SEI Podcasts

Conversations in Artificial Intelligence, Cybersecurity, and Software Engineering

Visibility Through the Cloud with Network Flow Logs

Featuring Ikem Okafo and Dr. Tim Shimeall as Interviewed by Dan Ruef

Welcome to the SEI Podcast Series, a production of the Carnegie Mellon University Software Engineering Institute. The SEI is a federally funded research and development center sponsored by the U.S. Department of Defense. A transcript of today's podcast is posted on the SEI website at sei.cmu.edu/podcasts.

Dan Ruef: As we all know, organizations are increasingly adopting cloud deployments for their flexibility and cost savings. The shared security model utilized by cloud service providers removes some of the need for system administration and security but leaves the monitoring of hosted applications and resources to the organization. Cloud flow logs are a valuable source of data to support these security responsibilities and attain situational awareness. The SEI has a long history of support for flow log collection and analysis. In early 2025, we released open source capabilities to optimize cloud flow log collection and analysis for Azure and AWS.

Welcome to the SEI Podcast Series. My name is Dan Ruef. I am the technical manager of the Network Situational Awareness Group in the SEI CERT Division. Joining us today to talk about how you can use cloud flow logs to attain cyber situational awareness in the cloud is Ikem Okafo and Dr. Tim Shimeall, also from the SEI CERT Division. They have been leading our research in this area. Welcome to both of you. Thank you for joining me.

Tim Shimeall: Thank you.

Ikem Okafo: Thanks, Dan.

Dan: We always start our podcast by asking each of our experts to share some interesting parts of their background and explain how they got to be here with us, sharing their information.

Tim: Well, my background prior to the SEI was academic, teaching out in Monterey, California at the Naval Postgraduate School. I did a visit here at the SEI back in the 1998 timeframe and found it so attractive that I changed Monterey to beautiful Pittsburgh and moved out here and have been enjoying life here in Pittsburgh at the SEI since. I have done a number of interesting projects including monitoring networks as the century changed in 2000 and exploring the network security at the 2002 Olympic Games and working some of the work on insider threat investigation that the SEI has been doing, including having lunch with an astronaut, which was actually pretty cool, at Johnson Space Flight Center. Since 2002 I have been working with the network flow log work, principally with the SEI SiLK [the System for Internet-Level Knowledge] tool suite. In recent years, I have been focusing more on cloud flow logs and the analysis that can be done as part of the shared security model.

Dan: All right. Ikem, how about you?

Ikem Okafo: I don't think I am quite as interesting as Tim was. But prior to joining the SEI, I worked as a DevSecOps engineer for about four and a half years, setting up cloud infrastructure and monitoring. Outside of that, I do have quite a wide array of interests. I play multiple musical instruments. Currently I found a new fascination for signal processing, things like EEGs and things like that, which is really fun. Then for the past two years or so, I've been trying to go into the realm of quantum computing because I find that interesting. However, the one thing I can tell you is after about two years of learning quantum computing, I still don't understand a thing. It is not very intuitive. Here at the SEI, I work as a security analyst working a lot with cloud flow logs and network flow logs to say the least. I have done some AI agent research: AI agent vulnerability related to MCPs- model context protocol. Yes, I think that is a summary.

Dan: All right. Thanks. For our audience members who are new to this topic, can you give us an overview of cloud flow logs and their role in securing a network?

Tim: Well cloud flow logs, like other flow logs, are a log of the connections made to and from a network past some sensor location. The goal there is to simply have a reliable record of what traffic has passed across the network, a pretty inclusive record. Not just things that are wrong, but things that are normal. That supports a variety of different uses including usage monitoring. *What are my top consumers? What are my top producers? Are there any misuse of resources that is going on?* It supports service verification. *Are my critical services still running as expected? Are there unexpected delays?* Those kinds of things and threat detection as well as network forensics.

Ikem: To add to what Tim said, in summary, cloud flow logs enable you to understand the behavior of your network. What is going on? Open ports, closed ports, your traffic flow. It helps with understanding when things are normal versus anomalies, which is a key reason why you want to actually collect these logs.

Dan: Great. Thank you. It sounds like there are many similarities between cloud flow log analysis and usage and the traditional networks that we have been monitoring for decades. What are the key differences between collecting flows on an organizational or on premises network and the cloud?

