On Attack Causality in Internet-Connected Cellular Networks

Patrick Traynor, Patrick McDaniel and Thomas La Porta
Systems and Internet Infrastructure Security Laboratory
Networking and Security Research Center
The Pennsylvania State University
University Park, PA 16802
{*traynor, mcdaniel, tlp*}*@cse.psu.edu*

Abstract

The emergence of connections between telecommunications networks and the Internet creates significant avenues for exploitation. For example, through the use of small volumes of targeted traffic, researchers have demonstrated a number of attacks capable of denying service to users in major metropolitan areas. While such investigations have explored the impact of specific vulnerabilities, they neglect to address a larger issue - how the architecture of cellular networks makes these systems susceptible to denial of service attacks. As we show in this paper, these problems have little to do with a mismatch of available bandwidth. Instead, they are the result of the pairing of two networks built on fundamentally opposing design philosophies. We support this a claim by presenting two new attacks on cellular data services. These attacks are capable of preventing the use of high-bandwidth cellular data services throughout an area the size of Manhattan with less than 200Kbps of malicious traffic. We then examine the characteristics common to these and previous attacks as a means of explaining why such vulnerabilites are artifacts of design rigidity. Specifically, we show that the shoehorning of data communications protocols onto a network rigorously optimized for the delivery of voice causes that network to fail under modest loads.

1 Introduction

The interconnection of cellular networks and the Internet significantly expands the services available to telecommunications subscribers. Once limited to basic voice services, these systems now offer data connections at the lower end of broadband speeds. Accordingly, devices attached to such networks are capable of engaging in applications ranging from traditional voice communications to streaming video. While initial uptake of these services has been slow [1, 18], notable advances in connection speed and an expanded set of supported devices (e.g., laptops) are beginning to spur substantial acceptance and usage.

The transformation of these systems from isolated providers of telephony to Internet-attached general purpose communication networks has already been marred by concerns of inadequate security. As connections between such systems and external data networks have developed, a number of researchers have noted weaknesses in the telecommunications infrastructure. For example, our previous work on targeted text messaging attacks demonstrated the ability to deny service to large metropolitan areas with the bandwidth available to a single cable modem [16,47]. While these and a host of other exploits [39, 44] have explored the impact of specific attacks against cellular networks, they have all failed to answer a larger question: "*How does the architecture of cellular data networks inherently make them susceptible to denial of service attacks?*" Unexpectedly, the answer to this question has little to do with bandwidth constraints. Instead, these vulnerabilities are the result of the conflict caused by connecting two networks built on fundamentally opposing design philosophies.

In this paper, we argue that low-bandwidth denial of service attacks in telecommunications networks are artifacts of incompatibility caused by interconnecting systems built with two differing sets of design requirements. While the merits of independent "smart" and "dumb" architectures have been widely debated, none have examined the inherent security issues caused by the connection of two mature systems built on these opposing design tenets. To support our assertion, we present two new vulnerabilities in cellular data services. These attacks specifically exploit connection setup and teardown procedures in networks implementing the General Packet Radio Service (GPRS). Through a combination of analysis and simulation, we characterize the impact of such attacks on legitimate voice and data services in the network. We then use these new attacks, in combination with previously discussed vulnerabilities, as demonstra-

ble evidence that the translation of traffic between these two network architectures is the root of such problems. Through this, we seek to develop a larger sense for *why* such attacks are possible, even in the presence of a cellular network with hypothetically infinite bandwidth. Ultimately, by understanding causality, the discovery of future vulnerabilities is vastly simplified.

In so doing, we make the following contributions in this work:

- **New Vulnerability Analysis:** We identify and develop a realistic characterization of two new vulnerabilities in cellular data networks. These exploits target specific components of the expensive connection setup and teardown procedures and can prevent legitimate use of data services. While the partitioning of voice and data flows in such networks is designed to protect each traffic type from the other, our attack on setup mechanisms demonstrates that optimizations made for efficiency can result in the disruption of voice services.
- **Implications of Combined Design Philosophies on Security:** We use the body of available vulnerabilities as the basis for an analysis to determine the underlying cause of such denial of service attacks. Consequently, we show that these problems are not necessarily the result of poor protocol design but are instead deeply rooted in opposing architectural assumptions.

The remainder of this paper is organized as follows: Section 2 offers a brief overview of our previous work on targeted SMS attacks to prime the reader with additional data points; Section 3 presents and offers an initial analysis for our newly discovered vulnerabilities; Section 4 uses monitoring of deployed cellular networks and simulation to support the conclusions made in the previous section; Section 5 coalesces the previous attacks on cellular networks as data points in our larger argument; Section 6 offers a discussion of techniques to address such problems; Section 7 provides related work; Section 8 offers concluding thoughts.

2 Prior Work - Text Messaging Attacks

We present a high-level overview of our previous attacks on text messaging [16, 47]. With some five billion messages sent each month in the United States alone [28], this service has become one of the premier streams of revenue for cellular network operators. To encourage widespread use, providers have opened a significant number of gateways between the Internet and their networks. Whether through email, instant messaging applications or even a provider's website, it is possible to exchange asynchronous communications with cellular subscribers. The ability to communicate across such networks, however, is not without potential consequences.

A cellular network[1] must perform multiple tasks before delivering a text message. The network first conducts a series of lookups to determine the location of the destination device. The device must then be awoken from an energy-saving sleep state and authenticated. A connection can then be established and the incoming text message delivered. Critical to this process is the *Standalone Dedicated Control Channel* (SDCCH), which is responsible for the authentication and content delivery phases of text messaging. With a bandwidth of 762bps [6], this constrained channel is shared by the setup phases of both text messaging and voice calls. Consequently, by keeping the SDCCH saturated with text messages, incoming legitimate voice and text messages can not be delivered by the network. Understanding this, an adversary attempting to exploit this system can use web-scraping and feedback from provider websites to create "hit-lists" of targeted devices. By sending traffic to these targeted devices at a rate of approximately 580Kbps, the adversary would be able to deny service to all of Manhattan.

Attack mitigation techniques, ranging from queue management to resource allocation strategies on the air interface, were then shown to diminish much of the impact of such attacks. While successful, these countermeasures did not consider the use of cellular data services such as GPRS to alleviate targeted text messaging attacks. Logically, delivering data traffic over separate, higher bandwidth links should provide the most complete solution to this problem. However, as we show in the next section, it is possible to disrupt cellular data services with less bandwidth than was used in the original SMS attack.

3 New Vulnerabilities in Cellular Data Services

We present two new denial of service (DoS) vulnerabilities in cellular data services. These attacks use a relatively small amount of traffic to exploit connection setup and teardown mechanisms. We use publicly available specifications to provide an initial characterization of these attacks and as a means of demonstrating the potential for the interruption of data services in major metropolitan areas.

3.1 Network Architecture

Before a GPRS/EDGE[2] network provides any services to a mobile device user, a series of attachment and authentication procedures must take place. On power-up,

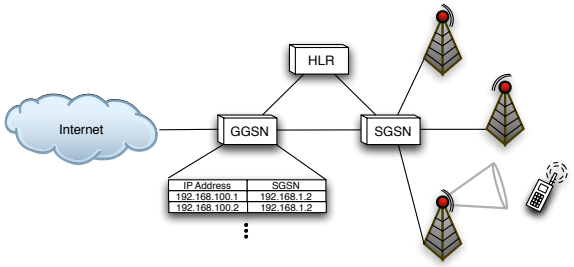

Figure 1: A high level network architecture for cellular data networks.

