate that just identifies it uniquely. So you're establishing your user identity and your machine identity, and the access proxy can then pull

information about the machine plus pull information about the user and make a calculation of trust right then and there. Do you have enough to get to, in this case, your HR web tool?

[0:17:41.3] JM: So as we discussed earlier, we can't assume that the internal network is safe. We need to take measures to remove trust from the internal network. So we've discussed how the process works if you're external to the network, if you're sitting in Starbucks. So if I'm sitting in a Google building, on a Google Wi-Fi network, does that same authentication process proceed in the same way?

[0:18:07.0] MS: Yeah, it's exactly the same. That's the beauty of it. So from a user perspective, from a Google employee perspective, they actually go through the exact same flow whether they're at work, they're at home, they're at a hotel. It doesn't matter, and that's helpful.

[0:18:21.6] JM: Of course. I mean, that simplifies the surface area of design choices you have to make.

[0:18:26.7] MS: It simplifies support too.

[0:18:29.8] JM: Is there a big deal, like the internal support system for somebody who can't login to something?

[0:18:36.4] MS: Well, let's say you took a fairly large company and you change the way they do security, that could create a tremendous load on your support team, because the way people do things day-to-day has just completely changed. It's sort of pulling the rug out from under them, and that could create a lot of confusion, especially if you have non-technical folks who might be, say, in a sales org instead of an engineering org. So we thought a lot about the supportability of both moving to this new system, but any kind incremental change along the way, we wanted to make sure that we could automate, that we made a self-serve, that we could even provide self-remediation wherever possible, and that if we did have to get the support team to do a fix or to un-break someone, that we had created the right tools for them so that they didn't have to go searching or hunting down, "Hey, what do I do when someone walks up and has this kind of an error?"

Supportability was something we definitely thought about and it ties into that whole user experience point that I mentioned earlier, which is you could deploy something that makes your company more secure, but if you've just hurt everyone's productivity, they're going to be mad at you and they're going to be getting less done. So what's the price of the security? There is a similar price around supportability that we were very conscious of and worked on optimizing.

[0:20:02.3] JM: What you just alluded to is the fact that — Maybe we haven't really made this explicit earlier, but this idea BeyondCorp, the security model was very successful at Google and now it's been evangelized and, in some cases, productized and other people are adopting it. I'd like to get into that a little bit later, but to talk more about — A little bit more about the engineering.

You've mentioned this access proxy term. So when I use an enterprise application at Google, all of my communications to external and internal clients go through an access proxy. What is an access proxy?

[0:20:45.1] MS: It is an intermediary that is going to allow or deny your request to get to a service or a resource. In many ways, it's actually similar to the scalable servers that we used to deploy Google services worldwide. So we have what we call Google frontends, which are a mix of load balancer, denial of service mitigation and global consistence scale. So we have these deployed all over the place so that when you go to bring up google.com or Gmail or something like that, you don't have to worry about going across the country and hitting some backend there. You've got something close to you that is ready to give a response, and a consistent response.

So what we did is we took the same frontends that we used for big scalable services, like Google search, and adapted them to become these access proxies. So we added some logic on top and we got a lot of things for free, like that load-balancing and denial of service protection. So you can think of them really as just a server that is the internet-facing, the public-facing portion of the "corporate network". So it doesn't exist anymore, but anything that would be internal to Google is protected by these access proxies, which the access is an important part of their name. They are deciding, "Do you get access to this or not?" So they're making that decision each time you make a request.

[SPONSOR MESSAGE]

[0:22:29.2] JM: The octopus, a sea creature known for its intelligence and flexibility. Octopus Deploy, a friendly deployment automation tool for deploying applications like .NET apps, Java apps and more. Ask any developer and they'll tell you that it's never fun pushing code at 5 p.m. on a Friday and then crossing your fingers hoping for the best. We've all been there. We've all done that, and that's where Octopus Deploy comes into the picture.

Octopus Deploy is a friendly deployment automation tool taking over where your build or CI server ends. Use Octopus to promote releases on prem or to the cloud. Octopus integrates with your existing build pipeline, TFS and VSTS, Bamboo, Team City and Jenkins. It integrates with AWS, Azure and on-prem environments. You can reliably and repeatedly deploy your .NET and Java apps and more. If you can package it, Octopus can deploy it.

It's quick and easy to install and you can just go to octopus.com to trial Octopus free for 45 days. That's octopus.com, O-C-T-O-P-U-S.com.

