computers are bad

https://computer.rip - me@computer.rip - fax: +1 (505) 926-5492

2023-08-07 STIRred AND SHAKEN

In a couple of days, I pack up my bags to head for DEFCON. In a rare moment of pre-planning, perhaps spurred by boredom, I looked through the schedule to see what's in store in the world of telephony. There is a workshop on SS7, of course [1], plenty of content on cellular, but as far as I see nothing on the biggest topic in telecom security: STIR/SHAKEN.

I can venture a guess as to why: STIR/SHAKEN is boring. So here we go!

The Nature of Circuit Switching

Understanding today's robocalling problem requires starting a long time ago. Taking you all the way back to the invention of the telephone would be a little gratuitous, but it is useful to start our discussion with the introduction of direct distance dialing in 1951. In that year, the first long-distance call was completed based only on the customer dialing a number. Over the following decades direct distance dialing became more common and fewer telephone users had to speak to an operator to have a long-distance call established. Today, it's universal.

Handling dial calls over long distance trunks is a bit complicated, though. For local calls, handling was relatively simple. The other customer was connected to the same exchange that you were, so the exchange just needed to be able to detect your dialing and select the correct local loop corresponding to the number you dialed. Step-by-step (SxS) switches have been handling this problem since the turn of the 20th century. For long distance calls, though, the recipient will not be on the same switch---they'll be on a foreign exchange.

To establish a long-distance call, your exchange needs to connect (via a trunk line) to the exchange of the person you are dialing. Most of the time there is no direct connection available between the two exchanges, so they need to call each other through a tandem switch, sort of a telephone exchange for telephone exchanges.

If you are familiar with the general concept of pulse dialing and SxS switches, you might already see why is tricky. When you dial on an SxS switch, the dial in your telephone moves in a sort of lockstep with the selector mechanism in the telephone switch. Each digit is regarded separately. The switch doesn't know any "history" of your dialing: each selector switch just knows that you reached it somehow and it is responsible for connecting your call to the next step in the switch architecture. It doesn't know what digits you dialed before, and it won't know what digits you dialed after. It just advances one step for each pulse, as long as it's connected.

The SxS switch is the main origin of the film and television myth of the 60-second telephone trace. "Hawaii Five-0" (the original) is one of the few television shows I know of to accurately depict the drama of a telephone trace on an SxS switch, with the cops trying to keep the killer on the phone as an exchange technician with a clipboard rushes around the aisles of switch frames writing down the positions of the selector switches. "Tracing" a call really involved tracing, backtracking through the switch from the local loop of the callee to the local loop of the caller, one selector at a time.

The calling process between local exchange switches and tandem switches is sort of similar, except that direct dialing requires some form of memory. When you dial a number, your local exchange switch matches it against patterns and determines that it needs to be sent to a tandem switch. It connects your call through a trunk to the tandem, but you've already stopped dialing, so there are no pulses for the tandem to use. For the tandem switch to know what to do, your local exchange switch must store the number you dialed, and then dial it again over the trunk... essentially, you dial your local exchange switch, your exchange switch dials the tandem, and the tandem (assuming it has a connection to the destination) dials the local exchange switch on the other end.

There are a surprising number of distinct ways this has been implemented but I will use the general term outpulsing, most descriptive of a scheme where each switch sends pulses to the next just like your rotary phone. In practice, a more common system (prior to digital signaling) was "multifrequency" or MF signaling. MF is not to be confused with dual-tone multifrequency or DTMF, which is based on MF but different.

This communication between switches, required for long distance calls and today even on local calls for functions like local number portability, is known as signaling. Outpulsing and MF are examples of in-band signaling, where the signaling is sent through the same path as the voice conversation it's setting up. Today, in-band signaling has largely been replaced with out-of-band signaling, such as SS7. In these schemes, switches have separate data connections that they use to carry signaling that may be associated with a call, or may be non-call-associated and purely a data transaction.