Tim: Well, principally, Dan, when you are doing a cloud-hosted service, you are not owning or controlling the hardware that is involved. You don't see the network interface card. You don't have the direct control over what is going on. That is both a benefit and a drawback. The benefit for cloud hosting is that it allows you to have more expansive usage as your network usage increases, and to also focus very specifically in on what is of security interest to your organization, which tends to be applications and data, those kinds of things rather than open ports or connections or those kinds of things. Some of the security weaknesses that would be of concern in traditional networks really kind of drop out because the hosting service handles it, including things like scanning or incomplete connections. Those are generally not visible on the cloud flow logs because that is not something you are going to be charged for. That is not something you are going to get involved with. There are also visibility differences in the sense that I see perspective based on successful connections rather than on attempted connections that were not successful on linking it to my network on my cloud infrastructure. I know Ikem has done some work looking at various cloud hosting options. Maybe

he could follow up with some of that.

Ikem: Yes. In addition to everything Tim just said, another key difference is cost. You have to pay for the generation of these logs. You have to pay for the transportation of these logs to the storage bucket or storage blob. You have to pay for that storage itself. If you want to actually use those logs in a useful manner, you either have to enable traffic analytics, which is a cloud vendor specific tool, or you have to do those optimizations yourself, which in and of themselves those compute also cost money. These are things that typically deter or discourage organizations from collecting flow logs or actually doing meaningful things with them.

Dan: OK. And Tim, you mentioned that the cloud service providers take care of plenty of things. Do they take care of making sure that cloud flow logs cover all of your resources?

Tim: No. That is something you have to configure. Generally, you have to specify what resources are going to be covered within the cloud flow log and exactly how you want to deal with that. What they do control is what the format and content to those logs are. There are only certain vendor supported options which are available, and you really can't exceed the vendor supported options without a lot of effort instrumenting it. The results are likely to be inconsistent if you try and avoid the vendor-supported options.

Dan: OK. So you are saying we can't assume that when you spin up a new resource in the cloud that it will be covered by flow log collection like it is on premises when you plug in a new laptop connected to the network and under the router.

Tim: Correct.

Dan: We have to keep track of all of that. What risks are likely detected using flow logs? Are those risks and things that we want to look for and do analysis very different from the on-prem traditional collection?

Tim: No. There are a lot of risks which are in common between cloud and on-prem, and those would be covered by both sets of logs or available through both sets of logs. The most obvious of which is service crashes or is shut down. That will be visible in your cloud flow logs by connection failures and the lack of presence of successful connections. Data volume abuse, in the sense that somebody is making far too many connections or far more

connections than is typical, and there is a sudden change in behavior in that regard, that would be visible through your cloud flow logs. An example would be exfiltration of data using DNS [domain name system] information. Typically, suddenly a large increase in DNS traffic because the attacker is now using subtly different DNS connections to pump data out of the network and into their collection framework. We mentioned top talkers. Yes, overall usage monitoring in terms of who is talking most and who is receiving the most data across the network, and whether that matches your usage profile that you are applying or if something is going astray. And clearly also after-incident forensics.

Dan: Makes sense. It sounds like the cloud is almost just computers that can still be hacked and be misused in similar ways as on premises with just different infrastructure surrounding it.

Ikem: Absolutely, yes.

Dan: When we are looking for those risks amongst our resources in the cloud, are there ways to break up or categorize or group the different resources beyond just, *It is in my cloud*.

Ikem: Various cloud vendors or cloud providers do things a certain way. For instance, Azure has these things called network security groups[NSGs], which are basically filters that are attached on a either per subnet basis or to your NICs- your network interface cards. You can visualize or collect logs on traffic flowing through your NSGs. Then you can also, which is something relatively new, about a year or so old, is a VNet flow logs, where you can enable collection of logs on a VNet scale. If you were to enable logs on an NSG, for instance, you would have to do that for every single NSG within your infrastructure, which you might miss some of them. If you have a very broad or complicated infrastructure, you could actually, you could miss some things and therefore not collect the full volume of logs. I think Azure, or rather AWS, does things a little differently with your VPC flow logs, which isn't quite exactly like Azure does it. This is something that will come up sometime during our conversation where there is a lack of standardization in terms of log collection or flow log generation, rather.

Dan: OK. So, the flow logs can be organized similar to how your network is organized. They can be groups. You can know where the flow logs came from, from particular Vnets, network security groups or VPCs.

Ikem: Correct. Correct. More specifically, the NICS they actually came from,

network interface cards they came from.

Dan: Oh, wow, OK. You can really drill down into what you want to focus on. So now that we know what you could look for and what the benefits are, and to some extent how to set them up, what are tips to getting started with these cloud flow logs? Because the files stay in the cloud, they're generated from the cloud. How do I say, *Please monitor and give me flow logs*.