[INTERVIEW CONTINUED]

[0:24:00.7] JM: This is handled by the access control engine, which sits inside the access proxy, and if I understand correctly, there's a feedback loop between the data that's gathered from a user's device, and the user's session, and the user's ongoing trust level and this access control engine, because if a user's trust level decreases, then their access might decrease. Can you describe — Tell me whether or not that's correct, and then describe the way that the access control engine works?

[0:24:34.9] MS: Sure. So what's can happen, let's say I'm running my laptop, everything's great, and then there is a new critical security patch from my operating system. Well, until that patch has been applied, I might be in a more vulnerable state. My machine might be less healthy, less safe, so the access control engine is going to get a new rule, say, from some of our platforms, operations or security teams saying, "If someone wants the highest level of trust, they need to have this version of security patch." Since I no longer have that — Since I don't have that patch

yet, I won't be able to earn that high level of trust. So when I make a request and the access proxy sees, “Max is looking for something that requires a very high level of trust, does he meet the criteria?” and is looking at the inventory data to see how trustworthy is my computer is the host that I'm using for this session right now.

As those policies change, because of patches, because of malware discoveries, because of new vulnerabilities or just policies that we change internally, the rules that the access control engine is testing are going to give a different answer for how much trust can Max earn right now.

[0:25:58.5] JM: The way that like an internal application implements its level of access or specifies its level of access is like if I'm building an internal application, if I understand this correctly, I have an access control language where I can specify my level of access. So the level of access they want users to access it through. So for example like if I have a job board within Google, probably have a pretty open level of access, but if you're talking about the internal —

[0:26:29.8] MS: Code review.

[0:26:30.4] JM: Yeah, code review, exactly for the search, like the core search algorithm, then that's probably something you want a high level of access control for. Can you describe — Do have that picture right? Is it the responsibility of an application owner to define the level of access that somebody can authenticate it with?

[0:26:48.4] MS: Yeah, although they’re usually working with teams in security and in an operation. So they might have a sense for, “This is a tool that shows me the menu of the café. So it doesn't really require much trust.” Versus this is a tool that lets me look at critical bugs. So it does require a high level of trust.”

It's not super detailed, so we don't want to create a lot of extra work. The team then might be deploying a new internal service. They're going to need to specify at least something there to say, “Here are the requirements for someone to get in.” So that the access proxy has a criteria to test against. But it doesn't need to be very complicated and it doesn't need to go into a lot of detail if that's not necessary for their service. Sometimes it’s going to be more based on a person type, right? “Max isn't in finance, so he can't get to some of the financial databases or

financial dashboards." That's just a factor of who I am or what user groups I'm in, and that would be a good example of something they might put in that access control engine for that, say, financial dashboards service.

[0:28:00.7] JM: I'm not sure how much detail you can talk about this or to what degree you're familiar with it, but I think we've talked about the finer points of BeyondCorp pretty well at this point. I'd love to know, like if you could describe kind of the end-to-end way that I'm a user, I'm just engaging with a Google network throughout the day, whether internally or externally or some combination of the two, and throughout the day data is being piped into the system that is going to evaluate my trust over time, and that level of trust is going to propagate to the access proxy, and then the access proxy is going to update whether or not I'm able to access various systems. Could you maybe go into a little more detail on the bigger picture of how the system works?

[0:28:48.8] MS: I'm not sure I understand. In terms of the day-to-day interactions bigger picture or what you mean?

[0:28:54.3] JM: So like I am walking around and I'm doing these different things with my different devices and the data is going into the system and maybe you can just like walk through an example of some events that might happen. For me as a user throughout the day, which are going to update the internal amount — The internal data lake. I don't know if it's a data lake that's being aggregated about these different devices, and then that's making its way to the access proxy or — I don't know. Is there any more detail to be explored there that would be worthwhile?

[0:29:26.3] MS: I mean, there's a bunch of different data points, some of them around where did you sign in or which devices are you using. I don't think that a day-to-day sort of normal situation is going to have a lot of massive changes. The reason that we would change someone's trust is because probably something strange happened, "Oh! Jeff was in California one day, but then all of a sudden was in India the next day or China." That might be a little weird, or that a new vulnerability was discovered, and now you are out of compliance not because of something you did, but because something you didn't do. So what we want to do is

make sure you don't get access to critical or confidential stuff until you're back in compliance with our security policy.

