sandvine
Intelligent Broadband Networks

Internet Traffic Classification

A Sandvine Technology Showcase

Contents

Executive Summary 1
Introduction to Internet Traffic Classification ... 2
Sandvine's Traffic Classification Technology 3
The Global Internet Phenomena Program .. 4
Technical Foundation 4
Overcoming Routing Asymmetry 4
Stateful Awareness 5
Correlating across Flows and Sessions 5
Looking inside Tunnels and Encapsulation .. 5
Traffic Identification 5
Signatures 6
Trackers 6
Analyzers 6
Traffic Attributes 7
Traffic Measurements and Metrics 8
User-Defined Measurements 8
SandScript Constructs 8
Encryption, Obfuscation, and Proxies 9
Respecting Content Privacy 10
SandScript's Unique Qualifications 10
Specific Examples 10
Conclusion 16
Additional Resources 16
Invitation to Provide Feedback 16

Executive Summary

Accurate traffic identification and insightful measurements form the foundation of network business intelligence and network policy control. Without identifying and measuring the traffic flowing on their networks, CSPs are unable to craft new subscriber services, optimize shared resource utilization, and ensure correct billing and charging.

Many techniques exist to identify traffic and extract additional information or measure quantities, ranging from relatively simple to extremely complex; in general, advanced techniques that can provide the most comprehensive information and actionable utility are processor-intensive and are therefore only available on best-of-breed deep packet inspection (DPI) and policy control platforms. So-called embedded solutions typically make do with simplistic approaches that are prone to service- and revenue-impacting errors.

To ensure the industry's highest accuracy, largest breadth of completeness, and most comprehensive measurements, Sandvine makes the industry's highest investment in traffic identification.

Additionally, our policy control platform delivers fundamental functionality that is required in order to recognize traffic in modern high capacity data networks: specifically, our platform overcomes routing asymmetry, delivers stateful awareness across multiple sessions, and inspects within tunneled and encapsulated traffic.

Furthermore, the flexibility and versatility of Sandvine's policy definition language, SandScript, enables identification and measurements even when encryption and obfuscation measures are in place.

Introduction to Internet Traffic Classification

Accurate traffic identification and insightful measurements form the foundation of network business intelligence and network policy control. Without identifying and measuring the traffic flowing on their networks, CSPs are unable to craft new subscriber services, optimize shared resource utilization, and ensure correct billing and charging.

First and foremost, CSPs must understand their use cases (now and in the future), as these determine tolerance for accuracy (completeness, false positives, false negatives). It is likely less of a problem if reports show information that is wrong by a small margin, but it can be catastrophic (and very public) if subscriber billing/charging is incorrect or management policies are applied to the wrong traffic.[1]

Traffic classification goes beyond identification (i.e., determining what the traffic is) and extends into extracting information (e.g., video resolution, media type, CDN of origin, etc.) and measuring characteristics (e.g., duration, QoE, etc.); however, not all solutions are created equal.

Many techniques exist to identify traffic and extract additional information or measure quantities, ranging from relatively simple (e.g., regular expressions) to extremely complex (e.g., stateful trackers and analyzers); in general, advanced techniques that can provide the most comprehensive information and actionable utility are processor-intensive and are therefore only available on best-of-breed DPI and policy control platforms. So-called embedded solutions typically make do with simplistic approaches.

For an extensive examination of traffic classification techniques, the trade-offs involved, and the challenges that must be overcome, please refer to the Sandvine whitepaper *Identifying and Measuring Internet Traffic: Techniques and Considerations*.

[1] This excerpt (with bold text for emphasis) from an embedded DPI platform's public documentation illustrates the importance of understanding solution accuracy: *"(This technique) is a method to analyze network traffic such that all the traffic is analyzed by the generic behavior of each flow. (This platform) supports behavioral traffic analysis for P2P (Peer-to-Peer), VoIP (Voice over IP), Upload and Download. If the generic behavior of protocols is detected and traffic classified correctly using behavioral analysis, lesser amount of unknown traffic flows can be seen.* ***These behavioral detections must not be used for charging purposes. Important: This feature is...meant only for statistical purposes (not for charging purposes).***"

Sandvine's Traffic Classification Technology

To confidently manage the network, CSPs must trust that the traffic identification upon which business intelligence, subscriber billing, and policy enforcement are based is accurate.

Sandvine's philosophy towards traffic identification is to focus on accuracy first, and completeness second. That is, we will not sacrifice accuracy (i.e., we will not accept false positives) to reduce the amount of traffic that is unrecognized. Simply put, false positives are unacceptable, as they can have a disastrous impact across a range of use cases.

That said, we routinely see traffic recognition rates upward of 95%[2].

Our philosophy towards traffic measurements is to go far beyond bytes and measure what matters, including direct measurements like duration, discrete events, and round-trip time, and also calculated metrics like quality of experience.

To ensure the industry's highest accuracy, Sandvine makes the industry's highest investment in traffic identification. All development and testing is performed entirely in-house, and a custom-built lab environment replicates real-world conditions as much as is possible. Furthermore, a collection of probes distributed in networks around the world provides two vital functions:

1. Anomaly Detection: These probes continually monitor regional traffic profiles for shifts/deviations from the statistical norm, in order to automatically identify changes in format for existing traffic (e.g., a sudden increase or decrease in BitTorrent traffic indicates a change in the protocol), and to identify new/emerging sources of traffic (e.g., an increase in the amount of unrecognized traffic). When an anomaly is detected, the engineering team is notified within hours of the incident and updated detection algorithms and techniques can be released within 24 hours.
2. Real-World Testing: These probes allow us to test new and updated Loadable Traffic Identification Packages (LTIPs) in the real world, before they are released to customers.

