www.pipelinepub.com
Volume 19, Issue 1

Cybersecurity: Finding One Needle in Many Haystacks

By: **Mark Cummings, Ph.D.**, **Bill Yeack CSE**

How is it that after all these billions of dollars of investment, the bad guys can still get in? The good news is that today's cybersecurity technology blocks 99 percent of attacks.The bad news is that there is such a massive volume of attacks that the remaining one percent that gets through can create a financial and operating nightmare.

For example, in the first half of 2022, there were 236 million ransomware attacks reported during which the bad guys got in. There were likely many more that have not become public. This means that adequate protection relies on quickly finding and neutralizing the one percent. Unfortunately, the indications of an attack are a tiny needle in the vast collection of haystacks of data. The larger the collection of haystacks, the longer it takes to find these needles. Making the collection of haystacks smaller dramatically reduces the chances of finding them at all.

There are two paths emerging to solve this problem:

1. New central site hardware based on emerging massively powerful chips
2. Non-centralized security orchestration architectures

The optimal solution is likely to be a combination of these two paths. This combination is especially important in light of the fact that an increase in hardware horsepower is likely to also be used by attackers. Both emerging solutions are briefly described below.

The haystacks

The power of our interconnected world has created amazing advantages. As a result, mid- to large-size organizations have networks of thousands of phones and PCs, hundreds to thousands of servers, tens of private clouds, tens of public clouds, millions of apps and applets, billions of accesses to the apps, and millions of emails, web accesses, app accesses, and so on per day. Metadata on this can be quite large. Buried in this metadata (many haystacks from all of these many sources) is one piece of data that shows abnormal behavior (the needle). Typical tools in use today gather all of the metadata possible, store it in one place—and then seek to find that needle. The more data, the more likely the data is to include the needle. But the more data there is, the longer it can take to find those needles that are the symptoms of a breach.

How large can the haystacks get? Industry observers have told me that users of one industry-leading tool have generated so much cybersecurity metadata that they have had to create a second data center the size of their primary operational data center to store it. Thus, there is also additional overhead in collecting, transporting, securing, and storing this information.

With current AI technology, this search process is done through pattern recognition. Inference engines look for patterns that they have been trained to find. There are problems with both the time to train and the time to find. Finding is done with inference engines running on trained models. With general purpose clouds, it can take hours, days, or weeks to find the needle. And that is if the attacks are still following the same pattern the system has been trained to detect, end user or app behavior has not changed, and the overall system has not changed.

This is the challenge—while industry insiders are telling me that, in some cases, the time to catch a successful attack before it takes over your whole network is now less than 20 minutes. This is part of the reason we have seen so many successful ransomware attacks. A senior tech executive at Google, responsible for a Google AI accelerator chip, told me that with their industry-leading chip, it took two weeks to train a system. That is two weeks after all the training data and all the other resources including programmers were available. Unfortunately, sophisticated attackers are changing their attacks frequently. These changes happen as fast as every hour, with some in just minutes.

Because of current processing times for both training and inference, AI systems are not able to respond as quickly as we would like. In addition, identification of symptoms of a breach is only part of the job. Sometimes these symptoms are not caused by a breach, but by something else—thus are false positives. These AI systems leave it to

others to filter false positives, and if true positives, stop the breach and fix the problems it caused. Leaving it to others to complete the process of identification and remediation cuts across the rest of an organization's cybersecurity tools. It is not unusual for a mid- to large-size organization to have 150 or more separate cybersecurity tools in use. Each works well within its own sphere—but not with others, particularly when they come from different vendors. Staff are left with the job of manually tying all these together. In addition, they must determine what is a real breach,

not a false positive; what remediation is required; and what manual intervention in technically challenging, complex, volatile systems to actually perform for this remediation. This leads to lots of security holes that attackers take advantage of.

Emerging innovative hardware

Large-scale use of AI systems was initially targeted at selling products. This involves determining which consumers were good targets for a particular product, and what, when, and how to most effectively present content to get them to buy. Inference engines were built on existing racks of X86 and ARM processors (CPUs) deployed in central site cloud systems. Timely response was important. Systems needed to figure out what to show a person visiting a web site before they would click and go to another page. Existing processors were slow and consumed a lot of expensive electricity.

As gaming technology evolved, graphic processing units (GPUs) for gaming PCs were developed. People began using the GPUs to accelerate inference. Over time, the CPU and GPU vendors started customizing their chips for AI. NVIDIA became one of the leading GPU vendors and started offering an AI chip in the $1,500 range. Such chips demonstrated that there was a very attractive market opportunity for AI accelerator chips.

Today there are approximately 40 chip startups addressing the AI market, with still others in stealth mode. Some are focused on other segments such as autonomous vehicles. The others tend to be focused on central site AI. In general, they each have an innovative architecture that seeks to increase speed and decrease power consumption very dramatically. Below are four examples.