This fundamental model of the telephone system has not changed. To connect a call from one location to another, a circuit must be established through multiple switches. To establish the circuit, each switch in the path signals to the next switch to indicate where it is trying to reach. This is sort of like the packet switching schemes more familiar to the computing industry, but there is an important difference, a result of the circuit-switching nature of the telephone system: a phone call need only go one direction. As each switch speaks to the next they are establishing a circuit, which will remain established and carry voice both directions. Each switch in the path only needs to know the destination, none need to know the source, as the connection that direction is already open.

It's sort of like Tor. The basic concept of Tor, onion routing, is that with encryption of signaling information each node only needs to know the node before it and the node after. In the telephone system, each switch does know the final destination of the call (there is no encryption to protect this information from switches earlier in the path), but only one switch knows the origin of the call: the caller's local exchange switch. The other switches just know which inbound trunk the call arrived on.

Well, it's not really quite that simple, but to be honest, it's almost that simple. Over time signaling between switches has expanded to convey more and more information. The most important payload is the destination number. But something that at least looks like source information has been added not once, but twice.

First, there is the matter of billing. The telephone system is concerned first with establishing phone calls, but a close second is billing for them. As calls pass between carriers, even within the Bell System with its separate operating companies, carriers further along in the path need to know who they should bill for carriage. When you place a long distance call (assume with me that you pay for long distance), equipment on your local exchange switch records the call's source, destination, and duration in order to bill you for the time.

But there may be other carriers involved in connecting the call, and they want their share too.

ANI

Other carriers in the calling path will also record the call's details so that they can bill the originating carrier for their fractional portion of the rate (this sub-billing of long distance calls is one of the reasons telephone rate regulation is complex). In the era of manually connected long-distance calls, the operators at the different exchanges would speak with each other and give, among other things, an account identifier for the source of the call. After all, in a manual exchange it is the operator who does the signaling. Direct distance dialing requires automatic operation, though, so signaling had to be enhanced to convey the billing information. The solution is called ANI, or Automatic Number Identification.

ANI is basically what it sounds like: signaling is extended to include a number for the calling party, so that each exchange on the way can record it as part of the billing record (they already know the destination number and the duration, since they have to participate in establishing and maintaining the call). But it's very important to understand the origin of ANI. A lot of people hear of ANI and assume it to be a system that tells you the phone number of the caller. That's only really correct in sort of an incidental way. The real purpose of ANI is to give a billing account for the calling party, and telephone companies, for convenience, used telephone numbers to keep their customer ledgers.

ANI does not necessarily identify the caller, it only identifies who should be billed for the call [2]. Notably, the ANI concept does not span the diverse nature of the telephone network today. VoIP gateways and, in general, anyone carrying calls by means other than the POTS or the Plain Old Telephone System have other ways of identifying customers and tracking and billing for calls. This obviously includes VoIP, but also includes most cellular calls today as carriers have transitioned to GSM and now IP architecture [3].

Internally, these systems don't use ANI at all, instead using the accounting features of their own protocols. When delivering calls to the POTS, they may provide the same ANI for every call (just to identify that they are the carrier to be billed) or an ANI that isn't a dialable number but just a phone number they have held to use as a billing ID. It may reflect their internal accounting organization much more than it does the calling party.

Here's something that's key to understand: ANI is a feature of the conventional POTS telephone system, originating with electromechanical switches and present in today's TDM and packet telephone switches. It is not a feature of the telephone system at large. Consider that most European countries never used ANI, instead using one of a couple of other approaches to the same problem, and the dominant form of telephony in the US today (cellular) is derived from European technology.

VoIP and cellular carriers are inconsistent about how they use and handle ANI, and provide ANI on calls into the POTS only for compatibility with POTS billing systems. ANI on calls from POTS to cellular or VoIP carriers may be totally discarded, depending on the specific setup. It's often frustratingly inconsistent; like some others I run a VoIP line that intentionally captures ANI information since it can be useful to understand how payphones are set up (besides the calling number, ANI identifies the type of caller, including whether or not it's a payphone). I specifically hunted for a VoIP provider that conveys ANI and still find that it's missing on a lot of calls. One likely factor is that a surprising number of surviving payphones are now cellular, and cellular carriers generally don't use ANI.