In terms of LTIP update frequency, Sandvine's customers typically fall into one of two categories:

1. Those who update their LTIP version once per month, using each month's major release pack
2. Those who update many times per month, using the minor packs that are released in lock-step with new traffic developments[3]

Updating the LTIP version on deployed devices does not impact service, and is a trivial task using our Control Center management interface[4]; in fact, Control Center lets CSPs update up to 250 Sandvine elements simultaneously[5].

[2] Our Global Internet Phenomena program lets us carefully monitor the rates of unrecognized traffic at our deployments around the world, and these rates are consistently less than 5% of traffic volume. The precise level in any network will vary, of course, based on local characteristics, management policies, and the frequency with which customers update their Loadable Traffic Identification Packs. That said, this average is based upon a mix of some customers – some of whom update frequently, and some of whom don't - and it's not uncommon to see a recognition rate of more than 97%.

[3] The number of minor releases per month varies based upon changes or lack thereof in global traffic profiles. On average, in 2014 and 2015 there were five minor release packs each month.

[4] You can learn about Control Center in general here: https://www.sandvine.com/platform/control-center/

[5] See for yourself in this video: https://www.youtube.com/watch?v=b0Y8J01IfW8

The Global Internet Phenomena Program

Sandvine's industry-renowned Internet traffic classification expertise is confirmed by the trust and authority granted by outside agencies to our Global Internet Phenomena reports[6] and our Internet Phenomena blog[7], which are based on traffic carried by Sandvine's global customers. This program puts our identification and measurements squarely in the spotlight and invites public scrutiny.

This program, and the behind-the-scenes efforts that are underway year-round, ensure that Sandvine has unparalleled visibility into global traffic profiles. For Sandvine's customers, this means that a protocol or application that emerges in a distant network is recognized before it begins to be carried in any significance elsewhere.

Technical Foundation

Before traffic identification signatures and techniques can even be applied, or in the course of applying such techniques, a number of fundamental technical requirements must be met, including the ability to overcome routing asymmetry, to be statefully aware across multiple sessions, and to inspect within tunneled and encapsulated traffic.

Overcoming Routing Asymmetry

By design, all broadband networks exhibit routing asymmetry of one form or another; that is, traffic packets relating to the same flow can take different routes through the network. This poses a significant problem for many traffic identification systems, because the identification engine does not see all packets associated with a particular flow[8].

A Sandvine deployment completely overcomes routing asymmetry by using a network processing unit (NPU) as the first point of examination for incoming packets within our Policy Traffic Switch (PTS)[9]; this NPU ensures that all packets relating to a particular flow, session, and subscriber are presented in order and symmetrically to only one processing core. In this manner, all identification and measurements are based on a complete packet flow.[10]

Furthermore, the NPU resolves routing asymmetry from the perspective of the processing core even in large deployments consisting of multiple points of data intersect (i.e., multiple, separate PTS units). That is, even when packets associated with a flow take different routes through the network and enter different PTS units, these packets are still inspected by a common processor core within the PTS cluster.[11]

The end result of these technologies is that Sandvine's traffic identification and measurements engine sees all traffic completely symmetrically.

---

[6] You can find the reports here: http://www.sandvine.com/trends/global-internet-phenomena/

[7] …and the blog here: http://www.internetphenomena.com/

[8] An extensive explanation of routing asymmetry, including detailed discussion about the challenges asymmetry poses for accurate identification and measurement of traffic, is available in the whitepaper *Applying Network Policy Control to Asymmetric Traffic: Considerations and Solution*.

[9] The PTS is our PCEF/TDF. More information is available here: https://www.sandvine.com/platform/policy-traffic-switch.html

[10] An explanation of how the NPU works is available in *Maximizing Performance with Processor and Core Affinity*.

[11] An explanation of this solution is available in *Policy Traffic Switch Clusters: Overcoming Routing Asymmetry and Achieving Scale*.

Stateful Awareness

Many types of traffic can only be positively identified if the recognition technology has complete awareness of protocol state.

For instance, consider the examples of FTP (this is also applicable to SIP, RTMP, RTSP, and many more). FTP includes a control channel that mediates the protocol data session. In this case, the identification solution must recognize the FTP control traffic and maintain a finite state machine to track the exchange of information between the two FTP endpoints. By examining the control traffic, the identification solution can detect on what ports the data transfer will occur. Without understanding the control traffic, it is impossible to distinguish the forthcoming FTP data from random traffic, as there is no recognizable information within those packets.

Sandvine's traffic classification technology includes statefully aware signatures, called *trackers*, that understand how distinct protocols work. Based on this understanding, the tracker can monitor control traffic flow to update state information and identify subsequent flows that arise.

Correlating across Flows and Sessions

In many cases, a positive identification is only possible if the recognition solution can correlate and apply signatures to the same asset across multiple transactions issued into the same, or different, connections.

For measurements, the need to consider multiple flows and distinct sessions is even more pronounced – for instance, video streams are often long-lived and split content into chunks that arrive through multiple connections. Only a solution that can link all of these connections into a single measurement can deliver meaningful information about video duration and quality.