What you do day-to-day is not likely to really change much. Your machine still has the same identity, still has the same MAC address, still has the same DHCP data. Not a lot of that is changing. Maybe if you get a new computer, something would change and you might need to go through a sort of a trust bootstrapping process where let's say your computer just explodes and you're not hurt, luckily, but you need a new laptop. You go to your tech support, you'll get a new laptop. The laptop doesn't come with any trust built-in. We don't inherently trust devices. It's going have a certificate so we know what it is, which laptop it is and it'll go through — It probably gets a couple software updates after you open it up, but then what you can do from, say, your phone, which does already have some trust, is you can say, "This device, this new laptop, this is mine. So I want to establish a higher baseline level of attempted trust, and you can kind of bootstrap yourself.

So that comes back to the how supportable is this system. So we built a lot of processes to allow you to recover from an incident like that without even needing help from someone in tech support or operations, and you could just go to an internal inventory portal on your phone or on your desktop, if you have one, and say, "I want to sign up for higher level of trust on this laptop," and it'll earn that level of trust if you meet the criteria, say, the machine's up to day, fully encrypted, etc. etc.

[0:31:56.5] JM: So if a new vulnerability emerged, such as heartbleed, and you wanted to enforce that users updated their systems in order to maintain access. Like you need to patch heartbleed if you want to have access to Google systems, what would be the process for pushing out that change to access policy?

[0:32:20.6] MS: If someone would make a code change, that would be changing one of the rules, probably adding a rule or maybe a couple , because it might have some operating system specific rules around what level of trust you can and can't turn. In a case like that, what we might want to do is actually redirect people to a self-remediation guide instead of just saying, "No. You don't have enough trust." And that can leave people, "What am I supposed to do? I can't get my work done." That leaves them confused and they all go to support and then you

have thousands of people on the support desk. Instead, what we would do is still limit the amount of trust they can get it if they're in a vulnerable state, but also allow them to access a self-remediation site to take the steps necessary on their own to install a patch, fix a bug or whatever the case may be. Sometimes it's just updating a browser or operating system, I remember going through a lot of flash zero days a couple of summers ago. It was like every week, right? It was just a new one.

So support team actually made a quick page. They just said, "Check my flash. Am I safe or not?" Because people didn't know. I didn't even know on a day-to-day basis whether I was safe or not, because it just felt like every week there was a new one, and so I'd just go to this page and it would say, "You're good to go. You can go along and do your work." or "No. You're not in a safe state. Here's what you need to do." Probably just restart your browser and it will auto update. But it removed a lot of confusion. It got people on their feet faster and it prevented tens of thousands of people from flooding our support desks.

[0:34:03.1] JM: BeyondCorp was not the first system of security that Google had internally. There was a migration to BeyondCorp. I think — Is it still ongoing or is the migration complete?

[0:34:18.1] MS: I think it's mostly complete.

[0:34:20.3] JM: Mostly complete.

[0:34:20.8] MS: It was hard. We were also in the same boat that most big companies are in. We had a VPN, we had a privileged corporate network and we didn't really allow much access if you are outside the network. So that migration was hard, because in addition to this urgency of improving our security posture, we also had a mandate to not break anyone, to not impair people's ability to do work, and that's a challenging set of constraints to manage simultaneously. So we had to get clever about automating the scale. This wasn't a crazy manual process for many, many, many hosts, but also doing that migration in a way that would be invisible.

I think it worked pretty well. So a key part of that was understanding our network traffic. So turn back the clock, we're in this previous state where you're either on the privileged corporate network, and then you can't get access to all the good stuff, or you're off of it and you have to

set up your VPN software, which nobody likes to do or you have to do some — Jump through some hoops to get to anything, or you don't have access.

What we would do is we built a second network, a separate VLAN, that wasn't unprivileged managed network. So it's managed in the sense that we let devices on there explicitly. We'd assign them dynamically when we thought that they be ready. It started with no devices at all. We just had this separate network that has no devices and no special access, pretty useless. But then what we do is we look at the traffic for an individual host, "Here's Max's laptop, and based on the agents we have on the machine, we can say what his Max gone to over the last month? What kind of things is he trying to do on his computer, on his laptop?"