CID

And then there is another service, similar in concept but very different in its purpose: Caller ID. Caller ID is intended to show the recipient of a call the identity, or at least phone number, of the caller. The "name" part of Caller ID (called CNAM) has always been a bit of a bust outside of POTS carriers, for reasons that could fill out another post. But it is quite reliable in providing a phone number.

Here's the problem: the CID is neither required nor expected to correspond to the origin of the call. There are lots of reasons for this, but consider a common one: in a large business, customer service may be performed by multiple call centers. When customer service calls you, they want the CID to give the main toll-free number you should use to reach customer service, not the specific call center that made the call. There are about a million variations of this same idea, that all come down to large organizations wanting to be able to call you from many places while still having their intended inward number appear on CID.

There are other scenarios as well. It's not unusual for companies with a lot of telephone traffic to have arrangements with multiple long-distance carriers and use whichever is cheapest for a specific call. They may not even have inward phone numbers with these carriers, and even when they do they don't want the CID number to be different depending on which carrier the call happens to be routed through. The same scenario exists in more modern systems; Google Fi uses multiple distinct cellular carriers and assigns "ghost numbers" to identify their customers on each of them. When a Google Fi customer makes a call, the CID should be that customer's primary phone number, not an internal number used for US Cellular provisioning.

All of this is to explain why the CID value on a telephone call is whatever the caller says it should be. Well, assuming the caller is something like a commercial customer with the technical ability to provide a CID value. This isn't really a bad thing, CID was never intended to tell you where the call was from... it was intended to tell you who the call was from, and how to call them back. The reality of the phone network is that that won't always match the origin of the call.

I think this whole thing is easier to explain by analogy to a technology more of us know the intimate details of, email. The CID value on a phone call is analogous to the "Reply-To" in an email, telling you how to return an email (or call) to the person. The ANI is analogous to a "Received" header, telling you some information about one hop in the process but not necessarily the first hop, and not necessarily enough to identify the originator.

Everything gets even more complicated when you consider the diversity of carriers. There are telephone carriers using multiple distinct technologies with their own signaling and billing infrastructure, and then there are other countries to contend with. Foreign countries often don't even have telephone numbers of the same length as North American numbers (or a fixed length at all for that matter), and so billing for international calls has always been a special case.

Spam, Robocalls, Scams, Etc.

This all works fine for the purposes of the telephone system. I mean, at least for a long time, it did. But have you noticed what's up with email lately? It seems that, given an open communications system, people will inevitably develop something called a "cryptocurrency" and badly want to make sure that you get in on something called an "ICO." The general term for this phenomenon is "spam," and the fact that it is only one letter away from "scam" is meaningful as the line between mere unsolicited advertising and outright crime is often razor thin.

In the email system, this problem has been elegantly solved by a system of ad-hoc, inconsistent, often-wrong heuristic classifiers glued to a trainwreck of different cryptographic attestation and policy metadata schemes that still haven't solved the problem. It is, perhaps, no surprise that the phone system is taking a generally similar approach.

Let's discuss a few differences between email spam mitigation and telephone robocall mitigation. The spam problem is arguably a bit harder for email because of the absolutely dizzying number of possible counterparties: every mail server out there. In the telephone system, the counterparties are limited to other telephone carriers, but that's still a very long list, to an extent that would probably surprise you. The FCC reports that there are 931 conventional telephone carriers and 1,787 interconnected VoIP carriers, and that's just the US. Most problematic traffic originates from overseas, where these numbers become far larger.

In the world of email, spam is nonetheless mitigated in part by completely blocking traffic from mail servers known to primarily originate spam. This is facilitated by a system of blacklists maintained by various companies and industry groups. One might wonder why the telephone system doesn't do the same? There are two things to consider.