Since the Sandvine deployment maintains core affinity, the same processing core will always see all flows and sessions associated with a particular 'piece' of traffic; additionally, signatures and measurements are applied across these distinct sessions, so all identification and measured quantities are completely accurate.

Looking inside Tunnels and Encapsulation

A significant portion of traffic that will be inspected for identification and measurement is contained within tunnels (e.g., GTP, GRE, L2TP, Q-in-Q, and IP-in-IP) or encapsulation (e.g., MPLS, EoMPLS, and VLAN).

Sandvine's traffic classification and measurement technology works even in the presence of tunneling and encapsulation by removing the flow headers, performing the inspection, and then reapplying the same headers. This support for encapsulated and/or tunneled traffic is critical for deployment flexibility and enables accurate, comprehensive traffic classification even when tunnels and encapsulation are present.

Traffic Identification

To implement network policy control use cases, the Sandvine Policy Engine[12] expresses policy instructions with an event-driven policy language called SandScript[13]. SandScript allows CSPs to link

[12] The Policy Engine is the 'brain' of the platform: https://www.sandvine.com/platform/policy-engine.html
[13] More information about SandScript is available here: http://www.sandvine.com/technology/sandscript.html

conditions to actions in real-time, and information provided by the traffic identification technology is immediately available as conditions in policy evaluation.

To identify traffic, Sandvine uses three main techniques: *signatures*, *trackers*, and *analyzers*. Sandvine's traffic identification is achieved without false-positive prone port-based dependencies or 'suspected' categories.

Signatures

Signatures identify traffic based on observed patterns within, across, and between both packets and flows. With Sandvine's implementation, this can even include mathematical operations and decryption.[14] However, as described previously, sometimes it isn't possible to positively identify traffic without prior knowledge of the flow, which is where *trackers* come into play.

Trackers

Trackers are somewhat 'protocol-aware', in that they don't just identify traffic but also understand at some level how it works. Relating to the previous example of FTP, a *tracker* monitors the control traffic to learn where the data transfer is going to appear.

However, even *trackers* aren't always sufficient for some advanced use cases that require more in-depth real-time analysis of a flow; these use cases require *analyzers*.

Analyzers

Analyzers are completely protocol-aware: not only do they understand the language of the protocol, but an *analyzer* can extract detailed protocol-specific fields (e.g., the name of a particular OTT video provider) and directly interact with the protocol.[15]

Importantly, the use of *analyzers* goes beyond simply identifying traffic and extends into enabling advanced measurements, including the counting of individual videos, messages, downloads, etc.

Sandvine's use of *analyzers* is a good example of how a traditional 'signature count' (i.e., of signature library size) is more or less useless in assessing the completeness of traffic identification. For instance, a single RTMP *analyzer* can extract infinite distinct RTMP content providers, obviating the need for distinct recognizers for each video service.

Another benefit of this abstracted approach is that there is no need for a library update – new providers are automatically extracted as they are observed.

DNS Analysis

Many CSPs have discovered the shortcomings of using lists of IP addresses as aids in traffic identification, typically to enable application- or service-based subscriber tiers (e.g., for zero-rating Facebook usage). Unfortunately, this approach has many flaws; for instance: IP ranges require frequent updates (e.g., daily, hourly, real-time), and all but guarantee false positives and false negatives.[16]

---

[14] So, our use of *signature* should not be confused with a simple regular expression match.

[15] For instance, Sandvine's HTTP redirection capability (used by CSPs for a variety of use cases, including subscriber notification and regulatory content filtering) works by actually interacting with the HTTP traffic and modifying it by sending an HTTP 403 redirect. Similarly, *analyzers* make it possible to redirect video over RTSP (for instance, to a video cache or video optimization solution) by sending a "302 Moved Temporarily" response and closing the initial RTSP connection via a "Connection: close" header.

[16] Many CSPs have rolled out 'Unlimited Facebook' plans based on IP ranges, only to lose massive amounts of revenue.

A partial solution that is in place in many embedded systems is to rely on DNS traffic to populate these IP lists. Essentially, the solution pays careful attention to DNS traffic and recognizes requests for specific domains; the solution then waits for the response and extracts the IP address provided and adds this IP to a list.

When done correctly, this approach is quite useful, but it is not without pitfalls. For instance, not all DNS servers are trustworthy (e.g., malware will often point subscribers to off-net malicious DNS servers that trick subscribers into going to identity theft sites), and a DNS-based solution alone will not know when to remove 'expired' IP addresses from the list.

With encryption becoming widespread, Sandvine's traffic recognition is increasingly incorporating DNS analysis as a technique, but without the common pitfalls. For instance, DNS analysis is never relied upon as the sole identifying mechanism and not all DNS servers are trusted equally.

Traffic Attributes

Beyond simply identifying traffic, *analyzers* can extract additional data from the traffic itself; this information is then available for business intelligence purposes, and can also be used in real-time policy decisions and enforcement (for instance, traffic associated with a particular video provider or mobile device model can be selectively redirected through a video optimization solution).

In Sandvine nomenclature, this information takes the form of an *attribute*.

Think of the English meaning of the word attribute: "something belonging to or characteristic of a person/thing". It's the same with traffic classification, except the subject at hand is "a characteristic of network traffic".