As we move more of these services behind the access proxy into that BeyondCorp world so that the test of, "Is Max authorized to see this thing?" is not a test of network, but a test of identity and device trust. As we shifted over gradually the services, we could then shift over hosts and people who were only accessing services that fit the new model. So we automated all of this. We monitor "What is Max doing on his laptop for the last 30 days?" and then replay it, a software simulation of the same traffic on the unprivileged network. Could Max have access to all those things in the last month if you were not on the privileged network, if you are on the normal no special privileges as if it's his home Wi-Fi network? If the answer to that is yes, I can — I'm going to notify Max, but I'm going to also — I love talking myself in the third person, excuse me. I'm going to give this person a heads up that we're shifting you so that you have instructions for, "Oh my God! Something stopped working. I can go back to the old way." But I can just shift Max over, shift someone over to the new unprivileged network and they shouldn't even see any change at all.

[0:37:37.5] JM: Was there an internal team that was charged with doing this migration or was it decentralized with various teams all doing something — Like different teams who are responsible for different applications being responsible for moving their own application to BeyondCorp?

[0:37:56.1] MS: You're right in both counts. The applications movement was the responsibility of those application owners, and so that was happening earlier. Moving the people was a different step and there was a team for all this, but it also ebbed and flowed constantly and it would may

be pull in more people from the platform operations teams or expertise we needed to understand how Linux versus Windows versus Mac versus Chrome OS were going to behave differently, where we'd pull in more people from a security team.

So this team was cross functional and fluctuating in size. So you did have a core team, but you also had that team making requests, demands, sort of negotiations with other teams, especially those application owners or the network operations teams and similar to do the work that they needed to do in order to enable this new model. It's a mix of the two, but there is definitely a core team that was managing the build out of this new context-based access trust system and the team that was managing the whole migration effort. How do we make it a good user experience? How do we make it not terrible support load? How do we build that simulator accurately so we could test is this traffic going to work, or if something goes wrong. It was another really interesting lessons learned that I don't think we anticipated at the beginning. Let's say we move someone over. So that whole process I was just explaining. Shift them to the non-privileged network, but then something does go wrong. Let's say they have a yearend reconciliation process that involves a specific tool. So we didn't catch it in our last 30 days, because they really only use it every December and January, and we shifted them over in June. So then December rolls around, they try to access their yearend tool and it's busted. They get a forbidden error because the access proxy, for some reason, is incompatible with this tool yet.

We need to have a way to figure out what happened. Why did someone get a know when they should have got yes? Now it's not necessarily safe to tell that individual, because they could be an attacker and we can't really tell, or they could be a malicious agent on a legitimate employee's machine, but we do need our support teams to have the tools necessary to debug, to troubleshoot and figure out, "Did this access get denied because they're in the wrong group or because their OS is out of date or something else?" That was a tool that we had to build in order to scale, but I don't think that we necessarily anticipated that from the get-go, although it proved to be really vital to scaling this across all Google employees and services.

[SPONSOR MESSAGE]

[0:40:59.7] JM: DigitalOcean is a reliable, easy-to-use cloud provider. I've used DigitalOcean for years whenever I want to get an application off the ground quickly, and I've always loved the

focus on user experience, the great documentation and the simple user-interface. More and more people are finding out about DigitalOcean and realizing that DigitalOcean is perfect for their application workloads.

This year, DigitalOcean is making that even easier with new node types. A $15 flexible droplet that can mix and match different configurations of CPU and RAM to get the perfect amount of resources for your application. There are also CPU-optimized droplets, perfect for highly active front-end servers or CICD workloads, and running on the cloud can get expensive, which is why DigitalOcean makes it easy to choose the right size instance, and the prices on standard instances have gone down too. You can check out all their new deals by going to do.co/sedaily, and as a bonus to our listeners, you will get $100 in credit to use over 60 days. That's a lot of money to experiment with. You can make $100 go pretty far on DigitalOcean. You can use the credit for hosting or infrastructure and that includes load balancers, object storage. DigitalOcean Spaces is a great new product that provides object storage and, of course, computation.

Get your free $100 credit at do.co/sedaily, and thanks to DigitalOcean for being a sponsor. The cofounder of DigitalOcean, Moisey Uretsky, was one of the first people I interviewed and his interview is really inspirational for me, so I've always thought of DigitalOcean is a pretty inspirational company. So, thank you, DigitalOcean.