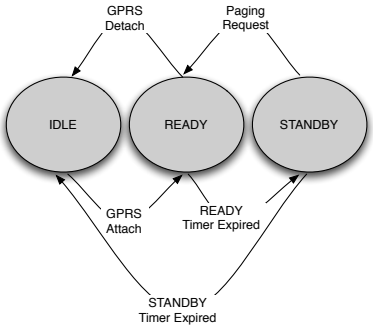

Figure 2: A state transition diagram for mobile devices, including transition functions.

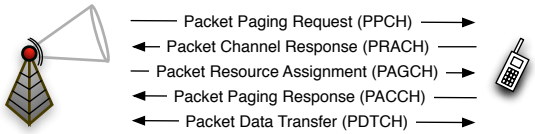

Figure 3: When the first packet of a session arrives at the base station, the host must be paged and then assigned logical resources. The messages and channels used to accomplish this are shown above.

a device (e.g., mobile phone) transmits a *GPRS-attach* message to the network. The base station forwards this message to the attached *Serving GPRS Support Node* (SGSN), which authenticates the user's identity with the help of the *Home Location Register* (HLR). The HLR supports both voice and data operations in the network by keeping track of information including user location, availability and accessible services. When this process completes, the mobile device has a virtual connection with the network.

In order to exchange packets with external networks, the mobile device must then establish a *Packet Data Protocol* (PDP) context with the network. The PDP context is a data structure stored in the SGSN and the *Gateway GPRS Support Node* (GGSN) and is responsible for mapping billing information, quality of service requirements and an IP address to a user device. While many phones do not currently automatically establish a PDP context on power-up, the trend towards doing so (e.g., email-capable phones and GPRS-equipped laptops) is rapidly increasing. As cellular providers move into the broadband Internet market, such numbers will continue to expand rapidly.

Having been authenticated and registered, a mobile device is capable of exchanging packets with hosts internal and external to the cellular network. At some time after attachment, a packet originating from an Internet-based host and destined for a mobile device arrives at the GGSN. The GGSN compares the destination IP address to those of established PDP contexts and, upon finding the corresponding entry, forwards the packet to the corresponding SGSN. The SGSN begins the process of connection establishment and wireless delivery. Figure 1 highlights this network architecture.

The final hop of packet delivery occurs over the air interface. The details of this step, however, depend upon the current state of the device. As power has traditionally been a concern in this setting, mobile devices are not constantly listening for incoming packets. To accommodate this constraint, devices operate in one of three states: IDLE, STANDBY, and READY. Devices in the IDLE state are unregistered with the network and therefore unreachable. In the power-saving STANDBY state, in which the vast majority of time is spent, devices periodically listen for network "wake up" messages known as pages. Upon receiving a page from the network, the device transitions into the READY state. In this state, a device constantly monitors the air interface for incoming packets. When packets are not received for a number of seconds, devices transition back into the STANDBY state to conserve power. These three states and the transitions between them are shown in Figure 2.

On the arrival of the first packet in a flow, the SGSN begins the process of locating the targeted device. If the destination device is not currently in the READY state, the base station nearest to the device is unknown to the network. Accordingly, the SGSN creates paging messages to be sent from a number of base stations. Upon receiving a paging request, a base station transmits a message to multiple sectors (i.e., service areas) over the *Packet Paging Channel* (PPCH). Whether due to interference or sleep cycles, the paging process typically re-

quires multiple iterations. If the targeted device is awake and hears its temporary identifier in a paging message, it attempts to alert the network of its presence by responding on the *Packet Random Access Channel* (PRACH). The base station receiving this response alerts the SGSN that the destination device has been located. The network then responds on the *Packet Access Grant Channel* (PAGCH) with a message containing a list of *Packet Data Traffic Channels* (PDTCHs) that should be monitored for incoming data. The device acknowledges receiving this message over the *Packet Associated Control Channel* (PACCH). At the end of this setup, as illustrated in Figure 3, the network can then route traffic directly to the READY state device. Note that the above channels are largely complementary to channels used for voice signaling (the naming convention, minus the "Packet" prefix, is the same). Because running two sets of control channels leads to the underuse of limited spectrum, the standards documents indicate that it is acceptable for voice and data control channels to be shared [3, 7].

3.2 Packet Multiplexing on the Air Interface

Data services have been available from cellular networks for a number of years. Like voice telephony, these circuit-switched services required that a single endpoint monopolize a channel for the entire duration of its connection to the network. Regardless of whether this connection was used to constantly stream content or intermittently deliver packets, the provider charged the end user for the entire duration of the connection. Accordingly, demand for such inefficient services was not great. GPRS overcomes these limitations by multiplexing multiple traffic flows over individual links. Accordingly, it is possible to serve a large number of users on a single physical channel concurrently and only charge them for the packets they exchange.

GPRS provides data service by building on the timeslot structure of GSM. Specifically, a contiguous piece of radio spectrum is subdivided into equal timeslots. When assigned a timeslot, a user exerts temporary control over a small piece of the air interface. To provide the illusion of continuous control, sets of eight timeslots are grouped into a frame so that each can be serviced once every 4.615ms. This sampling across timeslots creates physical channels, upon which voice, data and control traffic can be delivered. When used for data, these physical channels are referred to as *Packet Data Channels* (PDCHs). Each set of 52 frames creates larger units known as multiframes. These multiframes are subdivided into 12, four-timeslot blocks, with logical channels then mapped onto each block. The remaining four timeslots in a multiframe are used for time synchronization and signal strength

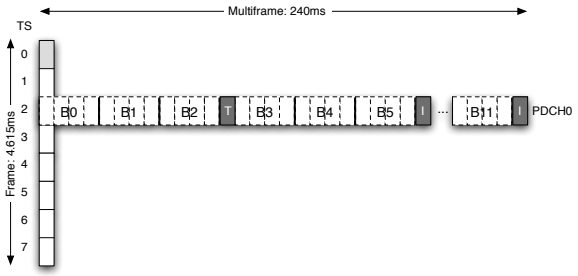

Figure 4: Each timeslot in a GPRS TDMA frame is used to create physical channels called Packet Data Channels (PDCHs). Every 52-frame time period creates a multiframe, which is divided into twelve bursts of four. Each group, or bursts, holds a single logical channel. The specific allocation of these channels is dependent on the network. The remaining timeslots are used for time synchronization and idle measurement.

measurement periods. For example, in Figure 4, block $B0$ may function as a PPCH and blocks $B1$, $B4$ and $B7$ may be used as PDTCHs [3] [7].

When the first packet in a flow arrives at a base station for a user in STANDBY mode, the paging method described above occurs. As part of connection establishment, the flow receives a unique MAC layer label known as the *Temporary Flow Identifier* (TFI). Every subsequent packet belonging to the *Temporary Block Flow* (TBF) is marked with this TFI so that a targeted mobile device knows which packets to decode. When the base station has no more packets to send to the destination mobile device, the TBF and its associated TFI expire and can be reused by other flows in the immediate area. Upon TBF expiration, the mobile device returns to the STANDBY state.

3.3 Exploiting Teardown Mechanisms

Because the process of locating, paging and establishing a connection between the network and an end device is expensive, the immediate expiration of a TBF is impractical. For example, minor variations in packet interarrival times would force a system as described above to frequently relocate, repage and reestablish connectivity with users. Accordingly, networks implement a delayed teardown of resources. This means that devices remain in the READY state and retain their TBF for a number of seconds before the network attempts to reclaim its logical resources. When a packet is delivered to the user, the network sets a timer[4], which is reset to its default value on the arrival of each additional packet. The standards recommend a timer value of approximately five seconds [2]. Given that the connection establishment process requires roughly the same amount of time, such a value is entirely reasonable.