First, the telephone industry does. The FCC maintains a blacklist of telephone carriers that have been found to take inadequate measures to prevent abuse (or more often intentionally facilitate abuse), and directs other carriers to block all traffic from them. One could argue that the FCC is overly conservative in their process for adding carriers to this list, but there are reasons.

Second, we have to consider that the telephone network is considerably deeper than the email network. What I mean by this is that, although email was designed to facilitate multi-hop routing, it is rare for email to actually pass through multiple distinct organizations en route (it often passes through multiple distinct MTAs, but these are all devices or service providers used by one of the two organizations involved). In the telephone system, multi-carrier routing is common, and all but universal for international traffic. The inevitable result is mixing of genuine and abusive traffic from the same carrier.

So why don't telephone carriers just block traffic from carriers that they receive robocalls from? For one, they are legally prohibited from doing so. This is under the doctrine of common carriage. If telephone carriers were allowed to pick and choose which carriers they would accept calls from, larger carriers would be able to use this as negotiating leverage to obtain unreasonably favorable terms from smaller carriers. This was a very real problem in the telephone system before common carriage rules were implemented, and it is a problem in the internet today, leading to the ongoing debate over ”net neutrality.”

But there are also practical reasons. Blocking traffic is known to come at risk. Anyone who has administered mail servers with a significant number of users will know that blacklists are far from foolproof and some popular blacklists are very prone to listing mailservers that originate any spam whatsoever, even from a single compromised user account. Institutional mail administration can seem to be roughly half trying to get back off of blacklists, and shore up outbound spam detection to avoid the next incident... but as we know, heuristic spam detection doesn't work very well.

The problem is even more acute in telephony, as savvy telephone spam operations intentionally get their traffic mixed with genuine traffic, ensuring that a carrier cannot block the origin wholesale without losing calls their customers actually wanted to receive. The typical strategy is to originate calls in a foreign country with relatively lax telephone regulation, India has long been a top choice since it offers both a loose regulatory environment and plenty of English speakers. A telephone spam operation need only find a commercial telephony provider with poor oversight and interconnection to a major national telephone carrier, and then the robocalls are being introduced from a carrier with millions of customers generating genuine traffic. These often arrive through low-cost international gateway service that route calls via VoIP, popular not only to scammers but everyday users seeking lower international calling rates.

The core problem is mixing: telephone carriers receive abusive traffic mixed in with genuine traffic, and they have few ways to determine what's what. Some would suggest that foreign-originating calls with US CID numbers are inherently suspect, but did you know that India has an inexpensive labor market and a tremendous number of English speakers? When I said that customer service call centers need to have the correct inbound number appear on CID, a lot of those call centers are overseas! There are several other reasons as well for foreign calls to legitimately come with US CID numbers.

To really sort out spam, carriers need a consistent, reliable way of determining what carrier a call actually came from. Not just the carrier that handed the call to them, but the carrier that originated it. Hey, email has a thing sort of like that, DKIM. What if someone did DKIM for telephones?

Someone did. It's called STIR/SHAKEN.

STIR/SHAKEN

STIR/SHAKEN stands for Secure Telephony Identity Revisited/Signature-based Handling of Asserted information using toKENs. Yes, it is a tortured acronym. STIR and SHAKEN are actually two separate standards, but they fit together closely. STIR comes from an RFC and describes headers for VoIP traffic. SHAKEN comes from two industry groups and describes how to encode the same headers into SS7 messages. In other words, STIR/SHAKEN are the same logical protocol defined for VoIP and POTS, respectively.

STIR/SHAKEN describes a cryptographic attestation, which is attached to a call by a carrier (ideally the originating carrier) and signed with a private key belonging to that carrier. Through the magic of public-key cryptography, subsequent carriers that handle the call can verify the STIR header. In practice, STIR is based on JWTs---a STIR header is basically just a JWT with a few standard fields. Those fields are the destination phone number, the source phone number, and the type of attestation.