[INTERVIEW CONTINUED]

[0:43:06.7] JM: That simulation process that you referred to a couple of times, so the process of building that was you recorded a bunch of network traffic and then you were able to replay it to make sure that a BeyondCorp implementation was not going to mis-authenticate people or prevent people from logging in who needed the login. Can you talk more about that simulation?

[0:43:28.9] MS: Yeah, that basically what it did. So there were a couple pieces to that. Part of it was around simulating traffic in a wide scope and a macro scale so that we could see — If you picture in the middle of the migration process, we're moving over many, many services and we can use that logging and the analysis of that logging to prioritize what's the next thing that's going to move. If we know that a thousand people need this financial reconciliation tool, but only a dozen people need this travel planning tool, we're going to start with the one that impacts

more people. So being able to log and analyze and simulate the traffic across many different people allowed us to make good priority decisions.

On an individual machine, we were able to also put a local agent on there that could essentially do that test before they moved over to the unprivileged network. So let's say that I'm in a pre-migration state. So I'm still connecting to the privileged corporate network, which is why I'm able to get to all my services. That agent could look at my traffic and simulate, "What if Max had been on the unprivileged network? Would he be able to get to the same services and backends and tools?" And we were able to turn that agent from monitoring to an enforcement mode. So you could actually say, "Make it so that he's on the unprivileged network or just simulate and log what would've happened." That was to make sure we didn't break a lot of people's workflows when we did migrate them and also to learn more about what are people using.

Sometimes in this process, we actually discovered services that we thought were done, that we thought were turned down, and they were just happily chugging along still without anybody really knowing about them, and that's kind of scary. So it was good to do that analysis also to discover things that maybe shouldn't have still been running but were — And we only found that out because of the large amounts of traffic that we looked at and that we monitor across our network.

[0:45:48.9] JM: It reminds me of a show I did pretty recently with a company called [inaudible 0:45:53.4] that does large-scale analysis of traffic flowing through the internet so that companies who become [inaudible 0:46:01.8] customers can find out about like how their traffic flow — Like servers that they don't even know about, like orphaned servers, which turned out to be like if your company gets to a reasonable size, you probably have some orphaned infrastructure that you are forgetting about.

[0:46:19.2] MS: Yeah. I guarantee it. I'm sure. So that was a great thing to learn as we are doing some of this analysis. There's a mix of local individual machine monitoring and playback, but also large aggregate logs monitoring and playback, and that way we could also migrate people with more confidence. We could say, "Hey, everybody on this finance team, all of them use these three tools and these three tools all work, and everybody's good to go. So we can shift over a group of people and if they have similar workflows and similar profiles based on

what kind of things they access, we can make that decision as a group and speed up the migration." It turns up that migration was about as much work as the design of the thing. Change is hard.

[0:47:07.9] JM: And was a lot of that diplomacy and writing documentation and explaining that the people, evangelizing.

[0:47:15.5] MS: All of the above had. You got to do all of them, but it's also building the automated systems. We didn't want to have to manually click through hundreds of thousands of hosts. So we needed to, instead, that whole traffic logging and simulation, that was to allow automated migration. So if we see, "Everything runs fine on the unprivileged network." No one had to push a button to then migrate that machine. That just kicks off an automated process, it sends an email to them, queues them up for seven days later if they're going to migrate, and we'd check in. So all of that can be automated so that the operations team is instead responding to weird corner cases or escalations or firefighting instead of the sort of manual toil of hands-checking another host to see if it could go over.

[0:48:08.9] JM: This sounds also like it was a pretty good way of polishing what eventually became this process of other companies adopting BeyondCorp, which I'd love to discuss more. I guess, what was the sequence of events? I know BeyondCorp, this paper was originally published the first one. When the first one published? Like three years ago? Longer?

[0:48:30.2] MS: Yeah, 2014. I guess we're 2018 now, huh?

[0:48:32.7] JM: Yeah. Oh, wow!

[0:48:33.4] MS: 2014 was the first paper, and only published a few more in 2016 and 2017, and I'm working on another one right now that I hope will be out before the end of 2018.

[0:48:43.8] JM: Oh, what's the focus of that one?

[0:48:46.5] MS: I don't want to talk about it too much yet, because I need to get a couple of more internal ducks in a row, but I want to keep telling you a story. I love to get feedback from

people, and I often will go to a conference and talk about this model and talk about both what we did at Google, but then what we've enabled people to do on our cloud platform, because I have a lot of conversations with leaders of IT and information security at big companies and they say, “This sounds awesome, Max. How do I do it? Can I buy that?” and for a long time I had say, “I'm sorry. I don't have anything to give you. I can tell you why we did it and how we did it, but it's a lot of work. It took us many years with many, many teams and hundreds of person hours and not a lot of companies have that to invest.”