Because TFIs are implemented as a 5-bit field, an adversary capable of sending 32 messages to each sector in a metropolitan area can exhaust logical resources and temporarily prevent users from receiving traffic. Targeted devices would not need to be infected or controlled by the adversary; rather, hit-list generation techniques similar to those discussed in our previous work [16] could be used to locate hosts able to receive traffic. If this task can be repeated before the TBF timers expire, a denial of service attack becomes sustainable. In order to more explicitly characterize the bandwidth requirements, we model such an attack on Manhattan using well known parameters [35, 48]. Given an area of 31.1 miles^2 and a sector coverage area of approximately 0.5 and 0.75 miles^2, Manhattan contains 55 sectors. Using a READY timer of 5 seconds and 41 byte attack packets (i.e., TCP/IP headers plus one byte), the delivery of legitimate data services in Manhattan could be prevented with the attack shown below:

$$\begin{aligned} Capacity &\approx 55 \text{ sectors} \times \frac{32 \text{ msgs}}{1 \text{ sector}} \times \frac{41 \text{ bytes}}{1 \text{ msg}} \times \frac{1}{5 \text{ sec}} \\ &\approx 110\ Kbps \end{aligned}$$

The exhaustion of all hypothetical TBFs may not be necessary given current usage and deployed hardware. As the current demand for voice services far outpaces cellular data usage, only a small percentage of physical channels in a sector are used as PDCHs. Because GPRS/EDGE are not extremely high bandwidth services, allowing 32 individual flows to be concurrently multiplexed across a single PDCH would be detrimental to individual throughput. Accordingly, often only a subset of the 32 TBFs (4, 8 or 16 [26,33]) are usable. The maximum number of concurrent TBFs in a sector is therefore $min(d * u, 32)$, where d is the number of downlink PDCHs and u is the maximum number of users per PDCH. While the number of PDCHs can be dynamically increased in response to rising demand for data services, networks typically hold unused channels to absorb spikes in voice calls. It is therefore unlikely that all 32 TBFs will be available at all times, if ever. A more realistic approximation of the bandwidth required to deny access to data services is given by:

$$\begin{aligned} Capacity &\approx 55 \text{ sectors} \times \frac{4 \rightarrow 16 \text{ msgs}}{1 \text{ sector}} \times \frac{41 \text{ bytes}}{1 \text{ msg}} \times \frac{1}{5 \text{ sec}} \\ &\approx 14.1 \rightarrow 56.4\ Kbps \end{aligned}$$

The brute-force method of attacking a cellular data network in a metropolitan setting is simply to saturate all of the physical channels with traffic. Even at their greatest levels of provisioning, the fastest cellular data services are simply no match against traffic generated by Internet-based adversaries [39, 45]. Such attacks, obvious by the sheer volume of traffic created, would likely be noticed and mitigated at the gateways to the network. However, with knowledge of the interaction between different network elements, it is possible for an adversary to launch a much smaller attack capable of achieving the same ends. A basic understanding of the packet delivery process provides the requisite information for realizing this attack.

Given a theoretical maximum capacity of 171.2 Kbps per frequency and as many as 8 allocated frequencies per sector, an adversary attempting the brute-force saturation of such a system would instead need to generate the volume of traffic as calculated as:

$$\begin{aligned} Capacity &\approx 55 \text{ sectors} \times \frac{171.2 \text{ Kbps}}{1 \text{ frequency}} \times \frac{8 \text{ frequencies}}{1 \text{ sector}} \\ &\approx 73.56 Mbps \end{aligned}$$

By attacking the logical channels instead of the raw theoretical bandwidth, *an adversary can reduce the amount of traffic needed to deny service to a metropolitan area by as much as three orders of magnitude.* Note that networks implementing EDGE, which can provide three times the bandwidth of a GPRS system, would experience the same consequences given the same volume of attack traffic.

3.4 Exploiting Setup Procedures

If connections to an end host must repeatedly be reestablished, the interarrival time between successive packets becomes exceedingly large. Delaying resource reclamation is therefore a necessary mechanism to ensure some semblance of continuous connectivity to the network. This latency, however, is not simply the result of the time required for a user to overhear an incoming paging request. To better understand setup cost, we examine a network in which resource reclamation occurs immediately after the last packet in a flow is received.

Of particular interest to such an analysis is the performance of the common uplink channel, the PRACH. Because this channel is shared by all hosts attempting to establish connections with the network, the PRACH inherently has the potential to be a system bottleneck. To minimize contention, access to the PRACH is mediated through the slotted-ALOHA protocol. Given a channel divided into timeslots of size t and time synchronization across hosts, end devices attempting to establish connections transmit requests at the beginning of a timeslot. In so doing, the network reduces the amount of time during which collision can occur from $2t$ in the random access case to t. While slotted-ALOHA offers a significant improvement over random access, its throughput remains

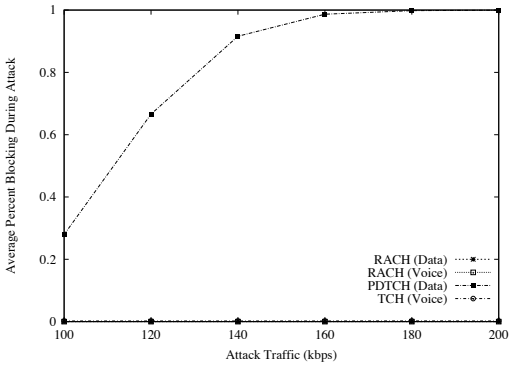

Figure 5: Blocking of legitimate traffic for varying attack traffic loads. Note that blocking only occurs on the PDTCH. These loads represent the entire attack bandwidth used across Manhattan.

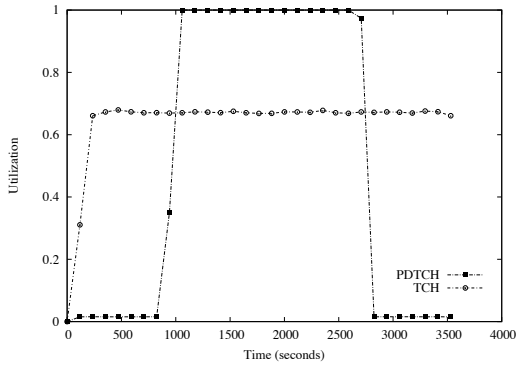

Figure 6: TFI utilization for a Manhattan-wide attack at 200Kbps. Actual PDTCH utilization (not shown) is virtually zero because of infrequent arrivals for these established flows.

low. Given a traffic intensity of G messages per unit time, the normalized throughput γ of slotted-ALOHA is:

$$\gamma = Ge^{-G}$$

The maximum theoretical utilization of channel implementing slotted-ALOHA is 0.368. In reality, however, this value is significantly lower. As the number of incoming connection establishment requests increases, so too does the need for retransmission due to collision. The throughput of such a system therefore typically stabilizes at a point far below this optimum value. Given a large number of paging requests, potentially caused by the immediate reclamation of resources as described above, the throughput of this already constrained channel would be severely degraded. Accordingly, the rate at which responses to connection establishment requests will pass through this channel is much lower than the available bandwidth. Because the behavior of the PRACH is highly unstable and affected by feedback (i.e., retransmissions due to collision), we leave the characterization of specific traffic volumes necessary to cause blocking to the next section.

4 Attack Characterization

In order to better characterize the observations made in the previous section, we extend the GSM simulator from our previous work [47] to include support for GPRS data services. The parameters of this simulator were set by information from a variety of sources. The means by which these parameters were chosen are discussed in the Appendix.