Esperanto's architecture uses 1,000 processors on a chip to generate lots of processing horsepower to speed up AI inference while keeping power consumption low. According to company materials, Esperanto's chip includes working silicon and rack-mounted systems that interconnect large numbers of chips. The chips interface to the most widely used types of inference models.

Groq's architecture breaks down processors into their constituent parts, then arranges a matrix of them in a bucket brigade fashion. Argonne National Laboratories says this dramatically speeds up inference. According to Groq, their architecture allows very efficient data stream-oriented processing that can apply a single data input stream to many models. Like Esperanto, Grog incorporates working silicon and rack-mounted systems and the chips interface to the most widely used types of inference models. Furthermore, Grog offers software development tools end users can use.

Cerebras' architecture uses wafer scale integration—a single piece of silicon the size of a dinner plate. This dinner plate consumes 20,000 watts of power, so it needs a large, sophisticated cooling system. The company appears to be focused on training AI models, where they can reduce training times from weeks to minutes.

Abacus Semiconductor Corporation's (ASC) architecture is focused on making the interconnect between processors more efficient. They describe how current processors in multi-chip arrays spend 95 percent of their time handling communication between processors or waiting for results to come back from another processor. Reducing these delays can make ASC systems far faster and more power-efficient than the others. According to their materials, they have interfaces that support common AI models, but also common general purpose cloud apps, and high-performance scientific apps. They do not say that they have working silicon.

Attackers and new hardware

The creation of new chips introduces the possibility of creating two new types of threats. First, the new chips themselves will be subject to attack. It is possible that the designers have taken precautions against that, but there is always the possibility that the new architectures will open up new vulnerabilities. Possibly more significant is

that it is likely that attackers, especially state-sponsored attackers, will acquire these new chips. Those working on quantum encryption have already anticipated increasing horsepower being available to crack codes. So, that is not likely to be a problem. Other ways that innovative attackers might use the new architectures are not so clear; however, it would be prudent to assume that there will be attempts to do so.

Non-centralized security orchestration

Another approach to finding the needles is to not have haystacks in the first place. Don't gather all the data about the network in a central site and then try to find the needles. Instead, put specially architected intelligence out in the network. Each location only looks at local data. Locations cooperate with each other in a simplified fashion to determine and track normal behavior and find deviations that indicate a breach. They also work together to determine the best way to stop the damage and repair it—and then do so. This approach has been designed, demonstrated, and patented but not yet productized.

Attackers and non-centralized security orchestration

Attackers have been quick to adopt the latest technology to automate attacks and so, they are likely to use orchestration technology internally. A greater risk, though, is that the non-centralized system itself will likely become a target. Recently, we have seen several very large, successful attacks based on compromise of a particular piece of infrastructure. The reason this has happened is that the infrastructure vendors did not understand and recognize the risk. Thus, from the beginning they did not design in adequate protection. In the case of this non-centralized orchestration, specific steps have been taken to strengthen its skin, and to quickly find and neutralize intrusions.

A combination is optimal

Each of these technologies has valuable strengths. But these strengths lie in different complementary areas—thus, combining them will create the optimal outcome.

Non-centralized orchestration can very quickly find attack symptoms that are localized in space and time. They also are well-positioned to execute remediation. This is because they focus on collecting data from their local area and not on holding large amounts of historical data. In addition, they have local access to control mechanisms.

Centralized systems collect data from the entire network and hold a lot of historical data. Because of this, they have trouble quickly finding attacks with localized symptoms. But they do well in finding attacks that are slow to develop, whose symptoms spread widely in the network. An example of such is 'deny, deny, admit.'

For example: In one type of a deny, deny, admit attack, an attacker attempts to log on in Berlin and is denied. Then, tries in London and is denied. Finally, tries in San Francisco, and is admitted. The San Francisco system doesn't know that the attacker has attempted and failed in two other cities.

Both a non-centralized and centralized system can find this pattern. A central site system with data from all three cities already has all the data necessary to find this kind of pattern. With hardware acceleration and frequent downloads of data from the field, it may be able to quickly find such an attack. A non-centralized orchestration system node in San Francisco has to ask for data from outside its locality. It can do this, but maybe not as efficiently as the central site system. On the other hand, the non-centralized orchestration system is in the best position to perform automated remediation.

Thus, an optimal implementation would be a combination of the two.

Cybersecurity technology innovation

The cybersecurity challenge doesn't stand still. Attackers have been increasing the number, variety, and rate of change of attacks. Trying to find the attacks that get through our outer defenses has become a data overload problem—like finding a needle in a bunch of haystacks. Fortunately, technology available to counter these attacks has also been moving forward. An emerging class of semiconductor processors that can act as AI accelerators can be combined with non-centralized orchestration to quickly find these needles and remediate the attacks.

Disclosure: the authors have relationships in both the semiconductor and orchestration spaces.