So what we did is we took the fundamental philosophy behind BeyondCorp and a lot of the things we learned and built and turn that into identity-aware proxy, which is an access proxy on Google cloud platform that does many of the same things. Checks identity against a list or a group to figure out, “Do you grant access to this cloud application or not?” That way, another company can start using the same model with very little work.

[0:50:04.6] JM: These companies who want to go BeyondCorp, are there specific problems that they're experiencing, like people opening a Trojan horse PDFs or something like that? Being victim to phishing scams? What are the kinds of problems that these companies have that entices them to go BeyondCorp?

[0:50:27.1] MS: Yeah, all of the problems. Spear phishing is still a really effective way to compromise the upper levels of pretty much any kind of company. So they want to go because they also have seen how the work that we’re doing has changed. By we, I mean just people, people in technology. Employees aren’t just sitting at the office. They are working from home, they are working from the road. They're not just using a computer. They’re using a tablet or a phone.

They’re also frequently hiring employees that might be vendors or contractors that should really only have access to one specific tool, and in that old model, the castle and moat we were talking about, if you're inside the castle, you have access to all the rooms. So it's very hard to restrict a contract employee to just a specific set of tools in that all or nothing model.

[0:51:22.7] JM: This has been built into a Google cloud platform product. So it's IAP. Is that what you call it?

[0:51:32.1] MS: Yeah, identity-aware proxy.

[0:51:33.6] JM: Right. Okay. So what was the process of — I did a show but a BigQuery a while ago and one thing that was funny about that show was the guest, Jordan, mentioned that they thought it would be straightforward to productize BigQuery. It turned out to be a lot more difficult than they anticipated. I guess there was some internal couplings and other issues that just made it harder to productize than they anticipated. What was it like doing the productization of the access proxy, the BeyondCorp system?

[0:52:05.4] MS: I mean, it's definitely been gradual. IAP really just launched this past summer, summer of 2017. So if you think about how much time it's had to mature versus BeyondCorp, which really started at Google in, say — I don't know, 2011 or so, I think. There's a lot more that we can build into IAP and we're working on it very actively. I'd BeyondCorp is sort of wrapping up the shift to Google and the way that we've changed access to be context-based instead of network-based.
IAP is starting down that road and adding a lot of exciting stuff this year. It's also getting a lot of great feedback from customers who are using it to start the same model internally and then giving us feedback on what else they want to do with it.

So your question around why are people wanting to do it. I think people are starting to see that it is safer. That *t's* safer to grant access in a more granular way instead of that binary, you're in or you're out kind of model that a firewall VPN world gives, and they want to have more defense in-depth. They want to have just giving access to the thing you need right now as supposed to everything inside all at once.

[0:53:26.5] JM: Principle of least privilege.

[0:53:27.6] MS: Exactly. Yeah. There are so many different jargony phrases around zero trust networking, lease privilege, [inaudible 0:53:33.1]. They all kind of boil down to this same model, which is I'm going to give you just the trust that you can earn now that you need to look at

something you want as supposed to a whole bunch of automatic trust that I don't really monitor very well.

[0:53:50.9] JM: For these companies who are doing the BeyondCorp migration that don't necessarily have the Google infrastructure or the Google knowledge of computer systems, do you have any suggestions or tips for intelligently planning such a migration and maybe setting milestones along the way or big snafus to be aware of?

[0:54:17.1] MS: Two of the things that were challenging for us and that are going to be, I think, challenging for a lot of our companies, is getting your people and your devices in a really good shape. So the people part is really important for trust that's based on identity. Is this someone in engineering? Is the someone in finances? Is this someone in the US or in France or in China? A lot of companies don't have the consistent and reliable information about their people and the groupings of their people, and without that, you can't really have any identity-based trust.

So having a good handle on who has what kind of job? What kind of access does this job need? It's surprisingly hard in some places to really answer that; who's in engineering? Whose in finance? Kind of a question, and you need that if you're going to be using those determinations to grant or deny access to certain tools.