4.1 Modeling Attacks on Teardown Mechanisms

To demonstrate the exploitation of delayed resource teardown, we simulate a GPRS network under varying traffic loads. Although the full complement of TBFs may not be available in all real deployments [26, 33], we conservatively allow for up to 32 concurrent flows. When in use, each TFI is held for exactly five seconds unless a new packet arrives. While it is possible for a single device to obtain multiple TFIs, we assume that all incoming flows for a given destination share a single TBF [4]. Because of observations made on deployed networks, both voice and data setup requests share a number of control channels. We therefore replace data control channels with their voice equivalents (i.e., RACH instead of PRACH).

Legitimate voice and data calls were modeled as Poisson random processes and generated at rates of 50,000 and 20,000 per hour, respectively, across Manhattan. The duration of these flows are also generated in a similar fashion with means of 120 and 10 seconds, respectively. These values represent standard volumes and exhibit no blocking. Attack flows, each consisting of a single packet, are also modeled by a Poisson random process with rates ranging from 100-200 Kbps. Each run, of which there were 1000 iterations for each attack load, simulated an hour of time with attacks occupying the middle 30 minutes.

Figure 5 shows the blocking rates of legitimate traffic caused by an attack on the delayed teardown mechanism. At a rate of 160 Kbps or greater, the ability to use cellular data services within Manhattan is virtually nonexistent. The amount of traffic required to execute such an attack is slightly greater than the estimation of a perfect

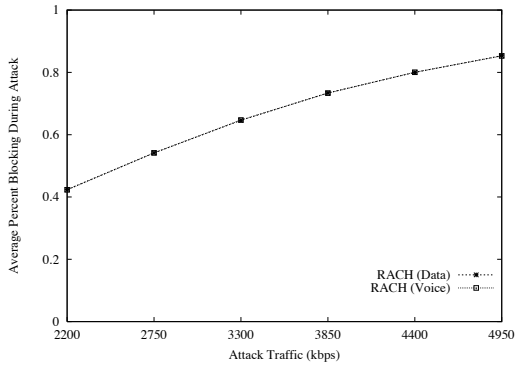

Figure 7: Blocking caused when immediate resource reclamation is enforced on data sessions. Notice that because both voice and data flows use the RACH, increased data requests cause voice blocking. No blocking was observed on other channels.

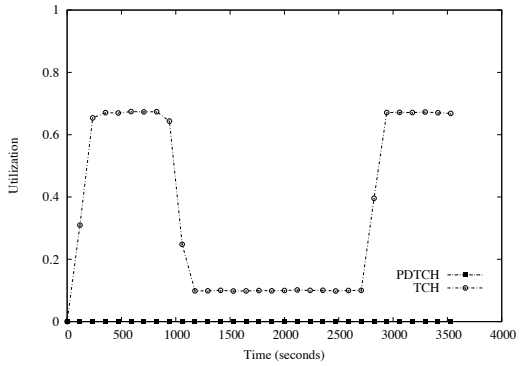

Figure 8: The impact of RACH congestion on voice calls. Notice that during the attack phase, voice call blocking on the RACH causes a significant under utilization of traffic channels.

scenario in Section 3.3 due to the exponential interarrival rate used to generate packets. However, because this more realistically represents the nature of packet delivery in a network given the presence of other traffic, it offers a more accurate characterization of the attack. In spite of having the potential to deliver large volumes of traffic once flows are established, these results demonstrate that use of cellular data services can in fact be denied with less bandwidth than was used in the targeted text messaging attacks [16, 47].

Figure 6 offers additional insight into the attack by providing the utilization profile for a number of channels. Most importantly, only the PDTCHs operate at capacity during the attack. This utilization represents the state of virtual resources, not channel bandwidth. None of the channels responsible for delivering voice, most critically the *traffic channels* (TCHs), are measurably affected by the increase in data traffic. Note that this is deliberate as cellular data services such as GPRS are designed to completely separate voice and data services.

4.2 Modeling Attacks on Connection Setup

To characterize the impact of frequent connection reestablishment on a cellular data network, we simulate a variety of traffic levels in the presence of immediate resource recovery. Specifically, when the base station no longer has packets to send for a particular flow, the targeted device returns to the STANDBY state. Except for delayed teardown procedures, all network settings and conditions including legitimate traffic volumes and interarrival patterns, remain the same. Attacks in this scenario, each of which occurs according to a Poisson random distribution, range from 2200-4950 Kbps spread across all of Manhattan. As in our previous experiments, each attack traffic level was run for 1000 iterations.

Figure 7 shows the blocking rates for legitimate traffic on a number of channels. Unlike the attack in the previous section, in which PDTCH blocking occurred because of TBF exhaustion, no loss of packets was observed on the PDTCHs. In spite of this, the results of these simulations confirm a more significant vulnerability - both voice and data flows experience blocking on the RACH. Although such networks strive to separate voice and data traffic, the dual use of control channels allows misbehavior in one realm to affect the other. Generating just over 3 Mbps of traffic for the entire city of Manhattan, an adversary is capable of blocking nearly 65% of all traffic - voice and data. For a network in which a blocking probability of 1% is typically viewed as unacceptable, such an attack represents a serious operational crisis.

Figure 8 provides further information about the impact of the 4950Kbps attack on voice and data services. The most notable consequence of this attack is observable in the nearly 80% decrease in TCH utilization. The near zero utilization of PDTCHs offers an explanation to the lack of blocking observed in the previous figure - the majority of legitimate traffic is being filtered out before it can ever be delivered by the PDTCHs. Accordingly, a network using the settings described above is subject to attacks capable of denying both voice and data services.

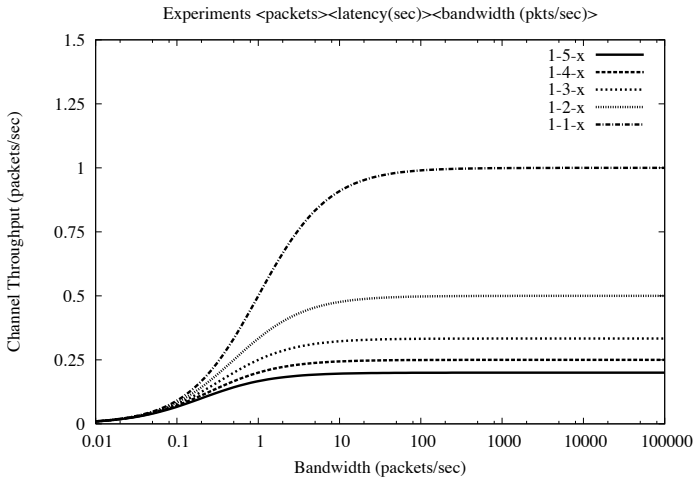

Figure 9: Given a connection establishment latency and the size of requests (in packets), we examine the impact of varying bandwidth on system throughput. When the available bandwidth allows for the virtually instantaneous delivery of requests, system throughput plateaus. This result indicates that bandwidth is ultimately not the bottleneck in this system. (log-scale)

5 The Meeting of Conflicting System Design Philosophies

At first glance, the differences between each of the attacks on cellular networks appear stark. Targeted text messaging attacks fill and maintain a low-bandwidth control channel at capacity. Adversaries attacking cellular data services exhaust virtual resources or take advantage of access protocol inefficiencies. In reality, all of these vulnerabilities are remnants of a conflict between the design philosophies of telecommunications and traditional data networks. Specifically, they are the result of contrasting definitions of a flow and the role of networks in establishing them. To make such a claim more concrete, we begin by demonstrating how a pair of seemingly adequate techniques for mitigating the above attacks fails to do so.