Attributes provide a mechanism for associating data with a subscriber or with a subscriber's sessions. SandScript can be used to define attributes and associate them with one or more subscribers. The attributes that are attached to subscribers can be used to trigger policy-defined management of the subscribers and their flows, which is very powerful. Some sample attributes include:

- service tier
- video codec
- IP addresses
- audio codec
- MAC addresses
- client device
- client device type
- content (e.g., video, audio) provider
- operating system
- browser
- media stream type
- session protocol
- media container
- transport protocol
- video resolution

The freeform nature of the SandScript policy language lets CSPs use attributes alone or in combination to enact highly specialized network policy control. For instance, *attributes* are used extensively to

detect device tethering and to identify the presence of multiple devices behind a household NAT. For a detailed explanation of how Sandvine's traffic classification technology delivers device awareness, please read *Policy Control for Connected and Tethered Devices*, available at www.sandvine.com.

Traffic Measurements and Metrics

Measurements and metrics go beyond identifying traffic and extracting additional information by actually assigning quantity to observations. A trivial example is volume: the number of bytes associated with a particular designation of traffic. However, volume is only scratching the surface.

Measurements, which are simply counts of observed characteristics, can add immense value for CSPs. From a business intelligence standpoint, *measurements* give a level of understanding that simple byte counts just cannot. Additionally, by linking in real-time to policy control, *measurements* enable remarkable policy control use cases. For instance, real-time, topology-aware measurements of access network round-trip time can be used to trigger precise congestion management policies.[17]

Other available *measurements* with Sandvine include:

- Duration of video or audio streams
- Counts of the number of events
- Tracking of "top" items (e.g., most frequently requested URLs, most popular video providers, etc.)
- Summations (e.g., adding up a number of observed or measured values)

Measurements can be defined and invoked in SandScript as part of any policy control implementation. Additionally, to simplify the customer experience, Sandvine has published a number of Policy Packs[18] that include comprehensive sets of *measurements*; these Policy Packs can simply be imported into and referenced by policy.

Metrics take *measurements* one step further by calculating a value based on observations. Two prominent examples available to Sandvine customers are voice-over-IP quality of experience (QoE) and video QoE[19].

User-Defined Measurements

In addition to the *measurements* that are available in SandScript and the Policy Packs, there is a generic measurement type that can be used to implement any kind of user-defined counter or statistic.

SandScript Constructs

Although we are perhaps now extending beyond the intended scope of this technology showcase document, it is worthwhile to note that SandScript includes constructs that aid with the 'processing' of measurements. For instance, SandScript includes the concepts of:

- 'Top': find out the top instance of observed/measured values (e.g., to identify the websites with the most hits)
- 'Sum': aggregate observed/measured values (e.g., to sum up all video minutes consumed on the iPhone 6)

[17] The network congestion management use case is worthy of an entire technology showcase by itself: *The QualityGuard Congestion Response System*.
[18] For instance, the "Network Business Intelligence (NBI) Policy Pack".
[19] This one also gets its own technology showcase: *Video Quality of Experience Score*.

- 'Histogram': counts observed/measured values into bins of defined sizes (e.g., explore the distribution of video durations)

Encryption, Obfuscation, and Proxies

Information can be hidden, guarded, or unavailable for many reasons, and three widespread mechanisms in the context of traffic classification are:

- Encryption: both to ensure content privacy (i.e., encoding information such that it can only be read by an authorized party) and as an obfuscation measure
- Obfuscation: hiding or disguising information to prevent detection[20]
- Proxies: primarily to compress data or increase performance (by reducing latency)

A fairly detailed explanation of the most common encryption and obfuscation technologies, and proxy services, is included in the whitepaper *Identifying and Measuring Internet Traffic: Techniques and Considerations*.

In general, the impact of each of these mechanisms can be summarized as follows:

- **Encryption:** It is important for CSPs to keep in mind that encryption does not mean something is undetectable or unidentifiable, it just means that the content is private. Because most encrypted traffic relies on accepted standards (e.g., IPSEC, TLS), it is generally easy to detect, although capabilities do vary by solution vendor.[21] Furthermore, encryption rates are already high (for instance, YouTube and Netflix are both encrypted, and together account for more than 50% of network traffic in many regions) and are getting higher[22], thanks to efforts like the Electronic Frontier Foundation's (EFF) HTTPS Everywhere initiative[23]. Encryption really only impacts solutions that depend on content information, like program titles or MIME types.
- **Obfuscation:** Applications and services that provide obfuscation are relatively rare, as the general need or motivation for obfuscation isn't widespread. Since there are no standards, identification of obfuscated traffic is case-by-case, and typically involves many detection mechanisms working in tandem. Because of this computational requirement, dedicated solutions are superior to embedded solutions.
- **Proxies:** In general, the proxies themselves can be identified, but identification within the proxy tunnel is not practical. While this reality could be a point of concern for CSPs, it's important to keep in mind that proxies are only useful in very limited scenarios. For instance, for all the attention that SPDY receives, it will only be useful for a tiny subset of Internet users and a tiny percentage of overall traffic. [24] Additionally, as more websites and services adopt TLS/SSL, even on networks where it makes sense to use SPDY its use will decline.

[20] Obfuscation is also used to protect user identity (for instance, by using something like TOR), but this paper is addressing *traffic* identification, not *user* identification.

[21] For instance (as is explained a little later in this document), the "server_name" field is visible in TLS, but exists at a variable offset. As a consequence, solutions with hardware fast-paths for TLS traffic will struggle, as they typically lack the flexibility to handle non-fixed offsets.