Ikem: First you want to start off with identifying your key assets. *What are my key assets within our cloud infrastructure that we absolutely need to monitor and understand what is going on with them*? After that and collecting those logs, you actually want to make sure you are, in fact, analyzing those logs. You don't want to wait for an incident to occur before you go through your logs. You have to either go through them monthly, quarterly, whatever it is. But you have to go through them to understand the behavior, or what you are seeing to avoid incidents occurring in the first place. So be proactive, not reactive.

Tim: Following on from Ikem's points, there is also the sense that you really have to have an understanding of what kind of output you are most interested in. Are you strictly interested in understanding threats and the response to threats that some analysis that you will be doing, or are you more interested in understanding, *I just want to make sure my service is reliable. It is running. I don't care necessarily why it stopped running, but if I need to restart it, I need to know that fast.* So more of a real-time processing of the flow logs. Its streaming options, which would be exported out of the cloud, could be of great interest.

Dan: All right. Once someone decides or determines what they want to focus on and what they want to monitor, how can they go in and just get some sample logs or get started with looking at how can we get the cloud to generate such flow logs?

Ikem: I am going to go from an Azure perspective here. Azure has this service called Network Watcher. That is where you go in to trigger your NSG flow logs or VNet flow logs. As part of that configuration, you have to specify a storage blob that those flow logs get dropped into. You can trigger that for an hour or so. They are typically collected at layer four of the transport layer. Azure specifically samples it or collects those data in a one-minute interval or at a one-minute interval, right? You can trigger for ten minutes or whatever it is to get a sample or an idea of what those logs look like to begin with and then start to formulate your plan optimization wise, because, as we will

discuss at some point, there is actually some formatting issues with those logs in terms of usability.

Tim: From Google's perspective, Google has a network management software package that you have to go through and configure how you want the network logs to do in place. It is very much an interactive user interface that the network managers can make use of, to configure where the logs are going to be generated and where the data is going to be collected on. For AWS, there are several different collection options which are available. They have various configuration options for dealing with that. You would have to determine which configuration options support your needs the best.

Dan: OK. So it sounds like there are many, many knobs and switches that can be toggled to fine tune what you want. But to get started, just picking what you would like to monitor and telling the cloud where to put some of those logs for a period of time is at least a way to get them to see the logs to get started?

Ikem: Correct. Yes.

Dan: Once folks have gotten started, and they have begun to collect and look at their flow logs, what are some collection and processing issues that could come up with those cloud flow logs?

Ikem: One of the biggest issues we have seen is the usability issues. Sorry about that. Some of these logs come in a nested JSON [Java script object notation] format that make analysis very difficult. Part of what we did here at the SEI was to create a script that consolidates those logs, flattens those nested JSON structures to make analysis easier. And I will pass it on to Tim to talk more about the script.

Tim: Well, the scripts basically make the security relevant aspects of the flow logs much more visible. It is not dealing with the accounting-level materials in terms of *OK, which account is this associated with, which charging environment is there*? It is looking more at what services, what ports, what transfer has occurred, and making those more available to analysts as they go. There are several different ways that we can use these scripts, and we will get into that a little bit later. But the scripts themselves just help raise the visibility of the aspects that a lot of organizations are interested in for security analysis. One of the other issues that we come into mentioning the need to transform some of these scripts to improve usability, is you need to make sure you have enough processing power to handle the data volume that is going to

come across. You are logging every single connection coming into your cloud hosting environment. That is a lot of connections or could be a lot of connections. You want to make sure that you are able to have sufficient processing to keep up. Cloud helps you there by the capability of spinning up additional resources. But that again is going to be a cost magnifier. You are going to need to have a balance point of what we collect, how much we collect, those kinds of things in order to provide for effective processing and collection of logs. There are a few host-centric issues. There are some visibility limits as to what we know about each individual host. They tend to be represented as instances rather than particular addresses or particular hosts. Largely because cloud lets you rehost things rather rapidly and expand things rather rapidly. Those kinds of aspects are really hard to track analytically where if you tie it to an instance, it is a little bit more persistent related.

Dan: Are there any storage efficiency problems that you could run into with these logs or ways to make it more efficient so you can store more of it?