The most obvious approach to addressing the data attacks described in Section 3 is to expand the range of possible TFI values. Unfortunately, as mentioned earlier, these limitations are necessary given the bandwidth available to GRPS/EDGE networks. The use of 32 (or fewer) concurrent flows per sector is a requisite concession for providing basic levels of connectivity between the network and end devices. In order for an increased pool of identifiers to have a meaningful effect, the bandwidth available to data services would also need to be significantly increased. This combination of approaches is actually implemented in 3G cellular networks such as UMTS [8]. However, even these networks suffer from the high cost of connection establishment (i.e., delivering the first packet in a flow).

A session establishment period lasting a few seconds represents only a small fraction of the total lifetime for a connection persisting for a number of minutes. Given the limited amount of spectrum allocated to cellular providers, such infrequently used channels predictably occupy as little space as possible to avoid wasting bandwidth. Because the duration of a packet flow may not provide sufficient time over which such an expense can be amortized, the minimal allocation of bandwidth to connection establishment may in fact create a system bottleneck. To capture the impact of additional bandwidth on connection setup, we offer a simple model of request throughput for a sector as follows:

$$Throughput = \frac{\#\ Packets}{Setup\ Latency + \frac{\#\ Packets}{Bandwidth}}$$

If the expense associated with connection establishment was the result of inadequate resources, an increase in bandwidth should alleviate much of this cost. Such a scenario would be equivalent to increasing the size of the smallest link in a traditional data network to improve end-to-end throughput. However, the calculated effects of increased bandwidth on overall throughput are extremely limited in this setting. Because connection establishment exchanges contain fixed-length messages and not the variably sized packets of data delivery, the presence of additional bandwidth does little to improve performance after each channel can send paging requests instantaneously. As is shown in Figure 9, the limit of system throughput as bandwidth approaches infinity becomes:

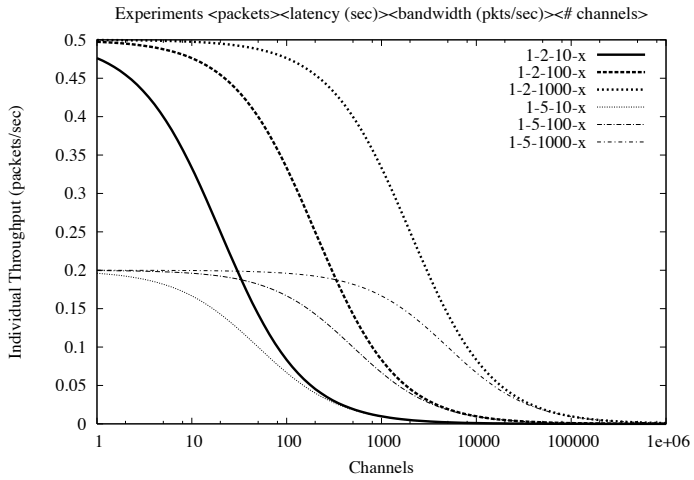

Figure 10: Increasing the number of channels can improve overall system throughput. However, individual throughputs and connection setup times react inversely. Reducing the expense of connection establishment must therefore come from a reduction in connection setup latency. (log-scale)

$$\lim_{BW \to \infty} Throughput \quad = \quad \frac{\#\ Packets}{Setup\ Latency}$$

Increasing system throughput can, for this reason, be accomplished in one of two ways. In the first, the number of channels over which connections can be sent could be increased. Such a change would allow many more connection establishment requests to be sent in parallel. While increasing the throughput of the system as a whole, this approach would prove detrimental to individual users. As shown in Figure 10, subdividing a fixed bandwidth into additional channels intuitively reduces the throughput of a single user. Adding extra channels could also potentially create elevated contention for the shared uplink channel (RACH). More importantly, increasing the throughput of the system does not necessarily reduce cost with respect to delay experienced by individual users. Therefore,

> *Decreasing the cost of connection establishment in a cellular data network is not a matter of increasing bandwidth but rather the reduction of connection setup latency.*

The concept of connection establishment is considerably different in cellular and traditional data networks. In the case of the former, the network must page, wake, and negotiate with a targeted device before ultimately delivering traffic. Whether due to misaligned sleep cycles, missed paging messages or congestion, this set of operations can require more than five seconds before being able to transmit data. As discussed in Section 3, these concessions are made because the network assumes that end devices are limited both in terms of power and computational ability. True packet-switched networks provide no such services; rather, higher layers in the protocol stack implement functionality as needed. In general, each packet is treated as an individual entity and is simply forwarded to the next logical hop. Whether it is wired or wireless in nature, there is no connection to be established from the perspective of the network[5]. Nodes responsible for routing packets do not assume that their next hop neighbors have any specific abilities other than moving the packet closer to its intended destination. Accordingly, connection setup latency is more accurately depicted as propagation delay from the viewpoint of these networks. Given that the delay of propagation time and connection establishment differ by many orders of magnitude, the underlying cause of low-bandwidth attacks on cellular data networks becomes more clear.

The vulnerable components in both the targeted text messaging and cellular data service attacks are those mechanisms responsible for translating traffic from one network architecture to another. While a data network simply forwards individual packets as they arrive, a cellular data network interprets the first packet in a flow as an indicator of more traffic to come. Rather than simply forward that packet to its final destination, the network dedicates significant processing and bandwidth re-

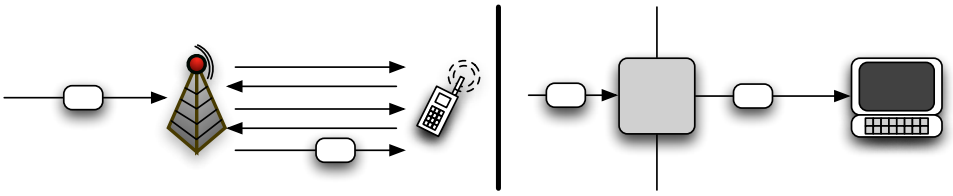

Figure 11: A comparison of the cost of delivering a single packet in cellular and traditional data networks. In the cellular data case (left), a significant amount of delay is added because of connection establishment procedures, whereas the router in the traditional setting (right) simply forwards the packet to the final hop.

sources to ensure that the end device is ready to receive data. This assumption is valid in traditional telephony because of the nature of voice communication. Except for cases of an immediate hangup, sessions are guaranteed to contain multiple "packets" of information. Data communications, however, do not necessarily share this characteristic. Any protocol or application generating packets separated by a number of seconds (e.g., instant messaging programs, session keep-alive messages, applications implementing Nagle's algorithm [34]) violates this model. Whether it is embodied by text messages or data traffic, the amplification of a single incoming packet into a series of expensive delay inducing setup operations is the source of such attacks. Figure 11 reinforces this conclusion by comparing generalizations of the two architectures.

Connection establishment in cellular and traditional networks are so different because the philosophies upon which these systems are based are incompatible. The notion that the middle of a network provide only a limited set of simple functions is at the core of the end-to-end principle [42]. By making no assumptions about the context in which a packet's contents will be used, the network is free to specialize in a single task - moving data. Services not used by all applications, including reliable delivery, content confidentiality and in-order arrival, become the responsibility of higher layers of the protocol stack in the end hosts. The concentration on sending packets allows networks built according to the end-to-end principle to be flexible enough to support new application types and usage models as they emerge. Telecommunications networks are built on the opposite model. Hard service requirements, especially for real-time interaction, forced the network to provide the majority of service guarantees. Because the functionality of the network was once limited to voice applications, telecommunications systems could be tightly tailored to a specific set of constraints. The inclination to build a network in such a manner was addressed by the original end-to-end argument:

> *"Because the communications subsystem is frequently specified before the applications that use the subsystem are known, the designer may be tempted to "help" the users by taking on more function than necessary." [42]*

Because these specialized networks implement more functionality than is absolutely necessary, they exhibit *rigidity*, or the inability to adapt to meet changing requirements or usage [15]. Rigidity in design causes such systems to enforce assumptions appropriate for one subset of traffic on all others. The treatment of each packet as part of a larger flow is one embodiment of such inflexibility. This rigidity is also apparent when examined from the perspective of evolving end devices. For example, many laptops now contain hardware supplying access to cellular data networks [21, 37]. Regardless of their ability to implement services at higher layers of the protocol stack or their access to power, these end devices are forced to transition between STANDBY and READY states simply because such behavior is mandated by the network. Devices connecting via 802.11 could simply trade off the overhead associated with paging at the cost of additional power use. This point is made more obvious when put in the context of home or office LANs supported by a cellular backhaul connection. The network would require such systems to participate in the process of location determination and connection establishment in spite of their lack of mobility. By building assumptions and services into the network itself, the system as a whole is made less flexible. When conditions change and assumptions fail to hold, the rigidity of cellular data systems causes them to break.