[22] *Global Internet Phenomena Spotlight: Encrypted Internet Traffic* quantifies the proportion of encrypted Internet traffic on global networks as of April 2015 and extends recent observations into projections for the future.

[23] You can learn more about this initiative here: https://www.eff.org/https-everywhere

[24] Wondering why? SPDY primarily exists to eliminate TCP round-trip time latency on mobile networks; in reality, this latency is only a problem in 3G environments. Further, SPDY only applies to browser traffic, and even then only to non-SSL traffic. So, ultimately, SPDY will be used for non-SSL browser-based traffic on 3G networks.

Respecting Content Privacy

While many of Sandvine's competitors have attempted to delve into traffic content (i.e., going beyond simply identifying "this is Netflix" and measuring its characteristics to instead say "this is Netflix and is episode 4 of House of Cards"), Sandvine has never had such an interest. This is an important distinction, for at least two reasons:

1. Sandvine respects users' privacy: We seek to identify traffic, its attributes, and its measured characteristics, because those are needed to achieve CSP use cases including business intelligence, service creation, traffic optimization, and network security. Revealing the precise content of a traffic flow does not advance any of these use cases.
2. Content intelligence solutions are incredibly vulnerable to proprietary encryption[25]: Most content providers have an incentive to protect details of their service usage, and will actively take measures to prevent third-parties from extracting and revealing this information. For CSPs, this reality means that a content intelligence solution bought today can be rendered mostly useless tomorrow[26]. In contrast, most content companies have no incentive to prevent mere identification of their traffic – they just don't want anyone investigating more deeply into the exact content itself.

SandScript's Unique Qualifications

SandScript provides flexibility and versatility that typical rules-based systems and generic deep packet inspection (DPI) solutions cannot match. As such, it is uniquely able to combine and apply a wide range of techniques towards identifying (and extracting information from and measuring) traffic.

In general, identifying encrypted and obfuscated traffic usually involves a collection of techniques, including using stateful *trackers* to prime for upcoming flows, *analyzers* to extract particular fields (that can be used directly, in combination, or in mathematical operations), decryption, deciphering, certificate inspection, behavioral heuristics, and other mechanisms[27].

This level of flexibility gives Sandvine's traffic identification incredible versatility – traffic recognition techniques can evolve on the fly and can be updated without any impact to traffic flow. Practically, this means that the precise combination of factors used to identify particular traffic can be adjusted as that traffic itself changes with new version updates.

Additionally, this flexibility simply is not available in 'embedded' DPI systems (i.e., those built into packet gateways like GGSNs) or from vendors who just license a DPI library and load it onto hardware.

Specific Examples

The following sub-sections discuss real-world examples in which Sandvine's traffic identification and measurements have been applied to encrypted and obfuscated traffic.

[25] This is also why no network-based DRM enforcement system has ever been successful.

[26] This has already happened to at least one vendor's product, impacting all CSPs who had purchased it. Interestingly, the vendor's website continues to promote capabilities that are no longer available, even though the lack of availability is publicly acknowledged and received widespread coverage.

[27] As a general note to the reader: watch out for solutions based on IP ranges or even DNS inspection. IP ranges require real-time updates, and essentially guarantee both false positives and false negatives. DNS inspection is often proposed as a solution to these problems (i.e., "let's extract IP addresses from DNS responses"), but must be done correctly and comprehensively to be a useful solution.

Identification within SSL/TLS and HTTPS

Secure Sockets Layer (SSL) and Transport Layer Security (TLS) are cryptographic protocols designed to provide secure communications, and are used extensively in applications where security is required (e.g., banking, VPNs, etc.). HTTP Secure (HTTPS) adds the security capabilities of SSL/TLS to HTTP communications. HTTPS is technically not a protocol by itself, as it simply HTTP on top of SSL/TLS.

Several years ago, we recognized the need to develop traffic identification technologies that would preserve the value of policy control in a world with large amounts of SSL/TLS traffic[28]. Since then, we have introduced and continue to refine technology that allows us to identify the provider of SSL/TLS traffic, which enables the use cases in which our customers are interested.

Even when traffic is encrypted with SSL/TLS, there are still many fields available (e.g., `server_name` - see Figure 1[29]) that can be observed while the connection is being established and the certificate[30] is being exchanged. In combination with other factors and techniques, this information makes it possible to identify the provider of SSL/TLS content.

Of course, SSL/TLS prevents content inspection and ensures content privacy, but (as stated above) that is of no consequence to Sandvine.

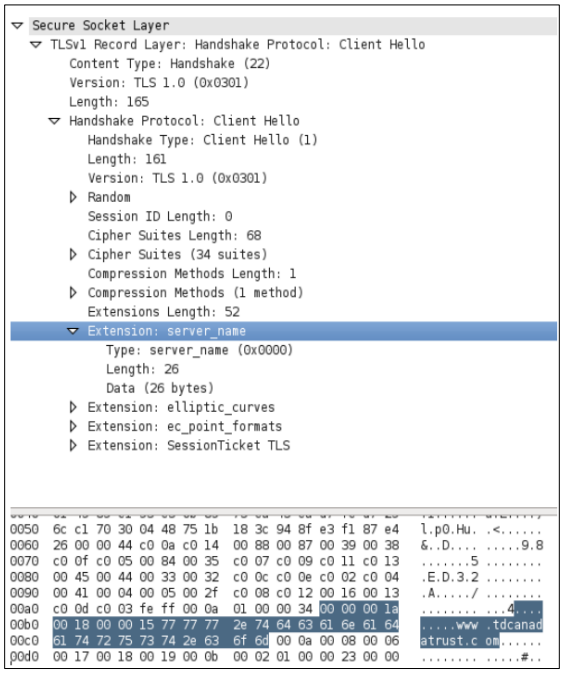

Figure 1 - Packet capture showing the `server_name` and the corresponding value.