Tim: JSON format is voluminous. It is a text format. So you are not using some of the efficiencies you can gain by storing things in binary format, for example. The flip side is that there are a lot more flexible tools around to handle it, because there are a lot of JSON-formatted tools that you can load data into and get different results as you process it. Because it is so voluminous, you want to be looking at compression issues and compression capabilities, which again adds to your processing load. You also want to think about how long you are going to retain the logs for in order to keep things managed. The need depends on whether or not you are really interested in what is happening now on your network or what is happening historically. The wrinkle there is you want is now to include the time period, which you are likely to detect an attack, and experience has shown that can be months. You need to allow for sufficient storage to store several months' worth of data rather than just what is happening this week and storing this week. And that, again, adds to the storage load. I am sure there are other encoding issues there that could be improving the storage capability. That is still somewhat being experimented on.

Ikem: Agreed. And there are certain fields also within the collected logs that are not necessarily needed for your security for security purposes, which could be pulled out to furthermore consolidate those logs and improve storage.

Dan: Do we know if the cloud service providers are using traditional network

sensors with efficient flow packaging like Zeke or YAF [Yet Another Flowmeter], or if they are generating the data off of their own mechanisms.

Tim: The strong impression is they are generating data off their own mechanisms. A lot of the hallmarks that would show Zeke-encoded or YAF encoding just are not present. It is not using any standard format or even any commonly accepted format across the industry. There are variations between various cloud flow providers in terms of how they collect and what they collect that are moving forward that way. It is a service, and the service is depending on the service offerings of the various cloud service providers.

Dan: What are the differences between those logs from the various cloud providers? Because like you said, they are not using a standard traditional sensor. What are those differences, and how do we see that, and how does it affect any analysis or storage?

Tim: Well, the most obvious difference is some providers collect one piece of information, and other providers do not. As I recall, Google only collects TCP flow logs rather than looking at other network services that might be implemented on UDP. There may be ties to how the data is encoded in the sense that if you look at AWS or Google it is collecting things on a connection-by-connection basis. If you look at Azure, it is connecting things on an event-based basis, saying, *OK, this this activity is beginning. This activity has continued throughout an event at this time. This activity has ended at this time*. And it only gives you the data volumes at continue and end periods. So it is very different sort of data encoding and data storage methods being used, principally related to their collection infrastructure. There are varying time intervals which are available. You can configure time intervals saying, *I want to collect data in five-second buckets, in one-minute buckets, in five-minute buckets*, *maybe even as long as ten-minute buckets* and saying, *Show me what the activity during that period, those periods of time.* One of the big risks is you get issues of where they are encoding things in ways that make it very difficult to look back months and say, *What entities were actually active*? So. since instances come and go, the fact that instance X was active back in September doesn't necessarily mean the instance Y isn't experiencing the same connections here in November. You have to look at ways to more persistently identify hosts and connections and services across the time interval of interest within the analysis.

Dan: OK, so the different cloud service providers have different formats of their flow logs and sometimes different contents of their flow logs.

Tim: Yes.

Dan: What approaches have been tried to address these differences, especially in hybrid clouds where you can combine cloud flow log collection with on-premises flow log collection?

Tim: Well, the simplest approach is you simply have different analytical pipelines. You are making analyses from the initial data based on its data format and its configuration, and you are just running through an idiosyncratic analysis method for that particular format and content and coming up with as close as you can the results that you are interested in. And you are doing that separately for each of the log streams from each of the cloud service providers. The advantage is you are using as much of the data available from the cloud service provider as possible. The disadvantage is when it comes time to consolidate across your network and say, *What is happening across my network*? The pieces don't line up all together. You get pieces which overlap and pieces which don't, and that can make the results very difficult to interpret. We actually wrote a blog post recently on this subject where we went through some of these various options.

A couple other options are we have separate analyses to produce the exact same format of results, so that brings things a little bit more into conjoining processes. You again have separate data and separate analysis to get to the data. But the results are presented in a common result format so that you can more easily draw and say, *OK, AWS shows us this. Azure shows us that. Google shows us this. Let's put them together and get a common view*.

The final approach is you come up with a composite view of the data itself. Before you do any analysis, you transform the data into a common shared format and then go forward with a shared analytical pipeline and shared results from there. That allows you to make use of, say, more standard formats such as IPFIX [Internet Protocol Flow Information Export]. Transform the nested JSON into a more IPFIX type format and then apply IPFIX sensitive tools moving forward on that. Ikem, you have also explored some vendor approaches, isn't that correct?

Ikem: Correct. There are several third-party vendors that provide tools or services that could be used to better, I guess, address these differences between log formats collected from various cloud providers. But you have to pay attention to the contracts. Their capabilities. Theirlimits. They are not all perfect. Actually, they are not perfect. So pay attention.