6 Constructing Robust Cellular Data Networks

Addressing the specific attacks detailed in this paper may be realistic in the short term. Optimized paging techniques [9, 25] may help to reduce search time and its resulting delay. As was done with the SMS attacks [47], techniques from queue and resource management could be used to mitigate blocking on the RACH. The move to 3G and a significantly larger pool of identifiers would re-

duce the practical likelihood of virtual resource exhaustion. While such methods would indeed mitigate many of the example vulnerabilites discussed in this work, a strategy for building robust cellular data systems based on constant patching would ultimately fail. All of the above solutions merely treat the symptoms of a larger problem. Accordingly, as long as there is a disconnect between the ways in which data is delivered in cellular and traditional data systems, exploitable mechanisms will exist. Such mechanisms need not be limited to the wireless portion of the network; rather, any component of the core network involved in establishing a session will be vulnerable.

The larger issue discussed in this paper, that of vulnerability caused by the exchange of traffic across two incompatible networks, will not be easily solved. Genuinely addressing this problem will require notable changes to the interaction between cellular data networks and end devices. Once such technique might require a significant increase of location awareness on the side of the network. Between the generation of paging lists and bandwidth used in multiple sectors, significant processing resources and time are spent finding a device each time a connection establishment occurs. Instead of knowing that a device is serviced by a potentially large set of base stations, an improved system might require location update information from a device each time it moves between sectors. Used in concert with much shorter sleep cycles, such an improvement to location knowledge may make the elimination of paging possible. This approach, however, would have a serious impact on resources in both end devices and the network. From the user perspective, increased monitoring and interaction with the network would negatively impact battery life. In the case of the latter, the overhead needed to process such an increase in messaging would also affect network performance. A more radical approach would be to replace cellular data services with a new high-bandwidth wireless protocol. Instead of necessarily sharing bandwidth and timeslotting schemes with voice communications, this new protocol would be assigned to a separate portion of the spectrum. In so doing, designers of the new data system would not be constrained by any of the rigidity forced upon current cellular data networks. In addition to technical tradeoffs, this solution would also need to deal with the complexities involved in spectrum allocation - reducing its viability for the forseeable future.

These solutions are not an endorsement of any technology or architecture over another. Instead, they are simply the product of an observation of the impact on availability caused by interconnecting diametrically opposed methods of system design. Being beholden to a specific architecture and failing to understand the problems caused by linking such networks are in fact the causes of the rigidity seen in this system. It is highly unlikely that similar thinking will correct the problem.

7 Related Work

Representing perhaps the oldest functioning digital systems, telecommunications networks have evolved significantly since their inception over 100 years ago. While the nature of these systems themselves has transformed from manually configured and static to automated and mobile, many consumer behaviors have remained largely unchanged. Specifically, the frequency and duration of user calls have become largely predictable behaviors. System designers have used these anticipated conditions to optimize resource allocation throughout their networks. The degree to which telecommunications networks are tailored to such behavior quickly becomes obvious in the presence of unexpected changes to network usage. For example, the explosion in use of dial-up modems in the early 1990s caused widespread congestion because users were remaining connected for longer than expected time periods. Temporary fluctuations or surges, such as those seen minutes after the attacks on September 11th 2001, often render telecommunications networks unusable [35]. Such systems do not gracefully degrade under increased traffic volumes; rather, they often cease to provide service to the vast number of subscribers.

Recognizing this, our previous work focused on the ability to recreate the consequences of such high-traffic denial of service events through the use of low-bandwidth attacks. Using targeted loads of text messages, we were able to demonstrate the ability to deny voice and SMS service to major metropolitan areas with the bandwidth available to a cable modem [16]. We later characterized these attacks through simulation and measurement and discussed the tradeoffs inherent to a number of mitigation strategies [47]. Serror et al. [44] offered additional insight by exploring attacks on call paging channels. Ricciato [39] provided a general discussion of the potential to flood data channels in next generation networks with traffic generated by Internet-based pathogens. Raccic [36] and Mulliner [32] then examined attacks on MMS. While by no means the only methods of causing service outages, these attacks are the first to address the potential for denial of service made possible by the connection between cellular networks and the Internet.

Denial of service attacks have been studied in a variety of other contexts. Websites ranging from DNS roots [17], search engines [40] and software vendors [19] to online casinos [10] and news services [41] have all been temporarily disabled by overwhelming volumes of

traffic. Real-world processes and resources connected to the Internet, including banking networks, emergency services [30] and even postal delivery [13] have also been subjected to such attacks. In response, significant work has been undertaken to classify [29] and alleviate [22–24, 43, 46, 49–52] such problems. Unfortunately, none of these solutions have been widely deployed.

The debate over which network architecture is more resilient against such problems has raged for nearly 30 years. Advocates of the "smart" network, which is embodied by centralized control and decision-making, argue that this architecture provides the ability to prevent such overloading from occurring [31]. Supporters of "dumb" network architectures, which are built around the end-to-end principle [11, 12, 38, 42], contend that placing such control in the network itself dampens the ability to perform its intended task - routing packets. While both approaches have their tradeoffs, the discussion of the consequences of connecting systems that deal with transferring information in fundamentally different ways has not been addressed from the perspective of security.

8 Conclusion

Efforts to address recently discovered vulnerabilities in cellular networks have focused on treating symptoms instead of the disease. Attempts to solve individual exploits have been largely ad-hoc and, in their efforts to mitigate specific problems, create significant additional complexity and vulnerabilities in these systems. Without an understanding of why such attacks are happening, this cycle of vulnerability discovery and patching will continue indefinitely. The problems presented in this and other papers are artifacts of a larger architectural mismatch. Specifically, in spite of a concerted effort to support packet-switched traffic, cellular data networks are still, at their essence, circuit-switched systems. Because of this inflexibility, any mechanism responsible for connection establishment in these networks is vulnerable to a low-bandwidth denial of service attack.

We arrive at this conclusion by making the following contributions:

- Although conventional wisdom suggests that increased bandwidth provides robustness against such attacks, we use two new vulnerabilities to demonstrate that low bandwidth denial of service attacks can prevent legitimate access to cellular data services. In so doing...
- ... we demonstrate that a mismatch of bandwidth between cellular data networks and the Internet is not the cause of such attacks. Instead, they are the result of the contrasting ways in which "smart" and "dumb" networks treat flows. From this...
- ...we show that in their uniform treatment of all flows, regardless of size or duration, cellular data networks exhibit design *rigidity*. By building significant assumptions about the behavior of traffic into the network itself, such systems are made brittle in the face of changing conditions.