It is worth noting, however, that just because the `server_name` field is available does not mean that vendor capabilities are equivalent. In this case, the `server_name` field occurs at a variable offset, so solutions that rely on hardware-based inspection have difficulty reading the field. Software-based solutions, like Sandvine's, have no problem at all.

Addressing Proxies

Network proxies create another challenge that must be overcome by traffic classification solutions, as these proxies act as intermediaries that can disguise the origin and content of traffic. While there is sufficient overlap that mutually exclusive designations are difficult, in general one must consider data compression proxies, proxy applications, and web proxies[31].

- **Data Compression Proxies:** These are proxy services that provide data compression to users (with the intent of reducing bandwidth usage). The precise practical impact varies based upon the location of the proxy. If it is within or upstream of the CSP's network (see Figure 2), then positive identification of traffic is possible, but if the proxy is below the CSP's network (for

[28] For instance, the increase of SSL traffic associated with a consumer shift to cloud services was highlighted in the 1H2012 Global Internet Phenomena Report

[29] This image was the top search result for "wireshark+ssl+server_name", and comes from security researcher Michael Dundas; many more examples are available online with similar searches.

[30] As in, the public key certificate.

[31] There is also the emerging idea of a "Trusted Proxy" that would enable a third-party 'man-in-the-middle' (e.g., CSP) to decrypt traffic on-the-fly, perform an operation (for instance, implement video optimization) and then re-encrypt. The latest IETF draft for this concept is available here: http://tools.ietf.org/html/draft-loreto-httpbis-trusted-proxy20-01

instance, on a mobile device), then it is only possible to determine the identity of the proxy provider, and that a proxy is in use.

- **Proxy Applications:** These are applications that can be installed on (typically mobile) devices to provide users with privacy[32] and more efficient data usage. Similarly, add-ons/plug-ins or configurations can instruct web browsers to use certain optimization protocols or techniques. In most cases, it is only possible to identify the provider of the proxy application, but not the traffic itself.
- **Web Proxies:** Web proxies are a subset of proxies, and are typically intended to provide anonymity on the web and to provide a means around geographic restrictions (for instance, to get access to the U.S. Netflix library from another country). Web proxies generally do not provide encryption[33]. With a web proxy, it is usually possible to identify the traffic, as the web proxy is upstream of the CSP network (see Figure 2).

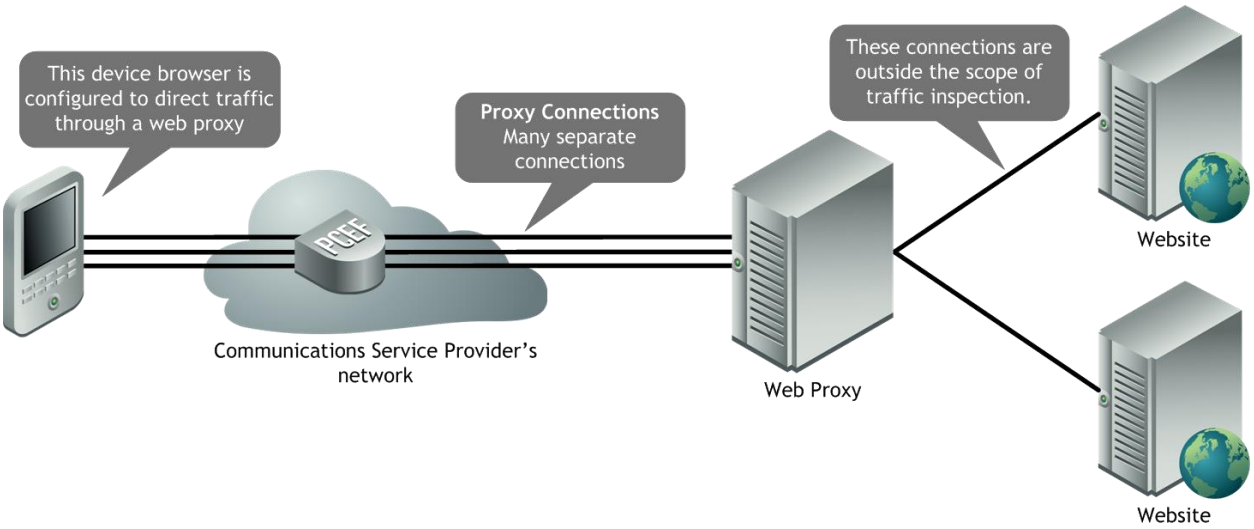

Figure 2 - Simplified view of a device using a web proxy (the same diagram applies to a data compression proxy). In this deploymenet, the proxy exists upstream of the CSP network, so the CSP can still inspect traffic.

Differentiating Internet Applications and Services[34]

Many innovative subscriber services, such as fixed-price unlimited application usage, depend upon the ability to identify a specific collection of services; however, it is not uncommon for a single web service provider to offer many services (e.g., email, cloud storage, IM, etc.), only some of which are included in the fixed-price plans.