There are three types of STIR/SHAKEN attestations, called A, B, and C. An A attestation is a statement from the original carrier that the call came from one of their customers and the carrier knows that they are entitled to use the STIR "from" telephone number attached (which must match the CID number). A B attestation states that the originating carrier knows the call came from one of their customers, but they don't have knowledge of the customer's entitlement to the source number. A C attestation is the fallback---it states that the originating carrier got the call from somewhere, but they don't know anything about it.

Keys for STIR/SHAKEN are distributed through a public-key infrastructure very similar to that used for TLS. Certificate authorities issue certificates to telephone carriers that give the carrier's public key and identifying information. That way, any STIR/SHAKEN attestation can be verified to have originated with a specific carrier as a legal entity. You can now know exactly which carrier a call came from.

STIR/SHAKEN immediately solves one problem. For any call with an A or B attestation, the originating carrier is known. If the call is abusive, you know exactly which carrier should get in trouble. It also takes a step towards solving another problem: the CID number should match the STIR/SHAKEN number, and if there is an A attestation you know that the carrier is promising the customer is really entitled to that phone number (e.g. they pay the carrier to hold that number for their inbound calls).

Unfortunately, there isn't currently a way to link a phone number directly to a STIR/SHAKEN certificate. That is, an A attestation is a promise from the carrier that the customer is authorized to use the phone number, but there's no way to actually check if that carrier is in a position to make that claim about the phone number in question. There is a system in development to address this issue (that basically provides a database to correlate phone numbers with STIR/SHAKEN carrier certificates), but it's also not that big of a problem in practice as any carrier issuing an improper type A attestation can easily be identified and shamed (actually, FCC policy is that they will be fined and, probably more significantly, their certificate will be revoked).

STIR/SHAKEN is a huge step forward because it facilitates two things:

- The carriers that originate abusive traffic can be more easily identified so that a regulatory agency can bring penalties
- Specific source carriers can be blocked regardless of the path the call took through other carriers, based on the STIR/SHAKEN attestation.

FCC mandates for STIR/SHAKEN require not only that carriers attach attestations to calls, but also that they validate the attestations and block calls where the attestation is invalid or belongs to a blacklisted carrier.

The bad news

So why are there still spam calls?

Unfortunately, STIR/SHAKEN is far from universal. The FCC made STIR/SHAKEN implementation mandatory for US telephone carriers as of June 30, 2022. That was over a year ago, but the FCC issued numerous exemptions to small and rural carriers with difficulty affording the required equipment (remember that telephone switches can have fifty year service lives and there is some very old equipment still in use), and besides, the FCC mandate applied only to the United States.

A May 3, 2023 report from TransNexus estimates that only a bit over a quarter of phone calls terminating in the US bear STIR/SHAKEN attestations. Fortunately more and more carriers are adopting STIR/SHAKEN, but despite the "mandatory" deadline there is still a long ways to go. Many have criticized the FCC for being far too slow in enforcing attestations, but to be fair, the FCC is acutely sensitive to the fact that rural and small-market telephone carriers are often barely above water, and suddenly imposing costly requirements could lead to a minor crisis as smaller telephone carriers run out of money.

STIR/SHAKEN is also imperfect. TransNexus finds that calls with a type B attestation are actually more likely to be robocalls than those with no attestation at all. In a way, this makes sense, as these calls are apparently coming from carriers who do not keep track of customer entitlement to phone numbers. The problem is that that's a rather common situation, for example because of customers using multiple carriers for outbound calls to get optimal rates. There is a silver lining here, though. Those carriers placing attestations on robocalls are putting themselves at risk, as those attestations are tools for action against them.