Dan: It sounds like with the different formats and the different contents, you either convert to the same at the beginning or combine your separate analyses at the end. But somewhere along the line, things have to be put together for a single picture.

Ikem: Correct.

Dan: What other data from cloud providers could be useful for enhancing security and combining with these flow logs that we are now collecting and analyzing.

Tim: That brings up the whole area of data enrichment. So can we tie an external endpoint with a particular location? Can we try a service not just saying this is web service, but is this web service to ESPN, to a local school district, or to a critical partner? And the security attention we want to attach to those various classes of websites was very different. We are much more interested in making sure our critical partner gets rapid response than we are with our ESPN connection. Unless of course you are an ESPN event provider or something like that. There are probably service-specific logs that are being collected as well. This is particularly true for software-as-a-service hosting. That may be all that you have. Rather than going with cloud flow logs, you are going with service specific logs and then lining it up with what usage monitoring you have available to you. There is connection-based logging, specifically where you are looking and seeing what authentication is taking place, a user identity is associated with this connection, so you can more closely identify not just the what is going on, but the who is going on and where that is possible, that could be handled. I am not sure what other logs we want to discuss. Ikem, you have some thoughts?

Ikem: I can chime in. Yes, things like DNS logs, Route 53 logs. From AWS, you know S3 bucket logs or storage blob logs and CloudTrail as well, which more is dealing with IAM [Identity and Access Management] like, API authentication and things like that, which give you a more holistic view of your whole cloud environment. You can kind of couple these logs together with cloud flow to actually make more sense or extract more information from your logs. Yes, everything Tim said and yes.

Dan: Wonderful. Are there other tools or resources available, for our audience who want to learn more about your work in this area?

Tim: Most definitely. Ikem, Dan and others have collaborated on a recent article in *CrossTalk,* which is an open-source journal. They talk about this

particular area and a lot of the common threats that we talk about it. Certainly the blog post that I mentioned that has appeared recently, is also supporting. There are a lot of web resources on flow record analysis in general and cloud flow record analysis. The SEI in particular has within its library, an archive of many past presentations on cloud flow analysis, some by the vendors, some by user organizations, particularly associated with conferences in the past.

Dan: Ikem, you mentioned the scripts and capabilities to optimize the flow logs. Where can those be found?

Ikem: Oh, yes. Those scripts can be found at tools.netsa.cert.org. And we have two for Azure specific flow logs. One for NSG, one for VNet flow logs as well as a script for optimizing VPC flow logs as well, which is an AWS-specific service. Then we also have certain other tools like SiLK and YAF.

Tim: Frankly, there is a lot of information on cloud resources and on traditional flow analysis on that site. My understanding is it gets about hits a month, in terms of activity. It is being used and is affecting the community that is out there, but it is still also under active development. There is a lot of work that's being done to research new approaches.

Dan: Wonderful. And we will be sure to include all those links in the transcript, because URLs are always difficult to remember. All the links to blogs, capabilities, documentation will be made available very easily. What can we bring you back to talk about in a few months? Where are you hoping this work goes or where are advancements needed in this area?

Tim: There are a lot of directions in which to improve things. Certainly, as we gain experience in using the logs and in talking with adopting organizations about their needs, there is going to be revisions in terms of software and experiences based on that. Some of this is going to be very domain specific. Talking about various manufacturing technologies, talking about various production technologies, plugging it better into a DevSecOps view and some of the adaption that is made there. There is also a lot of questions about how do you improve the efficiency. Do all need to be treated the same, or can we gain storage efficiency by reducing some of the less informative connections into the network? Some of this may also start adopting some of the some of the AI views, particularly if you're talking a lot of information and boiling down a lot of information, to isolate out the significant parts of it is something large language models typically are pretty good at, or machine learning models are particularly very good at. As you tell them, and point

them, and give them more insight about *this is what I'm interested in.* Starting to exploit some of those technologies in terms of more advanced approaches. I can't guarantee that's going to be the next few months, but it is probably going to be the foreseeable future.

Dan: Seems like a logical evolution. Tim, Ikem, thank you for joining us today. We look forward to bringing you back for future episodes and hear updates about the work and where things are moving forward. Finally, a reminder to our audience that our podcasts can be found on SoundCloud, Spotify, Apple Podcasts, and of course, the SEI's own YouTube channel. If you like what you see today, give us a thumbs up. Thank you for joining us.