Recently, to enable a new plan in a prepaid-dominant market, Sandvine needed to separate the instant messaging web service from many other services provided by the same entity. Complicating matters, all web services traffic to and from this entity was carried over HTTPS.

[32] In reality, instead of achieving privacy, users are actually granting complete visibility to the proxy provider (e.g., Google, Opera, etc.), who have business interests in expanding their visibility into user activity beyond their own hosted services.
[33] For a web proxy to provide encryption, it would need its own certificate, but that would prevent the consumer device from verifying the signatures of the websites and services to which the device is ultimately connecting
[34] Zero-rated and unlimited Facebook usage plans are increasingly popular, and many readers might have taken recent note that Facebook now supports TOR (The Onion Router). What is the impact to CSPs? Likely none. Facebook's support for TOR is to enable Facebook usage in regions where it is restricted or privacy is in doubt; in this scenario, it is unlikely that there would be Facebook plans. Similarly, users savvy enough to use TOR for their Facebook sessions likely do not expect that Facebook traffic to be zero-rated, as it is deliberately obfuscated.

By using a combination of factors, including a number of pieces of information from the SSL certificate, we were able to distinguish the IM traffic, the email component, and other pieces to enable a valuable subscriber service.

Identifying Cloud-Hosted Voice-over-IP

Many CSPs are interested in understanding VoIP usage, in terms of the number of calls, total minutes, and, most importantly, subscriber quality of experience.

Recently, a popular VoIP client implemented a new cloud-based service that also applied SSL encryption to calls. A rudimentary recognition technique would fall prey to false positives by 'recognizing' the VoIP traffic as the actual cloud host; to accurately measure VoIP traffic on the network, this service needed to be separated.

Detailed examination of the VoIP traffic revealed a collection of fields that could be extracted by an *analyzer* and used in combination to positively identify this VoIP service, and preserve accurate business intelligence for the CSP; as a result, the CSP was able to continue to gain insight into VoIP usage on their network in terms of number of calls, total minutes, etc.

Separating Instant Messaging Capabilities

Instant messaging clients enable a wide range communications formats, including exchange of text, images, video files, and other attachments. By understanding the variation in these uses, CSPs can anticipate subscriber needs and craft precisely targeted services.

However, with encryption being implemented to protect subscriber privacy, it becomes more difficult to recognize the different behaviors and actions associated with IM clients.

Recently, a globally popular instant messaging client introduced full SSL encryption of messages; despite this change, Sandvine was still able to distinguish text, images, videos, and file attachments from other IM client traffic.

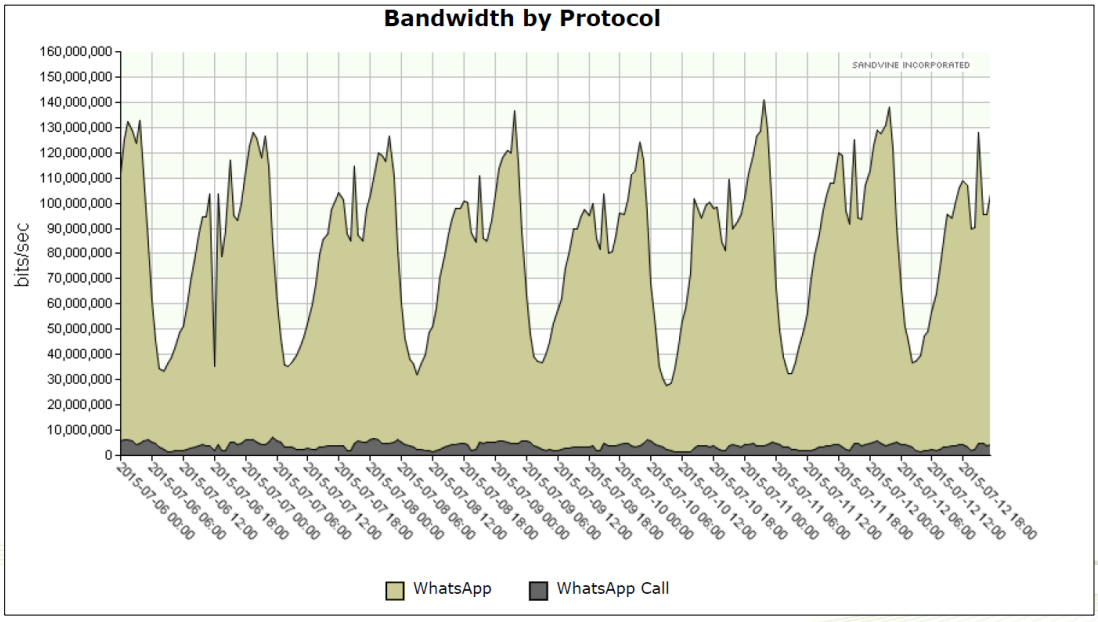

Figure 3 - This Sandvine Network Demographics report shows bandwidth attributable to WhatsApp text content and WhatsApp voice calls, even though WhatsApp is encrypted

Differentiating Quality of Encrypted Video

Many CSPs closely track detailed video statistics in order to anticipate subscriber demand and to detect shifts in quality that could indicate network or routing issues; video content encryption has the potential to undermine these efforts.