Then the second part is the devices. What devices are managed? What's on your network? What do you think you should trust in the future? If a company has gotten really loose, agile, laissez-faire about how people bring in devices or access these corporate tools from different devices, what I recommend is they take much firmer look at how am I managing my corporate devices? Which ones are in a good healthy state and which ones are unhealthy and how to make that determination?

Because if you just flip something on overnight that said, "Any unhealthy device can't get access." You're just going to break a huge portion of your company and you need to have a plan for how do I evaluate health? I think, first, just how do I get a full picture of inventory. That actually is surprisingly hard. But then, how do I tell what in my inventory is healthy and what isn't healthy and how do I move the unhealthy machines into a healthy state so I can give them more access?

[0:56:21.5] JM: What do you think of these vendors? I searched BeyondCorp on Google and I found a number of vendors who have built up consulting businesses around helping companies move to BeyondCorp. What do you think about this vendor ecosystem?

[0:56:38.9] MS: It's a hard problem. I'm glad to have more smart folks involved in solving this problem, because I'm doing this to try and make everybody more secure.I don't want more Equifax and neither do you. So if we can improve the security model that everybody uses, great. I think we're going to have a much better time being citizens of the internet because of it.

[0:57:03.7] JM: Does this extended my consumer login? Like when I, as a visited normal Gmail user, login to Gmail or Google Docs or whatever, am I using sort of BeyondCorp type system?

[0:57:18.6] MS: Not exactly. There are some overlapping components deep down in there around how much do I trust this login? Maybe if you've logged in from your friend's computer or a family member's computer instead of your normal laptop, you might get an extra challenge from Google saying, "Hey, can you confirm this is you," or depending on how you've set up your Google account.

So some of that logic is also built into the BeyondCorp logic, but this whole method of access and security is really for a corporate environment. You have some internal services, like in HR or a finance system and you need to protect those. So the consumer world has similar access security questions, but there is not this concept of, "You have a managed machine and I have this identity that's part of groups or organizational units or some work hierarchy," and so those components are specific to a company.

[0:58:19.9] JM: Okay. Just to wrap up. I noticed you have degrees in both CS and psychology. When does this psychology degree come in handy in your day-to-day life?

[0:58:31.4] MS: I was actually really lucky. I managed to find one degree that got both words on my diploma. So it was a fun major for me.

[0:58:36.9] JM: What was that degree?

[0:58:38.5] MS: It was computer science and psychology.

[0:58:40.0] JM: Oh, okay.

[0:58:40.9] MS: They had a combined major in my college.

[0:58:42.8] JM: Oh, fascinating.

[0:58:44.0] MS: Which is really cool. The psychology to me is important, because while computers can be very predictable in how they respond to a stimulus, people are often less so. When you're building, especially, security systems, most of the time social engineering is the most effective way to bypass security measures for any big a company. Spear phishing works by just tricking you into trusting something, because it looks trustworthy. So you need to understand not just how do the computer systems work, but how do people think or respond to them. Often they respond without a whole lot of thinking. It's just a gut reaction. It's a, "Oh yeah, I've seen this before. I just click on cancel button and then it works," and that's a dangerous pattern if it could be abused.

So when you talk about making a security system according "foolproof" or able to handle a wide variety of technically savvy or non-savvy people, I think that understanding of how people think and feel is really valuable.

[0:59:54.1] JM: Okay, Max. Well, great talking to you. I really appreciate you come on the show and I'm excited to see what develops at BeyondCorp.

[1:00:00.0] MS: Thank you very much.

[END OF INTERVIEW]

[1:00:04.5] JM: If you are building a product for software engineers or you are hiring software engineers, Software Engineering Daily is accepting sponsorships for 2018. Send me an email, jeff@softwareengineeringdaily.com if you're interested.

With 23,000 people listening Monday through Friday and the content being fairly selective for a technical listener, Software Engineering Daily is a great way to reach top engineers. I know that the listeners of Software Engineering Daily are great engineers because I talked to them all the time. I hear from CTOs, CEOs, directors of engineering who listen to the show regularly. I also hear about many newer hungry software engineers who are looking to level up quickly and prove themselves, and to find out more about sponsoring the show, you can send me an email or tell your marketing director to send me an email, jeff@softwareengineering.com.

If you're listening to the show, thank you so much for supporting it through your audienceship. That is quite enough, but if you're interested in taking your support of the show to the next level, then look at sponsoring the show through your company. So send me an email at jeff@softwarengineeringdaily.com. Thank you.

[END]