Addressing these issues can therefore come from one of two approaches. In the first, methods of safely translating traffic between packet- and circuit-switched networks could be developed. Alternatively, such networks could be redesigned to truly support packet-switched mechanisms. By genuinely separating voice and data, not only in the spectrum they occupy but also in the techniques through which they are delivered, robust cellular data networks could be constructed. In the absence of such changes, cellular networks will continue to remain vulnerable to low-bandwidth exploits.

Acknowledgments

This work was supported in part by Raytheon through a Wireless IR&D contract. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of Raytheon.

We would also like to thank Kevin Butler, William Enck, Joshua Schiffman and our anonymous reviewers for their invaluable comments.

Notes

[1] We use the GSM architecture to provide specific details in our explanation. Similar mechanisms exist in other cellular networks.

[2] Enhanced Data rates for GSM Evolution (EDGE) is largely equivalent to GPRS. The most significant difference is the use of a new wireless modulation technique known as 8-phase shift keying (8PSK), which allows higher data rates.

[3] Note the subtle difference in naming. PDTCHs are virtual channels that are run on top of physical PDCHs.

[4] This timer is referred to in the specifications as T3169 [2]. It is actually started when the counter N3101, which indicates the number of radio blocks that have passed since the last exchange with the targeted device occurred, reaches its maximum value. Our description above is meant to simplify the exact mechanisms for the reader without loss of precision.

[5] We consider connection establishment in terms of individual flows. Initial access to almost every network has a cost (authentication, etc). This startup cost, however, is amortized in both settings.

[6] At the time of this writing, Cingular Wireless had not yet been renamed AT&T.

[7] The voice network equivalent of the PRACH is employed due to the observed presence of dual-use control channels.

References

[1] 3G Newsroom. High speed mobile data driving uptake of PC cards. http://www.3gnewsroom.com/3g_news/mar_06/news_6855.shtml, 2006.

[2] 3rd Generation Partnership Project. General Packet Radio Service (GPRS); Mobile Station (MS) - Base Station System (BSS) interface; Radio Link Control/Medium Access Control (RLC/MAC) protocol. Technical Report 3GPP TS 44.060 v7.6.0.

[3] 3rd Generation Partnership Project. General Packet Radio Service (GPRS); Overall description of GPRS radio interface; Stage 2. Technical Report 3GPP TS 03.64 v8.12.0.

[4] 3rd Generation Partnership Project. GSM/EDGE Radio Access Network; General PAcket Radio Service (GPRS); Overall description of the GPRS radio interface; Stage 2. Technical Report 3GPP TS 43.064 v7.2.0.

[5] 3rd Generation Partnership Project. Physical layer on the radio path; General description. Technical Report 3GPP TS 04.18 v8.26.0.

[6] 3rd Generation Partnership Project. Technical realization of the Short Message Service (SMS). Technical Report 3GPP TS 03.40 v7.5.0.

[7] 3rd Generation Partnership Project. Technical Specification Group GSM/EDGE Radio Access Network; Multiplexing and multiple access on the radio path. Technical Report 3GPP TS 05.02 v8.11.0.

[8] 3rd Generation Partnership Project. Technical Specification Group Radio Access Network; Medium Access Control (MAC) protocol specification (Release 7). Technical Report 3GPP TS 25.321 v7.2.0.

[9] A. Abutaleb and V. O. Li. Paging strategy optimization in personal communication systems. *Wireless Networks*, 3(3):195–204, 1997.

[10] S. Berinato. Online Extortion – How a Bookmaker and a Whiz Kid Took On an Extortionist and Won. *CSO Online*, May 2005.

[11] S. Bhattacharjee, K. Calvert, and E. Zegura. Active Networking and the End-to-End Argument. In *Proceedings of the IEEE International Conference on Network Protocols (ICNP)*, 1997.

[12] M. Blumenthal and D. Clark. Rethinking the design of the Internet: the end-to-end arguments vs. the brave new world. *ACM Transactions on Internet Technology (TOIT)*, 1(1):70–109, 2001.

[13] S. Byers, A. Rubin, and D. Kormann. Defending Against an Internet-based Attack on the Physical World. *ACM Transactions on Internet Technology (TOIT)*, 4(3):239–254, August 2004.

[14] Cingular Wireless. Cingular Wireless. http://www.cingular.com/, 2007.

[15] D. Clark, J. Wroslawski, K. Sollins, and R. Braden. Tussle in Cyberspace: Defining Tomorrow's Internet. In *Proceedings of ACM SIGCOMM*, 2002.

[16] W. Enck, P. Traynor, P. McDaniel, and T. F. La Porta. Exploiting Open Functionality in SMS-Capable Cellular Networks. In *Proceedings of the ACM Conference on Computer and Communication Security (CCS)*, November 2005.

[17] R. Farrow. DNS Root Servers: Protecting the Internet. *Network Magazine*, 2003.

[18] M. Grenville. Stats & Research: 3GSM Visitors Low Users Of Mobile Data. http://www.160characters.org/news.php?action=view&nid=1950, 2006.

[19] C. Haney. NAI is latest DoS victim. http://security.itworld.com/4339/NWW116617_02-05-2001/page_1.html, February 5 2001.

[20] J. Hedden. Math::Random::MT::Auto - Auto-seeded Mersenne Twister PRNGs. http://search.cpan.org/~jdhedden/Math-Random-MT-Auto-5.01/lib/Math/Random/MT/Auto.pm. Version 5.01.

[21] Hewlett-Packard. HP to Drive Mobile Connectivity Around the Globe with Vodafone. http://www.hp.com/hpinfo/newsroom/press/2006/060706b.html, 2006.

[22] J. Ioannidis and S. Bellovin. Implementing Pushback: Router-Based Defense Against DDoS Attacks. In *Proceedings of Network and Distributed System Security Symposium (NDSS)*, February 2002.

[23] A. Juels and J. G. Brainard. Client Puzzles: A Cryptographic Countermeasure Against Connection Depletion Attacks. In *Proceedings of Network and Distributed System Security Symposium (NDSS)*, 1999.

[24] A. Keromytis, V. Misra, and D. Rubenstein. SOS: Secure Overlay Services. In *Proceedings of ACM SIGCOMM*, 2002.

[25] B. Krishnamachari, R.-H. Gau, S. B. Wicker, and Z. J. Haas. Optimal sequential paging in cellular wireless networks. *Wireless Networks*, 10(2):121–131, 2004.

[26] C. Lepschy, G. Minerva, D. Minervini, and F. Pascali. GSM-GPRS radio access dimensioning. In *IEEE Technology Conference (VTC Fall)*, 2001.

[27] C. Luders and R. Haferbeck. The Performance of the GSM Random Access Procedure. In *Vehicular Technology Conference (VTC)*, pages 1165–1169, June 1994.

[28] K. Maney. Surge in text messaging makes cell operators :-). `http://www.usatoday.com/money/2005-07-27-text-messaging_x.htm`, July 27 2005.

[29] J. Mirkovic and P. Reiher. A Taxonomy of DDoS Attacks and DDoS Defense Mechanisms. *ACM SIGCOMM Computer Communication Review*, 34(2):39–53, 2004.

[30] D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford, and N. Weaver. Inside the Slammer Worm. *IEEE Security and Privacy*, 1(4), July 2003.

[31] T. Moors. A critical review of 'End-to-end arguments in system design'. In *Proceedings of the IEEE International Conference on Communications (ICC)*, 2002.

[32] C. Mulliner and G. Vigna. Vulnerability Analysis of MMS User Agents. In *Proceedings of the Annual Computer Security Applications Conference (ACSAC)*, 2006.

[33] R. Mullner, C. F. Ball, K. Ivanov, and H. Winkler. Advanced quality of service strategies for GERAN mobile radio networks. In *IEEE Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2004.