With many major video services implementing content encryption, Sandvine developed techniques that preserve valuable business intelligence even when video is encrypted. As a simple example, consider YouTube content carried via HTTPS: even though the content is encrypted, it is still possible to distinguish between various levels of display resolution by carefully measuring bitrates and volumes (and accounting for the bursty nature of the stream delivery).

Measuring Streaming Traffic Bypass

To overcome regional licensing and content restrictions for primarily video and audio content (e.g., Netflix, Hulu, BBC's iPlayer, Amazon, sports events, music streaming services, etc.), some consumers have turned to web-based services:

- DNS Proxies: The most common form of bypass, these services change the DNS used by subscribers to a DNS that is purposely configured to ignore geographically segmented DNS responses. The result is that a content server believes a subscriber is located in a different country – one for which the content is licensed and can be served. These services are popular because they are low-cost, require little or no technical ability, and apply to all devices within a household.
- VPN Services: These services provide point-to-point tunneling in which the endpoint of the VPN is in a country for which content is licensed. These typically have a higher cost than DNS services, usually require some technical ability, and might be limited to a small number of devices.
- Plug-ins: These solutions combine a few technologies including VPN, DNS, and manipulation of URL (i.e., altering the request to source content from another country) for different streaming servers. While most of these solutions are browser plug-ins, they are also available for non-browser platforms (e.g., gaming consoles, smart TVs, etc.). Like DNS proxies, these are low-cost and don't require much technical ability on the part of the user, but they may not work for all services.

Providing comprehensive business intelligence into streaming usage requires recognizing streaming traffic when any of these techniques are in use.

To recognize and measure streaming usage when DNS proxies are in use, the Sandvine system identifies subscribers who are actively connected to a DNS service and who exceed a minimum traffic threshold that indicates that streaming is in progress. When these conditions are met, streaming traffic is measured. Identification of the streaming proxies themselves is achieved via a combination of manual research into popular and emerging services, and automated discovery techniques.

Counting streaming traffic carried via VPN tunnels is similar to the mechanisms used to differentiate encrypted video stream bitrates, with heuristics being leveraged to recognize streaming characteristics.

Many of the same techniques that detect DNS and VPN services are used to detect when a plug-in is in use. Interestingly, some plug-ins indicate the use of peer-to-peer type networks, so techniques that are applied to detecting encrypted P2P filesharing can be reapplied here.

Applying Machine Learning to Outbound Email Spam Identification

Email spam leaving the network remains a costly problem for CSPs: a few Trojan-infected users can cause an entire network to be blacklisted, impacting the ability of subscribers and business customers to send email.

Bayesian content inspection approaches to identifying email spam are, at worst, ill-suited to the rapid evolution of spam Trojans and, at best, overmatched by the sheer volume of outbound spam messages.

Since 2004, we have provided CSPs with outbound spam defense built on machine learning algorithms. These algorithms have two levels of measurement (absolute metrics, which are measured per 60-minute sliding window, and ratio metrics, which are measured per 5-minute sliding window) that examine 10 and five metrics, respectively.[35]

Used in combination, these metrics and algorithms allow CSPs both to identify email spammers and to automatically tune the sensitivity of the detection to strike the optimal balance between security and permissibility – all without ever looking at the actual email content.

[35] Example absolute metrics include: attempted messages, SMTP resets, SMTP sessions, unique HELO/EHLO names, and unique recipient domains. Example relative metrics include: attempted messages per period, recipients per period, and unique sender addresses per period.

Conclusion

Accurate traffic identification and insightful measurements form the foundation of network business intelligence and network policy control. Without identifying and measuring the traffic flowing on their networks, CSPs are unable to craft new subscriber services, optimize shared resource utilization, and ensure correct billing and charging.

Many techniques exist to identify traffic and extract additional information or measure quantities, ranging from relatively simple (e.g., regular expressions) to extremely complex (e.g., stateful trackers and analyzers); in general, advanced techniques that can provide the most comprehensive information and actionable utility are processor-intensive and are therefore only available on best-of-breed DPI and policy control platforms. So-called embedded solutions typically make do with simplistic approaches.

To ensure the industry's highest accuracy, largest breadth of completeness, and most comprehensive measurements, Sandvine makes the industry's highest investment in traffic identification.

Additionally, our policy control platform delivers fundamental functionality that is required in order to recognize traffic in modern high capacity data networks: specifically, our platform overcomes routing asymmetry, delivers stateful awareness across multiple sessions, and inspects within tunneled and encapsulated traffic.

Furthermore, the flexibility and versatility of Sandvine's policy definition language, SandScript, enables identification and measurements even when encryption and obfuscation measures are in place.

Additional Resources

In addition to the documents and resources mentioned already in this paper, readers with a particular interest in encryption should consider reading the GSMA's position paper *Network Management of Encrypted Traffic*. That paper makes specific recommendations for "technical architects with knowledge of the operator network traffic management functions".

Invitation to Provide Feedback

Thank you for taking the time to read this technology showcase. We hope that you found it useful, and that it helped you understand our traffic classification technology.

If you have any feedback or have questions that have gone unanswered, then please send a note to whitepapers@sandvine.com

Headquarters
Sandvine Incorporated ULC
Waterloo, Ontario Canada
Phone: +1 519 880 2600
Email: sales@sandvine.com

European Offices
Sandvine Limited
Basingstoke, UK
Phone: +44 0 1256 698021
Email: sales@sandvine.co.uk

sandvine®
Intelligent Broadband Networks