That's an interesting aspect of STIR/SHAKEN: by forcing carriers to sign the calls they hand on, it gives them a level of responsibility and liability for the contents of those calls. This has introduced a sort of KYC system for telephone carriers. Around the time of the mandatory STIR/SHAKEN rollout, a VoIP termination provider I had used for years suddenly demanded that I send copies of my passport, incorporation documents, and FCC filings. Carriers signing calls are getting more cautious about the kinds of customers they will accept, and recent FCC enforcement actions will probably accelerate this trend. It's a bit unfortunate in that the barrier to entry for hobby VoIP operations is getting higher and higher, but, well, that's just like email.

And that's sort of the point. The world of telephony spam mitigation is very comparable to the world of email spam mitigation, but a couple of decades behind. Carriers have already begun to introduce extensive heuristic spam detection for SMS, but the industry and FCC have been hesitant to go that route for telephone calls. Experience with SMS might be a reason why; I used to work for a company that sent a lot of SMS and we constantly struggled with carriers blocking our appointment and medication reminder messages, even getting to the point of "burning" a short-link domain name because of a major carrier blocking all messages that contained it without explanation. Heuristic detection really is imperfect, and while SMS might have relaxed reliability expectations people want phone calls to work every time.

So instead, the telephone industry is going the cryptographic attestation route. Email has done this as well. But we have to temper our expectations: extensive heuristic detection, blacklisting, and cryptographic attestation schemes have failed to completely tame the phenomenon of spam email. Telephone spammers are in good company: their colleagues in the email industry have kept it going, despite huge effort in opposition, for almost thirty years.

But the telephone industry clearly needs to move faster if they expect to reach even the level of success email providers have. Unfortunately, "Faster" and "The FCC" are not famously friendly. Many jump to the conclusion that the telecom industry is complicit in the situation, but it's a little more complex. Some major telecom industry associations actively support STIR/SHAKEN, and in general most telecom industry associations have lobbied the FCC to move more quickly on the robocall issue and to allow carriers greater latitude to take their own actions to mitigate the problem.

It's hard to clearly lay blame in this situation. For the FCC's part, it has moved extremely slowly, extending STIR/SHAKEN deadlines almost indefinitely until the federal legislature passed the TRACE act to force their hands. The telecom industry continues to acuse the FCC of lethargy in its response to the problem. At the same time, some of the largest telephone carriers have been some of the most resistant to implementation, arguing that it's unreasonable to impose the enormous cost on them.

This argument gains a bit of weight when you consider that many in the industry are skeptical of STIR/SHAKEN as a technical approach; it was developed by organizations that are mostly controlled by telecom equipment and software vendors rather than telecom carriers. The carriers seem to feel that STIR/SHAKEN is an inadequate approach to the problem with a severe case of design-by-committee, and the design of STIR/SHAKEN and the FCC's regulations around it are both unclear when applied to common real-world situations.

If you want a single source of the robocall problem, perhaps it is this: the telecom industry is fiercely profit motivated. Carriers stand to save money by not implementing STIR/SHAKEN, telecom equipment and software vendors stand to make money by forcing carriers to do so. Whether or not it actually addresses the problem is largely orthogonal to this basic dynamic.

[1] SS7 is very interesting, but I often complain that the security community has an excessive focus on it considering the rarity of actual exploitation of SS7. People talk about how you shouldn't use SMS 2FA because of problems with SS7; that's total nonsense. You shouldn't use

SMS 2FA because a thirteen year old will con your carrier into giving them access to your account.

[2] There is a bit of nuance here. It is possible to subscribe to ANI service on a trunk, which is usually done by businesses. It's also common for PSAPs, 911 call centers, to have ANI service as a way to determine the origin of calls. Both of these have become far less common as ANI has become less reliable. The modern E911 standard is a result of the fact that ANI is not capable of providing reliable caller identification.

[3] For the most part, Verizon is the only cellular carrier that still has traditional TDM telephone switches. Their days are presumably numbered now that Verizon has retired their legacy 3G service, which for historic reasons was far more based on traditional (American) telephone technology than AT&T's.