[34] J. Nagle. RFC 896 - Congestion Control in IP/TCP Internetworks. `http://www.ietf.org/rfc/rfc896.txt`, 1984.

[35] National Communications System. SMS over SS7. Technical Report Technical Information Bulletin 03-2 (NCS TIB 03-2), December 2003.

[36] R. Racic, D. Ma, and H. Chen. Exploiting MMS Vulnerabilities to Stealthily Exhaust Mobile Phone's Battery. In *Proceedings of the IEEE International Conference on Security and Privacy in Communication Networks (SecureComm)*, 2006.

[37] M. Reardon. ThinkPads to support Cingular 3G technology. `http://news.com.com/ThinkPads+to+support+Cingular+3G+technology/2100-1034_3-6017968.html`, 2006.

[38] D. Reed, J. Saltzer, and D. Clark. Active Networking and End-To-End Arguments. *IEEE Network*, 12(3):67–71, May/June 1998.

[39] F. Ricciato. Unwanted Traffic in 3G Networks. In *ACM SIGCOMM Computer Communication Review*, 2006.

[40] M. Richtel. Yahoo Attributes a Lengthy Service Failure to an Attack. *The New York Times*, February 8 2000.

[41] P. Roberts. Al-Jazeera Sites Hit With Denial-of-Service Attacks. *PCWorld Magazine*, March 26 2003.

[42] J. H. Saltzer, D. P. Reed, and D. D. Clark. End-To-End Arguments In System Design. *ACM Transactions on Computer Systems*, 2(4):277–288, 1984.

[43] S. Savage, D. Wetherall, A. Karlin, and T. Anderson. Practical network support for IP traceback. In *Proceedings of ACM SIGCOMM*, pages 295–306, October 2000.

[44] J. Serror, H. Zang, and J. C. Bolot. Impact of paging channel overloads or attacks on a cellular network. In *Proceedings of the ACM Workshop on Wireless Security (WiSe)*, 2006.

[45] S. Staniford, V. Paxson, and N. Weaver. How to 0wn the Internet in Your Spare Time. In *Usenix Security Symposium*, pages 149–167, 2002.

[46] A. Stavrou, A. Keromytis, J. Nieh, V. Misra, and D. Rubenstein. MOVE: An End-to-End Solution To Network Denial of Service. In *Proceedings of Network and Distributed System Security Symposium (NDSS)*, 2005.

[47] P. Traynor, W. Enck, P. McDaniel, and T. La Porta. Mitigating Attacks on Open Functionality in SMS-Capable Cellular Networks. In *Proceedings of the Twelfth Annual ACM International Conference on Mobile Computing and Networking (MobiCom)*, 2006.

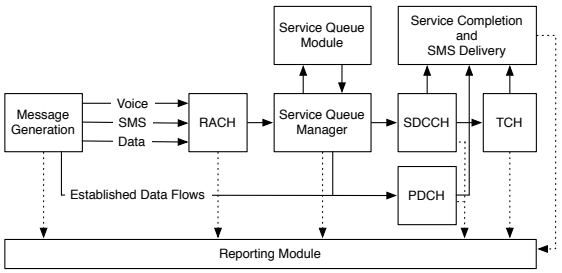

Figure 12: Simulator Architecture

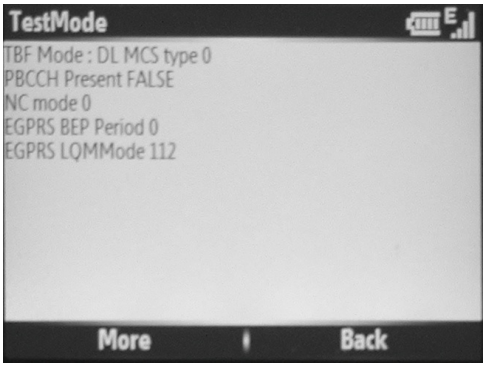

Figure 13: A Samsung Blackjack (SGH-i607) running in Field Test Mode provides operational data on the associated cellular network including channel configuration (shown here) and signal strength.

[48] United States Census Bureau. United States Census 2000. `http://www.census.gov/main/www/cen2000.html`, 2000.

[49] L. von Ahn, M. Blum, N. Hopper, and J. Langford. CAPTCHA: Using hard AI problems for security. In *Proceedings of Eurocrypt*, pages 294–311, 2003.

[50] L. von Ahn, M. Blum, and J. Langford. Telling humans and computers apart automatically. *Communications of the ACM*, 47(2):56–60, 2004.

[51] J. Wang, X. Liu, and A. A. Chien. Empirical Study of Tolerating Denial-of-Service Attacks with a Proxy Network. In *Proceedings of the USENIX Security Symposium*, 2005.

[52] B. Waters, A. Juels, J. Halderman, and E. Felten. New client puzzle outsourcing techniques for DoS resistance. In *Proceedings of ACM Conference on Computer and Communications Security (CCS)*, pages 246–256, 2004.

Appendix

Simulator Design

We extend the GSM simulator built in our previous work [47] to provide support for GPRS data service. In total, the project contains nearly 10,000 lines of code (an addition of approximately 2,000 lines) and supporting scripts. A high-level overview of the components is shown in Figure 12, where solid and broken lines indicate message and reporting flows, respectively. Traffic is created according to a Poisson random distribution through a Mersenne Twister Pseudo Random Number Generator [20], saved to a file and then loaded at runtime. The path taken by individual requests depends on the flow type. We focus on the data path as the behavior of SMS and voice messages were explained in the previous iteration of the simulator.

If the network has not currently dedicated resources to a flow on the arrival of a packet, it is passed to the RACH module. This random access channel is implemented in strict accordance with 3GPP TS 04.18 [5] and is tunable via `max_retrans` and `tx_integer` values. Messages completing processing in the RACH are then delivered to the Service Queue Manager module, which in turn redirects data packets to the PDCH module. If a TFI is available, the packet is assigned the virtual resource, timers are set to five seconds and the packet is then delivered according to a FIFO ordering. The arrival of additional packets in a flow resets the timers to their default values to maintain resource control. When timers expire, the network reclaims a TFI for use in the delivery of other flows. Packets arriving at the Message Generation Manager as part of an active flow bypass the connection setup phases of the network and move directly to the PDCH module.

The accuracy of simulation was measured in two ways. The components used by voice and SMS were previously verified using a comparison of baseline simulation against calculated blocking and utilization rates. With 95% confidence, values fell within ± 0.006 (on a scale of 0.0 to 1.0) of the mean. The simple nature of the PDCH module allowed verification of correctness through baseline simulations and observation.

Parameter Setting

When possible, we use settings found in currently deployed cellular data networks. However, such values are largely unpublished or unavailable to the general population. To find this information, we ran a Samsung Blackjack (SGH-i607) attached to the Cingular Wireless network [6] [14] in Field Test Mode. This mode of operation effectively turns a phone from a communications device to a network auditing platform. In addition to reporting the identification and signal strength read-

ings of nearby base stations, Field Test Mode provides network deployment information including channel allocation and layout. Accordingly, use of this mode of operation is typically restricted; however, access codes and device firmware upgrades are readily available online. As is shown in Figure 13 and of particular interest to properly modeling the behavior of real networks, the field `PBCCH Present FALSE` indicates that voice and data control traffic use the same channels. This configuration, as previously discussed, is permitted by the standards [7] and effectively minimizes the amount of spectrum reserved for control information. Such a setting is believed to be common across the majority of provider networks. From these observations, the establishment of voice and data connections occurs over shared control channels in our simulations.

Other parameters are set using additional literature. For example, the RACH [7] is optimally set to reduce the probability of request blocking by allowing up to the maximum of seven retransmissions per request by the base station [27].